Contributions to Statistical Methods for Non-standard Data: Data Depth and Formal Concept Analysis

Hannah Frederike Caecilie Blocher

München 2025

Contributions to Statistical Methods for Non-standard Data: Data Depth and Formal Concept Analysis

Hannah Frederike Caecilie Blocher

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig–Maximilians–Universität
München

vorgelegt von
Hannah Frederike Caecilie Blocher
aus Reutlingen

München, den 03.03.2025

Erstgutachter: Prof. Dr. Thomas Augustin
Zweitgutachter: Prof. Dr. Gerd Stumme
Drittgutachter: Prof. Dr. Volker Schmid
Tag der Einreichung: 03.03.2025
Tag der mündlichen Prüfung: 26.05.2025

Acknowledgement

I want to express my heartfelt gratitude to everyone who supported me throughout the past years in completing this dissertation. Special thank you goes to

... Thomas Augustin for your invaluable support, advice, encouragement and time. In particular, I am deeply grateful for your guidance and unwavering support in reorganizing my dissertation studies such that they truly suited me.

... Gerd Stumme and Volker Schmid for serving as my external reviewers, and to Christian Heumann and Fabian Scheipl for steering my examination committee.

... my coauthors Georg Schollmeyer, Christoph Jansen, Julian Rodemann, Malte Nalenz and Thomas Augustin. I deeply appreciate the fruitful discussions, insightful tips and ideas you shared with me. In particular, I want to thank Georg Schollmeyer for encouraging me to start my dissertation studies and for providing me with the necessary background knowledge by giving the lecture on formal concept analysis and supervising my master's thesis.

... all former and current members of the working group: Thomas Augustin, Cornelia Fütterer, Polina Gordienko, Lea Höhler, Christoph Jansen, Gilbert Kiprotich, Dominik Kreiß, Malte Nalenz, Julian Rodemann, Georg Schollmeyer, and Patrick Schwaferts. Thank you for all the encouragements, supports, meal breaks and laughs we have shared.

... all my colleagues at the department of statistics. I am grateful for the enjoyable lunches, coffee breaks, and the supportive atmosphere. Special thanks to Conni, Katharina, and Daniel for sharing your office space, and to Elke and Brigitte for the professional management of the administrative aspects of the department.

... Stanislav Nagy for the insightful discussions on depth functions and to Gerd Stumme for the engaging conversations on formal concept analysis. These exchanges not only deepened my understanding of each topic but also enabled me to combine them effectively.

... the Evangelisches Studienwerk Villist e.V. Thank you for believing in me and my project, for every single wonderful weekend in Villigst (or elsewhere) and also the opportunity to help organizing the *Sommeruniversität 2025*.

... the LMU Mentoring program at Faculty 16, and in particular Anne-Laure Boulestix, for advice regarding my dissertation journey, as well as the financial support with equipment and conferences.

... the following companies and corresponding developers for inventing and implementing helpful software programs: R (for providing a programming language), RStudio (for an integrated development environment for R), Gurobi (for solving

binary linear programs), Github (for providing a distributed version control), LaTeX (for typesetting documents), DeepL (used for language editing) and ChatGPT by OpenAI (used for language editing).

... family and friends, whose endless phone calls, meals, laughter, etc., provided the energy and perspective I needed to complete this dissertation, reminding me of what truly matters.

... my partner Nicco for your unwavering support and for always believing in me. Thank you for everything!

Zusammenfassung

Die vorliegende kumulative Dissertation befasst sich mit der Entwicklung statistischer Methoden für sogenannte Nicht-Standard Daten. Der Begriff "Nicht-Standard Daten" umfasst Datenstrukturen, die nicht in klassische statistische Datenformate passen, wie beispielsweise Daten, deren Einzelbeobachtungen partiell geordnete Mengen oder gemischt ordinal-numerisch-räumliche Beobachtungen sind. Klassische statistische Methoden basieren jedoch meist auf der (impliziten) Annahme, dass die Daten einem der klassischen Datenformate entsprechen. Nutzer*innen stehen daher vor dem Dilemma, entweder ungeeignete Methoden anzuwenden, wodurch die resultierenden Analysen und Interpretationen potenziell verfälscht werden, oder ganz auf eine Analyse zu verzichten, weil die zugrunde liegende Datenstruktur keine direkte Anwendung klassischer Methoden erlaubt. Diese Dissertation trägt zur Lösung dieses Problems bei, indem sie explizit für Nicht-Standard Daten flexible, nicht-parametrische Methoden entwickelt.

Das obige Dilemma wird in der Dissertation auf zwei verschiedene Weisen adressiert. Der primäre Ansatz basiert auf dem Konzept der Tiefenfunktionen und der formalen Begriffsanalyse. Die formale Begriffsanalyse betrachtet nicht die konkreten, einzelnen Datenwerte, sondern die daraus entstehenden Relationen zwischen den Datenelementen. Tiefenfunktionen messen die Zentralität beziehungsweise Abgelegenheit eines Punktes relativ zu einer Verteilung oder Stichprobe und klassischerweise sind sie für normierte Vektorräume definiert. Diese Dissertation erweitert mit Hilfe der formalen Begriffsanalyse das Konzept der Tiefenfunktionen auf Nicht-Standard Daten. Um die Begriffe Zentralität und Abgelegenheit für diese Daten zu spezifizieren, werden strukturelle Eigenschaften eingeführt, die sowohl die Idee der Tiefenfunktionen für normierten Vektorräumen auf Nicht-Standard Daten übertragen, als auch die relationale Struktur der Daten berücksichtigen. Auf Basis dieser strukturellen Eigenschaften werden zwei konkrete Tiefenfunktionen vorgestellt: die ufg-Tiefe und die generalisierte Tukey-Tiefe. Die Tiefenfunktionen für allgemeine Nicht-Standard Daten werden im Anschluss auf die spezielle Datenstruktur der partiell geordneten Mengen angewendet. Hier besteht der Grundraum aus allen möglichen partiellen Ordnungen über einer festen Menge von Elementen. Durch diese Einschränkung auf partiell geordnete Mengen können die beiden Tiefenfunktionen gezielt anhand weiterer, für diese Daten spezifischen Eigenschaften analysiert und deren Berechnung optimiert werden. Darüber hinaus wird die ufg-Tiefe für partiell geordnete Mengen auf Benchmarking Probleme im Bereich des maschinellen Lernens angewendet.

Im Gegensatz zu dem obigen Ansatz, der allgemein auf Nicht-Standard Daten anwendbar ist, verfolgt der zweite Ansatz einen fokussierteren Zugang: Dieser konzentriert sich auf gezielt eine spezifische Art von Nicht-Standard Daten, nämlich auf Daten mit gemischter Messskala. Solche Daten treten beispielsweise im Bereich des Benchmarkings durch die gleichzeitige Betrachtung mehrer Gütemaße und Datensätzen auf. Die entwickelten statistischen Methoden basieren auf Präferenzsystemen und der generalisierten stochastischen Dominanz und stellen für Zufallsvariablen, die in Räume mit gemischten Messskalen abbilden, (regularisierte) statistische Tests auf deren stochastischer Ordnung vor.

Summary

This cumulative dissertation develops statistical methods for so-called non-standard data. Non-standard data, like mixed categorical-numerical-spatial data and partial order-valued data, comprises data structures that cannot be represented by conventional statistical data formats. However, classical statistical methods often (implicitly) assume a conventional data structure. Hence, applying these methods to non-standard data implies distorting the inherent structure of the data, which leads to possibly invalid analyses. As a result, applicants face the dilemma that, on the one hand, using standard statistical methods risks misrepresenting and therefore misinterpretation of the data, while, on the other hand, preserving their true structure often makes these methods unusable. This dissertation contributes to overcome this dilemma by introducing flexible and non-parametric methods that are applicable across a large variety of data types.

The dissertation tackles the above dilemma through two approaches: The first approach combines the concept of data depth functions and formal concept analysis. Instead of focusing on the data values directly, as common in statistics, formal concept analysis is concerned with the relationship between the data elements. Depth functions quantify the centrality or outlyingness of data points with respect to a data cloud or an underlying distribution and are typically defined for normed vector spaces. This dissertation generalizes depth functions to non-standard data by building on the relational structure provided by formal concept analysis instead of the data values directly. As a first step, structural properties are introduced to formalize the notion of centrality, outlyingness and depth for non-standard data. These structural properties adapt existing ideas on defining depth functions from normed vector spaces to non-standard data and incorporate the relational structure of the data represented by formal concept analysis. Building on this notion, this dissertation introduces two depth functions for non-standard data: the ufg-depth and the generalized Tukey-depth. Afterwards, this general framework for non-standard data is discussed for the special case of partial order-valued data, where the underlying ground space is the set of all partial orders on a fixed set of items. This concrete case of non-standard data allows to analyze the depth functions under additional aspects, such as further properties and improvement of computation. Besides the theoretical evaluation, the ufg-depth for partial order-valued data is applied to benchmarking problems in the field of machine learning.

In contrast to the first approach, which is concerned with all types of non-standard data, the second approach takes a more focused perspective and considers the concrete case of data with mixed scales of measurement. Such data naturally occur in contexts like benchmarking, where multiple performance measures and multiple data sets are used simultaneously. For these data, the dissertation introduces a (regularized) statistical test on the stochastic order of random variables in space with mixed scale of measurements, which build on preference systems and generalized stochastic dominance.

Contents

List of Figures

List of Tables

Contributions of the Thesis

This dissertation is composed of the following thematically sorted contributions, Throughout the rest of this dissertation, they are referred to as Contribution 1 to Contribution 7.

1. Hannah Blocher and Georg Schollmeyer (2025). "Data Depth Functions for Non-standard Data by use of Formal Concept Analysis". In: *Journal of Multivariate Analysis* 205, 105372

2. Hannah Blocher and Georg Schollmeyer (2024). *Union-free Generic Depth for Non-standard Data*. ArXiv:2412.14745. URL: https://arxiv.org/abs/2412.14745. (last accessed: 02.03.2025)

3. Hannah Blocher, Georg Schollmeyer, and Christoph Jansen (2022). "Statistical Models for Partial Orders based on Data Depth and Formal Concept Analysis". In: *Information Processing and Management of Uncertainty in Knowledge-based Systems*. Ed. by Davide Ciucci, Inés Couso, Jesús Medina, Dominik Ślęzak, Davide Petturiti, Bernadette Bouchon-Meunier, and Ronald Yager. Cham: Springer, 17–30

4. Hannah Blocher, Georg Schollmeyer, Malte Nalenz, and Christoph Jansen (2024). "Comparing Machine Learning Algorithms by Union-free Generic Depth". In: *International Journal of Approximate Reasoning* 169, 109166. (Invited Paper for the ISIPTA 2023 Special Issue)

5. Julian Rodemann and Hannah Blocher (2024). "Partial Rankings of Optimizers". In: *The Second Tiny Papers Track at ICLR 2024, Tiny Papers @ ICLR 2024*. Ed. by Tom Burns and Krystal Maughan. Vienna: OpenReview.net

6. Christoph Jansen, Georg Schollmeyer, Hannah Blocher, Julian Rodemann, and Thomas Augustin (2023). "Robust Statistical Comparison of Random Variables with Locally Varying Scale of Measurement". In: *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*. Ed. by Robin Evans and Ilya Shpitser. Pittsburgh: PMLR, 941–952

7. Christoph Jansen, Georg Schollmeyer, Julian Rodemann, Hannah Blocher, and Thomas Augustin (2024). "Statistical Multicriteria Benchmarking via the GSD-Front". In: *38th Conference on Neural Information Processing System (NeurIPS*

2024). Ed. by Amir Globerson, Lester Mackey, Angela Fan, Ulrich Paquet, Jakub Tomczak, Cheng Zhang, and Lam Nguyen. Vancouver: OpenReview.net

Short Summary of Each Contribution

Contribution 1: This contribution develops depth functions based on formal concept analysis. We provide a general notion of centrality and outlyingness for non-standard data by introducing 13 structural properties. Furthermore, we introduce a generalization of the Tukey depth, see Tukey 1975, and analyze it based on the introduced structural properties. Thus, this contribution provides statisticians with concepts to discuss centrality and depth for more types of data than are currently covered.

Contribution 2: Building on Contribution 1, we introduce a second concrete depth function: The union-free generic depth function (ufg-depth for short). This depth function is a generalization of the well-known simplicial depth on $\mathbb{R}^d$, see Liu 1990. We analyze it using the structural properties defined above in Contribution 1 and under further computational aspects. In addition, we provide two applications of the ufg-depth to specific non-standard data types: mixed categorical-numerical-spatial data and hierarchical-nominal data. All in all, this contribution develops and analyzes a further depth function for non-standard data. The corresponding R-code allows users to easily apply the ufg-depth on their data.

Contribution 3: This contribution provides a formal context representing partial order-valued data. By applying this formal context to a modified version of the generalized Tukey depth of Contribution 1, we introduce a specific depth function on partial order-valued data. Moreover, this contribution discusses the notion of unimodality by using the structural property of quasiconcavity of Contribution 1. We use all these ideas to define a unimodal location-scale model for partial order-valued data. Additionally, we provide a sampling algorithm for partial orders.

Contribution 4: The ufg-depth of Contribution 2 is applied to the formal context for partial order-valued data introduced in Contribution 3. Using this further structure, this contribution analyzes the ufg-depth in terms of how the partial order structure of the data is represented by this depth function. Furthermore, we provide a detailed application of the ufg-depth in the field of comparing (machine learning) algorithms based on different data sets and multiple performance measures. We apply the ufg-depth to the partial order-valued data given by the OpenML benchmark suite, see Vanschoren et al. 2013,

and the UCI suite, see Kelly et al. 2017. In addition, we point out the differences to other benchmark approaches and provide a detailed application framework for users.

Contribution 5: This contribution is a natural continuation of Contribution 4. It applies the ideas of Contribution 4 to benchmark suites concerning optimizers, see Schneider et al. 2019; Wu et al. 2023; Hansen, Auger, Ros, et al. 2010. The focus here is on optimization problems that are analyzed based on multiple performance measures. Unlike Contribution 4, Contribution 5 addresses not only the performance order of the optimizers, but also the test functions that produce the optimizer orders.

Contribution 6: We propose a statistical test that builds on preference systems and generalized stochastic dominance order, developed in Jansen, Schollmeyer, and Augustin 2018. Furthermore, we introduce a robust and regularized version of the test statistic. Thus, this contribution provides statistical inference methods to analyze multidimensional data with differently scaled dimensions. By providing an efficient algorithm, we enable the demonstration of the method on a poverty, finance and medicine analysis.

Contribution 7: Building on Contribution 6, this contribution presents two statistical tests for analyzing the dominance structure of algorithms based on multiple data sets and multiple, differently scaled performance measures. These tests are also presented in a regularized and robustified version. Utilizing the algorithm developed in Contribution 6, we apply this method to two benchmark suites: the OpenML suite and the PMLB suite, see Vanschoren et al. 2013; Olson et al. 2017.

Declaration of the Author's Specific Contributions

Fruitful collaboration with several co-authors resulted in all of the articles presented in this dissertation. By referring to each article separately, the contributions of each author are clarified below. Since all these collaborations were very close, it is necessary to go into detail in order to describe the contributions. This should not obscure the fact that each part was intensively analyzed by all authors.

Contribution 1: Hannah Blocher had the management and coordination responsibility and wrote most of the article. She was responsible for the structure of the article. She contributed the main part of embedding the properties into already existing discussions about data depth functions and presenting mathematical relations of the properties. The analysis of the generalized Tukey depth was mainly done by her. A very intensive and close collaboration with Georg Schollmeyer led to the writing of this article. In particular, this includes the definition of the depth functions as well as the properties and theorems. It is therefore necessary to be more precise to divide the contributions.

Development of the Depth Functions and Properties: The general definition of the (empirical) depth function stems from Hannah Blocher, the (empirical) generalized Tukey depth from Georg Schollmeyer. The structural properties have been developed by both authors. To be more precise: Hannah Blocher had the idea for Properties (P2), (P3), (P4), (P5), (P9), (P10) and (P11), whereas Georg Schollmeyer had the idea for Properties (P6), (P7), (P8), (P12) and (P13). Property (P1) stems from both authors.

Statement and Proof of the Theorems: Hannah Blocher stated Theorems 2, 3 (Part 4), 4, 7 and 8. She also contributed in the ideas behind the running texts except for Sections 5.4, 6.4, Example 4 of Section 3, and Remark 7 and 8. The idea for the proof for Theorems 2, 3, 4, 6, 7 and 8 also stems from Hannah Blocher. Hannah Blocher realized the proofs of Theorem 2, 3, 4, 5, 6, 7 and 8. Georg Schollmeyer stated Theorems 1, 5, 6, 9, 10 and for Theorems 1, 5, 9 and 10 Georg Schollmeyer had the idea of the proof. Furthermore, Georg Schollmeyer realized Theorems 1, 5, 9 and 10. The ideas for Sections 5.4, 6.4, Example 4 of Section 3, and Remark 7 and 8 are from Georg Schollmeyer. Part 3 of Theorem 3 was stated and proven by both authors.

Analysis of the generalized Tukey depth: The analysis of the generalized Tukey depth was mainly done by Hannah Blocher, except for the above mentioned theorem statement (Theorem 6), remarks (Remarks 7 and 8) and Subsection 6.4, which were done by Georg Schollmeyer.

Writing: Georg Schollmeyer wrote Section 5.4 and 6.4, Example 4 of Section 3, the explanation to the generalized Tukey depth in Section 4 and Theorem 1, including the proof. Hannah Blocher wrote Sections 1, 2, 3 (except for Example 4), 4 (except for the explanation of the generalized Tukey depth), 5 (except for Subsection 5.4), 6 (except for Subsection 6.4) and 7. All graphics were done by Hannah Blocher.

Review: Both authors contributed to the review of the manuscript and responded to the editor and reviewers. Hannah Blocher took the lead.

Contribution 2: Hannah Blocher led the conceptualization, methodology development, original draft writing, supervision, and project administration for this contribution. She developed the ufg-depth and analyzed it based on the structural properties outlined in Contribution 1, except for the consistency property (which was co-analyzed with Georg Schollmeyer) and the universality properties (analyzed by Georg Schollmeyer). Hannah Blocher also conducted the analysis and computation of the ufg-depth for the gorilla nesting site data, while Georg Schollmeyer performed these tasks for the GGSS data. Hannah Blocher wrote the majority of the contribution, with the exception of Sections 4.4 and 5.2, which were written by Georg Schollmeyer. It is important to note that every part of this contribution was the result of extensive collaboration and discussion between both authors.

Contribution 3: Hannah Blocher was responsible for the management and coordination of this article. Conceptualization and methodology were mainly done by Hannah Blocher and Georg Schollmeyer. All ideas were discussed by all authors, as was the writing.

To be more precise: Sections 1, 2 and 7 were written jointly by all three authors, where Christoph Jansen contributed most parts of Sections 1 and 7. Hannah Blocher wrote the majority of Sections 3, 4, and 6. She stated and proved Lemma 1 and developed the algorithm for sampling a partial order based on this lemma. The properties from Definition 1 were developed and discussed by all authors. The problems of classical methods described in Section 2 were extensively discussed by all authors and the solution via the concrete scaling method from Section 4 was developed by Hannah Blocher. Georg Schollmeyer wrote most parts of Section 5. He developed the generalizations of the Tukeys depth, the peeling depth and the enclosing depth, as well as the proposal for the weighting within the modified Tukeys depth. All authors contributed to the revision of the article in several discussion rounds.

Contribution 4: Hannah Blocher developed the idea of the ufg-depth. She stated and proved most of the theorems. Major parts of the development of the algorithms and their implementation were carried out by her. She wrote most of the article. Georg Schollmeyer, Christoph Jansen and Hannah Blocher discussed the definition of ufg-depth and all its properties in detail. All authors supported the analysis with intensive discussions and contributed with detailed proofreading and help with the general structure of the article. Hannah Blocher, Christoph Jansen, and Georg Schollmeyer were involved in revising the manuscript and responding to the reviewers.

To be more precise: Hannah Blocher wrote all parts except for Section 1.1 (written by Christoph Jansen), Section 1.2 (written by Georg Schollmeyer), Section 2 (written by Christoph Jansen), the comparison to Jansen, Nalenz, et al. 2023 in Section 7.1.2 (written by Christoph Jansen and Georg Schollmeyer), Section 7.2.1 (written by Malte Nalenz and Georg Schollmeyer) and the conclusion (written by Christoph Jansen and Georg Schollmeyer). Furthermore, Hannah Blocher claimed and proved Lemma 1, Corollary 2, Theorem 3 (1st and 2nd part), Theorem 4 (lower bound), Lemma 6, Corollary 8 and 9 and Theorem 10. Georg Schollmeyer made the claim and proof of the upper bound of Theorem 4. The claim about the connectedness in Theorem 3 (Part 3) was done by Georg Schollmeyer. Hannah Blocher and Georg Schollmeyer provided the proof of Theorem 3. Georg Schollmeyer provided the claim and proof of Theorem 5. The claim of Theorem 7 was done by Georg Schollmeyer and Christoph Jansen. Georg Schollmeyer proved Theorem 7. Hannah Blocher implemented the test whether a subset is an element of $\mathcal{S}$, Page 5 of the contribution. Georg Schollmeyer contributed with the implementation of the connectedness property. Malte Nalenz provided the data set and performed the data preparation.

Contribution 5: Hannah Blocher and Julian Rodemann contributed equally to this contribution. Using the CRediT roles for authorship as reference: Supervision, writing, data curation, investigation, formal analysis, validation, and review were contributed equally by both orders. Project administration was mainly done by Julian Rodemann. Both authors contributed to revising the article. Their individual contributions are described in more detail below.

The idea of applying the union-free generic depth for partial order-valued data to partial orders describing the performance of optimizers evaluated on a benchmark suite comes from Julian Rodemann (Concept). All three benchmark suites considered were proposed by Julian Rodemann (Resources). Hannah Blocher introduced the ufg-depth for partial order-valued data in Contribution 3, where she wrote and ran the R code for this contribution (Methodology, Software). Julian Rodemann wrote the first draft of the abstract and the introduction. The part on depth functions was added by Hannah Blocher. Together, the authors stated the contribution of the article at the end of the introduction. Hannah Blocher wrote the method section. In the results section, the embedding of the result into the theory of optimization and the characteristics of the benchmark suite DeepOBS was done by Julian Rodemann. Hannah Blocher analyzed the results and the general meaning of the depth values. Both authors together wrote

the outlook/result. Appendices A, B, and C were written by Hannah Blocher. Appendix D was written by Julian Rodemann. The first draft and idea for Appendices E and F was written by Julian Rodemann. The final version of Appendices E and F underwent several changes made by both authors. Note that each of these parts is based on intensive discussion between the two authors, with both authors revising each part several times.

Contribution 6: The idea and most parts of the original draft of the article come from Christoph Jansen. Georg Schollmeyer supplied parts of Proposition 7 and Proposition 8, including parts of the idea for their proof. He also discussed and drafted parts of the aspects related to regularization. Hannah Blocher wrote Section 8.1 and Appendix B, and developed the algorithm that allows the tests to be computed on such a large scale. The R code for performing the permutation-based tests in all three applications as well as the GitHub repository are also due to Hannah Blocher. Julian Rodemann wrote parts of the introduction to Section 6, most parts of Section 8.2 as well as Appendix D. Furthermore, he designed all plots in the article. The idea for Figure 3 in the main article as well as Figures 2 and 4 in the appendix was jointly developed by Christoph Jansen and Julian Rodemann. Julian Rodemann proposed the applications on finance and medicine. Thomas Augustin wrote some parts on related literature and parts of the introduction to Section 7. All authors contributed to revising the article in several discussion rounds.

Contribution 7: The idea and most parts of the original draft of the article come from Christoph Jansen. Section 5 was jointly written by Julian Rodemann, Christoph Jansen, and Hannah Blocher (in this order). The sufficient condition of a finite VC-dimension and the proof of Theorem 3.6 are mostly due to Georg Schollmeyer. Corollary 5.1 was jointly delivered by Georg Schollmeyer and Christoph Jansen. The idea of the dynamic versions of the tests in Section 4 were jointly developed by Christoph Jansen and Georg Schollmeyer. Building on the algorithm developed by Hannah Blocher in Contribution 6, the R code for performing the permutation-based tests in both applications as well as the GitHub repository are due to Hannah Blocher. She also had the idea of using computing time as ordinal performance measure. The analysis of the OpenML result in terms of correctness was also carried out by her. The R code for enabling the benchmark analysis of the PMLB experiments, in particular the hyperparameter tuning and integration of compressed rule ensemble learning (CRE), is due to Julian Rodemann and Georg Schollmeyer. Georg Schollmeyer contributed the idea of using feature- and class robustness as ordinal performance measures. Figures 2, 3, 4, 6, 7 and 8 were created by Julian Rodemann, whereas Figures 1 and 5 were created by Christoph Jansen. Thomas Augustin has contributed with discussions and reflections on the ideas and detailed proofreading. All authors contributed to revising the article.

Chapter 1

Motivation and Introduction: Why are Further Methods Needed?

This cumulative dissertation is concerned with *non-standard data* which cannot be naturally represented by conventional statistical data types. Such data occur naturally across numerous fields, for example occupation data can be nominal data with an inherited hierarchical order structure or preference rankings may contain incomparable parts. Such structure complexities place non-standard data beyond the scope of classical statistical frameworks, see Section 2.1 for details. To ensure useful results and interpretations, statistical methods need to take these unique characteristics into account.

In this first section, I present the current methodological gap in statistical methods for non-standard data. Based on this, I outline the idea behind the presented approaches and state the aim of this dissertation.

1.1 Motivation: Challenges in Statistics for Non-standard Data

Many statistical methods are built on the assumption that the underlying data conforms to specific structures, such as being nominal, ordinal, interval, ratio-scaled, or existing within a normed vector space. These assumptions are a foundation of both classical approaches, see, e.g. Fahrmeir et al. 2016, and modern machine learning methods, e.g. support vector machines or neural networks, see e.g. Mohri et al. 2018; Adcock et al. 2022; Chen 2024. However, real-world data often fail to match these assumptions. That leaves practitioners with a difficult choice: either they impose assumptions that may distort results and interpretations, or abstain from statistical analysis entirely.

This dilemma particularly occurs for non-standard data. Current approaches analyzing non-standard data often naively assume that the data can be represented by conventional data structures and with this introduce the risk to distort and invalidate the statistical interpretations, as shown in Behzadi et al. 2020; Dutcă et al. 2018; Julian 2001. Furthermore, as noted in Blocher, Schollmeyer, and Jansen 2022, determining

which assumptions to impose can itself be a complex and ambiguous process. This lack of statistical methods for analyzing non-standard data creates a crucial gap in methods that provide meaningful statistical analysis without compromising the true nature of their data.

1.2 Aim of this Dissertation and Outline of the Approaches

This dissertation addresses the above gap by developing statistical methods focusing on non-standard data. These methods preserve the natural, pre-given structure of the data and avoid to (implicitly) add assumptions that could distort results and interpretations. The dissertation pursues this goal through two approaches. The first approach, called *gsd approach* from now on, defines methods specifically for one kind of non-standard data types. The second *fca-depth approach*, which forms the core of this work, focuses on defining flexible and robust methods applicable on a wide variety of non-standard data.

The foundation of the fca-depth approach is to combine *depth functions* and *formal concept analysis* (*fca* for short), see Section 3.1. Formal concept analysis provides a systematic tool to transform diverse data types into a unified representation by grouping data points based on their relationships. This representation focuses on the relational structure between the data points instead of the single data values. Building on the representation via formal concept analysis, a notion of depth functions as a measure of centrality or outlyingness for non-standard data is introduced by developing structural properties. With this, the well-established concept of depth from $\mathbb{R}^d$ is extended to non-standard data. Having defined the notion of depth, the next question to be addressed is that of concrete mapping rules of a depth function. This work provides two concrete depth functions: the generalized Tukey depth and the union-free generic depth (a generalization of the simplicial depth in $\mathbb{R}^d$). They are analyzed in terms of their structural properties to represent centrality. This dissertation thus enables the development of non-parametric statistical methods which are not defined specifically for one particular data type, but for many different types simultaneously.

In addition to this general definition of depth, the dissertation discusses the two proposed depth functions for concrete non-standard data. With three separate contributions, see Section 3.2, the focus is on partial order-valued data where each single observation is an entire partial order that is regarded as an holistic entity and not as a collection of pairwise observations. In addition, mixed spatial-categorical-numerical data and hierarchical-nominal data are covered. First, a grouping system based on formal concept analysis is developed for each data type and the underlying quantities of the representation are analyzed. Afterwards, the two depth functions are applied to real-world data and the resulting center-outward order of the data elements is examined.

The dissertation also explores the gsd approach, focusing on multidimensional data with mixed scales of measurement. This smaller component of the dissertation proposes robust statistical tests for ordering random variables that are not purely numeric or ordinal. Therefore, this work uses *preference systems* and *generalized stochastic dominance* (*gsd* for short) to reflect more complex data types, presented and further developed by

Christoph Jansen, see Jansen 2018; Jansen 2025.

In conclusion, this dissertation provides applicants and researchers with novel methods for analyzing the wide spectrum of non-standard data. By maintaining the data intrinsic structure, the methods developed ensure valid statistical analysis and interpretation across a broad range of real-world applications.

Chapter 2

Methodological Background and Related Literature

Before presenting each contribution in detail, the following section provides an overview of the background topics that are essential for this dissertation. The relevant material is summarized without proof, referring to further literature.

2.1 Non-standard Data

Stevens 1946 introduced four different levels of measurement: nominal, ordinal, interval and ratio.[1] Later, higher-dimensional data have often been analyzed under the assumption that their structure aligns with a normed vector space. To some extent, these levels of measurement are still considered to be exclusive to this day. As a result, all data types are embedded into this already existing framework and therefore many classical statistical approaches, as summarized in works like Fahrmeir et al. 2016, are built upon these assumptions. Even newer methods frequently adhere to those, see e.g. Mohri et al. 2018; Adcock et al. 2022; Chen 2024. However, a substantial variety of data structures fall outside these categories and occur naturally across numerous fields.

To address this, my dissertation introduces the term *non-standard data* to encompass data types that do not conform to the classical levels of measurement or vector space structures. This terminology emphasizes the diversity of these data types and their departure from traditional statistical assumptions. Below, I outline three specific types of non-standard data addressed in this dissertation and refer to relevant literature to contextualize them.

Partial Order-valued Data: A *partial order* P on a set Ω is a subset of $\Omega \times \Omega$ that is reflexive (i.e. for all $\omega \in \Omega$ $(\omega, \omega) \in P$ is true), transitive (i.e. if $(\omega_1, \omega_2), (\omega_2, \omega_3) \in P$ then

[1] Note that further terms are used in the literature: categorical is an umbrella term for nominal and ordinal; and cardinal/numeric is an umbrella term for ratio and interval. In what follows, we use the terms as they are used in the research area under discussion (i.e., in Contributions 6 and 7, we use the term cardinal).

$(\omega_1, \omega_3) \in P$), and anti-symmetric (i.e. if $(\omega_1, \omega_2) \in P$ and $(\omega_2, \omega_1) \in P$ then $\omega_2 = \omega_1$). In contrast to *total orders*, partial orders allow for elements to be incomparable. An order that is only transitive and reflexive is called *preorder*.
This dissertation focuses on *partial order-valued data*, which posses two main characteristics. First, the ground space is the set of partial orders. This means that each observation represents one entire partial order that must be treated as a unified, holistic entity rather than merely as a collection of independent pairwise comparisons. Second, the partial nature of these orders, i.e. two items x_1 and x_2 are incomparable, is considered to be a valid and meaningful observation, not simply an indication of missing information of an observation that is interpreted as a total order.
This perspective is contrary to most existing methods dealing with partial order-valued data, see the discussion in Blocher, Schollmeyer, and Jansen 2022. For example, Lebanon and Mao 2007 assume that any observed incomparabilities arise from a missing data mechanism rather than reflecting the true nature of the data. Many statistical models for partial order-valued data are derived from models originally developed for *total order-valued data*. Hence, these models (implicitly) assume that the partial orders represent total orders, as can be seen by the adapted Mallows model to top-k orders, see Mallows 1957; Chierichetti et al. 2018, or in the generalized Bradley-Terry model, see Bradley and Terry 1952; Davidson and Beaver 1977. Moreover, often approaches, such as the method presented by Dittrich et al. 1998, analyze partial orders solely through independent pairwise comparisons, thereby neglecting the essential holistic structure of the partial order.

Hierarchical-nominal Data: *Hierarchical-nominal data* possess an intrinsic structure where categories are organized into multiple levels of hierarchy. For example, occupations, as categorized in the GGSS data set, see GESIS 2018, are sorted into four hierarchical levels. Hence, elements sharing categories up to the third level are more closely related than those differing at lower levels.
Existing analyses often overlook the hierarchy, as shown by Dutcă et al. 2018; Julian 2001 in studies of biomass allometric models, leading to errors such as overconfident predictions. In this dissertation, formal concept analysis is employed as a framework that respects and utilizes the hierarchical structure to ensure valid statistical modeling.

Mixed-typed Data: *Mixed-typed data* arise as Cartesian products of sets with differing data structures. For example, in benchmark studies, performance measures may include both numeric and ordinal-scaled variables. A similar example is mixed spatial-numeric-nominal data that frequently appear in spatial statistics.
Representing such data sets naively as interval-valued or numeric can introduce interpretation problems, as discussed by Behzadi et al. 2020. In this dissertation, mixed-typed data are analyzed using formal concept analysis, see Section 2.2, or preference systems, see Section 2.5, to preserve their inherited structure.

2.2 Formal Concept Analysis

Formal concept analysis (*fca* for short) explores relationship structures within data sets. Instead of considering the single data values themself, formal concept analysis reveals the connection among the data elements. This is achieved by defining formal concepts that consist of data elements sharing a set of attributes, and provide us an order of these formal concepts. By focusing on the relational structure, formal concept analysis offers the possibility to represent a data set in different ways and it provides users to obtain a clear understanding of the relation between the data and, in particular, of the underlying assumptions. Moreover, it connects real-world application problems with the lattice theory,[2] building on a well-studied mathematical framework for analyzing the practical problems.

Due to its data representation and its connection to the rich lattice theory, formal concept analysis has a wide range of topics for which it is used. One example is Kotelnikov and Milov 2018 where classification problems have been addressed. Furthermore, a large number of papers focus on the topics of data mining or knowledge discovery in databases in combination with formal concept analysis, see Poelmans et al. 2010 for a summary until 2010. Another example is the discussion on online news discourses in Horn et al. 2024; Draude et al. 2024. Moreover, questions within the area of machine learning are discussed with the help of formal concept analysis. Khatri et al. 2024 considers interpretability and Dudyrev et al. 2023 investigates ensemble learning via decision trees. Dürrschnabel and Priss 2024 utilize formal concept analysis to discuss the existence of regular Euler diagrams. Additionally, the theory has been extended to fuzzy formal concept analysis, for example, see Brito et al. 2018.

The basis of formal concept analysis theory is the *formal context* (G, M, I), where G is a set of *objects*, M is a set of *attributes*, and I is a *binary relation* between G and M. We express $(g, m) \in I$ by *the object g has the attribute m*, see Ganter and Wille 2012, p. 17. In this way, any object can be classified as either having an attribute or not having it. This is a formalization of a cross table. As an example, we consider the seven contributions of my dissertation as a set of objects, see Table 2.1. As attributes we choose topics like 'data depth (dd)' and define an object to have the attribute if the contribution covers that topic.

The attributes describing my dissertation studies are binary, in the sense that an object can either have or not have it. In general, we have many-valued attributes rather than just binary ones. By using *conceptual scaling* where each many-valued attribute is represented by a set of binary attributes, we again get a formal context. For example, when observing numerical data points, we use the attributes '$\geq r$' and '$\leq r$' for each $r \in \mathbb{R}$. This conceptual scaling is called *interordinal scaling*, see Ganter and Wille 2012, p. 42. Note that the interordinal scaling only takes the order into account and not the metric distance information. This is because two attributes that have exactly the same objects are redundant in the sense that one of these attributes is sufficient to describe

[2] Note that I am referring to order theory here and this should not be confused with lattices within group theory.

	dd	*fca*	*spatial data*	*hierarchical data*	*poset*	*sd*	*ml*
Contr. 1	×	×					
Contr. 2	×	×	×	×			
Contr. 3	×	×			×		
Contr. 4	×	×			×		×
Contr. 5	×	×			×		×
Contr. 6					×	×	×
Contr. 7					×	×	×

Table 2.1: Formal context representing the seven contributions of this dissertation. dd is data depth function, fca is formal concept analysis, poset is partial order, sd is stochastic dominance and ml is machine learning for short.

the relation between the objects and the other one can be dropped. On the one hand, some information is lost with the interordinal scaling method, but on the other hand, it is robust at different scales. This underlines the user-friendliness of formal concept analysis, since the observation of scale invariance of the interordinal scaling is directly apparent and not indirectly assumed by embedding the data set into another space.

Using the above formalization of a cross table, the following *derivation operators* describe the relation between the attribute and the object set, see Ganter and Wille 2012, p. 18:

$$\Psi : 2^G \to 2^M, A \to A' := \{m \in M \mid \forall g \in A\colon gIm\}, \tag{2.1}$$

$$\Phi : 2^M \to 2^G, B \to B' := \{g \in G \mid \forall m \in B\colon gIm\}. \tag{2.2}$$

Applying these derivation operators on the formal context given in Table 2.1, we obtain for example

$$\Psi(\{\textit{Contr. 2}\}) = \{\textit{dd, fca, spatial data, hierarchical data}\}, \tag{2.3}$$

$$\Phi(\{\textit{ml}\}) = \{\textit{Contr. 4 to Contr. 7}\}. \tag{2.4}$$

Thus, Ψ maps a set of objects to each attribute that each object has in the input. Φ works similar but has an attribute set as input.

Using these two derivation operators, one can define a *formal concept* (A, B) consisting of a so-called *extent* $A \subseteq G$ and *intent* $B \subseteq M$, where $\Psi(A) = B$ and $\Phi(B) = A$ hold. The set of all formal concepts and the formal context have a one-to-one correspondence. Furthermore, the set of all formal concepts is a complete lattice which now connects to lattice theory. A *complete lattice* is a partial order $(V, \leq)$ in which each subset $v \subseteq V$ has an infimum and a supremum. Now, the set of formal concepts is a complete lattice with the *subconcept-superconcept-relation*, see Ganter and Wille 2012, p. 20

$$\big((A_1, B_1) \leq (A_2, B_2) :\Longleftrightarrow A_1 \subseteq A_2 \big) \quad \Leftrightarrow B_1 \supseteq B_2. \tag{2.5}$$

Going back to the formal context representing my dissertation studies, see Table 2.1, the Hasse diagram in Figure 2.1 states the complete lattice of all formal concepts. There

we have, for example,

$$(A_1, B_1) = (\{\textit{Contr. 2}\}, \{\textit{dd, fca, hierarchical data, spatial data}\}) \tag{2.6}$$
$$(A_2, B_2) = (\{\textit{Contr. 1 to Contr. 5}\}, \{\textit{dd, fca}\}) \tag{2.7}$$

which are two formal concepts with $(A_1, B_1) \leq (A_2, B_2)$.

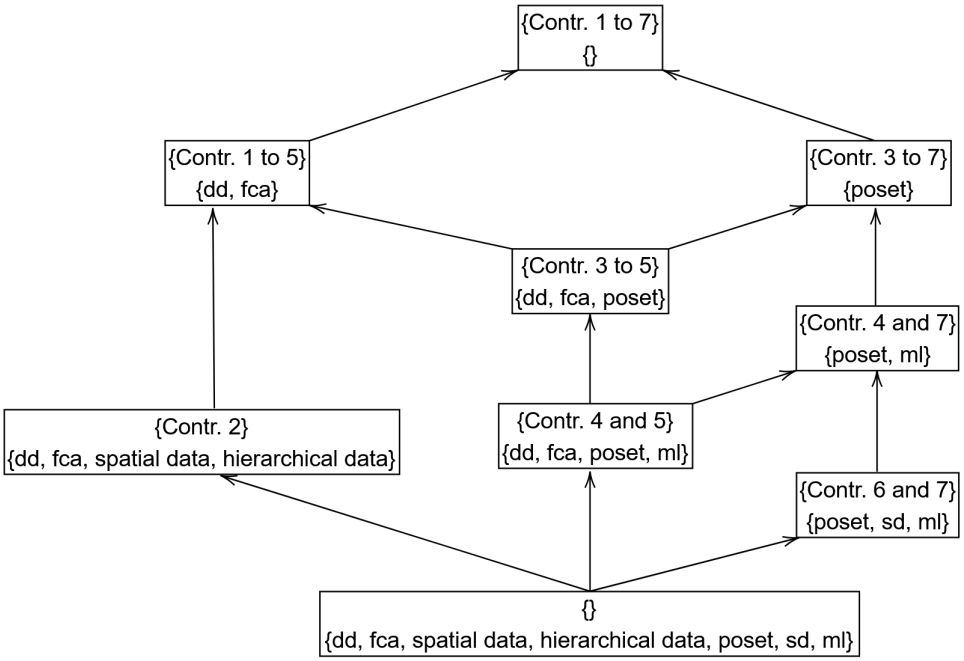

Figure 2.1: The Hasse diagram represents my dissertation studies and corresponds to the formal context given by Table 2.1.

Another way to represent the formal context is to consider only the set of extents. This gives us a *closure system* together with the corresponding *closure operator* $\Phi \circ \Psi$, see Ganter and Wille 2012, p. 8 and 18. A closure system is a family of sets which contains the entire space (here: G) and is closed under arbitrary intersections. As an example, one can think of the convex sets on $\mathbb{R}^d$ as closure system and the corresponding convex closure operator maps each subset onto the smallest convex set containing the subset.

We can even go a step further and use the fact that any closure system can be uniquely described by a *family of implications*, see Ganter and Wille 2012, p. 79. We say that $A \subseteq G$ implies $B \subseteq G$ ($A \to B$ for short) if and only if $(\Phi \circ \Psi)(A) \supseteq (\Phi \circ \Psi)(B)$. Again, one can think of the set of convex sets on $\mathbb{R}^d$. In this case, the edge points of a triangle imply the entire triangle itself given the convex closure operator. Considering the formal context in Table 2.1, 'Contr 1'-object implies the 'Contr. 2' to 'Contr. 5'-objects.

Some of these implications are redundant, in the sense that they follow semantically from others. For example, the information of implication {'Contr. 1', 'Contr. 2'} → {'Contr. 1', 'Contr. 2', 'Contr. 3'} is already given by implication {'Contr. 1'} → {'Contr. 1' to 'Contr. 5'}. All these semantic redundancies, called inference axioms, are summarized in Armstrong 1974; Maier 1983: For subsets $A, B, C, D, A_1, A_2, B_1, B_2 \subseteq G$, the *reflexivity* states that by default the trivial implications hold (i.e. $A \to A$ is true). The axiom of *augmentation* considers the union of the premise (i.e. $A_1 \to B$ implies $A_1 \cup A_2 \to B$) and the axiom of *additivity* the union of the conclusion (i.e. $A \to B_1$ and $A \to B_2$ imply $A \to B_1 \cup B_2$). The axiom of *projectivity* holds iff $A \to B_1 \cup B_2$ implies $A \to B_1$. The last two axioms discuss the transitivity of implications. The axiom of *transitivity* denotes the classical version (i.e. $A \to B$ and $B \to C$ imply $A \to C$) and the second, the axiom of *pseudotransitivity*, is slightly different (i.e. $A \to B$ and $B \cup C \to D$ imply $A \cup C \to D$).

In summary, we can now describe the data set by a formal context, a lattice, a closure system, and a family of implications. All of these representations have a (one-to-one) correspondence and give us further structure to define depth functions for a variety of non-standard data types.

Recall that in the beginning of this section we introduced formal concept analysis as applied lattice theory. The theoretical connection to lattice theory was described above. One reason for calling it "applied" is due to the representation and visualization of a data set as a complete lattice. By visualizing the data set in this way, we gain a transparent and clear understanding of the data set, especially when combined with the formal context that enforces appropriate attributes. This is due to the fact that an object can either have an attribute or not, which then leads to a good way to discuss and understand the data set. As an example, consider the lattice of my dissertation studies represented by the Hasse diagram in Figure 2.1. The connections between the single projects can be easily seen. For example, one can see that Contr. 1 only occurs in combination with Contr. 2 to 5. This shows that Contr. 1 discusses data depth and formal concept analysis in general and Contributions 2 to 5 are more specific and the thematically follow-up articles. Similarly is Contribution 3 the basis for Contribution 4 and 5. Moreover, the formal concept ($\{$*Contr. 3 to Contr. 7*$\}, \{$*poset*$\}$) shows that partial orders are the connection between the fca-depth approach (Contr. 1 to 5) and the gsd approach (Contr. 6 and 7). This is also reflected in my contributions to Contribution 6 and 7 as I supported these articles by introducing an algorithm based on partial orders.

2.3 Data Depth Function

Functions that measure centrality and outlyingness of multivariate data are called *(data) depth functions*. In $\mathbb{R}$, quantiles provide a natural way to define (data) depth functions with the median(s) being the center. When Tukey generalized quantiles to $\mathbb{R}^d$, the discussion of depth functions in $\mathbb{R}^d$ started. Since then, several depth functions have been proposed, such as the Tukeys depth, see Tukey 1975, the simplicial depth, see Liu 1990, the projection depth, see Zuo and Serfling 2000, or the zonoid depth, see Dyckerhoff

et al. 1996. This development was accompanied by a theoretical discussion on the general notion of centrality and outlyingness in $\mathbb{R}^d$ by developing desirable properties for depth functions in $\mathbb{R}^d$, see Zuo and Serfling 2000, p. 463f.

Over time, the scope of applications of depth functions grew as well. For example, Rousseeuw et al. 1999 used Tukeys depth function for outlier detection and Donoho and Gasko 1992 as a robust estimate of location. Since the depth functions give us an order, Li and Liu 2004; Dehghan and Faridrohani 2024 defined non-parametric tests based on depth functions. Multivariate scale and symmetry tests are defined by Dykerhoff 2002. Patil and Baidari 2019 used depth functions to estimate the number of clusters in a data set.

In general, a depth function is defined by

$$D : \mathbb{R}^d \times \mathscr{P} \to \mathbb{R}_{\geq 0} \tag{2.8}$$

with $\mathscr{P}$ being a set of probability distributions on $\mathbb{R}^d$. If a sample is drawn, we obtain the empirical depth function by using the empirical probability distribution $P_n \in \mathscr{P}$.

To illustrate the idea, we discuss the simplicial depth defined by Liu 1990. Let $\mathscr{P}$ be a set of probability measures on $\mathbb{R}^d$, then the simplicial depth $D(x, P)$ of a point $x \in \mathbb{R}^d$ is the probability $P \in \mathscr{P}$ that a random $d+1$ simplex[3] contains x. Thus, the simplicial depth is given by

$$D : \mathbb{R}^d \times \mathscr{P} \to [0, \infty[, (x, P) \mapsto P(x \in \Delta\{X_1, \ldots, X_{d+1}\}) \tag{2.9}$$

where $X_1, \ldots, X_{d+1} \overset{i.i.d.}{\sim} P$ and $\Delta\{x_1, \ldots, x_{d+1}\}$ is the $d+1$ simplex with vertices $x_1, \ldots, x_{d+1} \in \mathbb{R}^d$ (with i.i.d. being independent and identically distributed for short). The empirical simplicial depth function is obtained by using the empirical probability measures $\mathscr{P}_n$ instead of $\mathscr{P}$.[4]

By the year 2000, several different depth functions had been defined. Therefore, Zuo and Serfling 2000; Liu 1990 introduced desirable properties that a depth function can have to compare the existing depth functions. Based on these properties, they specified a general notion of depth functions and the meaning of centrality in $\mathbb{R}^d$. Further discussions on different depth functions and their properties are provided by, for example, Mosler and Mozharovskyi 2022. These properties relate to the definition of centrality and outlyingness as well as to other desirable structures like quasiconcavity. The following list provides a selection of these desirable properties given by Zuo and Serfling 2000, p. 463f, Liu 1990 and Mosler and Mozharovskyi 2022. Let $D(\cdot, \cdot)$ be a depth function based on Definition (2.8). Let $x \in \mathbb{R}^d$ and $P \in \mathscr{P}$. A depth function D should be

- *affine invariant* (the depth function is invariant under affine changes of the coordinate system). This means that for any non singular matrix $A \in \mathbb{R}^{d \times d}$ and $y \in \mathbb{R}^d$, $D(x, P) = D(Ax + y, P_{A\cdot+y})$, where $P_{A\cdot+y}$ is the transformed probability of P with $P_{A\cdot+y}(\Omega) = P(\{A \cdot x + y \in \mathbb{R} \mid x \in \Omega\})$ for Ω being measurable.

[3] A k simplex in $\mathbb{R}^d$ with $1 \leq k \leq d+1 < \infty$ consists of k affine independent points representing the edges and the smallest convex set containing those edges.

[4] Note that the more common empirical simplicial depth is based on an U-statistics and slightly differs from this definition. However, Nagy 2023 showed that the two expressions are asymptotically equivalent.

- *maximal at the center* (if the probability function has a unique center of symmetry, then the depth function has its maximum value at this center). This is true if $c = \arg\max_{z \in \mathbb{R}^d} D(z, P)$ for c being the center of the probability measure P.
- *quasiconcave* (the elements in $\mathbb{R}^d$ that have a depth value $\alpha > 0$ or higher form a convex set). For every $\alpha > 0$ the contour set $\{x \in \mathbb{R}^d \mid D(x, P) \geq \alpha\}$ is convex.
- *vanishes at infinity* (the depth function converges to zero if the norm of the point sequence converges to infinity), i.e., $\lim_{\|z\| \to \infty} D(z, P) = 0$.

Besides the focus on $\mathbb{R}^d$, the idea of depth functions has been adapted to further data types. Chakraborty and Chaudhuri 2014 discussed centrality and outlyingness on infinite dimensional spaces, Bolívar et al. 2023 for text data and Geenens et al. 2023 for abstract metric spaces. However, the limitation of these approaches is the assumption of a predefined data structure and therefore they are not applicable to the large variety of non-standard data which we tackle here by using formal concept analysis.

2.4 Related Literature on Combining Depth Functions and Formal Concept Analysis

In this section, we present the four publications that relate to the connection between depth functions and formal concept analysis. Cardin 2012 introduced a quantile approach based on complete lattices. A generalization of Tukey's depth function for complete lattices was introduced in Schollmeyer 2017a and related to formal concept analysis in Schollmeyer 2017b. Hu et al. 2023 defined a concrete outlier measure that is based on formal concept analysis. To my knowledge, these four publications are the only ones in which depth functions and complete lattices/formal concept analysis have been combined.

2.5 Stochastic Orders in Decision Theory

Decision theory is concerned with the problem of rational decision making. It is studied in numerous research fields such as philosophy, econometrics, psychology or statistics. In the following we focus on statistical decision theory, a branch that investigates decision making under uncertainty in the presence of statistical knowledge. In this dissertation, we utilize a stochastic order approach to identify those options which are in some sense optimal.

Modern decision theory has its beginnings with Ramsey 1931 and von Neumann and Morgenstern 1947 and is dominated by the aim to formalize and axiomatize decision making, see Peterson 2017, Chapter 1.5. The field gained further traction in statistical contexts with Savage's work, see Savage 1954, which remains a cornerstone of the discipline. Since then, solving statistical problems by embedding them into decision theory is a well established approach and has been further discussed by, e.g., Berger 1985 or Miescke and Liese 2008.

At its core, statistical decision theory builds on a basic setup for decision making, see, e.g., Peterson 2017: An *agent* has to decide between a set of known *acts* $\mathcal{X}$. Each act leads to different *consequences* $\mathcal{C}$ depending on which *state of the world* $S \in \mathcal{S}$ is actual. Formally, each act is a function $X : \mathcal{S} \to \mathcal{C}$ with $X \in \mathcal{X}$. The aim of the agent is now to find a set of acts $\mathcal{G} \subseteq \mathcal{X}$ that are optimal based on the agents preference of the consequences $\mathcal{C}$ and available information on the state of the world $\mathcal{S}$.

As an example, consider the decision of whether to continue reading this dissertation or not. In this scenario, the acts are "continue reading" or "stop reading" and the state of the world reflects the quality of the dissertation: low, medium, or high. The consequences of these acts depend on the quality of the dissertation. For instance, if the quality is high and the reader continues reading, he*she might enhance both knowledge and further curiosity on this topic. Conversely, if the quality is low and he*she continues reading, the benefit might be limited to being able to discuss the content with the author. If the reader stops reading, the consequence across all states of the world is the same: having time for other research activities. The decision problem is described by the decision matrix in Table 2.2. Each row represent one act, while each column represents one state of the world. The entries in the table represent the consequence when act X is chosen and state of the world S is true.

	quality is low	*quality is medium*	*quality is high*
continues reading	able to discuss with the author	enhances knowledge	enhances curiosity and knowledge
stops reading	time for other research	time for other research	time for other research

Table 2.2: Decision matrix describing the decision problem for the reader to continue reading this dissertation.

A first step to approach this decision problem is to formalize the preference of the agent on the possible consequences. For example, the reader might prefer gaining curiosity and knowledge over simply having more time for other tasks. These preferences can be formalized using *utility functions*, which assign numerical values to each consequence, making it easier to compare and evaluate them. To obtain a consistent decision problem we assume that the preference structure is at least a preorder and hence there exists a function $u : \mathcal{C} \to [0, 1]$[5] that preserve the order structure of the agent's preference. This means, if only ordinal information on the agent's preference is known, then $\mathcal{U}$ consist of all functions $u \in \mathcal{U}$ that have $u(C_1) \leq u(C_2)$ if and only if $C_1 \leq C_2$ is true. In the other case, where we have full cardinal information about the reader's preference, $\mathcal{U}$ equals one function u. In a second step, we assume that the reader has not read the dissertation already, so that the quality of the dissertation is unknown to the reader. Hence, he*she has to make a decision under uncertainty. However, maybe the reader is able to assign probabilities π to the state of the world. These probabilities represent his*her beliefs on

[5]These functions can be defined more generally with mapping on $\mathbb{R}$. However, we restrict ourselves here to $[0, 1]$ as this is the version we focus on in Contribution 6 and 7.

the quality of the dissertation based on the text he*she has already read.

Based on this, we use *stochastic dominance* to solve the decision problem, see Kamae et al. 1977; Lehmann 1955. We state that one act Y dominates another X if its expected utility $\mathbb{E}_\pi[u \circ Y]$ is at least as high as $\mathbb{E}_\pi[u \circ X]$ across all possible utility functions $u \in \mathcal{U}$ and strictly higher for at least one utility function. Hence, all acts that are not stochastic dominated by another act are called *optimal*. Note that, of course, for each $u \in \mathcal{U}$, $u \circ X$ and $u \circ Y$ need to be measurable and integrable.

In the case of the reading decision, we assume that the reader assigns equal probabilities to the states of the world and has the following total preference order of the consequences

"able to discuss with the author" $<$ "time for other research" $<$ "enhances knowledge" $<$ "enhances curiosity and knowledge".

Moreover, we assume that the reader wants to increase the distinction between the consequences. This can be achieved by selecting only functions $u : \mathcal{C} \to [0, 1]$ where the utility difference between two consequences is larger than a fixed value γ. Here, we assume $\gamma = 1/5$, i.e. for all $u \in \mathcal{U}$ and $C_1, C_2 \in \mathcal{C}$ with $C_1 \neq C_2$ we have $|u(C_1) - u(C_2)| > 1/5$. In this scenario we derive that the act "continues reading" stochastically dominates the act "stops reading". If the reader removes the restriction on the utility functions, then the two acts are incomparable as for some utility functions "continues reading" has higher expectation and for others "stops reading".

A more complex decision problem example is the selection of a machine learning algorithm for a classification task, see Contributions 7. Here, the decision involves choosing between different algorithms, each evaluated on several data sets using a selection of performance measures such as computation time and accuracy. These measures represent the consequence when choosing an algorithm based on a data set as actual state of the world. Since the measures often involve mixed scales, with some being ordinal and others cardinal, a more sophisticated approach than above is needed.

Therefore, we use Christoph Jansen's method for representing more complex decision problems introduced in his dissertation, see Jansen 2018, and extended in his habilitation, see Jansen 2025. The basis is the new concept of a *preference system* that incorporates both ordinal and cardinal information through two relations. The triple $\mathcal{A} = [A, R_1, R_2]$ is called a preference system iff $A \neq \emptyset$ and $R_1 \subseteq A \times A$ is a preorder on A and $R_2 \subseteq R_1 \times R_1$ is a preorder on R_1. The first relation R_1 represents the ordinal information where $(a_1, a_2) \in R_1$ if a_1 *is at least as desirable as* a_2. R_2 captures the cardinal information where $((a_1, a_2), (b_1, b_2)) \in R_2$ states that *exchanging* a_2 *with* a_1 *is at least as desirable as exchanging* b_2 *with* b_1. Therefore, R_2 describes the degree of improvement between two ordered consequences (a_1, a_2) in comparison to two other ordered consequences (b_1, b_2). Similar to the classical stochastic dominance approach above, a preference system is consistent if there exists a non-empty set of utility functions $\mathcal{U}$ with $u : A \to [0, 1]$ such that for all $u \in \mathcal{U}$ we have $u(a_1) \geq u(a_2)$ iff $(a_1, a_2) \in R_1$ and $u(a_1) - u(a_2) \geq u(b_1) - u(b_2)$ iff $((a_1, a_2), (b_1, b_2)) \in R_2$. Building on this, Jansen, Nalenz, et al. 2023; Jansen, Schollmeyer, and Augustin 2018 introduced the *generalized stochastic order*. Let $\mathcal{A} =$

$[A, R_1, R_2]$ be a consistent preference system with utility functions $\mathcal{U}$ and let $(S, \mathcal{S}, \pi)$ be the probability space that represents the state of the world. Then for two acts $X, Y : S \to A$, we say that X $(\mathcal{A}, \pi)$-dominates Y if and only if $\mathbb{E}_\pi[u \circ X] \geq \mathbb{E}_\pi[u \circ Y]$ for all $u \in \mathcal{U}$. Similarly to the above, we also have measurability and integrability assumptions. This definition induces a preorder on the set of acts/random functions. This stochastic order extends the concept of stochastic dominance to situations involving the above example of algorithm selection using data with mixed scales of measurements, making it possible to rank acts, even in complex decision problems.

Chapter 3

About the Contributing Material: Relations, Summaries, and Outlooks

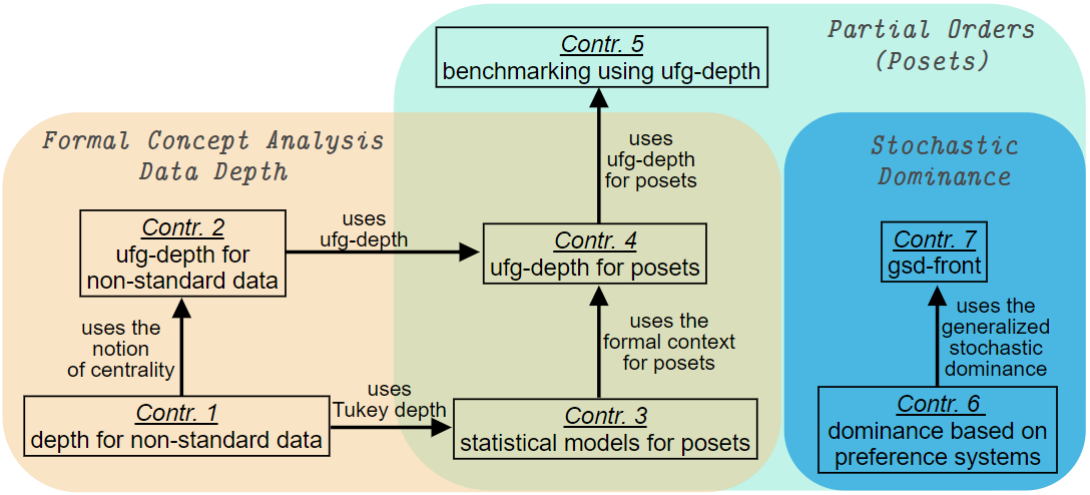

Figure 3.1: The diagram represents my dissertation and the connection between the contributions.

This cumulative dissertation explores approaches to define statistical methods for non-standard data that take the underlying data structure into account. In the following, I present all publications that are part of this dissertation. I outline the main contributions of each publication, critically reflect on them, and provide comments and possible further research questions. Furthermore, I present the connections between all contributions. Figure 3.1 provides an overview of all contributions and their links.

I begin with presenting the main fca-depth approach of my dissertation. As can be seen in Figure 3.1, each contribution in this part is related to Contribution 1, which is why

I start with discussing this publication. Since Contribution 2 also provides a perspective on depth functions for all kinds of non-standard data, I continue with its presentation. I then discuss the depth functions defined specifically for partial order-valued data. In order to clarify the connections between the contributions, I will start again from the bottom up. Finally, in Section 3.3 I explain the smaller gsd approach of my dissertation.

As is common in statistics, all following publications are produced in collaboration. Therefore, I will refer to 'we' in the following to make it clear that these contributions are made in collaboration. A detailed presentation of the individual contributions can be found in the beginning of this dissertation. To contribute to open science, the code of each contribution is freely available at https://github.com/hannahblo (last accessed: 02.03.2025). In particular, we developed an R package, see https://github.com/hannahblo/ddandrda (last accessed: 02.03.2025). Finally, please note that, unless otherwise indicated, all references apply exclusively to the main body of this dissertation and do not refer to the respecting pages of the contributions in the attachments.

3.1 Depth Functions using Formal Concept Analysis

This section fills the theoretical gap caused by the fact that depth functions are only defined for specific data types with a presupposed data structure, e.g. normed vector spaces. The new depth functions are defined in general terms and are not tailored to any particular type of data sets. By utilizing formal concept analysis and thereby considering the relational structure of the data, rather than the raw data value, this approach gives a fresh perspective on method designing in statistics.

3.1.1 Contribution 1: Structural Properties and Generalized Tukey Depth

> Hannah Blocher and Georg Schollmeyer (2025). "Data Depth Functions for Non-standard Data by use of Formal Concept Analysis". In: *Journal of Multivariate Analysis* 205, 105372

By establishing a general notion of centrality and outlyingness for data using formal concept analysis, this contribution provides a framework for defining and analyzing depth functions for non-standard data. The key contribution lies in introducing a general mapping rule for depth functions based on a formal context and presenting structural properties that formalize the notion of centrality for non-standard data. These structural properties enable a systematic classification of depth functions for non-standard data, as demonstrated through the generalized Tukey depth. Thereby, this contribution not only provides the theoretical understanding of depth, centrality and outlyingness of non-standard data, but also lays the groundwork for future research.

A central aspect of the framework is the new approach of defining depth functions using a formal context rather than directly relying on raw data values, as is common in most statistical methods. By representing data as objects within a formal context, the presented method makes use of the inherent relational structure between data elements through the extent set of the formal context. This perspective not only introduces a different way of conceptualizing depth functions, but also achieves an important balance between two competing objectives: it allows for broad applicability across diverse data types while still preserving and respecting the inherent structure of the data.

The basis is the definition of a general mapping rule for depth functions that is flexible and broad enough to cover a variety of concrete mapping rules for all kind of non-standard data. It is a natural generalization of the depth function definition for $\mathbb{R}^d$, see Definition (2.8) from Section 2.3:

$$D : G \times \mathcal{K} \times \mathcal{P} \to \mathbb{R}_{\geq 0} \tag{3.1}$$

with $\mathcal{K} \subseteq \{\mathbb{K} \text{ formal context} \mid G \text{ is object set of } \mathbb{K}\}$ and $\mathcal{P}$ being a set of probability measures on G.[1] The empirical depth function equals the above definition with empirical probability measure as input. Assume that in the definition the components of the mapping rule, the data G and the probability measures $\mathcal{P}$ are fixed, then we can obtain very different depth functions for G by using different scaling methods and get different formal contexts. Hence, the scaling method has a huge influence on the resulting depth function.

Building on this definition, we introduce 13 structural properties that clarify the notions of centrality and outlyingness for non-standard data. These properties follow the style of those established for $\mathbb{R}^d$ by, e.g., Liu 1990; Zuo and Serfling 2000 and Mosler and Mozharovskyi 2022, see Section 2.3. By formulating these properties in the context of formal concept analysis, we provide a systematic foundation for defining and analyzing depth functions beyond normed vector spaces.

The structural properties and their relations are illustrated in Figure 3.2, which is a copy of Figure 1 in Contribution 1 (page 54). Some of these properties extend existing concepts from $\mathbb{R}^d$, while others introduce new aspects specific to formal concept analysis. Note that several properties defined for $\mathbb{R}^d$ can be transferred directly. For example, the affine invariance property of Zuo and Serfling 2000, stating that a depth function should be independent of the coordinate system, can be naturally translated to the requirement of a depth function to solely depend on the resulting extent set rather than on specific scaling methods. This is formalized as the *invariance on the extents property (P1)*. Similarly, the notion of quasiconcavity, which ensures that depth contours form convex sets in $\mathbb{R}^d$ is adapted by requiring these contours to be extent sets of the formal context, as captured by the *quasiconcavity property (P7ii)*.

Other properties, such as vanishing at infinity, do not have a straightforward translation, as they rely on Euclidean concepts like norms and unboundedness. Instead, additional structural properties capture characteristics of the data that emerge naturally

[1] Note that here measurability assumptions are needed, for further details see Section 4 of Contribution 1.

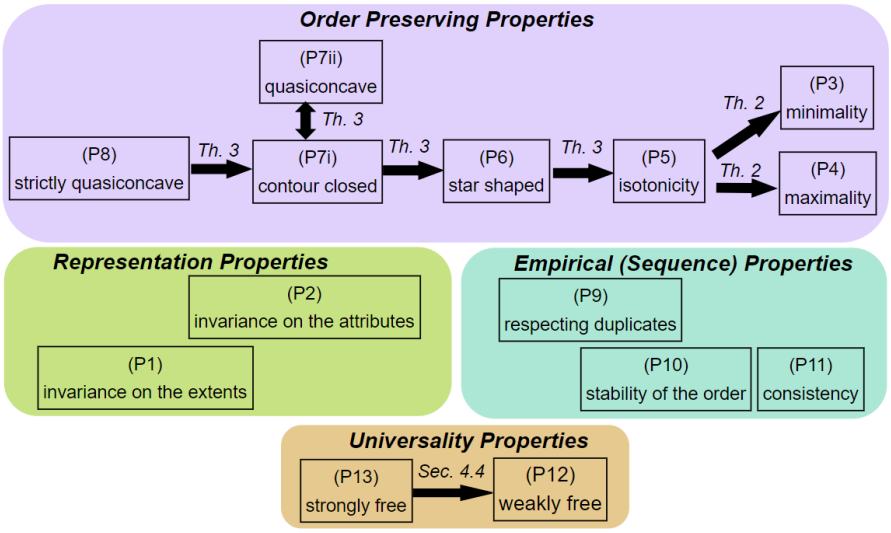

Figure 3.2: Overview of the structural properties together with their mathematical connections. It is a copy of Figure 1 in Contribution 1, page 54.

by applying formal concept analysis. Unlike $\mathbb{R}^d$, where centrality and outlyingness are typically denoted by an underlying probability measure, non-standard data represented by a formal context can contain centrality and outlyingness structures that are inherited from the relationship structure given by the extent set. For example, an object having every attribute belongs to all extents. From the perspective of formal concept analysis, this makes the object a central object. This concept is formalized as the *maximality property (P4)*.

The 13 structural properties can be grouped into four main categories. The first category, consisting of two representation properties, ensures that the depth function preserves the structure given by the formal context, e.g. invariance on the extents property (P1). The second category focuses on how the centrality-outwards order given by the depth function reflects the structure imposed by the formal context, e.g. maximality property (P4) or quasiconcavity property (P7ii). The third category, including Properties (P9) to (P11), addresses the empirical depth function under aspects like *consistency*, see Property (P11). The last category focuses on the richness of the depth function. While a constant depth function may technically satisfy many of the structural properties from the first three categories, it provides no meaningful information. To avoid this, the final two properties ensure that a depth function remains as informative as possible. They require that, while fulfilling a given structural property, the depth function is still flexible enough to mimic other depth functions that also satisfy the same property. Finally, we analyze the theoretical connections between these structural properties, as illustrated in Figure 3.2.

Following this theoretical foundation, we embed the generalized Tukey depth, see Tukey 1975; Schollmeyer 2017a; Schollmeyer 2017b, into our framework and characterize it using the structural properties introduced earlier. This generalization is achieved by adapting the idea of interordinal scaling to $\mathbb{R}^d$ and defining the half-spaces in $\mathbb{R}^d$ as attributes in a formal context. The generalized Tukey depth is given by

$$T_G : \begin{array}{l} G \times \varkappa_G \times \mathscr{P}_G \to [0,1], \\ (g, \mathbb{K}, P_G) \mapsto 1 - \sup_{m \in M \setminus \Psi(g)} P_G(\Phi(m)) \end{array} . \tag{3.2}$$

Our analysis shows that the generalized Tukey depth satisfies most of these properties. This is in line with previous studies on the Tukey depth in $\mathbb{R}^d$, see Zuo and Serfling 2000. However, we also observe that the generalized Tukey depth lacks the flexibility to mimic other depth functions for the formal context representing hierarchical-nominal data, see Example 8 of Contribution 1. This prevents the generalized Tukey depth to satisfy the *strongly freenness property* (P12).

Overall, this contribution provides the theoretical foundation that extends the notion of depth functions to non-standard data. Moreover, it is the starting point for robust, non-parametric statistical methods for non-standard data using formal concept analysis and depth functions.

3.1.2 Contribution 2: The Union-free Generic Depth

Hannah Blocher and Georg Schollmeyer (2024). *Union-free Generic Depth for Non-standard Data*. ArXiv:2412.14745. URL: https://arxiv.org/abs/2412.14745. (last accessed: 02.03.2025)

Several depth functions have been introduced for $\mathbb{R}^d$, see Section 2.3, but their adaptation to non-standard data, even with a presupposed data structure, remains limited. In this contribution we extend the well-known simplicial depth from Liu 1990 to the broader class of non-standard data and introduce the *union-free generic depth* (*ufg-depth*). Building on the structural properties established in Contribution 1, we analyze the ufg-depth in detail. Notably, while structural properties like quasiconcavity of ufg-depth do not universally hold, we introduce a refined variant, namely the *quasiconcave ufg-depth*, which is quasiconcave while offering greater flexibility and richerness than the generalized Tukey depth (see the universality properties in Contribution 1). Unlike Contribution 1, which is primarily theoretical, this work demonstrates the practical relevance of our approach through two concrete case studies: occupational data and mixed spatial-ordinal-numeric data. These examples illustrate not only the applicability of ufg-depth but also the broader potential of depth functions within the framework of formal concept analysis.

In $\mathbb{R}^d$, the simplicial depth of a point represents the probability that a randomly chosen $d+1$ simplex contains that point, see Section 2.3. To extend this concept beyond standard Euclidean space to non-standard data, we utilize that convex sets in $\mathbb{R}^d$ form a

closure system, see Example 1 of Contribution 2. By reformulating the definition of the simplicial depth in terms of this convex closure system, we establish a direct connection to formal concept analysis and pave the way for a more general depth function.

To construct this generalization, two key geometric notions must be adapted. First, the concept of "x lies in A" must be transferred to arbitrary closure systems. In $\mathbb{R}^d$, a point x is considered to "lie in" a set $A \subseteq \mathbb{R}^d$ if it belongs to the smallest convex set containing A, meaning that A implies x under the convex hull (closure) operator. Hence, this can be directly transferred to formal concept analysis by using the closure operator provided by the derivation operators. Secondly, simplices need to be redefined in the context of a more general closure system. A $d+1$ simplex is a special type of convex closure set. To generalize this, we define the *union-free generic family of implications* by removing all implications that are already given by other implications and the inference axioms (except for (pseudo-)transitivity). For convex sets, this leads to a set of implications where the premises correspond to edges of k simplices with $2 \leq k \leq d+1$.

With this, we introduce the union-free generic (ufg) depth which measures how frequently a data element is implied by a randomly selected ufg-premise, weighted according to the size of the premise, see Section 3.2 of Contribution 2. When applied to convex sets in $\mathbb{R}^d$ with a probability measure that is absolutely continuous to the Lebesgue measure, the ufg-depth equals exactly the simplicial depth. This confirms its consistency with the classical definition while significantly broadening its applicability.

We analyze the ufg-depth using the structural properties defined in Contribution 1. Simlilar to the simplicial depth in $\mathbb{R}^d$, the ufg-depth fulfills less structural properties than the generalized Tukey depth introduced in Contribution 1. In particular, the quasiconcavity property is not fulfilled. However, we develop an adjusted version: the quasiconcave ufg-depth. This depth function is quasiconcave and, in certain data scenarios, more flexible than the generalized Tukey depth. It means that, compared to the universality property, the quasiconcave ufg-depth can mimic a wider range of depth functions for hierarchical-nominal data than the generalized Tukey depth.

To illustrate the practical utility of the ufg-depth, we apply it to two real-world data sets: gorilla nesting sites in the Kagwene National Park, Cameroon, see Funwi-Gabga and Mateu 2012; Baddeley and Turner 2005, and occupation data from the German General Social Survey (GGSS), see GESIS 2018. These applications highlight how the ufg-depth provides deeper insights than conventional depth functions by respecting the data structure.

In our analysis of gorilla nesting sites, we discuss not only the spatial locations of nests but also elevation and vegetation of these locations. This perspective allows the ufg-depth to reveal patterns that a purely spatial analysis misses. For example, locations that appear central based on their spatial coordinates alone may still have a low ufg-depth value if they are outliers in terms of elevation and vegetation. This shows how the ufg-depth effectively captures centrality and outlyingness in a mixed-type data setting, as illustrated in Figure 4 (left) of Contribution 2.

The second application examines occupation data from the GGSS, where job classifications follow a hierarchical structure using the International Standard Classification of Occupations ISCO-08. Occupations are organized into increasingly specific subcategories, first distinguishing broad fields such as service occupations and subsequently refining them into specific roles like conductor or hairdresser. Classical approaches, such as identifying the classical modus, focus only on the most detailed classification level. In contrast, by including the hierarchical structure into the definition of the scaling method, the depth functions takes the information across all levels into account.

All in all, with the (quasiconcave) ufg-depth we introduce a further measure of centrality and outlyingness which can be even more flexible, i.e. in the sense of universality properties, than the generalized Tukey depth. Using Contribution 1, we provide a thorough analysis of this depth function. Moreover, the two examples highlight the flexibility and broad applicability showing how the ufg-depth respects the structure of different data types.

3.1.3 Outlook and Perspectives

These two contributions present the foundation of the depth-fca approach, which develops statistical methods for non-standard data. Together, they establish a theoretical basis, introduce two specific depth functions for non-standard data and offer two detailed examples. However, these contributions are only a starting point and many further research directions remain to be conducted.

Further Properties: A promising direction for further property definitions is to explore additional structural characteristics of depth functions. Although we presented several structural properties, there is still considerable scope to identify new characteristics that further define the depth functions. These properties might be specific to a particular type of data, for example the specific properties for partial ordered-valued data discussed in Contribution 4, or they might be general requirements that any depth function should meet.

Further general structural properties could explore more thoroughly how the formal context influences the behavior of a depth function. At present, only the invariance on the extents property addresses the situation where the mapping rule, object set, and probability measure remain the same, but the formal context differs. Although some desirable behaviors of the depth function under two different formal contexts are already covered by existing properties (as noted in Remark 1 of Contribution 1), a more detailed discussion could clarify the precise influence of the formal context on the depth function. One idea is to develop a property that is a localized version of the invariance on the extents property. For example, let G be the object set of two formal contexts that induce the two closure systems $\mathcal{E}_1$ and $\mathcal{E}_2$ on G. Let $\emptyset \neq A \subseteq G$ such that restricted to set A the two closure system equal, i.e. $\{E \cap A \mid E \in \mathcal{E}_1\} = \{E \cap A \mid E \in \mathcal{E}_2\}$. Then the resulting center-outward order of the elements in A should be identical, regardless of whether $\mathcal{E}_1$ or $\mathcal{E}_2$ is used in the depth function (under the assumption that the probability measure is preserved as well, similar to the assumption in Property (P1) of Contribution 1).

Another promising direction is to adapt the breakdown property from depth functions

in $\mathbb{R}^d$, see, e.g., Donoho and Huber 1983; Donoho and Gasko 1992. In its classical form, the breakdown property measures the minimum proportion of contamination in a sample that can cause the most central point to become arbitrarily large (with respect to a norm). For non-standard data, defining "arbitrarily large" can be very difficult. One approach to address this challenge is to examine how many sample elements must be contaminated for the central point to become the most outlying element. This idea is only in its preliminary stage, especially given that multiple definitions of the breakdown property exist, see Donoho and Huber 1983, and an intensive discussion is needed.

Further Depth Functions: Beyond the Tukey depth and simplicial depth, many depth functions have been proposed for $\mathbb{R}^d$. Generalizing these other depth functions to non-standard data is a natural next step. Depth functions that rely heavily on the normed vector space structure, such as the Mahalanobis depth that uses the Mahalanobis distance, see Mahalanobis 1936; Mosler and Mozharovskyi 2022, may be challenging to generalize to non-standard data. On the other hand, approaches like the convex hull peeling depth, see Barnett 1976; Eddy 1981, show promise in this regard. We already started to explore the convex hull peeling depth in the special case of partial order-valued data in Contribution 3. This depth function is defined recursively where one uses the convex layers to define the depth values. Hence, the definition is based on the concept of extreme points of convex sets in $\mathbb{R}^d$. This idea can be transformed to meet-distributive formal contexts, see Ganter and Wille 2012, p. 228ff. For formal contexts that are not meet-distributive, one could aim to define a further operator that behaves similarly, see Footnote 3 of Contribution 3. Note that the examinations done in Contribution 3 are only a starting point for the generalization and further investigation is necessary.

Finally, Hu et al. 2023 developed an outlyingness measure based on formal concept analysis introduced. Translating and analyzing this measure in a manner similar to the generalized Tukey depth from Contribution 1, is of interest. This can also be discussed for the quantile approach in Cardin 2012, see Section 2.4.

Statistical Inference: So far, our focus has been primarily on descriptive statistics, aside from the consistency property. Developing methods for statistical inference based on depth functions using formal concept analysis remains an open challenge. A first step could involve adapting two-sample tests. For example, building on the work of Li and Liu 2004; Dovoedo and Chakraborti 2014 one could define two-sample tests to evaluate whether the medians (elements with the highest depth value) of two distributions are equal. Initial steps might include the exploration of so-called depth-versus-depth plots, where the depth values between the two depth functions are compared. This step is essential as similar previous work, see Li and Liu 2004, only discussed these tests under the assumptions that the two distributions are identical except for translation or scale shift. However, for non-standard data the notions of "translation" and "scale shift" are not defined in general. Permutation tests, supported by detailed simulations similar to Dovoedo and Chakraborti 2014, could then provide robust test statistics. One-sample tests are even more challenging, as many existing tests rely on properties like symmetry in normed vector spaces. Here, an investigation of bootstrap methods may help overcome these limitations.

Further Applications and Scaling Methods: Applying the depth functions on specific kind of non-standard data provides concrete mapping rules for these data types that include the formal context information directly to the mapping rule. This is already done for the ufg-depth in the case of hierarchical-nominal or mixed ordinal-numeric-spatial data in Contribution 2 and partial order-valued data in Section 3.2. With this, applicants do not need to apply a scaling method in advance and therefore this makes the depth function accessible to broader range of potential users. Moreover, this provides valuable insights into the nature of the depth functions as one can define properties that should hold specifically for these kinds of data.

Further non-standard data that are of interest, for example, is the collection of musical pieces. These can be analyzed based on musical structure aspects such as harmony (chord sequence, key), rhythm (tempo, pattern repetitions) or melody (repetition of phrases) to identify outlying pieces from a given music period. Similarly, in archaeology, depth functions might help categorize graves of an excavated cemetery to detect atypical graves. To apply the depth functions to these non-standard data, one has to first develop a scaling method. Hence, a detailed discussion on what are important attributes/characteristics of these data is essential and, in particular, both applications would benefit greatly from collaboration with domain experts.

In addition to these important directions for further research, several more focused follow-up projects could be pursued. In Contribution 2, we briefly discussed a quasiconcave version of the ufg-depth. This is a two-step procedure that first computes the ufg-depth and then derives its quasiconcave variant. A direct next step would be to specify a concrete mapping function for this depth and perform a detailed analysis. Furthermore, while we touched on comparing the generalized Tukey depth with the union-free generic depth in Contribution 2, a comprehensive comparison across multiple data sets is still needed.

3.2 Depth Functions for the Concrete Case of Partial Order-valued Data

Building on the previous discussion of depth functions using formal concept analysis for non-standard data, we now focus on the special case of partial order-valued data. With this, we introduce statistical methods that are specifically designed to capture the intrinsic structure of this data. As explained in Section 2.1, our perspective on partial order-valued data has two key assumptions that set it apart form the classical approach. Firstly, we regard the incomparabilities as valid observations rather than as consequences of missing information. Secondly, each observation is viewed as a complete partial order that is a holistic entity and not merely a collection of pairwise comparisons.

3.2.1 Contribution 3: Statistical Model

Hannah Blocher, Georg Schollmeyer, and Christoph Jansen (2022). "Statistical Models for Partial Orders based on Data Depth and Formal Concept Analysis". In: *Information Processing and Management of Uncertainty in Knowledge-based Systems.* Ed. by Davide Ciucci, Inés Couso, Jesús Medina, Dominik Ślęzak, Davide Petturiti, Bernadette Bouchon-Meunier, and Ronald Yager. Cham: Springer, 17–30

This contribution provides a *unimodal location-scale model* for partial order-valued data that incorporates the incomparability part as well as the understanding of each partial order as an holistic entity. Therefore, we build on Contribution 1 by adapting the idea of the quasiconcavity to define unimodality, use the generalized Tukey depth and introduce a scaling method for partial order-valued data. Moreover, we present an algorithm for sampling from the model.

As already described in Section 2.1, well-known statistical models for partial order-valued data are based on statistical models for total rankings. Hence, the generalizations (implicitly) assume that the partial nature of the observations stem from a missing mechanism, see, e.g., Lebanon and Mao 2007. Others decompose a partial order into pairwise comparisons as done in Davidson and Beaver 1977. In contrast, we provide a model that indeed regards the incomparability of two items to be a precise observation and respects the structure of the partial order.

Therefore, we build on the generalized Tukey depth instead of the well-known approach to use a generalization of metrics. With this, the introduced statistical model on the set of all partial orders $\mathscr{P}$ on a fixed set of items $\mathcal{X}$ is given by

$$P(\{p\}) = C_\lambda \cdot \Gamma\left(\lambda \cdot \left(1 - D^\mu\left(p\right)\right)\right) \tag{3.3}$$

with $p \in \mathscr{P}$, Γ being a decreasing function, $\mu \in \mathscr{P}$ represents the median, D^μ a depth function with center at μ, λ the scaling parameter and C_λ the normalization constant.

Here, D^μ is a modified version of the generalized Tukey depth. Hence, the first step is to define a formal context that includes both of the above structure assumptions on partial order-valued data. The formal context $\mathbb{K} = (G, M, I)$ is given by $G = \mathscr{P}$ and the attributes $x_1 \geq x_2$ and $x_1 \not\leq x_2$ for all pair of items $(x_1, x_2) \in \mathcal{X} \times \mathcal{X}$. Since $x_1 \not\leq x_2$ and $x_2 \not\leq x_1$ can be both true for the same partial order, we included the knowledge on incomparability. Without the information of the non-existing dominance attributes, the underlying structure always implies all linear extensions of the partial order as they need to lie in the corresponding extent set. Hence, in this case, when quasiconcavity of Contribution 1 is fulfilled for the depth function, the depth function necessarily ranks the linear extensions of a partial order p higher (or at least cannot set them more outlying) than p. For this reason, we include the non-existing dominance attributes into the formal context.

Since we aim for a unimodal location-scale model, we have to slightly modify the generalized Tukey depth such that (I) the central partial order can be predefined and

(II) the model is unimodal. But first, we clarify the notion of unimodality, where we adapt the quasiconcavity definition of Contribution 1. A model (on a finite set of objects – here $\mathscr{P}$) is called *unimodal* if and only if for all $\alpha > 0$ the set of objects (here: partial orders) that have a higher probability than α defines an extent. Hence, the model is unimodal if and only if the depth function is quasiconcave.

To ensure part (I), we consider a localized version of the generalized Tukey depth of Contribution 1. In this localized version, the depth is maximal at a pre-given partial order which is the median μ. To achieve this, the attributes considered in the definition of the generalized Tukey depth are limited to those that hold for μ, in contrast to Contribution 1. The localized version is still quasiconcave, and therefore we fulfill part (II), however it is quite uninformative as it only has two values. Hence, we further modify it by integrating ideas of the convex hull peeling depth in $\mathbb{R}^d$. In $\mathbb{R}^d$ the convex hull peeling depth recursively removes extreme points. Here, we consider the reverse where we start from the center and recursively add extreme points. This provides us the necessary weights to modify the localized generalized Tukey depth such that the resulting depth function is informative, quasiconcave (and therefore the model unimodal) and the central partial order can be predefined.

To make this model practical, we propose a sampling algorithm based on the acceptance-rejection method. The algorithm generates a partial order and uses the number of possible paths reaching this partial order to adjust the acceptance probability. This ensures that the generated samples are drawn with the desired probability distribution defined by the statistical model.

In summary, we introduce a statistical model for partial order-valued data that explicitly includes the incomparability character of the partial orders and regards them as a holistic entity, in contrast to well-known approaches. Moreover, by clarifying the concept of unimodality the model becomes accessible via the localization and scale parameters. The provided sampling algorithm bridges the gap of this theoretical approach to users.

3.2.2 Contribution 4: Union-free Generic Depth – Properties and Application on Machine Learning Algorithms

Hannah Blocher, Georg Schollmeyer, Malte Nalenz, and Christoph Jansen (2024). "Comparing Machine Learning Algorithms by Union-free Generic Depth". In: *International Journal of Approximate Reasoning* 169, 109166. (Invited Paper for the ISIPTA 2023 Special Issue)

This contribution applies the ufg-depth of Contribution 2 on the formal context for partial order-valued data defined in Contribution 3. We discuss how the ufg-depth function captures essential characteristics of the data in detail. We show that the ufg-depth treats each partial order as a holistic entity and not as a composition of pairwise comparison. Moreover, it reflects both the comparability and incomparability part of the observations. A key contribution is the application of the ufg-depth to comparing machine learning

algorithms based on multidimensional performance measures across multiple data sets simultaneously. Therefore, we provide an algorithm that improves the naive computation of the ufg-depth. With a comparison to other methods, we demonstrate that the ufg-depth provides a reliable tool for algorithm comparison.

The question of how to compare the algorithm performances is a central topic in machine learning, see, e.g., Hansen, Auger, Brockhoff, et al. 2022; Hothorn et al. 2005. Most classical benchmark approaches rely on one single performance measure, e.g. accuracy or precision, and (implicitly) assume a total order of the algorithms based on their performance. Hence, they ignore that performance is often a latent value or multidimensional concept where the components of the performance measures might conflict. Moreover, these approaches solely provide one dominance structure that aims to represent the true underlying performance order of the algorithms based on multiple data sets. The ufg-depth addresses all these issues by allowing for incomparability and an analysis based on all performance measures simultaneously while describing an entire distribution on all possible performance structures.

First of all, the general data set-up is as follow: We aim to compare ℓ algorithms based on k performance measures and n data sets. We say that an Algorithm A is better than another Algorithm B based on one fixed data set if and only if the performance of Algorithm A is not worse than the performance of Algorithm B based on all performance measures and for at least one performance measure the performance of Algorithm A is even strictly better. When there exist two performance measures that contradict each other, then the two algorithms are incomparable. Hence, we obtain for every single data set one partial order.

Applying the ufg-depth with the formal context given in Contribution 3 and the empirical probability measure given by the n observed partial orders, we obtain a center-outward order of all possible performance structures of the algorithms. This now describes the distribution on all performance orders of the algorithms. In particular, we can denote partial orders that provide an atypical order and also those that are most central, i.e. whose performance order is the most supported/typical order based on the (empirical) distribution. Since the *ufg-depth on partial order-valued data* builds on the existence of non-dominance and dominance in a partial order, the ufg-depth does not have an inherent preference for total orders to be central.

In contrast to Contribution 2 where we analyzed the ufg-depth based on the structural properties, here we can exploit that the concrete mapping rule already includes the formal context and therefore only depends on the (empirical) probability measure. Hence, here we provide a narrower analysis of the resulting center-outward order. For example, we prove that the structure of the partial order-valued sample effects the depth function in a desirable manner, i.e., when a dominance (or incomparability) never occurs, the depth of every partial order containing following this dominance (or incomparability) is zero. Moreover, we show that the depth function cannot be broken down to pairwise comparisons, but reflects the holistic characteristic of the partial orders. We also exploit the additional provided structure to improve the computation of the ufg-depth, see Section 6 of the contribution.

Finally, we demonstrate the practical utility by applying it on two benchmark setups: UCI and OpenML, see Kelly et al. 2017; Vanschoren et al. 2013. The UCI analysis illustrates the difference between the ufg-depth and two further benchmark approaches, the extended Bradley-Terry model and the gsd-benchmark approach, see Davidson and Beaver 1977 and Jansen, Nalenz, et al. 2023. The main difference is that the ufg-depth provides a description of the distribution on the set of partial orders and not only one single order as a representative for the entire distribution.

The second example, based on OpenML, gives a detailed comparison of five classifiers based on four performance measures and 80 data sets. This outlines the applicability and utility of our introduced ufg-depth approach. For example, by obtaining not only the most central/typical performance order, but a measure on all partial orders, we could observe the strong support that Random Forests outperform Decision Trees for the considered data sets. Moreover, we discussed the selection of performance measures, given the counter intuitive result that the number of performance measures used cannot be reduced by considering the correlation between them and excluding one when another with a high correlation in it is already used.

Overall, with the focus on partial order-valued data instead of general non-standard data, we provide a new benchmark approach that incorporates incomparability. By considering an entire distribution instead of only one order, it differs substantially to well-known benchmark approaches.

3.2.3 Contribution 5: Comparing Optimizers

Julian Rodemann and Hannah Blocher (2024). "Partial Rankings of Optimizers". In: *The Second Tiny Papers Track at ICLR 2024, Tiny Papers @ ICLR 2024*. Ed. by Tom Burns and Krystal Maughan. Vienna: OpenReview.net

While Contribution 4 focuses on introducing the ufg-depth with a detailed analysis and comparison to other approaches, this tiny paper has a focus on applying the ufg-depth on benchmark problems of optimizers: BBOB, multi-objective evolutionary algorithms and DeepOBS, see Hansen, Auger, Ros, et al. 2010; Wu et al. 2023; Schneider et al. 2019.

One focus is now on the utility in the design and objective of benchmark suites. For example, we observe that even though we allow for incomparability, the existing dominance structure of the most central partial order falls apart quickly when taking more than one partial order, i.e. the k most central ones, into account. Hence, this shows that the general approach in benchmarking to provide a total order fails to represent performance of optimizers.

Moreover, we outline how the ufg-depth can support obtaining an diverse benchmark suite as the depth shows whether the benchmark suite leads to different performance orders or is limited to one specific structure. Additionally, we provide a short introduction to the general problem of aggregating orders, see Appendix F, and the questions considered in benchmarking, see Appendix E.

3.2.4 Outlook and Perspectives

This part already addresses aspects of the outlook on Contributions 1 and 2, see Section 3.1. In that section, we emphasized the importance of applying the introduced depth functions to concrete non-standard data – an approach exemplified here with partial order-valued data. Moreover, this special case of non-standard data suggests several promising directions for follow-up research projects.

Distinguishing between Indifference and Incomparability: In Contribution 4, we explain how to derive the partial order structure needed for benchmark problems. However, we do not address the situation where two algorithms have equal performance measures. In such cases, the algorithms should be considered indifferent rather than incomparable. Although Contribution 5 briefly mentions this issue, it does not explore the practical differences. However, since depth functions are generally defined on non-standard data, they can be easily adapted to account for indifference. One straightforward approach is to expand the set of objects G to include all preorders. Analyzing this adaptation and discussing how it affects the evaluation of benchmark problems is a logical next step for future research.

Comparison to Goibert et al. 2022: Goibert et al. 2022 develop a depth function specifically for total order-valued data and presented four properties that such a depth function should satisfy. These properties are tailored to total orders, for example the authors use a pseudo-metric based on the symmetric group of permutations that represent total orders. With this article in mind, there are now two further projects for our approach, the first one being the integration of their depth function into our framework. However, at first glance, this appears challenging because the depth introduced in Goibert et al. 2022 is based on a pseudo-metric, making it difficult to define a scaling method that mimics it. The other approach is to evaluate our ufg-depth function for partial order-valued data using the four properties the authors developed. To achieve this, one could either develop a new scaling method tailored for total order-valued data or restrict the object set G to total orders and then apply the scaling method introduced in Contribution 3. For example, by using Theorem 1 from Contribution 1, we can directly demonstrate that the "maximality at center property" from Goibert et al. 2022 holds for the generalized Tukey depth.

Handling of Missing Values: A missing value in a partial order means that a relation between two items, say x_1 and x_2, (or possibly more) was not observed. In such cases, the true relation could be either $x_1 \geq x_2$, $x_2 \geq x_1$ or the items x_1 and x_2 could be incomparable. However, thanks to the transitivity and antisymmetry of partial orders, the available data can sometimes provide enough information to determine the relation or at least limit the possibilities to two options.

How these missing values are handled depends on the depth function used. For example, in situations described by coarsening at random, see Heitjan and Rubin 1991, the generalized Tukey depth can just ignore the missing values as it only relies on the attributes and their individual proportions that are observed. In contrast, the ufg-depth uses a closure operator that must be adjusted. For more complex cases of missing values, upper and lower bounds for the depth value can be derived using the concept of cautious

data completion, see Augustin et al. 2014; Schollmeyer et al. 2023. However, note that a straightforward computation of these bounds can be computationally demanding.

Aggregation of Partial Orders: In Contribution 4, we used the most central partial orders as the order best representing the observed partial order-valued data. However, one could address the question of finding a representative for a set of partial orders, i.e. aggregating the partial orders, from the perspective of social choice theory. In Contribution 5, we briefly discuss the aggregation of total orders into a single total order. Arrow showed in Arrow 1951 that, under reasonable conditions such as non-dictatorship (meaning the aggregation must consider all orders rather than being dominated by a single order), this goal cannot generally be achieved. Potential further research that also addresses this limitation is the discussion of the aggregation of partial orders. Many of the conditions defined for total order aggregation can also be applied to the aggregation of partial orders. Besides the development of desirable conditions, it could be valuable to develop concrete aggregation methods that go further than simply taking the intersection of partial orders. Although the intersection indeed fulfills many desirable properties, it often results in the trivial partial order where all items are incomparable. This result might by quite uninformative and therefore an investigation in more advanced aggregation methods could be interesting.

Finally, it is important to apply the depth functions for partial order-valued data to a variety of concrete data sets in order to evaluate their strengths and weaknesses. While one further application has been presented by Arias et al. 2024, there remains a need to explore uses beyond the benchmark approach, such as the preference orders discussed in Dittrich et al. 1998. Moreover, the discussion on statistical inference outlined in Section 3.1.3 is equally relevant in this context.

3.3 Stochastic Dominance based on Preference Systems

The above two sections discuss the fca-depth approach. Now, we explore the gsd approach that provides a stochastic order for multidimensional data with mixed scales of measurement. Therefore, similar to Section 3.2, our focus here is on a specific type of non-standard data. We are interested in the order of random variables where we utilize generalized stochastic dominance and preference systems, see Section 2.5. In contrast to the above, we provide inferential statements on whether one random variable dominates another one rather than solely descriptive ones. Another difference is that we include also the cardinal information of the comparison.

3.3.1 Contribution 6: Dominance Between Two Random Variables under Mixed Cardinal-ordinal Information

Christoph Jansen, Georg Schollmeyer, Hannah Blocher, Julian Rodemann, and Thomas Augustin (2023). "Robust Statistical Comparison of Random Variables with Locally Varying Scale of Measurement". In: *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*. Ed. by Robin Evans and Ilya Shpitser. Pittsburgh: PMLR, 941–952

We introduce a statistical test for dominance between two random variables with locally varying scales of measurement. This approach builds on preference systems and generalized stochastic dominance to exploit the entire information in the data. To enhance the reliability of the test, we regularize and robustify the test using imprecise probability models. In the application, we focus on the special case of *mixed ordinal-cardinal data* and provide an efficient algorithm to compute the test statistic. We apply the test to three applications: (multidimensional) poverty analysis, finance and medicine.

To illustrate our approach, consider the application for poverty analysis using the GGSS data, see GESIS 2018. As poverty is considered to be multidimensional, see Sen 1985, we use income (numeric), health (ordinal) and education (ordinal) to measure poverty. Our statistical test evaluates the hypothesis that women are richer[2] than men, expressed as a comparison of two random variables with the null hypothesis $Y \leq X$ and the alternative hypothesis $Y \not\leq X$.

Intuitively, the aim is to control the probability of stating dominance, i.e. the alternative and null hypothesis are in reverse rolls. In the poverty example this means controlling for the probability that we falsely state that women are richer than men, or in other words, that women dominate men. However, that leads to the null hypothesis $Y \not\leq X$ that is too broad to define a meaningful test. Instead, we analyze the reverse and obtain $Y \leq X$ as null hypothesis and control for the statement that no-dominance exists. To mitigate the switch of the null and alternative hypothesis, one can consider the further null hypothesis $Y \geq X$ and therefore can control for both non-dominance statements.

Since we have data with local varying scales of measurement, the test statistic builds on preference systems and generalized stochastic dominance, see Section 2.5 and Jansen 2018; Jansen 2025. It is computed as the difference in empirical expectations, taking the infimum across all possible utility functions. The statistical test follows a permutation-based approach, assuming that the sample is i.i.d., as detailed in Pratt and Gibbons 2012.

To enhance the robustness of our test, we introduce a modified version that accounts for noise of the sample and potential deviations from the i.i.d. assumption. To address sample noise, we regularize the test statistic by restricting the set of utility functions

[2] In terms of higher income, education and better health. In this section, we always understand "poor" and "rich" within this context.

to those that detect a meaningful difference between two options. Specifically, we exclude utility functions that capture differences smaller than a threshold ε, which serves as a regularization parameter. This approach increases the test's discriminative power while still assuming i.i.d. sampling. To further improve robustness, we employ imprecise probability theory, see Augustin et al. 2014. Instead of relying solely on the empirical probability measure, we construct a credal set, which is a set of probability measures that account for potential deviations from the i.i.d. assumption. This allows us to handle cases where a proportion γ of observations deviates from the i.i.d. assumptions.

Computing the (regularized and robustified) test statistic is challenging as it involves an infimum over all possible utility functions. We develop an efficient algorithm for mixed ordinal-cardinal data as discussed in our poverty analysis. In this case, the computation of the test statistic can be translated to solve a linear program. The naive approach to compute the corresponding constraint matrix has a computational complexity of $\mathcal{O}(n^4)$, where n is the sample size. This high level of complexity comes from going through all pairwise comparison in A and R_1. However, in most practical cases, not everything in A and R_1 is comparable. Therefore, we propose an algorithm that pre-sorts the data and uses it to directly skip unnecessary comparisons. This drastically reduces the computation time of the constraint matrix for most practical cases and makes our method scalable.

Returning to the poverty example above, we apply the test to the GGSS data set with a significance level of 0.05. This results in the null hypothesis not being rejected for the unregularized test, meaning we cannot reject the null-hypothesis that women are richer than men. However, when we discuss the regularized test with $\varepsilon \in \{0.5, 0.75, 1\}$, we are able to reject the null hypothesis. When further accounting for deviations from the i.i.d. assumption, even a high regularization level of $\varepsilon = 1$ allows rejecting the null hypothesis only if at most one observation deviates from the i.i.d. assumption. It is important to note that rejecting the null hypothesis does not support the claim that women are poorer than men. It merely indicates that women are not significantly richer, leaving open the possibility that men and women are incomparable in terms of poverty.

All in all, this contribution presents a (robust and regularized) statistical test for analyzing stochastic dominance with mixed scales of measurement. By providing an efficient algorithm for computing the test statistic, we demonstrate its practical utility by three data examples.

3.3.2 Contribution 7: Comparing Machine Learning Algorithms using GSD-front

Christoph Jansen, Georg Schollmeyer, Julian Rodemann, Hannah Blocher, and Thomas Augustin (2024). "Statistical Multicriteria Benchmarking via the GSD-Front". In: *38th Conference on Neural Information Processing System (NeurIPS 2024)*. Ed. by Amir Globerson, Lester Mackey, Angela Fan, Ulrich Paquet, Jakub Tomczak, Cheng Zhang, and Lam Nguyen. Vancouver: OpenReview.net

By adapting Contribution 6 to benchmark problems, we provide statistical methods to gain inferential and robust statements on comparing (machine learning) algorithms across multiple performance metrics and data sets simultaneously. We introduce the concept of the *generalized stochastic dominance front (GSD-front)*, which identifies algorithms that are not outperformed by any other in the set. Additionally, we develop statistical tests to determine whether a given algorithm belongs to this GSD-front. To illustrate our approach, we apply it to machine learning algorithms from the OpenML suite, see Vanschoren et al. 2013, and the PMLB suite, see Olson et al. 2017.

Consider a scenario where we evaluate a new algorithm A against a set of existing state of the art algorithms $A_1, \dots, A_k$ with $k \in \mathbb{N}$ that all solve the same task. Our goal is to determine whether algorithm A improves the established algorithms, meaning it is not dominated by any of them. To achieve this, we define the GSD-front $gsd(\mathcal{C})$ with $\mathcal{C} = \{A, A_1, \dots, A_k\}$ that consists of all algorithms in $\mathcal{C}$ that are not strictly dominated by another algorithm in $\mathcal{C}$. This dominance builds on preference systems and the concept of generalized stochastic dominance, see Section 2.5, and therefore ensures the comparison based on multiple cardinal and ordinal quality metrics/performance measures simultaneously. Note that the definition of the GSD-front is in style of the Pareto-front.

On this basis, we introduce a statistical test that controls the probability to falsely stating that there exists no state of the art algorithm that dominates the new algorithm A. This corresponds to the null-hypothesis $H_0 : A \notin \text{gsd}(\mathcal{C})$ with the alternative hypothesis $H_1 : A \in \text{gsd}(\mathcal{C})$. Building on the statistical test of Contribution 6, we derive the *static test* to evaluate these hypotheses. Additionally, we propose the *dynamic test* to determine the largest possible subset of algorithms within which the new algorithm remains in the GSD-front. Both tests can be robustified against deviations from the i.i.d. assumption and sample noise by using the theory of Contribution 6.

We apply our statistical tests to algorithms and evaluations provided by the OpenML suite, see Vanschoren et al. 2013, that is already discussed in Contribution 4. As quality metrics we use accuracy and computation time (treated as ordinal). The considered null-hypothesis states that SVM does not belong to the GSD-front of SVM, CART, kNN, xGBoost, RF, GLMNet and LR. The static test shows that we cannot reject this null-hypothesis. However, according to the dynamic test, the SVM is not significantly dominated by kNN, xGBoost, RF and GLMNet and therefore lies in the GSD-front of SVM, kNN, xGBoost, RF and GLMNet. It is important to note that this result

remains significant even when 7 out of 80 data sets in the sample are contaminated. We provide a similar second application with algorithms from the Penn Machine Learning Benchmarking suite, see Olson et al. 2017, where we use accuracy and two measures on feature and class robustness (which are ordinal) as quality metrics.

In summary, we introduce a statistical framework for multi-criteria benchmarking that extends beyond descriptive comparisons to provide inferential insights, in contrast to Contribution 4 and 5. Additionally, we demonstrate its effectiveness through concrete benchmark examples.

3.3.3 Outlook and Perspectives

These two contributions provide inferential statements about the stochastic order of random variables with local scales of measurement. Additionally, we demonstrate how these statements can be applied to benchmark problems involving multiple performance measures at the same time. Building on this, there are a number of promising further research areas:

Expanding Computation to Larger and More Complex Data: In both contributions we focused on mixed ordinal and cardinal data, specifically one cardinal and two ordinal components. Currently, our provided code is designed for this three-dimensional data where the ordinal components have an underlying total order structure. However, the underlying theory supports more complex cases, such as higher-dimensional data or ordinal data with a partial order structures. Adapting the code to handle these cases is a natural next step. One challenge in this expansion is computational efficiency. Our preliminary analyses suggest that different data structures impact computation time in varying ways. For example, adding an additional ordinal structure that ensures that many elements are incomparable, can reduce the computation time. In contrast, introducing extra cardinal information can drastically increase computational demands. This observation suggests that exploring alternative data structures may lead to more efficient computation. Analyzing it in more detail could make our approach more practical for handling even larger data sets.

Survey on Benchmark Methods: Our literature review revealed a vast number of research articles on benchmark methods. However, a survey paper that categorized the different approaches is still lacking. Such a survey could provide valuable insights by classifying methods based on: Does the method build on pairwise comparisons, partial orders or total orders; is the method applicable to only one single quality metric or can it address multiple quality metrics simultaneously; on what kind of benchmark problems they are applied to; etc.

Besides the above two follow-up research directions, exploring how our approach performs in other scenarios is of interest. One potential application is evaluating algorithms for open-ended text generation based on a combination of human ratings (that are ordinal) and automatic evaluation metrics, such as perplexity, see Jelinek et al. 2005. Finally, the method can be further developed by considering regression methods that take meta-information of the data sets, e.g. the sample size, into account.

Chapter 4

Concluding Remarks

In this final section, I provide a general outlook and perspective on the methodological framework presented in this dissertation, complementing the more specific outlooks and perspectives in Section 3. Here, I emphasize the overall insight and outlook gained from this dissertation.

This dissertation addresses the dilemma faced by applicants working with non-standard data, see Section 1. Conventional statistical methods are developed for standard statistical data formats and, hence leave applicants dealing with non-standard data with two unsatisfactory options: Either applying conventional methods that may destroy the structure of the data and distort the result and interpretation, or abstain from statistical analysis completely. This dissertation contributes to overcome this dilemma through two different approaches: the fca-depth approach and the gsd approach.

The main focus of this dissertation lies in the fca-depth approach, which combines formal concept analysis and data depth functions to define statistical methods that respect the inherent structure of the data. Contributions 1 and 2 lay the groundwork by providing a systematic basis for discussing depth, centrality and outlyingness for non-standard data and introducing two depth functions applicable to all types of non-standard data. Contributions 3 to 5 build on this foundation, applying and examining the methodology in the specific case of partial order-valued data.

Beyond the immediate provision of statistical methods for non-standard data, these five contributions also demonstrate that formal concept analysis is a powerful theory to develop such methods. By using formal concept analysis, we generalized the concept of depth functions beyond their classical definitions. In particular, the comprehensible data representation given by formal concept analysis, which reveals the assumptions on the data and provides different representations of it (i.e. the formal context, the formal concepts or the formal implications), makes it extremely valuable for non-standard data. Moreover, it allows to define statistical methods on the relational structure of the data elements instead of directly using their concrete values. Hence, besides the natural follow-up research directions described in Sections 3.1.3 and 3.2.4, this dissertation underlines the potential of developing statistical methods from the perspective of formal concept analysis. A close collaboration between the formal concept analysis and statistics com-

munities could therefore lead to valuable advancements in statistical methods in general and especially for non-standard data.

The last two contributions explore the gsd approach that addresses also the challenge of analyzing non-standard data. Unlike the fca-depth approach, which applies broadly to various non-standard data types, the gsd approach examines one specific data structure, data with mixed scales of measurement. This focus on a specific type of non-standard data allows researchers working with such data to apply the statistical methods directly, without the need of extensive preliminary discussions on how to represent the underlying data structure. This is in contrast to the fca-depth approach that requires such considerations, as seen in Contribution 3 and 4 for partial order-valued data.

Both, the gsd approach and the discussion on the fca-depth approach in the special case of partial order-valued data, emphasize the importance of examining each non-standard data type individually. The gsd approach demonstrates that it is not necessary to first analyze non-standard data in a broad sense before focusing on specific types and instead, examining them separately can be equally effective.

In conclusion, developing statistical methods for non-standard data remains a complex challenge with many open research questions. However, this dissertation highlights that multiple strategies can be successful. By demonstrating the fruitfulness of these approaches, this dissertation hopefully encourages further research in statistical methods for non-standard data.

Further References

Adcock, Ben, Brugiapaglia, Simone, Dexter, Nick, and Morage, Sebastian (2022). "Deep Neural Networks are Effective at Learning High-dimensional Hilbert-valued Functions from Limited Data". In: *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*. Ed. by Joan Bruna, Jan Hesthaven, and Lenka Zdeborova. Virtual Conference: PMLR, 1–36.

Arias, Esteban Garces, Blocher, Hannah, Rodemann, Julian, Li, Meimingwei, Heumann, Christian, and Aßenmacher, Matthias (2024). *Towards Better Open-ended Text Generation: A Multicriteria Evaluation Framework*. ArXiv:2410.18653. URL: https://arxiv.org/abs/2410.18653. (last accessed: 02.03.2025).

Armstrong, William (1974). "Dependency Structures of Data base Relationships". In: *International Federation for Information Processing Congress*. Ed. by Jack Rosenfeld. Amsterdam: North-Holland Publishing Company, 580–583.

Arrow, Kenneth (1951). *Social Choice and Individual Values*. New Haven: Yale University Press.

Augustin, Thomas, Coolen, Frank, de Cooman, Gert, and Troffaes, Matthias (Ed) (2014). *Introduction to Imprecise Probabilities*. West Sussex: John Wiley & Sons.

Baddeley, Adrian and Turner, Rolf (2005). "Spatstat: An R Package for Analyzing Spatial Point Patterns". In: *Journal of Statistical Software* 12.6, 1–42.

Barnett, Vic (1976). "The Ordering of Multivariate Data". In: *Journal of the Royal Statistical Society: Series A (General)* 139.3, 318–344.

Behzadi, Sahar, Müller, Nikola, Plant, Claudia, and Böhm, Christian (2020). "Clustering of Mixed-type Data Considering Concept Hierarchies: Problem Specification and Algorithm". In: *International Journal of Data Science and Analytics* 10.3, 233–248.

Berger, James (1985). *Statistical Decision Theory and Bayesian Analysis*. New York: Springer.

Blocher, Hannah and Schollmeyer, Georg (2024). *Union-free Generic Depth for Non-standard Data*. ArXiv:2412.14745. URL: https://arxiv.org/abs/2412.14745. (last accessed: 02.03.2025).

— (2025). "Data Depth Functions for Non-standard Data by use of Formal Concept Analysis". In: *Journal of Multivariate Analysis* 205, 105372.

Blocher, Hannah, Schollmeyer, Georg, and Jansen, Christoph (2022). "Statistical Models for Partial Orders based on Data Depth and Formal Concept Analysis". In: *Information Processing and Management of Uncertainty in Knowledge-based Systems*.

Ed. by Davide Ciucci, Inés Couso, Jesús Medina, Dominik Ślęzak, Davide Petturiti, Bernadette Bouchon-Meunier, and Ronald Yager. Cham: Springer, 17–30.

Blocher, Hannah, Schollmeyer, Georg, Nalenz, Malte, and Jansen, Christoph (2024). "Comparing Machine Learning Algorithms by Union-free Generic Depth". In: *International Journal of Approximate Reasoning* 169, 109166. (Invited Paper for the ISIPTA 2023 Special Issue).

Bolívar, Sergio, Nieto Reyes, Alicia, and Rogers, Heather (2023). "Statistical Depth for Text Data: An Application to the Classification of Healthcare Data". In: *Mathematics* 11.1, 228–248.

Bradley, Ralph and Terry, Milton (1952). "Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons". In: *Biometrika* 39.3/4, 324–345.

Brito, Abner, Barros, Laécio, Laureano, Estevão, Bertato, Fábio, and Coniglio, Marcelo (2018). "Fuzzy Formal Concept Analysis". In: *Fuzzy Information Processing*. Ed. by Guilherme Barreto and Ricardo Coelho. Cham: Springer, 192–205.

Cardin, Marta (2012). "A Quantile Approach to Integration with Respect to Non-Additive Measures". In: *International Conference on Modeling Decisions for Artificial Intelligence*. Ed. by Vincenş Torra, Yasuo Narukawa, Beatriz Lopéz, and Mateu Villaret. Berlin, Heidelberg: Springer, 139–148.

Chakraborty, Anirvan and Chaudhuri, Probal (2014). "On Data Depth in Infinite Dimensional Spaces". In: *Annals of the Institute of Statistical Mathematics* 66, 303–324.

Chen, Zhengdao (2024). "Neural Hilbert Ladders: Multi-layer Neural Networks in Function Space". In: *Journal of Machine Learning Research* 25.109, 1–65.

Chierichetti, Flavio, Dasgupta, Anirban, Haddadan, Shahrzad, Kumar, Ravi, and Lattanzi, Silvio (2018). "Mallows Models for Top-k Lists". In: *Advances in Neural Information Processing Systems*. Ed. by Samy Bengio, Hanna Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett. Montréal: Curran Associates, Inc., 1–11.

Davidson, Roger and Beaver, Robert (1977). "On Extending the Bradley-Terry Model to Incorporate Within-Pair Order Effects". In: *Biometrics* 33.4, 693–702.

Dehghan, Sakineh and Faridrohani, Mohammad Reza (2024). "A Data Depth based Nonparametric Test of Independence between Two Random Vectors". In: *Journal of Multivariate Analysis* 202, 105297.

Dittrich, Regina, Hatzinger, Reinhold, and Katzenbeisser, Walter (1998). "Modelling the Effect of Subject-specific Covariates in Paired Comparison Studies with an Application to University Rankings". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 47.4, 511–525.

Donoho, David and Gasko, Miriam (1992). "Breakdown Properties of Location Estimates Based on Halfspace Depth and Projected Outlyingness". In: *The Annals of Statistics* 20.4, 1803–1827.

Donoho, David and Huber, Peter (1983). "The Notion of Breakdown Point". In: *A Festschrift for Erich Lehmann*. Ed. by Peter Bickel, Kjell Doksum, and J. Hodges. Belmont: Wadsworth, 157–184.

Dovoedo, Herve and Chakraborti, Subhabrata (2014). "Power of Depth-based Nonparametric Tests for Multivariate Locations". In: *Journal of Statistical Computation and Simulation* 85.10, 1987–2006.

Draude, Claude, Dürrschnabel, Dominik, Hirth, Johannes, Horn, Viktoria, Kropf, Jonathan, Lamla, Jörn, Stumme, Gerd, and Uhlmann, Markus (2024). "Conceptual Mapping of Controversies". In: *Conceptual Knowledge Structures*. Ed. by Inma Cabrera, Sébastien Ferré, and Sergei Obiedkov. Cham: Springer, 201–216.

Dudyrev, Egor, Kuznetsov, Sergei, and Napoli, Amedeo (2023). "Description Quivers for Compact Representation of Concept Lattices and Ensembles of Decision Trees". In: *17th International Conference on Formal Concept Analysis*. Ed. by Dominik Dürrschnabel and Dominigo Rodríguez. Cham: Springer, 127–142.

Dürrschnabel, Dominik and Priss, Uta (2024). "Realizability of Rectangular Euler Diagrams". In: *Conceptual Knowledge Structures*. Ed. by Inma Cabrera, Sébastien Ferré, and Sergei Obiedkov. Cham: Springer, 149–165.

Dutcă, Ioan, Stăncioiu, Petru Tudor, Abrudan, Ioan, and Ioraș, Florin (2018). "Using Clustered Data to Develop Biomass Allometric Models: The Consequences of Ignoring the Clustered Data Structure". In: *PLOS ONE* 13.8, e0200123.

Dyckerhoff, Rainer, Mosler, Karl, and Koshevoy, Gleb (1996). "Zonoid Data Depth: Theory and Computation". In: *International Conference on Computational Statistics*. Ed. by Albert Prat. Heidelberg: Physica-Verlag HD, 235–240.

Dykerhoff, Rainer (2002). "Inference Based on Data Depth by Rainer Dyckerhoff". In: *Multivariate Dispersion, Central Regions, and Depth. Lecture Notes in Statistics*. New York: Springer, 133–163.

Eddy, William (1981). "Graphics for the Multivariate Two-sample Problem: Comment". In: *Journal of the American Statistical Association* 76.374, 287–289.

Fahrmeir, Ludwig, Heumann, Christian, Künstler, Rita, Pigeot, Iris, and Tutz, Gerhard (2016). *Statistik: Der Weg zur Datenanalyse*. Berlin, Heidelberg: Springer.

Funwi-Gabga, Nebar and Mateu, Jorge (2012). "Understanding the Nesting Spatial Behaviour of Gorillas in the Kagwene Sanctuary, Cameroon". In: *Stochastic Environmental Research and Risk Assessment* 26.6, 793–811.

Ganter, Bernhard and Wille, Rudolf (2012). *Formal Concept Analysis: Mathematical Foundations*. Berlin, Heidelberg: Springer.

Geenens, Gery, Nieto-Reyes, Alicia, and Francisci, Giacomo (2023). "Statistical Depth in Abstract Metric Spaces". In: *Statistics and Computing* 33.2, 46.

GESIS (2018). *Allgemeine Bevölkerungsumfrage der Sozialwissenschaften ALLBUS 2014*. GESIS Datenarchiv, Köln. ZA5240 Datenfile Version 2.2.0. (last accessed: 02.03.2025).

Goibert, Morgane, Clemencon, Stephan, Irurozki, Ekhine, and Mozharovskyi, Pavlo (2022). "Statistical Depth Functions for Ranking Distributions: Definitions, Statistical Learning and Applications". In: *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*. Ed. by Gustau Camps-Valls, Francisco Ruiz, and Isabel Valera. Virtual Conference: PMLR, 10376–10406.

Hansen, Nikolaus, Auger, Anne, Brockhoff, Dimo, and Tušar, Tea (2022). "Anytime Performance Assessment in Blackbox Optimization Benchmarking". In: *IEEE Transactions on Evolutionary Computation* 26.6, 1293–1305.

Hansen, Nikolaus, Auger, Anne, Ros, Raymond, Finck, Steffen, and Pošík, Petr (2010). "Comparing Results of 31 Algorithms from the Black-Box Optimization Benchmarking BBOB-2009". In: *GECCO '10: Proceedings of the 12th Annual Conference on Genetic and Evolutionary Computation*. Ed. by Jürgen Branke. New York: Association for Computing Machinery, 1689–1696.

Heitjan, Daniel and Rubin, Donald (1991). "Ignorability and Coarse Data". In: *The Annals of Statistics* 19.4, 2244–2253.

Horn, Viktoria, Hirth, Johannes, Holfeld, Julian, Behmenburg, Jens, Draude, Claude, and Stumme, Gerd (2024). "Disclosing Diverse Perspectives of News Articles for Navigating between Online Journalism Content". In: *Nordic Conference on Human-Computer Interaction*. Ed. by Åsa Cajander, Mikael Wiberg, Annika Waern, Mikael Skov, Laia Vida, and Eike Schneiders. New York: Association for Computing Machinery, 1–14.

Hothorn, Torsten, Leisch, Friedrich, Zeileis, Achim, and Hornik, Kurt (2005). "The Design and Analysis of Benchmark Experiments". In: *Journal of Computational and Graphical Statistics* 14.3, 675–699.

Hu, Qian, Yuan, Zhong, Qin, Keyun, and Zhang, Jun (2023). "A Novel Outlier Detection Approach based on Formal Concept Analysis". In: *Knowledge-Based Systems* 268, 110486.

Jansen, Christoph (2018). *Some Contributions to Decision Making in Complex Information Settings with Imprecise Probabilities and Incomplete Preferences: Theoretical and Algorithmic Results*. Dissertationsschrift, LMU. URL: https://edoc.ub.uni-muenchen.de/22653. (last accessed: 02.03.2025).

— (2025). *Contributions to the Decision Theoretic Foundations of Machine Learning and Robust Statistics under Weakly Structured Information*. Habilitationsschrift, ArXiv:2501.10195. URL: https://arxiv.org/abs/2501.10195. (last accessed: 02.03.2025).

Jansen, Christoph, Nalenz, Malte, Schollmeyer, Georg, and Augustin, Thomas (2023). "Statistical Comparisons of Classifiers by Generalized Stochastic Dominance". In: *Journal of Machine Learning Research* 24.231, 1–37.

Jansen, Christoph, Schollmeyer, Georg, and Augustin, Thomas (2018). "Concepts for Decision Making under Severe Uncertainty with Partial Ordinal and Partial Cardinal Preferences". In: *International Journal of Approximate Reasoning* 98, 112–131.

Jansen, Christoph, Schollmeyer, Georg, Blocher, Hannah, Rodemann, Julian, and Augustin, Thomas (2023). "Robust Statistical Comparison of Random Variables with Locally Varying Scale of Measurement". In: *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*. Ed. by Robin Evans and Ilya Shpitser. Pittsburgh: PMLR, 941–952.

Jansen, Christoph, Schollmeyer, Georg, Rodemann, Julian, Blocher, Hannah, and Augustin, Thomas (2024). "Statistical Multicriteria Benchmarking via the GSD-Front". In: *38th Conference on Neural Information Processing System (NeurIPS 2024)*. Ed.

by Amir Globerson, Lester Mackey, Angela Fan, Ulrich Paquet, Jakub Tomczak, Cheng Zhang, and Lam Nguyen. Vancouver: OpenReview.net.

Jelinek, Frederick, Mercer, Robert, Bahl, Lalit, and Baker, Janet (2005). "Perplexity – A Measure of the Difficulty of Speech Recognition Tasks". In: *The Journal of the Acoustical Society of America* 62.1, 62–63.

Julian, Marc (2001). "The Consequences of Ignoring Multilevel Data Structures in Nonhierarchical Covariance Modeling". In: *Structural Equation Modeling: A Multidisciplinary Journal* 8.3, 325–352.

Kamae, Teturo, Krengel, Ulrich, and O'Brien, George (1977). "Stochastic Inequalities on Partially Ordered Spaces". In: *The Annals of Probability* 5.6, 899–912.

Kelly, Markelle, Longjohn, Rachel, and Nottingham, Kolby (2017). *UCI machine learning repository*. URL: https://archive.ics.uci.edu. (last accessed: 02.03.2025).

Khatri, Minal, Yin, Yanbin, and Deogun, Jitender (2024). "Enhancing Interpretability in Medical Image Classification by Integrating Formal Concept Analysis with Convolutional Neural Networks". In: *Biomimetics* 9.7, 421.

Kotelnikov, Evgeny and Milov, Viktor (2018). "Comparison of Rule Induction, Decision Trees and Formal Concept Analysis Approaches for Classification". In: *Journal of Physics: Conference Series* 1015.3, 032068.

Lebanon, Guy and Mao, Yi (2007). "Non-parametric Modeling of Partially Ranked Data". In: *Advances in Neural Information Processing Systems*. Ed. by John Platt, Daphne Koller, Yoram Singer, and Sam Roweis. Vancouve: Curran Associates, Inc., 1–8.

Lehmann, Erich (1955). "Ordered Families of Distributions". In: *The Annals of Mathematical Statistics* 26.3, 399–419.

Li, Jun and Liu, Regina (2004). "New Nonparametric Tests of Multivariate Locations and Scales Using Data Depth". In: *Statistical Science* 19.4, 686–696.

Liu, Regina (1990). "On a Notion of Data Depth Based on Random Simplices". In: *The Annals of Statistics* 18.1, 405–414.

Mahalanobis, Prasanta (1936). "On the Generalized Distance in Statistics". In: *Proceedings of the National Institute of Science (India)* 2.1, 49–55.

Maier, David (1983). *The Theory of Relational Databases*. Rockville: Computer Science Press.

Mallows, Colin (1957). "Non-null Ranking Models". In: *Biometrika* 44.1/2, 114–130.

Miescke, Klaus and Liese, Friedrich (2008). *Statistical Decision Theory: Estimation, Testing, and Selection*. New York: Springer.

Mohri, Mehryar, Rostamizadeh, Afshin, and Talwalkar, Ameet (2018). *Foundations of Machine Learning. Adaptive Computation and Machine Learning*. Cambridge, London: The MIT Press.

Mosler, Karl and Mozharovskyi, Pavlo (2022). "Choosing Among Notions of Multivariate Depth Statistics". In: *Statistical Science* 37.3, 348–368.

Nagy, Stanislav (2023). "Simplicial Depth and its Median: Selected Properties and Limitations". In: *Statistical Analysis and Data Mining: The ASA Data Science Journal* 16.4, 374–390.

Olson, Randal, La Cava, William, Orzechowski, Patryk, Urbanowicz, Ryan, and Moore, Jason (2017). "PMLB: A Large Benchmark Suite for Machine Learning Evaluation and Comparison". In: *BioData Mining* 10.36, 1–13.

Patil, Channamma and Baidari, Ishwar (2019). "Estimating the Optimal Number of Clusters k in a Dataset Using Data Depth". In: *Data Science and Engineering* 4, 132–140.

Peterson, Martin (2017). *An Introduction to Decision Theory*. Cambridge: Cambridge University Press.

Poelmans, Jonas, Elzinga, Paul, Viaene, Stijn, and Dedene, Guido (2010). "Formal Concept Analysis in Knowledge Discovery: A Survey". In: *Conceptual Structures: From Information to Intelligence*. Ed. by Madalina Croitoru, Sébastien Ferré, and Dickson Lukose. Berlin, Heidelberg: Springer, 139–153.

Pratt, John and Gibbons, Jean (2012). *Concepts of Nonparametric Theory*. New York: Springer.

Ramsey, Frank (1931). "Truth and Probability". In: *The Foundations of Mathematics and Other Logical Essays*. Ed. by R.B. Braithwait. London, New York: Taylor & Francis Group, 98–156.

Rodemann, Julian and Blocher, Hannah (2024). "Partial Rankings of Optimizers". In: *The Second Tiny Papers Track at ICLR 2024, Tiny Papers @ ICLR 2024*. Ed. by Tom Burns and Krystal Maughan. Vienna: OpenReview.net.

Rousseeuw, Peter, Ruts, Ida, and Tukey, John (1999). "The Bagplot: A Bivariate Boxplot". In: *The American Statistician* 53.4, 382–387.

Savage, Leonard (1954). *The Foundations of Statistics*. New York: John Wiley & Sons.

Schneider, Frank, Balles, Lukas, and Hennig, Philipp (2019). "DeepOBS: A Deep Learning Optimizer Benchmark Suite". In: *7th International Conference on Learning Representations (ICLR 2019)*. Ed. by Sergey Levine, Karen Livescu, and Shakir Mohamed. New Orleans: OpenReview.net.

Schollmeyer, Georg (2017a). *Lower Quantiles for Complete Lattices*. Technischer Report, LMU. URL: https://epub.ub.uni-muenchen.de/40448/. (last accessed: 02.03.2025).

— (2017b). *Application of Lower Quantiles for Complete Lattices to Ranking Data: Analyzing Outlyingness of Preference Orderings*. Technischer Report, LMU. URL: https://epub.ub.uni-muenchen.de/40452/. (last accessed: 02.03.2025).

Schollmeyer, Georg, Blocher, Hannah, Jansen, Christoph, and Augustin, Thomas (2023). *On the Analysis of Epiontic Data: A Case Study*. Short Paper at the 13th International Symposium on Imprecise Probabilities: Theories and Applications - ISIPTA 2023. URL: https://isipta23.sipta.org/accepted-papers/short-schollmeyer/. (last accessed: 02.03.2025).

Sen, Amartya (1985). *Commodities and Capabilities*. Amsterdam: North-Holland.

Stevens, Scott (1946). "On the Theory of Scales of Measurement". In: *Science* 103.2684, 677–680.

Tukey, John (1975). "Mathematics and the Picturing of Data". In: *Proceedings of the International Congress of Mathematicians Vancouver*. Ed. by Ralph James. Vancouver: Mathematics-Congresses, 523–531.

Vanschoren, Joaquin, van Rijn, Jan, Bischl, Bernd, and Torgo, Luis (2013). "OpenML: Networked Science in Machine Learning". In: *SIGKDD Explorations Newsletter* 15.2, 49–60.

von Neumann, John and Morgenstern, Oskar (1947). *Theory of Games and Economic Behavior, (2nd edition)*. Princeton: Princeton University Press.

Wu, Fei, Wang, Wanliang, Chen, Jiacheng, and Wang, Zheng (2023). "A Dynamic Multi-objective Optimization Method based on Classification Strategies". In: *Scientific Reports* 13.1, 15221.

Zuo, Yijun and Serfling, Robert (2000). "General Notions of Statistical Depth Function". In: *The Annals of Statistics* 28.2, 461–482.

Attached Contributions

Contribution 1

Hannah Blocher and Georg Schollmeyer (2025). "Data Depth Functions for Non-standard Data by use of Formal Concept Analysis". In: *Journal of Multivariate Analysis* 205, 105372

Journal of Multivariate Analysis 205 (2025) 105372

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Data depth functions for non-standard data by use of formal concept analysis

Hannah Blocher *, Georg Schollmeyer

Department of Statistics, Ludwig-Maximilians-Universität München, Ludwigstr. 33, 80539 München, Germany

ARTICLE INFO

AMS 2020 subject classifications:
primary 62H99
secondary 62G30

Keywords:
Conceptual scaling
Depth function
Formal concept analysis
Generalised Tukey depth
Non-standard data

ABSTRACT

In this article we introduce a notion of depth functions for data types that are not given in standard statistical data formats. We focus on data that cannot be represented by one specific data structure, such as normed vector spaces. This covers a wide range of different data types, which we refer to as non-standard data. Depth functions have been studied intensively for normed vector spaces. However, a discussion of depth functions for non-standard data is lacking. In this article, we address this gap by using formal concept analysis to obtain a unified data representation. Building on this representation, we then define depth functions for non-standard data. Furthermore, we provide a systematic basis by introducing structural properties using the data representation provided by formal concept analysis. Finally, we embed the generalised Tukey depth into our concept of data depth and analyse it using the introduced structural properties. Thus, this article presents the mathematical formalisation of centrality and outlyingness for non-standard data and increases the number of spaces in which centrality can be discussed. In particular, we provide a basis for defining further depth functions and statistical inference methods for non-standard data.

1. Introduction

Data depth functions generalise the concept of *centrality* and *outlyingness* to multivariate data and provide therefore a useful concept to define *nonparametric* and *robust statistical methods*. To achieve this, depth functions denote the center and outlying areas based on an underlying distribution or a data cloud. Moreover, they classify how near other areas are to the outlying or center ones. Since there does not exist one unique perspective on centrality, several different depth functions on $\mathbb{R}^d$ have been developed. Some examples are simplicial depth, see [1], zonoid depth, see [2], or Tukey depth, see [3,4]. To give a systematic basis on the notion of depth functions [5,6] set up properties to mathematically formalise the existing intuition about centrality and outlyingness in $\mathbb{R}^d$. In addition to $\mathbb{R}^d$, the scope of data types on which depth functions are defined increased in the last years. For example, in [7] the authors discuss properties of depth functions on functional data. [8] developed a depth function for total orders. [9] went the next step and defined depth functions for general metric spaces. In recent years, this concept has been used to construct statistical methods and to analyse different application settings. For example, [10,11] developed statistical nonparametric and robust tests using depth functions. Using the robust central-outward order [12] studied anomaly detection. In [13], data depth functions are used to visualise and describe features of hydrologic events given two real world streamflow data sets from Canada.

However, all these depth functions and analyses have in common that they build on spaces that have a strong underlying structure, like Banach spaces. Thus, to apply the concept of depth and resulting statistical methods, the data has to be embedded into statistical standard data formats, like numeric or spatial. The aim of this article is to consider depth functions without assuming one

* Corresponding author.
E-mail address: hannah.blocher@stat.uni-muenchen.de (H. Blocher).

https://doi.org/10.1016/j.jmva.2024.105372
Received 30 August 2023; Received in revised form 2 August 2024; Accepted 16 September 2024
Available online 19 September 2024

specific data type in advance. This includes data types like, e.g., set of partial orders or mixed spatial and ordinal data. We call data that cannot be embedded into statistical standard data formats *non-standard data*. To give a notion of centrality and outlyingness for non-standard data, a unified, flexible and general applicable data representation is needed. Therefore, we use the theory of *formal concept analysis* which transforms the data set into a *closure system* on the entire data set itself. In particular, it does not force the user to make assumptions about the data that are not necessarily true, simply for the sake of mathematical feasibility. We use this unified representation of the data to generally define depth functions for non-standard data. To clarify the notion of centrality and outlyingness for data represented via formal concept analysis we introduce *structural properties*. Thereby, we use the structure of the closure system to transfer the properties given by [5,6] and develop further properties representing the connections between the data points. In this context, we investigate that, unlike $\mathbb{R}^d$, non-standard data can have data elements that are naturally outlying or central based on the structure of the underlying space. Finally, we embed the *generalised Tukey depth* given in [14] into our concept of depth functions for non-standard data. Moreover, we use the provided structural properties to analyse the generalised Tukey depth. Note that, analogous to the generalised Tukey depth, the concrete outlier measures developed in [15] can also be analysed using the structural properties introduced here. These outlier measures are also based on formal concept analysis. Therefore, only a slight adaptation to the concept of depth introduced here is necessary.

This article is intended to give a systematic basis on depth functions for non-standard data. By introducing structural properties we give a notion of centrality and outlyingness. Furthermore, we provide a framework to analyse depth functions for non-standard data and start a discussion on the notion of centrality for such kind of data.

The structure of this article is as follows: We begin with concrete motivation examples. Since formal concept analysis is the basis of our general notion of a depth function, we give a short introduction to the theory of formal concept analysis in Section 3. Based on this, Section 4 introduces the general definition of depth functions for non-standard data by use of formal concept analysis. There we define the Tukey depth within our concept of depth functions. The next section states the structural properties and their restriction strength. Afterwards, we analyse the generalised Tukey depth using structural properties. In Section 7, we collect our concluding remarks.

2. Motivation

The purpose of this section is to present motivating examples to show that depth functions based on formal concept analysis cover a wider range of data types than the currently discussed depth functions known to the authors. Instead of using observations directly to define a depth function, we first apply formal concept analysis. This gives us a definition of depth functions that can be applied to a wide variety of data types simultaneously. Most importantly, it is simple for the practitioner to use, as it only relys on the definition of a cross-table, as will be seen in this section.

One of the goals of formal concept analysis is to discover relational connections between observations. Therefore, the observations are grouped according to the observed values. For example, consider the Titanic data set, where each passenger's age, sex and the passenger's class of travel is recorded. One group might be all passengers aged between 20 and 40 who are female. Another smaller group are all female passengers between the ages of 20 and 40 travelling in first class. In this case, the smaller group is more specific, e.g. in the smaller group the passengers also fulfil the condition that they all travel in first class. Formal concept analysis formalises the arrangement into different groups. The first step is to define a formal context that represents the data set. In a nutshell, a formal context is a generalisation of a cross-table.

To provide an intuitive introduction to the main concepts of formal concept analysis and to illustrate the motivation for defining depth functions based on formal concept analysis, we give now two examples. These examples also demonstrate the variety of different data types that can be represented by a formal context.

Example 1. For the Titanic data example provided by the Titanic machine learning competition, see [16], we get the formal context in Table 1 (this is just a snippet). Here each row represents an observation/passenger (i.e. called *objects* in formal concept analysis). The columns are binary *attributes* that can be either true or false. Transforming the non-binary observations (such as age, spatial values, etc.) into binary attributes is called *conceptual scaling*, see [17, p. 36ff]. Since the number of attributes can be infinite, this is a flexible, but also accessible approach to data representation. For classical data types there are already standard scaling methods, see [17, p. 36ff]. As an example, consider Table 1. The first two columns represent the sex of the passenger (f: female and m: male). The next three columns are the class level of the passenger (I, II and III). These two nominal observations are included via *nominal scaling*, see [17, p. 42]. The remaining columns represent the age of the passenger, e.g. g_1 is 34.5 years old. The corresponding attributes are the statements "$\leq x$" and "$\geq x$" for all $x \in \mathbb{R}$. This scaling method is called *interordinal scaling*, see [17, p. 42]. Note that this scaling method implies that there exists no object that has every attribute. The crosses in Table 1 indicate that the attribute applies to the passenger/observation. How different observation values are represented as a set of binary attributes depends on the scaling method used, see [17, p. 36ff].

We obtain the groups that represent the relationship between the observations by considering all possible combinations of attributes and summarising all the observations that agree on all these attributes. Note that these groups may overlap. They also have a certain mathematical structure, i.e. they show the implications between the passengers/observations. For example, in Table 1 we see that if g_3 and g_4 are in a group, we can directly imply that g_2 is also in the group, since there is no group with g_3 and g_4 that does not also contain g_2. By collecting all these implications, we can fully determine the arrangement of the groups given by the formal context. For further details see Section 3.

Table 1
This formal context contains five observations/passengers of the Titanic. Each row represents a passenger. The data is from the Titanic machine learning competition, see [16], with PassengerIds 892, 893, 894, 904 and 908. Note that the age 34.5 is due to the encoding notation. With "…", we denote that the attribute set is indeed infinite and not all attributes are denoted within the table. The same is true for the objects as this is only a snippet of the titanic data set.

	f	m	I	II	III	…	'≤23'	…	'≤24'	…	'≤34.5'	…	'≤35'	…	'≤47'
g_1		x			x						x	…	x	…	x
g_2	x				x										x
g_3	x			x											
g_4	x		x				x	…	x	…	x	…	x	…	x
g_5		x		x									x	…	x
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

	…	'≤67'	…	…	'≥67'	…	'≥47'	…	'≥35'	…	'≥34.5'	…	'≥24'	…
g_1	x	x	x								x	x	x	x
g_2	x	x	x				x	x	x	x	x	x	x	x
g_3		x	x		x	x	x	x	x	x	x	x	x	x
g_4	x	x	x											
g_5	x	x	x						x	x	x	x	x	x
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

We emphasise that the choice of binary attributes, as well as the decision which observation has an attribute and which does not, results in a concrete, easily accessible representation of the data. In particular, it exposes many implicit assumptions about the data. Moreover, the data are represented in such a way that we do not impose assumptions that are not present in the data, e.g. that the data can be embedded in a normed vector space. However, we want to point out that the choice of attributes can have a strong influence on the arrangement of the groups and therefore on the depth values. Instead of using interordinal scaling for the age component in the Titanic data set, another approach called ordinal scaling, which includes only the upper bound '$\leq x$', can be used, see [17, p. 42]. With this smaller formal context, age can only influence the groups through its upper bound. Thus, instead of looking at all passengers between the ages of, e.g., 20 and 40, with this smaller formal context, only all passengers with an age smaller than or equal to 40 are grouped without a lower bound.

We now use the structure given by the different groups and the resulting implications to give a notion of depth to the observations. In the Titanic example, we want to rank the five passengers according to how central or outlying they are compared to the other passengers. In the spirit of [18], in this paper we introduce properties that represent the notion of centrality and outlyingness. For example, Property (P1) states that if we use different scaling methods of the observations (i.e., use different binary attributes in the formal context) that lead to the same groups, the depth function should stay the same. As another example, consider the relationship between g_2, g_3, and g_4. Since any group that includes g_3 and g_4 also includes g_2, we have that g_2 has all the properties that g_3 and g_4 share. So we can say that g_2 is at least as specific as g_3 and g_4 combined. Following the idea of the quasiconcavity property in $\mathbb{R}^d$, see [6], Property (P7i and ii) states that the depth of g_2 must be at least as high as the minimum depth of g_3 or g_4. Other properties discuss that non-standard data may inherit a natural center-outward order, which is not present in $\mathbb{R}^d$. For example, Property (P4) states that an object lying in every group must be central, regardless of the probability measure.

Example 2. Another example that illustrates the wide range of data types that can be discussed using formal concept analysis is the set of partial orders. A data example for partial orders is given in [19], where the objects are partial orders of classifiers. In that article, we considered a set of classifiers applied to a set of data sets. These classifiers were evaluated for each individual data set by several performance measures. Then an algorithm i outperforms another algorithm j on a given data set if and only if there exists one performance measure that states that algorithm i is better than algorithm j and all other performance measures agree that algorithm i is not worse than algorithm j. This gives us a partial order for each single data set that represents the performance structure of the classifiers for that data set. Thus, a benchmark suite of 80 data sets results in 80 observed partial orders. Now, to represent this kind of data, in [14] we introduced a formal context for partial orders in general. Here, each row corresponds to a partial order. The columns of the formal context are the binary attributes needed to describe the partial order. In [14], the first $n \cdot (n-1)$, where n is the number of classifiers/items being compared, denotes all possible pairs where we have stated that one classifier dominates the other. The next $n \cdot (n-1)$ rows indicate the reverse, i.e. that one classifier does not dominate another. Note that including the second part as binary attributes is by no means straightforward. But as in [14,19] we decided to state that non-existence of dominance is indeed a precise observation and we do not want a push towards the linear extensions of the partial orders (which exists when deleting the second $n \cdot (n-1)$ columns), we have therefore included this part.

As can be seen in the two examples above, and also in Examples 4 and 5 below, defining a formal context gives us a flexible tool for evaluating data ranging from mixed nominal and ordinal observations to spatial data or partial orders. These examples demonstrate that formal concept analysis clearly reveals the relationships between data points, providing a simple and unified view of the underlying data structure. Finally, we want to highlight differences and connections between our approach and other existing approaches. In contrast to [8,9], we have completely turned away from a metric approach in our definition of depth functions. [8] discusses total orders and therefore a translation to our approach is straightforward, just restrict the space of all partial orders to the total orders, see Example 2. This is not so clear for [9] as the representation via a formal context depends on the concrete structure of the space. Example 5 gives an approach for $\mathbb{R}^d$ equipped with the Euclidean distance.

3. Formal concept analysis

This section briefly describes the parts of formal concept analysis (FCA) required for this article and illustrates them using the two examples above. It is based on [17]. For further readings, we refer to [20]. FCA was developed by Rudolf Wille, Bernhard Ganter and Peter Burmeister to build a bridge between mathematical lattice theory and applied users. It enables the analysis of relationships between the data points by representing the data in a unified and user-friendly manner.

The fundamental definition of FCA is the representation of a data set as a cross-table, see [17, p. 17].

Definition 1. A *formal context* $\mathbb{K}$ is a triple (G, M, I) with G being the *object* set, M the set of *attributes* and $I \subseteq G \times M$ a binary relation between G and M.

In our case, the objects are the data points and the attributes are characteristics of these data points. The relation I then states whether an object g has an attribute m, if $(g, m) \in I$, or not, if $(g, m) \notin I$. Thus, these attributes need to be binary-valued, whether they occur or not. Naturally, there exist characteristics of the data points which are many-valued, like sex or age. To include these many-valued characteristics as well into the formal context, we use *conceptual scaling methods*, see [17, p. 36ff] and Section 2, which transfers many-valued characteristics into a set of binary-valued attributes. By using scaling methods, we can represent a large variety of different data sets through a formal context. This allows us to transfer the most diverse data types into a uniform structure. Also, data sets which are not given in standard statistical data formats. We call such data *non-standard data*. Examples are given in Examples 1, 2, 4 and 5. The scaling method can also be used to reduce data complexity. This can lead to a conceptual scaling error, see for example [21].

Based on this user-friendly representation of the data set by a formal context, we can now define so-called *derivation operators*, see [17, p. 18]:

$$\Psi : 2^G \to 2^M, A \mapsto A' := \{m \in M \mid \forall g \in A : (g, m) \in I\},$$
$$\Phi : 2^M \to 2^G, B \mapsto B' := \{g \in G \mid \forall m \in B : (g, m) \in I\}.$$

The function Ψ maps a set of objects A onto every attribute which every object in A has. So, $A' = \Psi(A) = \cap\{m \in M \mid (g, m) \in I\}$ is the maximal set of attributes that every object in A has. The reverse, from attribute set to object set, is provided by the function Φ. We set $\Phi(\emptyset) = G$ and $\Psi(\emptyset) = M$. The composition of these two functions $\gamma := \Phi \circ \Psi : 2^G \to 2^G$ gives us then a family of sets which denotes the relationship between the data points. We have $\gamma(A) = \Phi \circ \Psi(A) = \cap_{m \in \Psi(A)}\{g \in G \mid (g, m) \in I\}$. In other words, every object set $E \subseteq G$ which is an element of the codomain of $\gamma = \Phi \circ \Psi$ is the maximal set of objects which have all the same attributes $\Psi(E)$ in common. Thus, the composition groups all those objects together that have the same attributes. With slight abuse of notation, for any $m \in M$ we write $\Phi(m)$ for $\Phi(\{m\})$. The same convention holds for $\Psi(g)$ and $\gamma(g)$ with $g \in G$.

Definition 2. The set $\gamma(2^G)$ is called the set of *extents*, see [17, p. 18]. The set $\gamma(A)$ is mentioned as *extent set*, shortly *extent*, of $A \subseteq G$.

To take advantage of this slightly different representation of the data set as a family of sets, we use that γ defines a *closure operator*, see [17, p. 8]. In what follows, we use the term *closure* always in the context of a closure operator or a closure system. When referring to a closed set based on a topology, metric, or norm we denote this by *topological(-ly) closed/closure*.

Definition 3. A closure operator $\gamma : 2^G \to 2^G$ is defined as a function on a power set to itself. A closure operator needs to be extensive (i.e. for all $A \subseteq G$, $A \subseteq \gamma(A)$), isotone (i.e. if $A \subseteq B \subseteq G$, then $\gamma(A) \subseteq \gamma(B)$) and idempotent (i.e. for all $A \subseteq G$, $\gamma(A) = \gamma(\gamma(A))$).

In particular, a closure operator always induces a *closure system* $\gamma(2^G)$. A closure system $\mathcal{S} \subseteq 2^G$ is a family of sets which contains the entire space (i.e. $G \in \mathcal{S}$) and any intersection of sets in $\mathcal{S}$ is again in $\mathcal{S}$ (for all $S \subseteq \mathcal{S}$ with $S \neq \emptyset$ we have $\bigcap_{s \in S} s \in \mathcal{S}$).

Note that there exists a one-to-one correspondence between the closure system and the closure operator. Since $\gamma = \Phi \circ \Psi$ is a closure operator, the set of extents is a closure system. Thus, the closure operator γ describes the closure system and vice versa. For more details on closure systems see [17, Chapter 0.3].

Example 3. Recall the Titanic Example 1. We obtain $\Psi(g_2) = \{\text{f}, \text{III}, \text{'}\leq 47\text{'}, \dots, \text{'}\leq 67\text{'}, \dots, \text{'}\geq 47\text{'}, \dots, \text{'}\geq 35\text{'}\}$. Therefore, we get $\gamma(g_2) = \{g_2\}$. One can show that the set of extents $\gamma\left(2^{\{g_1, g_2, g_3, g_4, g_5\}}\right)$ equals

$$\{\emptyset, \{g_1\}, \{g_2\}, \{g_3\}, \{g_4\}, \{g_5\}, \{g_1, g_4\}, \{g_1, g_5\}, \{g_2, g_5\}, \{g_2, g_3\}$$
$$\{g_1, g_4, g_5\}, \{g_1, g_2, g_5\}, \{g_2, g_3, g_5\}, \{g_1, g_2, g_4, g_5\}, \{g_1, g_2, g_3, g_5\}, \{g_1, g_2, g_3, g_4, g_5\}\}.$$

Let us take a closer look at how the closure operator describes the connection between the data points. This is the basic idea of how we will later use the closure operator to define structural properties for depth functions. As we pointed out above the closure operator groups data points together which have certain attributes in common. This means if $a \in \gamma(A) \backslash A$ the object a has all attributes which every object in A has as well. Thus, one can say that A *implies* a based on the relationship structure given by the formal context which is then included in the definition of γ. Therefore using the closure system or closure operator (both describe the same since they have a one-to-one correspondence) to define the structural properties of the depth function illustrates the relationship between the data points. For further details on implications see [17, Chapter 2.3]. Note that in [17] attribute implications are discussed and we focus here on object implications. Nevertheless, the concepts can be transferred to object implications.

Table 2
Minimal example of a hierarchical nominal scaling with two levels (1, 2) and two categories (a, b, respectively).

	a_1	b_1	a_1a_2	a_1b_2	b_1a_2	b_1b_2
a_1a_2	x		x			
a_1b_2	x			x		
b_1a_2		x			x	
b_1b_2		x				x

Example 4. Now, we introduce a further scaling method, the so-called *hierarchical nominal scaling*. This scaling method is inspired by the occupations of persons within a social survey. Usually occupations are categorised within a hierarchy of different levels, for example within the International Standard Classification of Occupations (ISCO) of 2008, see https://www.ilo.org/publications/international-standard-classification-occupations-2008-isco-08-structure (accessed: 24.09.2024). On a first level occupations are split into different categories ($a_1, b_1, \ldots$). For example category a_1 could be "Managers", category b_1 "Professionals", etc. Each of these categories is then split again on a more fine-grained level (Level 2) into further subcategories. In this case, category a_1 is split into a_1a_2 "Managers: Chief executives, senior officials, and legislators" and a_1b_2 "Managers: Administrative and commercial managers" and so on. Note that the Level 2 subcategories based on the first level b_1 split do not have to match those of the Level 2 splits based on the Level 1 a_1 split. Subsequently, the Level 2 categories are again subdivided into subcategories, and so on. Such data structure can be conceptually scaled in a natural way. For every level, we introduce attributes describing every single category based on the higher level classification, see Table 2. Here, for simplicity, we used only two levels with two categories, respectively.

Now, let us take a closer look at the extents given by Table 2 which are

$$\{\emptyset, \{a_1a_2\}, \{a_1b_2\}, \{b_1a_2\}, \{b_1b_2\}, \{a_1a_2, a_1b_2\}, \{b_1a_2, b_1b_2\}, G\}.$$

Furthermore, we obtain that $a_1a_2 \in \gamma(\{a_1b_2, b_1a_2\}) = \Phi\circ\Psi(\{a_1b_2, b_1a_2\}) = \Phi(\emptyset) = G$.

Example 5. In the next sections, we want to transfer the idea of depth functions from $\mathbb{R}^d$ to general non-standard data which are represented by a formal context. Before looking at this, we now show how one can represent the elements in $\mathbb{R}^d$ as objects of a formal context. We consider $\mathbb{R}^d$ together with the topology induced by the Euclidean norm. The scaling method introduced here is inspired by [22,23]. Let $G = \mathbb{R}^d$ be the object set and the attribute set $M = \{H \subseteq \mathbb{R}^d \mid H \text{ topologically closed halfspace}\}$. Define the relation I between M and G by $(g, H) \in I \Leftrightarrow g \in H$, and let $\mathbb{K} = (G, M, I)$ denote the corresponding formal context.

With this definition, let us consider $\gamma(2^G)$ induced by $\mathbb{K}$. Let $A \subseteq G = \mathbb{R}^d$, then $\Psi(A)$ are all halfspaces containing every object/point in A. Further on, $\gamma(A) = \Phi\circ\Psi(A)$ are then every object/point which lies in every halfspace in $\Psi(A)$. Thus $\gamma(A)$ is the intersection of all halfspaces in $\Psi(A)$ and therefore a topologically closed convex set. More generally, one can show that for every topologically closed convex set in $2^{\mathbb{R}^d}$ there exists a set $A \subseteq G = \mathbb{R}^d$ such that this convex set equals $\gamma(A)$. Thus, γ is the convex closure operator on $\mathbb{R}^d$ and the extent set $\gamma(2^G)$ is the set of all topologically closed convex sets.

More concretely, assume that $d = 2$ and $f \in \gamma(\{a, b, c\})$. This means that f lies in every topologically closed halfspace which contains also a, b and c. In other words, f shares the same attributes as a, b, c share. Thus, one can say that f is implied by a, b and c based on γ. If $e \notin \gamma(\{a, b, c\})$, there exists a halfspace which contains a, b and c but not e. To summarise, an object g lies in $\gamma(\{a, b, c\})$ if and only if g lies within the triangle given by the vertices a, b and c. This shows how the concrete definition of the closure system enhances the connection between single objects.

Finally, as we have now introduced several different data situations with different scaling methods, see Example 1 (nominal, numeric data), Example 2 (partial orders), Example 4 (hierarchical nominal data) and Example 5 (spatial data), we want to point out the importance of the scaling method used. First of all, deciding which observations (age, spatial, etc.) to include in the analysis is a general issue in statistics. At first sight, FCA may seem to add a further difficulty by requiring the definition of binary attributes. However, FCA provides a clear understanding of the relationships between observations. Thus, to check whether the scaling method represents the objective, one can look at all the extents as well as all the implications. If the groups do not represent an important fact about the data (e.g. the non-dominance part of the partial orders or the lower bound of the age value), then more attributes are needed. Conversely, if the extents are close to the power set of G, one should examine the data set to see if there really are no implications between the data points. If so, some sets of attributes need to be replaced or even deleted. While nominal, ordinal and interordinal scaling are commonly used scaling methods, it is important to discuss carefully for each data set how fine/coarse the arrangement of the groups should be and to consider several different scaling methods.

4. Definition of depth functions for non-standard data using formal concept analysis

Our aim in this section is to give a general definition of data depth functions for non-standard data using FCA. By representing the data points G via a formal context, with G being the object set, we obtain a unified structure that is not tailored to one specific data type. In particular, the newly provided depth function allows to analyse a large variety of different data types on centrality and outlyingness issues. With this, nonparametric methods can be developed for all these data.

The depth function presented here only specifies the domain and codomain but not the exact mapping rule. Thus, the structural properties presented later, see Section 5, can be seen as generic properties for this kind of depth functions.

Definition 4. We define a depth function using FCA by

$$D_G : G \times \varkappa_G \times \mathscr{P}_G \to \mathbb{R}_{\geq 0}$$

for a fixed set of objects G and a set of formal contexts $\varkappa_G \subseteq \{\mathbb{K} \mid G \text{ is object set of } \mathbb{K}\}$. $\mathscr{P}_G$ is a set of probability measures on G defined on a σ-field which contains all extent sets of formal contexts in $\varkappa_G$.

Thus, we compute the depth of an object set based on a probability measure and formal context representing the object relationships. We want to emphasise that G, $\varkappa_G$ and $\mathscr{P}_G$ depend on each other. Note that Definition 4 allows the restrictions of $\varkappa_G$ and $\mathscr{P}_G$ to a subset of all possible formal contexts or probability measures, which is sometimes necessary. For example, assume that $G = [0,1]$, then for every subset of G there exists a formal context such that this subset is an extent. Thus, in this case, a restriction to a subset of contexts is necessary if we want to allow the uniform distribution to be an element of $\mathscr{P}_G$. This follows from [24, Chapter 1.1] which shows that there cannot exist a probability measure on $[0,1]$ which formalises the intuition of volume and has the entire power set of $[0,1]$ as input. Another aspect is that restriction can lead to a set of formal contexts fulfilling additional structural requirements. With this, it can be possible to define mapping rules which are not possible in general. (See Section 5: Property (P8) does not hold for every formal context, but only for a subset.) Another example is given in [14,19] where we considered one single formal context on the set of partial orders, see also Example 2. There, we used the structure given by this concrete formal context to define the mapping rule. The same reasoning can be applied to the probability set $\mathscr{P}_G$. Thus, for a proper definition of the depth function, not only the exact mapping rule is important, but the considered formal contexts and probability measures as well.

A data set can be represented by different attribute sets and corresponding binary relations. Thus, Definition 4 can lead to many different depth functions even if the object set G, the probability $\Pr \in \mathscr{P}_G$ and a concrete mapping rule are specified. Hence, the choice of formal context and scaling method can have a huge impact on the depth values, i.e. this can also be seen in Section 5.

The empirical depth function corresponds to the depth function in Definition 4 with the empirical probability measure as input. To ensure that the empirical depth function is well defined we need to assume that every empirical probability measure $\Pr_G^{(n)}$ of every probability measure $\Pr_G \in \mathscr{P}_G$ is again an element of $\mathscr{P}_G$.

Definition 5. Let G be a set and $\varkappa_G \subseteq \{\mathbb{K} \mid G \text{ is object set of } \mathbb{K}\}$ a set of formal contexts on G. We assume that $\mathscr{P}_G$ consists of probability measures that are defined on a σ-field containing all extents of $\varkappa_G$. Furthermore, for every sample $g_1, \dots, g_n$ based on a probability measure $\Pr_G \in \mathscr{P}_G$, we have $\Pr_G^{(n)} \in \mathscr{P}_G$, where $\Pr_G^{(n)}$ is the empirical probability measure based on the sample. Then the empirical depth function for a sample $g_1, \dots, g_n$ with corresponding empirical probability measure $\Pr_G^{(n)}$ is given by

$$D_G^{(n)} : G \times \varkappa_G \to \mathbb{R}_{\geq 0}, (g, \mathbb{K}) \mapsto D_G(g, \mathbb{K}, \Pr_G^{(n)}).$$

Serving as an example, we consider the generalised Tukey depth, based on [22,23] and introduced in [14]. The Tukey depth on $\mathbb{R}^d$ of a point $g \in \mathbb{R}^d$, c. f. [3,4], is the smallest probability of a halfspace containing g. To build the bridge to FCA, we consider the formal context $\mathbb{K}$ with the object set $G = \mathbb{R}^d$, attribute set $M = \{H \subseteq \mathbb{R}^d \mid H \text{ topologically closed halfspace}\}$ and binary relation I with $(g, H) \in I$ if and only if $g \in H$, see Example 5. Based on this the Tukey depth for a point $g \in \mathbb{R}^d$ and probability measure $\Pr_{\mathbb{R}^d}$ on $\mathbb{R}^d$ can be written as

$$\inf_{H \in \mathcal{H}(g)} \Pr_{\mathbb{R}^d}(H) = \inf_{H \in \Psi(g)} \Pr_{\mathbb{R}^d}(H) = 1 - \sup_{H \in \Psi(g)} \Pr_{\mathbb{R}^d}(G \setminus H) = 1 - \sup_{H \in M \setminus \Psi(g)} \Pr_{\mathbb{R}^d}(H) = 1 - \sup_{m \in M \setminus \Psi(g)} \Pr_{\mathbb{R}^d}(\Phi(m)) \tag{1}$$

where $\mathcal{H}(g)$ is the set of all topologically closed halfspaces containing g and Φ, Ψ correspond to the derivation operators given by the formal context $\mathbb{K}$ defined in Example 5. The first and the last equality are translations into FCA language, where importantly the final right hand side does not involve the notion of a halfspace, which allows the generalisation of the Tukey depth to any arbitrary FCA setting. The third equality holds because $G \setminus H$ is a topologically open halfspace and the supremum does not change when considering topologically open halfspaces instead. The right hand side of Eq. (1) will now be the basis of the generalisation of Tukey depth to arbitrary formal contexts. Before we proceed, we shortly indicate, why we do not use directly the left hand side of Eq. (1): Generally, the topologically closed convex sets in $\mathbb{R}^d$ correspond to the extents within our approach to use FCA to define data depth functions. On the other hand, the topologically closed halfspaces of $\mathbb{R}^d$ have no general natural equivalent in FCA. On the left hand side, if we replace the family of topologically closed halfspaces with the family of all extents, we obtain a depth value of zero for all $g \in \mathbb{R}^d$ when, e.g., the probability measure is continuous w.r.t. the Lebesgue measure. This is of course unsatisfying. Moreover, note that unlike halfspaces, the complement of extents in FCA are generally not extents. This further separates the left and the right hand side of Eq. (1). Therefore, we take the right hand side of Eq. (1). As will be shown later in Theorem 6 of Section 6, taking the supremum over all halfspaces or taking it over all topologically closed convex sets in $\mathbb{R}^d$ gives the same result, which also translates to the generalised Tukey depth where the supremum over all extents and the supremum over all extents generated by one single attribute coincide.

Additionally, the supremum of the right hand side of Eq. (1) has a natural interpretation as a measure of outlyingness that can be expressed in the language of FCA, see [23, Section 2] and [22, Section 5]. In these articles, the author constructs special representative extents: For given $\alpha \in [0,1]$ we say that an extent is α-extensive if it contains at least a proportion of $\alpha \cdot 100\%$ of data points or probability mass. Of course, for given α there are many such extents but one can take for one α the intersection of all these extents. This intersection is again an extent which is then in a certain sense a representative extent w.r.t. a level α. The outlyingness of a point g is then given by the (empirical) probability mass of the most specific depth contour (i.e., the most specific representative extent that corresponds to the smallest possible α) that still contains g. A slightly different, but order-theoretically equivalent, definition is to take the least specific depth contour not containing g. This is exactly what is expressed by $\sup_{m \in M \setminus \Psi(g)} \Pr_{\mathbb{R}^d}(\Phi(m))$ in Eq. (1). With this motivation (and with the insight of Theorem 6) we get as generalised Tukey depth:

Definition 6. Let G be a set, $\varkappa_G$ a subset of formal contexts and $\mathscr{P}_G$ a subset of probability measures on G. Assume that $\varkappa_G$ and $\mathscr{P}_G$ are defined as in Definition 4. The generalised Tukey depth is given by

$$T_G : \begin{array}{l} G \times \varkappa_G \times \mathscr{P}_G \to [0,1], \\ (g, \mathbb{K}, \Pr_G) \mapsto 1 - \sup_{m \in M \setminus \Psi(g)} \Pr_G(\Phi(m)) \end{array}$$

where $\mathbb{K}$ defines the operator Φ and Ψ. We set $\sup_{t \in \emptyset} f(t) := 0$ for every function f.

Note that here, we take the supremum only over all extents generated by one single attribute, as this gives the same result as taking the supremum over all extents (c.f., Theorem 6) and is at the same time easier to compute. Additionally, from this definition it becomes clear that the (generalised) Tukey depth does not dependent on the dependence structure between the attributes, but only on the marginal distribution of the attributes. Observe that the term marginal is meant here as the probability of lying in a certain halfspace and should not be confused with the marginal distribution of points in $\mathbb{R}^d$ in the sense of the distribution of one coordinate. In the sequel, we will always refer to the marginal distribution as the distribution of the attributes.

Furthermore, note that $\varkappa_G$ is not restricted to any subset and, in particular, $\Pr_G$ is only restricted by $\varkappa_G$, see Definition 4. The second part of the mapping rule in Definition 6, $\sup_{m \in M \setminus \Psi(g)} \Pr_G(\Phi(m))$, corresponds to the supremum of the probabilities of the events which consists of all objects having an attribute the object of interest does not have. Thus, if $g \in G$ is an object that has all the attributes which occur, then g has a maximal depth of value one. Note, however, that there are usually no objects that have all attributes. For example, consider the spatial context described in Example 5, where the generalised Tukey depth corresponds to the well-known Tukey depth on $\mathbb{R}^d$. Thus, in this situation, the (generalised) Tukey depth is bounded by 0.5 and, in particular, strictly below 1. Another example is the interordinal scaling described in Section 2.

Now, we define the empirical version. Analogously to Definition 5, let $g_1, \dots, g_n$ be a sample of G based on $\Pr_G \in \mathscr{P}_G$. Then the empirical generalised Tukey depth function corresponds to T_G by inserting the corresponding empirical probability measure $\Pr_G^{(n)}$. This gives us:

Definition 7. Let G, $\varkappa_G$ and $\mathscr{P}_G$ be defined as in Definition 5. Let $(g_1, \dots, g_n)$ be a sample from G according to $\Pr_G \in \mathscr{P}_G$ with associated empirical measure $\Pr_G^{(n)}$. Then the empirical generalised Tukey depth is given by

$$T_G^{(n)} : \begin{array}{l} G \times \varkappa_G \to \mathbb{R}_{\geq 0}, \\ (g, \mathbb{K}) \mapsto 1 - \sup_{m \in M \setminus \Psi(g)} \sum_{\tilde{g} \in \Phi(m)} \Pr_G^{(n)}(\tilde{g}) = 1 - \frac{1}{n} \sup_{m \in M \setminus \Psi(g)} \#\Phi(m). \end{array}$$

As derived above, the generalised Tukey depth applied to the spatial context defined in Example 5 gives us the well-known Tukey depth on $\mathbb{R}^d$, see [3]. In particular, on $\mathbb{R}^1$ we get the ranking given by the classical quantiles, with the median being the central one and the depth values decreasing outwards.

Example 6. Recall Examples 1 and 3. Let us assume that the probability measure $\Pr_G$ in $G = \{g_1, g_2, g_3, g_4, g_5\}$ is uniform and the formal context $\mathbb{K}$ is given by Table 1. Then the generalised Tukey depths are

$$T_G(g_2, \mathbb{K}, \Pr_G) = 1 - \sup_{m \in M \setminus \Psi(g_2)} \Pr_G(\Phi(m)) = 1 - \frac{3}{5} = \frac{2}{5} = T_G(g_1, \mathbb{K}, \Pr_G) = T_G(g_5, \mathbb{K}, \Pr_G),$$

$$T_G(g_4, \mathbb{K}, \Pr_G) = 1 - \sup_{m \in M \setminus \Psi(g_4)} \Pr_G(\Phi(m)) = 1 - \frac{4}{5} = \frac{1}{5} = T_G(g_3, \mathbb{K}, \Pr_G).$$

Example 7. In Example 2 we briefly described the formal context in which partial orders are represented. See also [14,19]. Recall that the attributes represent the existence or non-existence of a dominance structure between two different algorithms i and j. Thus, each pair of algorithms i and j gives rise to four binary attributes: First, whether i dominates j or vice versa, and second, whether i does not dominate j and vice versa. As can be seen from the definition, the generalised Tukey depth only considers the marginal distribution of the attributes. Thus, based on the formal context describing partial orders, the generalised Tukey depth relies only on the probability of how often dominance between two classifiers occurs and how often it does not. More specifically, if we consider a sample, then the marginal distribution is given by the proportion of observed dominance between two classifiers, or the proportion of observed non-dominances. In the following, we omit the case where all partial orders have exactly the same probability, see [14] for details on this. This assumption implies that the partial order(s) with the minimum depth value are those that do not fulfil a (non-)dominance structure that has the highest proportion. There is no such simple answer for the maximum depth partial order. This is due to the transitivity assumption on the dominance structure. A partial order with highest depth value is one that summarises the most dominance and non-dominance structures that are most often observed.

Example 8. Recall the nominal hierarchical data structure discussed in Example 4. Note that the groups are already disjoint after the first partition and are then successively partitioned into smaller subsets. In the following we assume that the probability mass is not strongly focused on a small subset of possible observations, see the proof of Theorem 10 for details. Since the generalised Tukey depth includes only the maximum marginal probability of the attributes that are not true for the object of interest, we can observe that only the first partition is used for the generalised Tukey depth. Therefore, dropping all other attributes (representing the finer splits) and thus considering a smaller and less informative formal context does not change the generalised Tukey depth based on this nominal hierarchical formal context. In other words, the generalised Tukey depth ignores the further information of the data.

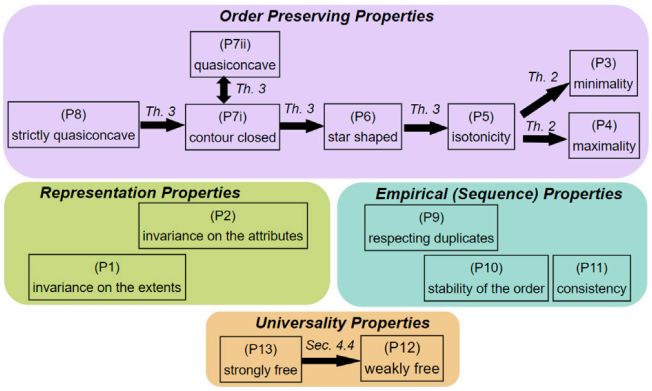

Fig. 1. Overview of the structural properties together with their mathematical connections. *Th* is Theorem and *Sec* is Section for short. This overview contains all direct implications between the properties. Further results like, e.g., limitations of the properties, see, e.g., Theorem 4, are not presented in the Figure.

Before presenting the structural properties, we define when two depth functions are isomorph and thus represent the same center-outward order.

Definition 8. Let G be a set and D_G and $\tilde{D}_G$ be two depth functions based on $\varkappa_G$ and $\mathscr{P}_G$, $\tilde{\varkappa}_G$ and $\tilde{\mathscr{P}}_G$ respectively, see Definition 4. Let $\mathbb{K} \in \varkappa_G$, $\tilde{\mathbb{K}} \in \tilde{\varkappa}_G$, $\Pr_G \in \mathscr{P}_G$ and $\tilde{\Pr}_G \in \tilde{\mathscr{P}}_G$. Then $D_G(\cdot, \mathbb{K}, \Pr_G)$ and $\tilde{D}_G(\cdot, \tilde{\mathbb{K}}, \tilde{\Pr}_G)$ on G are *isomorph* if and only if there exists a bijective and bimeasureable function $i : G \to G$ such that

$$D_G(g, \mathbb{K}, \Pr_G) \leq D_G(\tilde{g}, \mathbb{K}, \Pr_G) \iff \tilde{D}_G(i(g), \tilde{\mathbb{K}}, \tilde{\Pr}_G) \leq \tilde{D}_G(i(\tilde{g}), \tilde{\mathbb{K}}, \tilde{\Pr}_G)$$

is true for all $g, \tilde{g} \in G$. We call a bijective function *bimeasurable* if and only if i and i^{-1} are measurable w.r.t. the corresponding σ-fields. In what follows, $\cong$ denotes the isomorphism between two depth functions.

For simplicity of notation, we write D, $D^{(n)}$, T, $T^{(n)}$, $\varkappa$, $\mathscr{P}$, Pr and $\Pr^{(n)}$ instead of D_G, $D_G^{(n)}$, T_G, $T_G^{(n)}$, $\varkappa_G$, $\mathscr{P}_G$, $\Pr_G$ and $\Pr_G^{(n)}$ in the following if the underlying object set G is clear.

5. Structural properties characterising the depth function using formal concept analysis

After this general definition of depth functions based on FCA, we want to discuss some *structural properties* a depth function can have. With this, we tackle the question of what centrality is and which object is - in some sense - *closer* to the center than another object. In particular, we provide concepts to discuss centrality and outlyingness for different data types without necessarily presupposing one specific data structure. Thus, this section gives a starting point for a discussion on centrality, outlyingness and depth functions based on FCA. Furthermore, we define a framework upon which newly introduced depth functions can be studied and compared.

For normed vector spaces, there is already an ongoing discussion about this. The authors of [1,5,6,25] are concerned about these questions for data depth functions defined on $\mathbb{R}^d$. Furthermore, [7,26] discuss depth functions and centrality topics for functional data. All these examples have in common that they are based on a normed vector space and that the clarification of centrality and outlyingness is done by defining properties. For example, [5,6] demand that the depth of a point $x \in \mathbb{R}^d$ should converge to zero as the norm of x, i.e. $\|x\|$, tends to infinity. This reflects the intuition about outlyingness in unbounded spaces. In contrast to outlyingness, the definition of a center point does not immediately follow. Basically, every point in $\mathbb{R}^d$ can be the center. This follows from the fact that after translation, rotation, etc. the structure of a normed vector space does not change. The center generally seems to be only naturally specified in special cases. For example [5] assumes that for a probability measure that is symmetric around a point c, see [27], this point c should then be the center. In the case of a multivariate normal distribution, the center point is the mean vector. Since a depth function should represent a center-outward order, this point c needs to have a maximal depth value. Together with the center, one needs to discuss what it means that one point is further away from the center than another. In $\mathbb{R}^d$, this is achieved by the use of line segments, see [5,6]. A point p_1 which lies on the line segment between the center and a further point p_2 is said to be closer to the center than p_2. In other words, p_2 is more outlying than p_1. At the first glance, here we also use the normed vector space structure.

It follows from the above that the definition of the terms centrality and outlyingness seem to highly rely on the underlying distribution. In particular, the part defining the center point which is based on a symmetric distribution, see [27]. Slightly different to [5,27] is the approach of [6]. The properties given by [6] put emphasis on the structure of the underlying spaces and not on the probability measure. Thus, in contrast to e.g., [5,27], the definition of centrality of a point is more detached from the notion of symmetry/centrality of the underlying probability measure. Note that [6] gives an even stronger definition of depth functions. In the following, we discuss the term centrality and outlyingness from the perspective of the approach in [6].

In this article, we consider a space G where the structure of the data points is given by a formal context. In the style of [5] and [6] for depth functions in $\mathbb{R}^d$, we define *structural properties* of a depth function based on FCA to clarify the notion of centrality and outlyingness. These properties express characteristics of the data set when represented via a formal context and how these characteristics are included in the depth function. Some of the structural properties build upon the existing ones in $\mathbb{R}^d$ that we transfer to our new situation. To achieve this, we use that the set of all topologically closed convex sets is a natural closure system on $\mathbb{R}^d$. Since the extent set also defines a closure system, we obtain a direct translation to FCA by representing the properties by use of the convex closure system on $\mathbb{R}^d$. For example, we transfer the idea of a line segment to our data representation via a formal context, see Property (P6). Moreover, the "quasiconcavity" property, see [6, p. 19] has a natural translation to FCA by use of the closure system, see Property (P7i and ii). Other structural properties introduced discuss the possibility that the data itself may have a center-outward order. This inherited order is then represented by the formal context. For example, consider the extreme case that an object lies in every set of the closure system. This means that it has every attribute that is true for any element of the data set. Thus, this object is implied by every other object and is therefore more specific than any other object. So it should have maximum depth for any possible probability measure, see Property (P4). Of course, the other extreme case, where an object lies only in the extent corresponding to the entire set and therefore shares no attributes with the other data elements, should have minimal depth by default, see Property (P3). Note that these two extremes do not necessarily occur. For example, there is no such element in $\mathbb{R}^d$ together with the convex sets. All in all, one can say that the introduced structural properties are defined and discussed within two perspectives: The transfer of already existing properties in $\mathbb{R}^d$ and the new development of properties based on the theory of FCA, reflecting the inherited center-outward structure of the data.

In total 13 structural properties are presented which can be covered under four different categories: Representation properties, order preserving properties, empirical (sequence) properties and universality properties. An overview of the structural properties together with their mathematical connections can be seen in Fig. 1. In what follows, we fix the set of objects G and consider the depth function $D : G \times \varkappa \times \mathscr{P} \to \mathbb{R}_{\geq 0}$. Furthermore, we refer to Definition 4 when we say that D is a depth function. We use the term empirical depth function when considering a depth function in the style of Definition 5. Afterwards, in Section 6 we check these properties based on the introduced generalised Tukey depth, see Section 4.

5.1. Representation properties

Depth functions satisfying the following properties are structure preserving on G. This includes two parts: First, assume that we represent the data set by two formal contexts and that we consider two probability measures. Let us assume that between the corresponding extent sets and therefore the structure of the objects exists a bijective function. Furthermore, if additionally the probability values are preserved by this bijective function, then the depth function should be preserved as well. In other words, the order given by the depth function should not rely on the used scaling method unless it does substantially change the structure of the extent sets. Also, the influence of the probability measure is only based on the extent sets. In Section 2 we motivated this property by illustrating that the observation values are represented within the arrangement of the different groups of the passengers/observations. Thus, if neither the groups nor the probabilities of these groups change, then the depth value should also remain the same. This can be seen as an adaptation of the "affine invariance" property of depth functions in $\mathbb{R}^d$, see [5, p. 463]. There, the depth function equals the depth of the shifted version if the probability measure is shifted accordingly.

The second part of the representation properties considers the attribute values. Here, we say that the depth function preserves the structure if and only if two objects with the same attributes have the same depth value. From the perspective of FCA, two objects with the same attributes are duplicates. Therefore they should be assigned to the same depth value. In the formal context of the Titanic example, see Example 1, this means that two passengers of exactly the same age, sex and class of travel must have the same depth values.

(P1) *Invariance on the extents:* Let $\mathbb{K}, \tilde{\mathbb{K}} \in \varkappa$ be two formal contexts on G and let $\Pr, \tilde{\Pr} \in \mathscr{P}$ be two probability measures on G. If there exists a bijective and bimeasureable function $i : G \to G$ such that the extents are preserved (i.e. E extent w.r.t. $\mathbb{K}$ $\Leftrightarrow$ $i(E)$ extent w.r.t. $\tilde{\mathbb{K}}$) and the probability is also preserved (i.e. $\Pr(E) = \tilde{\Pr}(i(E))$), then

$$D(\cdot, \mathbb{K}, \Pr) \cong D(\cdot, \tilde{\mathbb{K}}, \tilde{\Pr})$$

is true.

(P2) *Invariance on the attributes:* For every $\mathbb{K} \in \varkappa$, $\Pr \in \mathscr{P}$ and $g_1, g_2 \in G$ with $\Psi(\{g_1\}) = \Psi(\{g_2\})$,

$$D(g_1, \mathbb{K}, \Pr) = D(g_2, \mathbb{K}, \Pr)$$

holds.

Remark 1. Property (P1) is the only property that directly discusses the connection between two different formal contexts. This is due to the fact that even just adding more attributes can strongly change the implications and the whole data structure representation, even though this procedure "only" adds more extent sets. Therefore, in general, it is not clear how the depth function should behave when more attributes are added. However, for example for interordinal scaling, which is "finer" than ordinal scaling, see Example 1, it can be shown that the following Property (P7i and ii) together with Property (P1) in many cases already cover the desired changes in the order of the data elements. More precisely, for ordinal scaling, Property (P7i and ii) ensures that the smallest

value has maximum depth, while for interordinal scaling it gives more flexibility and, depending on the underlying probability measure, any element can have the highest depth value. In particular, in many cases (e.g. uniform distribution) Properties (P1) and (P7i and ii) imply that the classical median has the highest depth. For the generalised Tukey depth, we always obtain the classical median as the highest depth for interordinal scaling.

5.2. Order preserving properties

While the representation properties ensure that similar structures on G lead to the same depth function, the next *order preserving properties* consider the obtained order by the depth function. These properties increase in their strength of restriction.

These properties are defined along the lines of "monotonicity relative to the deepest point" and "maximality at the center" properties, see [5, p. 463], and the "quasiconcavity" properties, see [6, p. 19], defined for $\mathbb{R}^d$. In contrast to $\mathbb{R}^d$ we neither have a norm nor a concept of translation and symmetry, but we can make use of the closure system and closure operator given by the extent set. When considering the properties defined for $\mathbb{R}^d$ in the context of the convex sets which define a closure system, we can build a bridge to FCA. In particular, the "quasiconcavity" property, see [6, p. 19], has therefore a natural adaptation. Nevertheless, the convex sets are a special case of closure sets: For example, the affine invariance is reflected in the set since by shifting the set of convex sets the family of sets does not change. This does not hold in general for closure systems. For example, there exist closure systems where one point occurs more often in the sets of the closure system than another point. We use the concept of a formal context where two natural extreme opposite characteristics of the objects can occur. The first one: If an object has every attribute then it lies in every extent set. Conversely, if an object has no attribute at all, then the only extent containing this object is the entire set G. Property (P3) and (P4) now ensure that these two opposite characteristics are also reflected in the depth function.

(P3) *Minimality:* Let $\mathbb{K} \in \varkappa, \Pr \in \mathscr{P}$. Further, let $g_{non} \in G$ such that for every extent $E \subsetneq G$ of $\mathbb{K}$ we have that $g_{non} \notin E$, then

$$D(g_{non}, \mathbb{K}, \Pr) = \min_{g \in G} D(g, \mathbb{K}, \Pr)$$

is true.

(P4) *Maximality:* Let $\mathbb{K} \in \varkappa, \Pr \in \mathscr{P}$. Assume there exists $g_{all} \in G$ such that for every extent E of $\mathbb{K}$ we have that $g_{all} \in E$. Then

$$D(g_{all}, \mathbb{K}, \Pr) = \max_{g \in G} D(g, \mathbb{K}, \Pr)$$

holds.

Note that a depth function which fulfils these two properties does not rely on the probability measure to set the values of the two extreme cases. On the other hand, there exist formal contexts such that the objects g_{non} or g_{all} do not exist at all. Recall Example 5 where we considered the spatial data.

Nevertheless, since we have now predefined the maximal depth value in specific cases, this has to be in line with adapting properties like "monotone on rays" or "quasiconcavity", see [6, p. 19]. In what follows we start with less restricting properties and increase their restriction strength slowly. We show that they imply Properties (P4) and (P3).

Property (P5) is inspired by the fact that in FCA, an object g_2 which lies in the closure of an other object g_1 implies that this object is more specific than g_2. In other words, g_2 has all attributes and possibly even more attributes than g_1. This is analogous to the assumption that $\gamma(\{g_1\}) \supseteq \gamma(\{g_2\})$ is true. Thus, Property (P5) says that an object g_2 must have a depth value as least as high as g_1.

(P5) *Isotonicity*: For every $\Pr \in \mathscr{P}$ and formal context $\mathbb{K} \in \varkappa$ with $g_1, g_2 \in G$ such that $\gamma_{\mathbb{K}}(\{g_1\}) \supseteq \gamma_{\mathbb{K}}(\{g_2\})$,

$$D(g_1, \mathbb{K}, \Pr) \leq D(g_2, \mathbb{K}, \Pr)$$

is fulfilled.

There exists a natural strengthening of the isotonicity property (P5) which leads to the adaptation of the "monotonicity relative to deepest point", see [5, p. 463], property in $\mathbb{R}^d$. For start, let us assume that the depth function is bounded from above. Furthermore, we assume that the depth function has its maximum at center $c \in G$. The "monotone on rays" in $\mathbb{R}^d$ definition in [6, p. 19] states that the depth of a point that moves further away from the center c on a fixed ray should decrease. Since we do neither have a norm nor a vector space, we cannot generally define what *further away* as well as *ray* means. Thus, we translate the definition to a setting using the convex closure operator instead. In this case, if a point/object $\tilde{g}$ lies on the line segment given by a further point g and the center c, then the depth of $\tilde{g}$ must be at least as high as the depth of g. In other words, when a point $\tilde{g}$ lies in the convex closure of another point g and the center, then we have a lower bound for the depth of $\tilde{g}$. Thus, the points/objects with depth values larger or equal to a fixed value form a starshaped set. This gives the name of the property. With the definition based on the convex closure system, we can easily transfer this property to FCA and obtain Property (P6).

(P6) *Starshapedness*: Let $\mathbb{K} \in \varkappa$ and $\Pr \in \mathscr{P}$. If there exists at least one center point $c \in G$ such that for all $g \in G$ and all $\tilde{g} \in \gamma_{\mathbb{K}}(\{c, g\})$ we have

$$D(\tilde{g}, \mathbb{K}, \Pr) \geq D(g, \mathbb{K}, \Pr),$$

then we call D starshaped.

In the above explanation, we assumed that the center point c has a maximal depth value. This assumption is not included in the definition of Property (P6) as the boundedness of the depth function, as well as the maximality at the center, follow directly. Let us fix $c \in G$ to be one center point. Since for every closure operator the isotonicity assumption (see Section 3, not to be confused with Property (P5)) holds, we get that for every object g we have $c \in \gamma(\{c, g\})$. Thus, by Property (P6) for every $g \in G$ we obtain

$$D(c, \mathbb{K}, \Pr) \geq D(g, \mathbb{K}, \Pr).$$

With this, c must have the highest depth value. Since the depth function maps to $\mathbb{R}$, this gives us also the upper bound of the depth function. Note that a center c is not always naturally given. Besides the spatial data, see Example 5, where one can use the probability measure to get a center, in the framework of FCA one can say that g_{all} in Property (P4) could be such a center. In contrast, in the example of the Titanic data, see Table 1, there is no passenger who is naturally a center. Each passenger has at least one component where their observation is not in the center (median for age and modus for sex and class of travel).

Furthermore, the starshaped property (P6) together with the invariance on the extents property (P1) has some implications on the point of maximal depth when the underlying probability measure has some symmetry property. As already indicated at the beginning of Section 5, because of the lack of a translation operation, etc., it is difficult to define symmetry in our setting. However, one can still define some notion of point symmetry, see the next theorem.

Theorem 1. *Let $\mathbb{K} \in \varkappa$ and $\Pr \in \mathscr{P}$. We assume that $\Pr$ is point symmetric around $s \in G$ in the following sense: There exists a bimeasurable involutory function $i : G \longrightarrow G$ (i.e. $i(i(g)) = g$ for all $g \in G$) such that*

1. *$\Pr(i(E)) = \Pr(E)$ for all extents $E \subseteq G$ and*
2. *for all $g \in G$ we have $s \in \gamma(\{g, i(g)\})$.*

Then for every depth function which fulfils Properties (P1) and (P6), the center of symmetry s of $\Pr$ has maximal depth.

Proof. Let c be one center point. First note that by the second assumption of this theorem we get $s \in \gamma(\{c, i(c)\})$. Because D is starshaped, we have $D(s, \mathbb{K}, \Pr) \geq D(i(c), \mathbb{K}, \Pr)$. Now, we use that there exists a bimeasurable involutory function i and that D is invariant on the extents (P1) to get $D(i(c), \mathbb{K}, \Pr) = D(c, \mathbb{K}, \Pr)$. With this, we conclude that $D(s, \mathbb{K}, \Pr) \geq D(c, \mathbb{K}, Pr)$. This proves that s has maximal depth. □

Note that object s of Theorem 1 is not necessarily a center point in the style of Property (P6).

These properties are written down in their order of strength. More precisely, Property (P6) implies Property (P5) and Property (P5) implies Properties (P4) and (P3), see Theorem 2. Thus, if Property (P6) is satisfied and if there exists an object g which has every attribute, then c must be one of the center point c discussed in Property (P6).

Theorem 2. *Let $\mathbb{K} \in \varkappa$ and $\Pr \in \mathscr{P}$. Let D be a depth function then the following implications hold for D:*

1. *Let $c \in G$ be a center object. If D satisfies Property (P6), then Property (P5) is true for D.*
2. *If D satisfies Property (P5), then Properties (P4) and (P3) are true for D.*

Proof. Assume that $\mathbb{K} \in \varkappa$ and $\Pr \in \mathscr{P}$. Let D be a depth function.

We begin by proving that Property (P6) implies Property (P5). Therefore assume that Property (P6) is true for D and let $c \in G$ be one center point. Further let $g_1, g_2 \in G$ such that $\gamma_{\mathbb{K}}(\{g_1\}) \supseteq \gamma_{\mathbb{K}}(\{g_2\})$. Then we get that $g_2 \in \gamma(\{c, g_1\})$ and therefore we have $D(g_1, \mathbb{K}, \Pr) \leq D(g_2, \mathbb{K}, \Pr)$.

The next step is to show that Property (P5) implies Property (P4). Assume that Property (P5) is satisfied and that g_{all} lies in every extent set. Then, we get for every $g \in G$ that $\gamma_{\mathbb{K}}(\{g_{all}\}) \supseteq \gamma_{\mathbb{K}}(\{g\})$ holds. Thus, $D(g_{all}, \mathbb{K}, \Pr) \geq \max_{\{g \in G\}} D(g, \mathbb{K}, \Pr)$ is true and Property (P4) follows.

Finally, we show that Property (P3) follows from Property (P5). Let $g_{non} \in G$ be an object which lies only in the entire set and in no other extent set. Furthermore, we suppose that Property (P5) is true for D. Since $\gamma(\{g_{non}\}) = G$ for every $g \in G$ we have $\gamma(\{g\}) \subseteq \gamma(\{g_{non}\})$. Due to Property (P5) we follow that $D(g_{non}, \mathbb{K}, \Pr) \leq D(g, \mathbb{K}, \Pr)$ for every $g \in G$. This gives us Property (P3). □

Remark 2. Let us point out some consequences of Property (P6). In contrast to [5, p. 463] we do not assume that the center c is unique. For example, the depth function is allowed to have a plateau at the highest point. In particular, we allow the depth function to be constant. Especially g_{non} can only be the center point if the function is constant, due to Properties (P3), (P4) and Theorem 2. Moreover, we get that when the depth function has at least two different values, then g_{non} must have the minimal value and g_{all} the maximal value. This allows us to specify a center point and an outlying point without relying on the other observed points/objects by defining the scaling method. Note that this situation does not occur often. For example, all scaling methods described in Section 3 do not have an object g_{non} and g_{all}.

Nevertheless, this stresses the importance of a meaningful and carefully chosen scaling method. Observe that the difference between having an attribute and not having an attribute is not symmetric and cannot be switched without eventually fundamentally changing the characteristic of the corresponding closure system.

The next order preserving properties are in the style of the "quasiconcavity" property, see, e.g., [6, p. 19]. In the context of $\mathbb{R}^d$ quasiconcavity states that the set of all points with a larger (or equal) depth value than $\alpha \geq 0$ needs to be a convex set. These

sets are called contour sets. In this case, there is a direct transfer to FCA by the extent set as a closure system. Therefore, we first have to define the contour sets within the theory of FCA. Let $\mathbb{K} \in \varkappa$ be a formal context and $\Pr \in \mathscr{P}$ be a probability measure. For $\alpha \in \mathrm{im}(D(\cdot, \mathbb{K}, \Pr))$ the contour set $Cont_\alpha$ is defined as follows

$$Cont_\alpha := \{g \in G \mid D(g, \mathbb{K}, \Pr) \geq \alpha\}.$$

Now, we say the depth function is countourclosed if every contour set is an extent set. This is stated in Property (P7i). Instead of considering the contour sets, an analogous statement for Property (P7i) is given in Property (P7ii). Here, we assure that the depth of an observation that lies in the closure of an input set is larger or equal to the infimum of the input. Property (P7ii) has a natural strengthening by assuming strict inequalities instead, see Property (P8).

(P7i) *Countourclosed*: For every formal context $\mathbb{K} \in \varkappa$, probability measure $\Pr \in \mathscr{P}$ and every $\alpha \in \mathrm{im}(D(\cdot, \mathbb{K}, \Pr))$ the contour set $Cont_\alpha$ is an extent of the formal context $\mathbb{K}$.

(P7ii) *Quasiconcave*: Let $\mathbb{K} \in \varkappa$ and $\Pr \in \mathscr{P}$. If for all $A \subseteq G$ and all $g \in \gamma_{\mathbb{K}}(A) \setminus A$ we have

$$D(g, \mathbb{K}, \Pr) \geq \inf_{\tilde{g} \in A} D(\tilde{g}, \mathbb{K}, \Pr),$$

we call D quasiconcave.

(P8) *Strictly quasiconcave*: Let $\mathbb{K} \in \varkappa$ and $\Pr \in \mathscr{P}$. If for all $A \subseteq G$ and all $g \in \gamma_{\mathbb{K}}(A) \setminus A$ we have

$$D(g, \mathbb{K}, \Pr) > \inf_{\tilde{g} \in A} D(\tilde{g}, \mathbb{K}, \Pr),$$

D is strictly quasiconcave.

Recall Example 1. We showed that a group containing g_3 and g_4 must also contain g_2. Property (P7i and ii) now formalises the statement in Example 1 that the depth of g_2 is at least as high as the minimum depth of g_3 and g_4. In Example 6 we saw that this is true for the generalised Tukey depth. In the next section we will prove that Property (P7i and ii) holds in general for the generalised Tukey depth, but that this is not true for Property (P8).

Again these properties have mathematical connections to each other. First of all, Property (P7i) and (P7ii) are equivalent. Secondly, Property (P8) is indeed stronger and implies Property (P7i and ii). Thirdly, if c is a center point with maximal depth, then with this c we can show that (P7i and ii) imply the starshaped property (P6). Moreover, Property (P7ii) implies Property (P5).

Theorem 3. *Let $\mathbb{K} \in \varkappa$ be a formal context and $\Pr \in \mathscr{P}$ a probability measure. Let $D(\cdot, \mathbb{K}, \Pr)$ be a depth function.*

1. *Statement (P7i) and (P7ii) are equivalent.*
2. *Property (P8) implies (P7ii).*
3. *If their exists $c \in G$ with maximal depth value, then Property (P7ii) implies (P6) with c being a center object.*
4. *Property (P7ii) implies (P5).*

Proof. Let $\mathbb{K} \in \varkappa$ and $\Pr \in \mathscr{P}$. Let $D(\cdot, \mathbb{K}, \Pr)$ a depth function. Note that the claim that Property (P8) implies (P7ii) is given immediately. Thus, we only show Part 1., 3. and 4. of Theorem 3.

1. First assume that Property (P7ii) is true and let $\alpha \in \mathrm{im}(D(\cdot, \mathbb{K}, \Pr))$ be arbitrary. We prove that Property (P7i) follows. Assume, by contradiction that $Cont_\alpha$ is not an extent set. Since γ is a closure operator and therefore idempotent, there exists $g \in \gamma(Cont_\alpha) \setminus Cont_\alpha$. Since Property (P7ii) is true, $D(g, \mathbb{K}, \Pr) \geq \inf_{\tilde{g} \in Cont_\alpha} D(\tilde{g}, \mathbb{K}, \Pr) \geq \alpha$ holds. This contradicts $g \notin Cont_\alpha$ and we get that Property (P7i) is fulfilled.
   For the reverse let (P7i) be true and let $A \subseteq G$ be arbitrary. We set $\alpha = \inf_{\tilde{g} \in A} D(\tilde{g}, \mathbb{K}, \Pr)$ and we know that $A \subseteq Cont_\alpha$. By (P7i) we know that $Cont_\alpha$ is an extent set. Since $\gamma(A)$ is the smallest extent set containing A and the set of extents is a closure system, we follow that $\gamma(A) \subseteq Cont_\alpha$. Thus the depth of every object $g \in \gamma(A)$ must be larger or equal to α which implies (P7ii).
3. Now, we assume that Property (P7ii) is true and we show that Property (P6) holds as well. Therefore, we assume that $c \in G$ has maximal depth value. Let $g, \tilde{g} \in G$ such that $\tilde{g} \in \gamma(\{c, g\})$ is true. Due to Property (P7ii) we get $D(\tilde{g}, \mathbb{K}, \Pr) \geq \min\{D(c, \mathbb{K}, \Pr), D(g, \mathbb{K}, \Pr)\} \geq D(g, \mathbb{K}, \Pr)$. The last inequality follows from the assumption that c has maximal depth value. This gives us Property (P6) with c being one center object.
4. Finally, assume that Property (P7ii) holds. Let $g_1, g_2 \in G$ with $g_1 \neq g_2$ such that $\gamma(g_2) \supseteq \gamma(g_1)$. Then the quasiconcavity property implies that $D(g_1, \mathbb{K}, \Pr) \geq D(g_2, \mathbb{K}, \Pr)$. This shows Property (P5). □

Property (P8), the strictly quasiconcavity, is indeed a very strong assumption. In particular, there exist formal contexts such that Property (P8) can never be fulfilled. An example can be found in [14] and Example 2. There, we discussed the special case of depth functions for partial orders and analysed Properties (P7i and ii) and (P8) in this case. We showed that the quasiconcavity cannot be fulfilled by the formal context of partial orders. Another example is given by Table 3 (left). Here, let $G = \{g_1, g_2, g_3\}$ and let $\Pr$ be an arbitrary probability measure. In contradiction, let us assume that Property (P8) is true for a depth function D on G. We have $\gamma(\{g_i, g_j\}) = \{g_1, g_2, g_3\}$ for all $i, j \in \{1, 2, 3\}$ with $i \neq j$. With this, we obtain that $\min\{D(g_1, \mathbb{K}, \Pr), D(g_2, \mathbb{K}, \Pr)\} < D(g_3, \mathbb{K}, \Pr)$. W.l.o.g. let $D(g_1, \mathbb{K}, \Pr) < D(g_3, \mathbb{K}, \Pr)$. Together with $\min\{D(g_2, \mathbb{K}, \Pr), D(g_3, \mathbb{K}, \Pr)\} < D(g_1, \mathbb{K}, \Pr)$ due to Property (P8), we have $D(g_2, \mathbb{K}, \Pr) < D(g_1, \mathbb{K}, \Pr) < D(g_3, \mathbb{K}, \Pr)$. However, this is a contradiction to $\min\{D(g_1, \mathbb{K}, \Pr), D(g_3, \mathbb{K}, \Pr)\} < D(g_2, \mathbb{K}, \Pr)$. Hence,

Property (P8) cannot hold for this context. In contrast, the quasiconcavity property (P7ii) allows equality which solves such problems. Note that in the case of the formal context given by Table 3 (left), Property (P7ii) leads to a constant depth function.

More generally, one can say that the strictly quasiconcavity assumption on D can never be satisfied if one formal context $\mathbb{K} \in \varkappa$ is an element of the following set:

$$\mathscr{C}^{\not P8} = \left\{ (G, M, I) \text{ formal context} \;\middle|\; \begin{array}{l} \text{exist } A, \tilde{A} \subseteq G \text{ such that } \#A < \infty, \#\tilde{A} < \infty \text{ and } A \cap \tilde{A} = \emptyset \\ \text{and } A \subseteq \gamma(\tilde{A}),\ \tilde{A} \subseteq \gamma(A) \end{array} \right\}.$$

Theorem 4 proves that for every $\mathbb{K} \in \mathscr{C}^{\not P8}$ there exists no strictly quasiconcave depth function. Here, we get a contradiction with the strict larger assumption. This case occurs naturally when the scaling method assigns to two different objects the same attribute values.

Theorem 4. *For every $\mathbb{K} \in \mathscr{C}^{\not P8}$ and every $\Pr \in \mathscr{P}$ there exists no depth function D such that Property (P8) is fulfilled.*

Proof. Let $\mathbb{K} \in \mathscr{C}^{\not P8}$. We assume that there exists a depth function D and a probability measure Pr such that Property (P8) is satisfied. Since $A \subseteq \gamma(\tilde{A})$ and $\tilde{A} \subseteq \gamma(A)$ we get $\forall a \in A \setminus \tilde{A} : D(a, \mathbb{K}, \Pr) > \inf_{\tilde{a} \in \tilde{A}} D(\tilde{a}, \mathbb{K}, \Pr)$ and $\forall \tilde{a} \in \tilde{A} \setminus A : \inf_{a \in A} D(a, \mathbb{K}, \Pr) < D(\tilde{a}, \mathbb{K}, \Pr)$. Furthermore, we assumed that $\#A < \infty, \#\tilde{A} < \infty$ is true. Thus, the infimum is attained in A and $\tilde{A}$. Let $a_m \in A$ be an argument of the minimum of $D(a, \mathbb{K}, \Pr)$, and analogously we set $\tilde{a}_m \in \tilde{A}$. Since $A \cap \tilde{A} = \emptyset$, we obtain that $\tilde{a}_m \in \tilde{A} \setminus A$ and $a_m \notin A \setminus \tilde{A}$. But with this we get $D(a_m, \mathbb{K}, \Pr) > D(\tilde{a}_m, \mathbb{K}, \Pr)$ and $D(a_m, \mathbb{K}, \Pr) < D(\tilde{a}_m, \mathbb{K}, \Pr)$. This cannot be true which contradicts the assumption of Property (P8) being true. □

We want to point out that A and $\tilde{A}$ being finite is crucial since in the proof we used that the infimum is attained and not only the largest lower bound of the sets. Secondly, the intersection of A and $\tilde{A}$ being empty is also necessary since else one can set the elements in the intersection to the minimal depth value. This is then following Property (P8).

While the importance of a meaningful scaling method for Properties (P3) and (P4) is easily seen, this comes also into account for the last properties. For example Property (P7ii) and (P8) state that the depth of an object g which is implied by a set A, i.e. $g \in \gamma(A)$, must have larger (or equal) depth than the minimal depth value of the elements in A. In the context of FCA, one can say that g contains all characteristics/similarities having the objects in A in common or even more. Therefore it is at least as specific (or more) than all the elements in A together. Thus, when defining the scaling method, not only the individual attribute values should be taken into account, but also which object combination is more specific than others.

All these order preserving properties aim at representing the underlying data structure in the depth function. This includes structures that are also valid for $\mathbb{R}^d$, e.g. quasiconcavity, but also the extreme cases where a natural center or outlying point is given by the data structure itself, see maximality and minimality property. This emphasises that non-standard data can be very different from $\mathbb{R}^d$, and these properties aim to preserve the different structures. In the next section, we focus on the set of probability measures $\mathscr{P}$ directly.

5.3. Empirical (sequence) properties

In this section, we consider the reverse of the above section. For a fixed formal context $\mathbb{K}$ we are interested in the behaviour of the depth function when the (empirical) probability measure changes. In what follows, we regard different empirical probability measures $\Pr^{(n)}$ induced by different samples $g_1, \dots, g_n \in G$ with $n \in \mathbb{N}$. We assume that $\Pr^{(n)} \in \mathscr{P}$ holds. Hence, in this section, we consider the empirical depth function, recall Definition 5.

The first two properties discuss how two empirical depth functions differ if the two corresponding samples differ in a specific manner. The first one considers the influence of duplicates in the sample in comparison to deleting the duplicates. In this case, the depth value of the duplicated object should be higher based on the sample where the duplicates exists. The second property studies the impact of one single sample element on the resulting center-outward order. Here we consider two empirical probability measures. The first empirical probability measure is based on a sample that has an object which greatly differs from the other object. The second empirical probability measure is based on the same sample but without this greatly different object. An object g_{diff} differs greatly from the other objects if this object has no attribute that any other object has. Or, equivalently, g_{diff} and a further object of the sample g_s are both elements of an extent E if and only if $E = G$ is true. Thus, this object g_{diff} should have no impact on the center-outward order of the other objects. Note that there are scaling methods that exclude the existence of an object g_{diff}. In the Titanic example, see Table 1, we used the interordinal scaling method to include the numerical observation. This tells us that for two objects/passengers there must be at least one attribute that both objects have in common. Conversely, if ordinal scaling is used instead in the Titanic example, object g_{diff} may occur.

(P9) *Respecting duplicates*: Let $\mathbb{K} \in \varkappa$. Let $g_1, \dots, g_n$ be a sample of G with $n \in \mathbb{N}$. Assume that there exist g_i and g_j in the sample with $i \neq j$ which have identical attribute set (so $\Psi(g_i) = \Psi(g_j)$). We set

- $\Pr^{(n)}$ to be the empirical probability measure of $g_1, \dots, g_n$ with $\Pr^{(n)} \in \mathscr{P}$, and
- $\Pr^{(n,-i)}$ to be the empirical probability measure $g_1, \dots, g_{i-1}, g_{i+1}, \dots, g_n$ (without g_i) and $\Pr^{(n,-i)} \in \mathscr{P}$.

Then

$$D^{(n,-i)}(g_j, \mathbb{K}) < D^{(n)}(g_j, \mathbb{K}).$$

(P10) *Stability of the order*: Let $\mathbb{K} \in \varkappa$ and let $g_1, \dots, g_n$ be a sample of G with $n \in \mathbb{N}$. Assume that there exists an $i \in \{1, \dots, n\}$ such that g_i is greatly different from all other objects $g_1, \dots, g_{i-1}, g_{i+1}, \dots, g_n$ of the sample. This means that the only extents which contain g_i and a further object g_j with $j \in \{1, \dots, i-1, i+1, \dots, n\}$ is G. Then the center-outward order of $g_1, \dots, g_{i-1}, g_{i+1}, \dots, g_n$ given by $D^{(n)}$ and $D^{(n,-i)}$ are the same. Recalling Definition 8 and the notation in Property (P9), we get

$$D^{(n)}(\cdot, \mathbb{K})\Big|_{\{g_1,\dots,g_{i-1},g_{i+1},g_n\}} \cong D^{(n,-i)}(\cdot, \mathbb{K})\Big|_{\{g_1,\dots,g_{i-1},g_{i+1},g_n\}}.$$

Let us take a look at Definition 5 of the empirical depth function. There, the empirical probability measure is used for the definition. Thus, we start with a discussion on how duplicates or greatly different object effect the empirical probability measure of the extents. First, let us consider the case of Property (P9) where two objects g_1 and g_2 are duplicates. This means that they have the same attributes and therefore lie in the same extent set. If the invariance on the attribute Property (P2) is fulfilled, then g_1 and g_2 have the same depth. Furthermore, the probability reflects the duplicates for every extent which contains g_1 (and therefore also g_2).

Let us now consider Property (P10) and let $g_1, \dots, g_n$ be a sample with $g_i = g_{diff}$ the greatly different object w.r.t. the rest of the sample. One can observe that there exist three distinctive cases on how the empirical probability measures differ on $E \subsetneq G$ extent:

- Case 1: $g_{diff} \in E$. Then $g_1, \dots, g_i, g_{i+1}, \dots, g_n \notin E$ and $\Pr^{(n)}(E) = 1/n \neq 0 = \Pr^{(n,-i)}(E)$.
- Case 2: $g_\ell \in E$ for some $\ell \in \{1, \dots, i-1, i+1, \dots, n\}$. Then $g_{diff} \notin E$ and $\Pr^{(n)}(E) = ((n-1)/n)\Pr^{(n,-i)}(E)$.
- Case 3: $g_\ell \notin E$ for all $\ell \in \{1, \dots, n\}$. Then $\Pr^{(n)}(E) = 0 = \Pr^{(n,-i)}(E)$.

Thus, Cases 2 and 3 show that the order of the extents which do not contain g_{diff} based on the probability measure $\Pr^{(n)}$ is the same as for $\Pr^{(n,-i)}$. Hence, the order structure given by the probability values of extents containing $g_1, \dots, g_i, g_{i+1}, \dots, g_n$ is not influenced by g_{diff}.

These are observations on the empirical probability measure, but we only generally defined the (empirical) depth function. Thus, the question if a depth function fulfils these properties is concerned with how these connections between the empirical probability distributions are included in the precise definition of the mapping rule.

Properties (P9) and (P10) focus on two empirical probability measures and how their difference influences the empirical depth function. We end this subsection by discussing the behaviour of the depth function based on a sequence of empirical probability measures. Let $(\Pr^{(n)})_{n\in\mathbb{N}}$ be a sequence of empirical probability measures, with $\Pr^{(n)}$ being given by an independent and identical distributed (i.i.d.) sample of $\Pr \in \mathscr{P}$ with size $n \in \mathbb{N}$. By assumption we have that for all $n \in \mathbb{N}$, $\Pr^{(n)} \in \mathscr{P}$. Property (P11) discusses the consistency based on the empirical depth function towards the (population) depth function.

(P11) *Consistency*: Let $\mathbb{K} \in \varkappa$ and $\Pr \in \mathscr{P}$ be a probability measure on G. Let $\Pr^{(n)}$ be the empirical probability measure of an i.i.d. sample $g_1, \dots, g_n$ of G with $n \in \mathbb{N}$ which is drawn based on $\Pr$. Then,

$$\sup_{g\in G} \mid D^{(n)}(g, \mathbb{K}) - D(g, \mathbb{K}, \Pr) \mid \to 0 \text{ almost surely.}$$

5.4. Universality properties

Finally, we introduce a notion of universality of depth functions. The main motivation is that in general there exists no strictly quasiconcave depth function, see Theorem 4. If one refrains from strict quasiconcavity, then a natural demand is to stick to quasiconcavity. However, the depth function that assigns every object the value zero is also quasiconcave but useless. The idea behind these properties is now the wish for a depth function to be *as strictly quasiconcave as possible*. While the following motivation is based on the strictly quasiconcavity property, one can generalise this idea to all properties defined above.

Before we introduce our notion of a richness of a depth function, we first indicate that defining this notion in a more naive way does not lead to an adequate notion of richness. For now, let the formal context $\mathbb{K}$ and probability measure $\Pr$ be fixed. An intuitive way to concretising the property of a quasiconcave depth function D of being as strictly quasiconcave as possible is to demand that there exists no other quasiconcave depth function E that is more strictly quasiconcave than D. This can be formalised by saying that E is more strictly quasiconcave than D if everywhere, where E violates strict quasiconcavity, also D violates strict quasiconcavity. By violations of the strict quasiconcavity property, we mean that only equality and not strict inequality is fulfilled. More precisely, a depth function E is more strictly quasiconcave than D if

$$\begin{aligned} quasiker(E(\cdot, \mathbb{K}, \Pr)) := \left\{ (A, g) \;\middle|\; A \subseteq G, g \in G \text{ such that } g \in \gamma(A) \setminus A,\ E(g, \mathbb{K}, \Pr) = \inf_{\tilde{g}\in A} E(\tilde{g}, \mathbb{K}, \Pr) \right\} \\ \subsetneq quasiker(D(\cdot, \mathbb{K}, \Pr)). \end{aligned}$$

In other words, there exists a pair $(A, g) \in quasiker(D(\cdot, \mathbb{K}, \Pr)) \setminus quasiker(E(\cdot, \mathbb{K}, \Pr))$ with $g \in \gamma(A)$ such that the depth function D violates for this pair the strictly quasiconcave property but E does not.

However, this definition leads to a problem. The following situation can occur: Assume we have a quasiconcave depth function D and $(A, g) \in quasiker(D(\cdot, \mathbb{K}, \Pr))$ as well as $(\tilde{A}, \tilde{g}) \in quasiker(D(\cdot, \mathbb{K}, \Pr))$. Then it may be the case that increasing the depth value $D(g, \mathbb{K}, \Pr)$ by some small amount ε leads to a depth function $E(\cdot, \mathbb{K}, \Pr)$ that is still quasiconcave and that now fulfils $quasiker(E(\cdot, \mathbb{K}, \Pr)) \subsetneq quasiker(D(\cdot, \mathbb{K}, \Pr))$. Now, assume that the same is true for increasing the depth of $\tilde{g}$ (without increasing the depth of g), but increasing the depth values of both g and $\tilde{g}$ at the same time is not possible without violating the quasiconcavity property. However, assume further that both the underlying probability measure $\Pr$ as well as the underlying context $\mathbb{K}$ are perfectly

symmetric w.r.t. g and $\tilde{g}$. In this case, it is not reasonable to arbitrarily increase one depth value to be more strictly quasiconcave. This means that according to the above definition of being as strictly quasiconcave as possible, there does not exist a possible and reasonable depth function that is as strictly quasiconcave as possible. This underlines the importance to emphasise that a depth function always inherently codifies both the structure of the underlying space as well as the concrete underlying probability measure.

We now introduce our notion of richness of a depth function. The following two universality properties are stated in a more general way such that they do not only capture the richness of a depth function w.r.t. the property of being quasiconcave. The introduced universality properties are also applicable to other properties of depth functions. Concretely, we take inspiration from a basic notion that is folklore in category theory. We adopt here the notion of an universal or a free object: Very roughly speaking, we say that a depth function D is free w.r.t. a set of some structural properties, if for every other depth function E that obeys these properties, we can obtain $E(\cdot, \mathbb{K}, \Pr)$ from D as a composition of $D(\cdot, \mathbb{K}, \tilde{\Pr})$ and an isotone function $f : \mathbb{R} \to \mathbb{R}$. Here, $\tilde{\Pr}$ is a suitable further probability measure. Informally, this means that we allow a meaningful depth function to have identical depth values if and only if the underlying probability measure and the structure of the space support this. In other words, a depth function is free if it is flexible enough to imitate every other depth function (with the same properties) by supplying it with a suitable probability measure $\tilde{\Pr}$. With this idea in mind we now state two different notions of freeness, one weak and one strong notion. Therefore, let $Q \subseteq \{(P1), \dots, (P10)\}$ be a set of properties from Sections 5.1, 5.2 and 5.3 and recall Definition 8. A depth function D is called

(P12) *weakly free* w.r.t. a family $\mathscr{P}$ of probability measures on the object set G and w.r.t. Q, if it satisfies every property from Q and if for every probability measure $\Pr$ and every depth function E on $\mathbb{K}$ that also satisfies all properties in Q, there exists a probability measure $\tilde{\Pr} \in \mathscr{P}$ and an isotone function $f : \mathbb{R} \longrightarrow \mathbb{R}$ such that

$$f \circ D(\cdot, \mathbb{K}, \tilde{\Pr}) \cong E(\cdot, \mathbb{K}, \Pr).$$

(P13) *strongly free* w.r.t. a family $\mathscr{P}$ of probability measures on the object set G and w.r.t. Q, if it satisfies all properties from Q and if for every $\varepsilon > 0$, there exists a class $\mathscr{P}^\varepsilon$ of probability measures from $\mathscr{P}$ with a diameter $d(\mathscr{P}^\varepsilon)$ lower than or equal to ε such that for every other depth function E on $\mathbb{K}$ that also satisfies all properties in Q and for every $\Pr \in \mathscr{P}$ there exists a probability measure $\tilde{\Pr} \in \mathscr{P}^\varepsilon$ and an isotone function $f : \mathbb{R} \longrightarrow \mathbb{R}$ with

$$f \circ D(\cdot, \mathbb{K}, \tilde{\Pr}) = E(\cdot, \mathbb{K}, \Pr).$$

Here, the diameter of a set $\mathscr{P}$ of probability measures is defined as $d(\mathscr{P}) := \sup_{P,Q \in \mathscr{P}} \sup_{A \in \Sigma} |P(A) - Q(A)|$ where Σ is the assumed underlying σ-field.

Remark 3. We say that a depth function D is richer than a depth function E if and only if we can modify D in such a manner that the center-outward order given by E can be reconstructed by D. This means that we can find a probability measure $\tilde{\Pr}$ and an isotone function f such that $f \circ D(\cdot, \mathbb{K}, \tilde{\Pr})$ equals E. Note that f only needs to be isotone and not bijective. Therefore, D must be at least as rich as E in distinguishing the individual objects of G. Finally, observe that $\tilde{\Pr}$ and f depend on $Q, \Pr, \mathbb{K}, D$ and E.

Within the notion weakly free we completely detach from $\Pr$ by allowing $\tilde{\Pr}$ to be any probability measure from $\mathscr{P}$. For the notion strongly free we restrict the allowed probability measures $\tilde{\Pr}$. Therefore, strongly free implies weakly free. The idea behind restricting the probability measure $\tilde{\Pr}$ is the requirement that a depth function should be able to detach itself to some degree from the underlying probability measure $\Pr$. More precisely, there exists a balanced measure $\Pr^*$ such that any structure within the depth values that is not due to the structure of the formal context can already be modified by an arbitrarily small modification of this balanced measure $\Pr^*$. The reason why we did not explicitly use such a balanced measure $\Pr^*$ in the definition is a technical one. The balanced measure does not need to be a probability measure (e.g., there is no uniform distribution on the real numbers). Therefore we work here with the set $\mathscr{P}^\varepsilon$ with arbitrary $\varepsilon > 0$.

Theorem 5. *For the hierarchical nominal context given in Table 2 of Example 4, there exists a depth function D that is strongly free w.r.t. quasiconcavity (and w.r.t. the family of all probability measures on G), namely the function D given by:*

$$D(g, \mathbb{K}, \Pr) = \begin{cases} 1 & \text{if } g = g^* := \arg\max_{\tilde{g} \in G} \Pr(\{\tilde{g}\}) \\ 1/2 & \text{if on the first level, } g \text{ has the same category as } g^* \\ 0 & \text{else .} \end{cases}$$

Here we assume that the argument of the maximum is unique. Otherwise we arbitrarily choose one of the arguments of the maximum.

Proof. First observe that D is indeed quasiconcave, because the contour sets are extent sets. The reason is that the images of γ are exactly the one element sets, the whole set G and the two element sets $\{a_1a_2, a_1b_2\}$ and $\{b_1a_1, b_1a_2\}$. Now, for $\varepsilon > 0$ define the class of probability measures $\mathscr{P}^\varepsilon := \{\Pr \mid \forall g \in G \text{ we have } \Pr(\{g\}) \in [1/4 - \varepsilon/4, 1/4 + \varepsilon/4]\}$. This class has in fact diameter lower than or equal to ε. Now, let $\varepsilon > 0$, let $\Pr$ be an arbitrary probability measure on G, and let E be an arbitrary quasiconcave depth function. Let further g be the object with the highest depth according to E. Without loss of generality we assume that $g = a_1a_2$. Additionally, we assume that there exists exactly one g with maximal depth (otherwise, arbitrarily choose from the objects with maximal depth). Then define $\tilde{\Pr} \in \mathscr{P}^\varepsilon$ by $\tilde{Pr}(\{g\}) := 1/4 + \varepsilon$ and $\tilde{\Pr}(\{\tilde{g}\}) := 1/4 - \varepsilon/3$ for $\tilde{g} \neq g$. If $\varepsilon > 3/4$, we set ε to $3/4$. Because E is assumed

to be quasiconcave, and since the contour sets are always nested, the contour sets of E are exactly the sets $\{a_1a_2\}, \{a_1a_2, a_1b_2\}$ and G. (For E it could also be the case that the set of contour sets is only a subset of these three sets). At the same time, due to construction, the contour sets of D are exactly the same which means that we can construct an isotone function $f : \mathbb{R} \to \mathbb{R}$ with $f \circ D(\cdot, \mathbb{K}, \tilde{\Pr}) = E(\cdot, \mathbb{K}, \Pr)$. Since in particular $\varepsilon > 0$ was arbitrary, D is in fact strongly free w.r.t. quasiconcavity and w.r.t. the family of all probability measures on G. □

Also for a finite hierarchical nominal context with more than two levels and with more than two categories in every level, one can show that there exists a depth function that is strongly free w.r.t. quasiconcavity and w.r.t. the family of all probability measures.

Remark 4. Note that while for example quasiconcavity implies starshapedness, freeness w.r.t. quasiconcavity generally does not imply freenness w.r.t. starshapedness. This is because a quasiconcave depth function D is not able to imitate a depth function E that is starshaped but not quasiconcave, if such a depth function E exists at all. In our concrete Example 4 of the hierarchical nominal context with two levels and two categories on every level, quasiconcavity is accidentally equivalent to starshapedness. For more than two levels this is not the case anymore.

6. Example: Generalised Tukey depth function

In Section 5, we generally introduced structural properties for depth functions based on FCA. One purpose of these structural properties is to give a systematic basis to analyse depth functions using FCA. Therefore, in this section, we take the opportunity to analyse the generalised Tukey depth given in Section 4. Before we begin with discussing the structural properties, we look at some general aspects of the generalised Tukey depth. Here and in the following, we assume that G is a set, $\varkappa$ a set of formal contexts and $\mathscr{P}$ a set of probability measures on σ-fields which contain every extent set given by $\varkappa$.

In Definition 6 the generalised Tukey depth is defined by using the extent sets of single attributes. Thus, in the computation one apparently only takes a proper subset of all possible extents into account, see, e.g., Examples 6 and 7. The reason for this is that considering all extent sets instead of only those induced by a single attribute does not change the depth value. Compare to Section 4 where this has already been mentioned.

Theorem 6. *Let G be an object set. For every formal context $\mathbb{K}$ with object set G and every probability measure $\Pr$ on G we get for the generalised Tukey depth function*

$$T(g, \mathbb{K}, \Pr) = 1 - \sup_{m \in M \setminus \Psi(g)} \Pr(\Phi(m)) = 1 - \sup_{B \subseteq M \setminus \Psi(g)} \Pr(\Phi(B)).$$

Proof. We have to show that

$$\sup_{m \in M \setminus \Psi(g)} \Pr(\Phi(m)) = \sup_{B \subseteq M \setminus \Psi(g)} \Pr(\Phi(B)). \tag{2}$$

Since $B \subseteq M \setminus \Psi(g)$ is a superset of $m \in M \setminus \Psi(g)$ we immediately get $\leq$ in Eq. (2). For the reverse inequality let $B \subseteq M \setminus \Psi(g)$ be arbitrary. Then for every $m \in B$ we get that $\Phi(m) \supseteq \Phi(B)$ and thus $\Pr(\Phi(m)) \geq \Pr(\Phi(B))$. Note that $m \in M \setminus \Psi(g)$. Together with the properties of a supremum, we get $\geq$ in Eq. (2). □

Remark 5. Analogously to Theorem 6 one can show that

$$T(g, \mathbb{K}, \Pr) = 1 - \sup_{m \in M \setminus \Psi(g)} \Pr(\Phi(m)) = 1 - \sup_{\substack{A \subseteq G \setminus \{g\} \\ A \text{ extent}}} \Pr(A)$$

is true for $g \in G$, $\mathbb{K}$ a formal context and probability measure $\Pr$. Again, we obtain $\leq$ immediately. For the reverse, assume for simplicity that the supremum is attained at extent $A \subseteq G \setminus \{g\}$. Since $g \notin A$ holds, there exists $m \in \Psi(A)$ such that $g \notin \Phi(m)$. In particular, we get for this m that $\Pr(\Phi(m)) \geq \Pr(A)$ is true which gives indeed equality above.

Remark 6. Theorem 6 shows that the generalised Tukey depth uses only the marginal distribution over the attributes. Thus, it is possible to change the incidence relation I and thereby change the dependency structures of the objects, but the generalised Tukey depth remains the same. Obviously, this statement is particularly true if the marginal distribution remains the same. However, since only the supremum of the extents sets of single attributes is used, even the marginal distribution can change and still the generalised Tukey depth remains the same. An example is given in Table 3 (middle and right). Here, we have $T(g, \mathbb{K}_1, \Pr) = T(g, \mathbb{K}_2, \Pr)$ for all $g \in \{g_1, g_2, g_3\}$ (with uniform probability measure on $\{g_1, g_2, g_3\}$), although in the second formal context, we have that g_2 implies g_3.

6.1. Representation properties

The first two structural properties, the representation properties, aim to ensure that the structure in the extent set is reflected in the depth function. The generalised Tukey depth function is based on the set of extents, see Theorem 6 and Remark 5. More precisely, for each g the depth is based on the probabilities of those extents which do not contain g. Since the function i in Property (P1) is bijective and preserves extents and probabilities, this implies that the depth values are preserved as well. This shows that Property (P1) holds. Furthermore, since two objects which equal in their attributes, always lie in the same extent sets, they have to have the same depth values. Thus, Property (P2) is true for the generalised Tukey depth.

Table 3
For the formal context on the left Property (P8) is never true (see Section 5.2). The middle ($\mathbb{K}_1$) and right ($\mathbb{K}_2$) formal contexts show the effect of the supremum in the definition of the generalised Tukey depth.

	m_1	m_2	m_3
g_1	×		
g_2		×	
g_3			×

	m_1	m_2	m_3	m_4
g_1	×			×
g_2		×	×	
g_3			×	×

	m_1	m_2	m_3	m_4
g_1	×			×
g_2		×	×	
g_3		×	×	×

Table 4
For the formal context on the left holds Property (P8) for the generalised Tukey depth. The formal context does not fulfil Property (P10) for the generalised Tukey depth.

	m_1	m_2	m_3	m_4
g_1	×	×	×	
g_2	×	×		×
g_3	×	×	×	×

	m_1	m_2	m_3
g_1	×		
g_2		×	
g_3		×	×

6.2. Order preserving properties

In Section 5, the order preserving properties are ordered in their strength. Thus, we prove that the contourclosed property (P7i) holds and with this further order preserving properties follow by Theorems 2 and 3.

Theorem 7. *The generalised Tukey depth fulfils Property (P7i).*

Proof. We prove Property (P7i) by contradiction. Assume that there exists a formal context $\mathbb{K}$ on an object set G and a probability measure Pr together with an $\alpha \in im(T(\cdot, \mathbb{K}, \Pr))$ such that $Cont_\alpha$ is not an extent set. This means that $\gamma(Cont_\alpha) \supsetneq Cont_\alpha$ since γ is a closure operator. Thus, there exists $g \in \gamma(Cont_\alpha) \setminus Cont_\alpha$. The generalised Tukey depth, see Definition 6, is based on the extent sets induced by a single attribute that the object g does not have. Since $g \in \gamma(Cont_\alpha)$ we know that for every attribute which g does not have there exists at least one $\tilde{g} \in Cont_\alpha$ which does not have this attribute either. Hence, we have for all $m \in M \setminus \Psi(g)$ there exists $\tilde{g} \in Cont_\alpha$ such that $1 - \Pr(\Phi(m)) \geq 1 - \sup_{\tilde{m} \in M \setminus \Psi(\tilde{g})} \Pr(\Phi(m))$. By definition of the contour set α, we get $T(g, \mathbb{K}, \Pr) = 1 - \sup_{m \in M \setminus \Psi(g)} \Pr(\Phi(m)) \geq \inf_{\tilde{g} \in Cont_\alpha} \left(1 - \sup_{\tilde{m} \in M \setminus \Psi(\tilde{g})} \Pr(\Phi(\tilde{m}))\right) = \inf_{\tilde{g} \in Cont_\alpha} T(\tilde{g}, \mathbb{K}, \Pr)$. This is contradicting $g \notin Cont_\alpha$ and we can conclude that Property (P7i) holds for every formal context $\mathbb{K}$ and probability measure Pr. □

Remark 7. Now we apply Theorems 2 and 3, see overview in Fig. 1, and obtain that since Property (P7i) is true for the generalise Tukey depth, the following properties hold as well: minimality property (P3), maximality property (P4), isotonicity property (P5), starshapedness property (P6) and the quasiconcavity property (P7ii).

In Section 5 we showed that there exist formal contexts such that the strictly quasiconcavity property (P8) does not hold for any depth function and any probability measure. Hence, here we are interested in formal contexts such that the generalised Tukey depth is strictly quasiconcave. A small example of a formal context such that the quasiconcavity is fulfilled is given in Table 4 (left).

To check this claim, we take a look at Property (P8) applied on the generalised Tukey depth. Thus, Property (P8) is satisfied if and only if for all $A \subseteq G$ we have for all $\tilde{g} \in \gamma(A) \setminus A$

$$1 - \sup_{m \in M \setminus \Psi(\tilde{g})} \Pr(\Phi(m)) > \inf_{g \in A} \left(1 - \sup_{m \in M \setminus \Psi(g)} \Pr(\Phi(m))\right) \Leftrightarrow \sup_{m \in M \setminus \Psi(\tilde{g})} \Pr(\Phi(m)) < \sup_{g \in A} \sup_{m \in M \setminus \Psi(g)} \Pr(\Phi(m)). \tag{3}$$

Note that $\leq$ in Inequality (3) follows by Property (P7ii). To check whether the generalised Tukey depth based on the formal context given in Table 4 (left) is strictly quasiconcave, we go through every subset $A \subseteq G$ with $\gamma(A) \setminus A \neq \emptyset$. This is only true for $A = \{g_1, g_2\}$ with $\gamma(A) = \{g_1, g_2, g_3\}$. The attributes of g_3 are a superset of the union of the attributes of g_1 and g_2 (i.e. $\Psi(\{g_3\}) \supseteq \Psi(\{g_1\}) \cup \Psi(\{g_2\})$). Observe that the attributes of g_1 and g_2 which maximise the supremum of the corresponding generalised Tukey depth function are attributes of g_3. Thus, when ensuring that there exists an attribute $m \in \Psi(g_3) \setminus \Psi(g_2) \cap \Psi(g_1)$ such that $\Pr(\Phi(m))$ is strictly larger than $\Pr(\Phi(\tilde{m}))$ based on those attributes $\tilde{m}$ which are not an element of $\Psi(g_3)$, the generalised Tukey depth is strictly quasiconcave. Note that it is sufficient that the depth of either g_1 or g_2 must be below the depth of g_3. Furthermore, we want to point out that we also have to assume that there does not exist a further object g_4 which is a duplicate of g_3, because then $\Psi(g_3) \setminus \Psi(g_2) \cap \Psi(g_1) = \emptyset$. With this, we get the strict part in Inequality (3). Based on the idea above, we can say that the generalised Tukey depth satisfies Property (P8) for every formal context which is an element of the following set:

$$\mathscr{C}^{P8} = \left\{ \begin{array}{l|l} (G, M, I) \text{ formal context without} & \text{for every subset } A \text{ and } g \in \gamma(A) \setminus A \text{ we have } \Psi(g) \supseteq \cup_{a \in A} \Psi(a) \text{ and} \\ \text{duplicates according to attributes} & \exists m \in \Psi(g) \setminus \cap_{a \in A} \Psi(a) : \Pr(\Phi(m)) > \sup_{m \in M \setminus \Psi(g)} \Pr(\Phi(m)) \end{array} \right\}.$$

6.3. Empirical (sequence) properties

The discussion about empirical (sequence) properties profits from the strong impact of the (empirical) probability measure on the definition of the generalised Tukey depth, see Definitions 6 and 7. With this we can immediately follow that Property (P9) is fulfilled by the generalised Tukey depth. The stability of the order property does not hold. Consider again the formal context given by Table 4 (right). Then g_1 is an outlier w.r.t. $\{g_2, g_3\}$. For the empirical generalised Tukey depth based on $\Pr^{(2)}$ for $\{g_2, g_3\}$ we get $T^{(2)}(g_3, \mathbb{K}) = 1 > 1/2 = T^{(2)}(g_2, \mathbb{K})$. When adding object g_1 to the probability measure $\Pr^{(3)}$ we get that $T^{(3)}(g_3, \mathbb{K}) = 2/3 = T^{(3)}(g_2, \mathbb{K})$. The order of g_2 and g_3 is not stable. Since this can always occur for small $n \in \mathbb{N}$, we cannot restrict the set of probability measures or formal contexts.

It remains to discuss the consistency property (P11). One can show that if the set of extents induced by single attributes $\mathcal{ES} := \{\Phi(m) \mid m \in M\}$ is a Glivenko–Cantelli class w.r.t. the underlying probability measure Pr, see [28], then the generalised Tukey depth is indeed consistent. This, for example, is given when $\mathcal{ES}$ has a finite VC dimension.

Theorem 8. *Let $\mathbb{K}$ be a formal context with G as object set. Let $Pr^{(n)}$ be the empirical probability measure given by an i.i.d. sample of sizes $n \in \mathbb{N}$ and based on probability measure* Pr. *When $\mathcal{ES}$ is a Glivenko–Cantelli class w.r.t.* Pr*, then $T^{(n)}$ is consistent and Property (P11) is fulfilled with*

$$\sup_{g \in G} \left| T^{(n)}(g, \mathbb{K}) - T(g, \mathbb{K}, \Pr) \right| \overset{n \to \infty}{\longrightarrow} 0 \quad \textit{almost surely}.$$

Proof. The proof is similar to the proof given by [4, p. 1816f]. Since we assume that $\mathcal{ES}$ is a Glivenko–Cantelli class w.r.t. Pr, we get $\sup_{E \in \mathcal{ES}} \left| \Pr^{(n)}(E) - \Pr(E) \right| \overset{n \to \infty}{\longrightarrow} 0$ almost surely. With this, the claim follows from

$$\sup_{g \in G} \left| 1 - \sup_{m \in M \setminus \Psi(g)} \Pr^{(n)}(\Phi(m)) - \left(1 - \sup_{m \in M \setminus \Psi(g)} \Pr(\Phi(m)) \right) \right| = \sup_{g \in G} \left| \sup_{m \in M \setminus \Psi(g)} \Pr^{(n)}(\Phi(m)) - \sup_{m \in M \setminus \Psi(g)} \Pr(\Phi(m)) \right|$$

$$\leq \sup_{E \in \mathcal{ES}} \left| \Pr^{(n)}(E) - \Pr(E) \right| \overset{n \to \infty}{\longrightarrow} 0 \text{ almost surely.} \quad \square$$

Remark 8. If $\mathcal{ES}$ has a finite VC dimension, then the convergence stated in Theorem 8 is additionally uniform over all possible underlying probability measures. A finite VC dimension of $\mathcal{ES}$ is given, for example, if M is finite.

Remark 9. We know that the generalised Tukey depth fulfils Property (P1). Let us assume that for a formal context $\tilde{\mathbb{K}}$ the set of all extents induced by one single attribute is not a Glivenko–Cantelli class w.r.t. some probability measure $\tilde{\Pr}$. Then, in some cases, one can define a second formal context $\mathbb{K}$ and probability measure Pr such that there exists a bijective and bimeasurable function i which preserves the extents and the probability measure. If the extents based on a single attribute given by $\mathbb{K}$ is a Glivenko–Cantelli class w.r.t. probability measure Pr, then we can transfer the consistency based on $\mathbb{K}$ onto $\tilde{\mathbb{K}}$. An example for this is given by the following two formal contexts: Let $G = \mathbb{R}^d$ and $\mathbb{K}$ be the formal context defined in Example 5. For $\tilde{\mathbb{K}}$ let $\tilde{M}$ be the set of all topologically closed convex sets and $\tilde{I}$ the binary relation with $(g, \tilde{m}) \in I$ if and only if g is an element of the corresponding convex set $\tilde{m}$. The extents equal again all topologically closed convex sets. Thus by applying Property (P1) we can replace $\tilde{\mathbb{K}}$ with $\mathbb{K}$ and can prove that also for $\tilde{\mathbb{K}}$ the generalised Tukey depth is consistent.

6.4. Universality properties

Let us end this discussion on the generalised Tukey depth by considering the universality properties. The next two theorems prove that the generalised Tukey depth is weakly free but not strongly free w.r.t. the quasiconcavity property. Therefore, recall the hierarchical nominal example, see Examples 4 and 8.

Theorem 9. *Let G be finite. Then the generalised Tukey depth is weakly free w.r.t. the family of all probability measures on G and w.r.t. the quasiconcavity property (P7ii).*

Proof. Let G be finite, let $\mathbb{K}$ be an arbitrary context with object set G and let Pr be an arbitrary probability measure on G. Moreover, let E be an arbitrary quasiconcave depth function. We set $E_1, \dots, E_K$ to be the unique increasingly ordered depth values of $E(\cdot, \mathbb{K}, \Pr)$ and let k_i be the number of objects with depth value E_i. Furthermore, let G_i denote the boundaries of the contour sets, i.e., the set of objects with depth E_i. We now have to construct a probability measure $\tilde{\Pr}$ and an isotone function f such that $f \circ T(\cdot, \mathbb{K}, \tilde{\Pr}) = E(\cdot, \mathbb{K}, \Pr)$.

To achieve this, we first consider properties this probability function needs to fulfil. Let $\tilde{\Pr} \in \mathscr{P}$ be a possible candidate. We set $p_i := \tilde{\Pr}(G_i)$ and $p_i^{min} := \min_{g \in G_i} \tilde{\Pr}(\{g\})$. Since E is quasiconcave, we have $g \notin \gamma(G_{\ell+1} \cup \dots \cup G_K)$ for all $g \in G_\ell$ with $\ell \in \{1, \dots, K-1\}$. This implies that for every $\ell \in \{1, \dots, K-1\}$ and every $g \in G_\ell$ there exists an attribute $m \in M$ with $(g, m) \notin I$, but $(h, m) \in I$ for all $h \in G_{\ell+1} \cup \dots \cup G_K$. Therefore, we get for $g \in G_\ell$

$$p_\ell^{min} \leq T(g, \mathbb{K}, \tilde{\Pr}) \leq \sum_{i=1}^{\ell} p_i. \tag{4}$$

Now, we construct probability measure $\tilde{\Pr}$ needed for Property (P12) in three steps: First set $\tilde{\Pr}(g) = 1$ for $g \in G_1$. Then recursively set $\tilde{\Pr}(g) = \sum_{i=1}^{\ell} \tilde{\Pr}(G_i) + 1$ for $g \in G_{\ell+1}$. Finally, we normalise $\tilde{\Pr}$ by $\tilde{\Pr}(g) := \tilde{Pr}(g) / \sum_{g \in G} \tilde{\Pr}(g)$ to get a probability measure. Due to construction and Inequality (4), it is ensured that $T(g, \mathbb{K}, \tilde{Pr}) < T(h, \mathbb{K}, \tilde{Pr})$ for $g \in G_\ell$ and $h \in G_{\ell+1}$. Thus, one can define function f by setting $f(T(g, \mathbb{K}, \tilde{\Pr})) = E_\ell$ for $g \in G_\ell$ and isotonically extending it to function $f : \mathbb{R} \to \mathbb{R}$. With this, we get $f \circ T(\cdot, \mathbb{K}, \tilde{Pr}) = E(\cdot, \mathbb{K}, \Pr)$ which shows that claim. □

Theorem 10. *For the context describing hierarchical nominal data given in Table 2 the generalised Tukey depth is not strongly free w.r.t quasiconcavity and w.r.t. the family of all probability measures on G.*

Proof. First note that the depth function D from Theorem 5 is quasiconcave. Furthermore, D is flexible enough to assign every arbitrary object g the highest depth 1. This implies that the object h with the same category on Level 1 as g has the second highest depth 1/2 and all other objects have depth 0. Now, the generalised Tukey depth $T(\cdot, \mathbb{K}, \Pr)$ needs to be as flexible as D to be strongly free w.r.t. quasiconcavity. Looking at Table 2, first observe that for object $g = a_1a_2$ to have the highest generalised Tukey depth, it is necessary that $\Pr(\{a_1a_2, a_1, b_2\}) \geq 0.5$. This is because if we have $\Pr(\{a_1a_2, a_1, b_2\}) < 0.5$, we can conclude that $T(a_1a_2, \mathbb{K}, \Pr) \leq 1 - \Pr(\{b_1a_2, b_1b_2\}) = \Pr(\{a_1a_2, a_1, b_2\}) < 0.5$. Since either $\Pr(\{b_1a_2\})$ or $\Pr(\{b_1b_2\})$ is smaller or equal to 0.5, we get that one of the objects b_1a_2 or b_1b_2 has a generalised Tukey depth of larger or equal to 0.5. Now, for object $h = a_1b_2$ to have the second highest depth, it is necessary that the corresponding supremum in Definition 6 is attained for an attribute that differs from $g = a_1a_2$. Because both g and h do not have attribute b_1, it is clear that attribute b_1 is relevant for the supremum within the generalised Tukey depth both for objects g and h. For h to have the second highest depth, it is therefore necessary that for h the supremum is attained for attribute a_1a_2. But for this it is necessary that

$$\Pr(\{a_1b_2\}) \geq \Pr(\{b_1a_2, b_1b_2\}). \tag{5}$$

Now, let $\varepsilon = 0.1$ and assume that T is strongly free w.r.t. quasiconcavity. Thus, we can find a family $\mathscr{P}^\varepsilon$ of diameter $\varepsilon = 0.1$ with the corresponding properties. Now take $\Pr^* \in \mathscr{P}^\varepsilon$ such that the generalised Tukey depth has the highest depth at $g = a_1a_2$ and for $h = a_1b_2$ the second highest depth. With the above considerations we can conclude first that $\tilde{Pr}(\{a_1a_2, a_1b_2\}) \geq 0.5$ and, thus, $\Pr^*(\{a_1a_2, a_1b_2\}) \geq 0.5 - \varepsilon = 0.4$. Additionally, because also object b_1a_2 could be the object with highest depth w.r.t. D and some probability measure Pr, an analogous argumentation implies that also $\Pr^*(\{b_1a_2, b_1b_2\}) \geq 0.5 - \varepsilon = 0.4$. But this, together with Inequality (5) implies that $\Pr^*(a_1b_2) \geq 0.4$. With analogous argumentations, one can also show $\Pr^*(\{g\}) \geq 0.4$ for all other objects. But this is a contradiction to $\Pr^*$ being a probability measure because $\Pr^*(G) = 0.4 * 4 = 1.6$. This provides the claim. □

7. Conclusions

This article introduced a general notion of depth functions for non-standard data based on FCA. This covers the analysis of centrality and outlyingness for a wide variety of different data types. In order to enable a discussion, especially about these non-standard data, which are not given in a standard statistical data format, we used FCA and introduced structural properties. In addition to adopting the properties discussed in $\mathbb{R}^d$, we also addressed the issue that non-standard data may inherit a central-outward order that should also be reflected in the depth function. We also provided a framework for analysing depth functions based on FCA. Finally, we used this framework to discuss and analyse the generalised Tukey depth. Building on this, there are several promising areas for further research, including (but not limited to)

Further concrete mapping rules for depth functions: In Theorem 5 we defined a depth function D which is strongly free w.r.t. Property (P8) and the formal context defined in Table 2. One can show that this depth function D only uses the probabilities of Level 1 to obtain the object g with the highest depth value. In further research, it is of interest to define mapping rules for depth functions which take more levels into account. We already introduced the union-free generic depth function for the special case of G being the set of partial orders, see [19]. This is an adaptation of the simplicial depth in $\mathbb{R}^d$, see [1], to the set of all partial orders. As shortly denoted there this can be generalised to general formal contexts. [15] give another examples for a depth function using FCA. The authors there propose concrete outlier measures based on formal concept analysis. Since there is a strong relationship between outlier measures and depth functions, the properties here can also be used to analyse this measure.

Larger scale application: In this article the focus lies on the general analysis of depth functions and the generalised Tukey depth. Hence, the examples discussed here aim to support and clarify claims of theoretical/structural properties of depth functions. In further research, a discussion based on the perspective of a larger scale application is of interest. For example, in [19] we applied the union-free generic depth function, which can be embedded into our concept of depth upon partial orders that represent the performance of machine learning algorithms.

Statistical inference: Building on this not only a descriptive analysis, but also statistical inference methods can be developed. Since depth functions describe in a robust and nonparametric manner the order of the data this can be used to define statistical tests and models. Analogously to the approach here, where we, among others, transferred already existing properties to our concept, this can be done for statistical inference methods, see [10] as starting point.

Analysis of generalised Tukey depth based on one specific scaling method: Here, we generally analysed the generalised Tukey depth function without focus on one specific scaling method. In [14] we discussed the special case of G being the set of partial orders and one scaling method. Since one fixed scaling method on a set G gives us more structure on the corresponding closure system an analysis can lead to more structural properties of the depth function.

CRediT authorship contribution statement

Hannah Blocher: Writing – original draft, Visualization, Supervision, Project administration, Methodology, Formal analysis, Conceptualization, Writing – review & editing. **Georg Schollmeyer:** Writing – original draft, Supervision, Methodology, Formal analysis, Conceptualization, Writing – review & editing.

Acknowledgements

HB and GS sincerely thank Thomas Augustin for the many helpful suggestions during the preparation of this article. They are also grateful to the two reviewers and the associated editor for their very important suggestions, which, among others, helped to improve the accessibility of the article. HB and GS gratefully acknowledge the financial and general support of the LMU Mentoring program. HB sincerely thanks Evangelisches Studienwerk Villigst e.V. for funding and supporting her doctoral studies.

References

[1] R. Liu, On a notion of data depth based on random simplices, Ann. Statist. 18 (1990) 405–414.
[2] R. Dyckerhoff, K. Mosler, G. Koshevoy, Zonoid data depth: Theory and computation, in: A. Prat (Ed.), COMPSTAT, Physica-Verlag HD, Heidelberg, 1996, pp. 235–240.
[3] J. Tukey, Mathematics and the picturing of data, in: R. James (Ed.), Proceedings of the International Congress of Mathematicians Vancouver, Mathematics-Congresses, Vancouver, 1975, pp. 523–531.
[4] D.L. Donoho, M. Gasko, Breakdown properties of location estimates based on halfspace depth and projected outlyingness, Ann. Statist. 20 (1992) 1803–1827.
[5] Y. Zuo, R. Serfling, General notions of statistical depth function, Ann. Statist. 28 (2000) 461–482.
[6] K. Mosler, Depth statistics, in: C. Becker, R. Fried, S. Kuhnt (Eds.), Robustness and Complex Data Structures: Festschrift in Honour of Ursula Gather, Springer, Berlin, Heidelberg, 2013, pp. 17–34.
[7] I. Gijbels, S. Nagy, On a general definition of depth for functional data, Statist. Sci. 32 (2017) 630–639.
[8] M. Goibert, S. Clémençn, E. Irurozki, P. Mozharovskyi, Statistical depth functions for ranking distributions: Definitions, statistical learning and applications, in: G. Camps-Valls, F. Ruiz, I. Valera (Eds.), Proceedings of the 25th International Conference on Artificial Intelligence and Statistics AISTATS 2022, in: Proceedings of Machine Learning Research, Vol. 151, PMLR, 2022, pp. 10376–10406.
[9] G. Geenens, A. Nieto-Reyes, G. Francisci, Statistical depth in abstract metric spaces, Stat. Comput. 33 (2023) 46.
[10] J. Li, R.Y. Liu, New nonparametric tests of multivariate locations and scales using data depth, Statist. Sci. 19 (2004) 686–696.
[11] S. Pawar, D. Shirke, Data depth-based nonparametric tests for multivariate scales, J. Stat. Theory Pract. 16 (2022) 1–21.
[12] P. Mozharovskyi, Anomaly detection using data depth: Multivariate case, 2022, arXiv https://arxiv.org/abs/2210.02851 (Accessed 24 September 2024).
[13] F. Chebana, T.B.M.J. Ouarda, Depth-based multivariate descriptive statistics with hydrological applications, J. Geophys. Res. 116 (2011) 1–19.
[14] H. Blocher, G. Schollmeyer, C. Jansen, Statistical models for partial orders based on data depth and formal concept analysis, in: D. Ciucci, I. Couso, J. Medina, D. Ślęzak, D. Petturiti, B. Bouchon-Meunier, R. Yager (Eds.), Information Processing and Management of Uncertainty in Knowledge-Based Systems, Vol. 1602, Springer, Cham, 2022, pp. 17–30.
[15] Q. Hu, Z. Yuan, K. Qin, J. Zhang, A novel outlier detection approach based on formal concept analysis, Knowl.-Based Syst. 268 (2023) 110486.
[16] W. Cukierski, Titanic - machine learning from disaster, 2012, https://kaggle.com/competitions/titanic (Accesses 24 September 2024).
[17] B. Ganter, R. Wille, Formal Concept Analysis: Mathematical Foundations, Springer, Berlin, Heidelberg, 2012.
[18] R. Serfling, Y. Zuo, Structural properties and convergence results for contours of sample statistical depth functions, Ann. Statist. 28 (2000).
[19] H. Blocher, G. Schollmeyer, C. Jansen, M. Nalenz, Depth functions for partial orders with a descriptive analysis of machine learning algorithms, in: E. Miranda, I. Montes, E. Quaeghebeur, B. Vantaggi (Eds.), Proceedings of the Thirteenth International Symposium on Imprecise Probability: Theories and Applications, Vol. 215, PMLR, Oviedo, 2023, pp. 59–71.
[20] C. Carpineto, G. Romano, Concept Data Analysis: Theory and Applications, Wiley, Chichester and Weinheim, 2004.
[21] T. Hanika, J. Hirth, Quantifying the conceptual error in dimensionality reduction, in: T. Braun, M. Gehrke, T. Hanika, N. Hernandez (Eds.), Graph-Based Representation and Reasoning - 26th International Conference on Conceptual Structures, Vol. 12879, Springer, Cham, Cham, 2021, pp. 105–118.
[22] G. Schollmeyer, Lower Quantiles for Complete Lattices, Technical Report, LMU, 2017, http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-40448-7 (Accessed 24 September 2024).
[23] G. Schollmeyer, Application of Lower Quantiles for Complete Lattices to Ranking Data: Analyzing Outlyingness of Preference Orderings, Technical Report, LMU, 2017, http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-40452-9 (Accessed 24 September 2024).
[24] T. Tao, An Introduction to Measure Theory, in: Graduate Studies in Mathematics, vol. 126, American Mathematical Society, Providence, RI, 2011.
[25] K. Mosler, P. Mozharovskyi, Choosing among notions of multivariate depth statistics, Statist. Sci. 37 (2022) 348–368.
[26] S. Durocher, S. Szabados, Curve stabbing depth: Data depth for plane curves, in: Y. Bahoo, K. Georgiou (Eds.), 34th Canadian Conference on Computational Geometry, pp. 121–128.
[27] Y. Zuo, R. Serfling, On the performance of some robust nonparametric location measures relative to a general notion of multivariate symmetry, J. Statist. Plann. Inference 84 (2000) 55–79.
[28] M. Dudley, E. Giné, J. Zinn, Uniform and universal Glivenko–Cantelli classes, J. Theoret. Probab. 4 (1991) 485–510.

Contribution 2

Hannah Blocher and Georg Schollmeyer (2024). *Union-free Generic Depth for Non-standard Data*. ArXiv:2412.14745. URL: https://arxiv.org/abs/2412.14745. (last accessed: 02.03.2025)

Union-Free Generic Depth for Non-Standard Data

Hannah Blocher*
Department of Statistics, Ludwig-Maximilians-University Munich
and
Georg Schollmeyer
Department of Statistics, Ludwig-Maximilians-University Munich

December 20, 2024

Abstract

Non-standard data, which fall outside classical statistical data formats, challenge state-of-the-art analysis. Examples of non-standard data include partial orders and mixed categorical-numeric-spatial data. Most statistical methods required to represent them by classical statistical spaces. However, this representation can distort their inherent structure and thus the results and interpretation. For applicants, this creates a dilemma: using standard statistical methods can risk misrepresenting the data, while preserving their true structure often lead these methods to be inapplicable. To address this dilemma, we introduce the union-free generic depth (ufg-depth) which is a novel framework that respects the true structure of non-standard data while enabling robust statistical analysis. The ufg-depth extends the concept of simplicial depth from normed vector spaces to a much broader range of data types, by combining formal concept analysis and data depth. We provide a systematic analysis of the theoretical properties of the ufg-depth and demonstrate its application to mixed categorical-numerical-spatial data and hierarchical-nominal data. The ufg-depth is a unified approach that bridges the gap between preserving the data structure and applying statistical methods. With this, we provide a new perspective for non-standard data analysis.

Keywords: (simplicial) data depth, formal concept analysis, non-parametric statistics, mixed categorical-numerical-spatial data, hierarchical-nominal data

*The authors gratefully acknowledge the funding and support of Hannah Blocher's doctoral studies by the Evangelisches Studienwerk Villigst e.V. and the LMU Mentoring Program.

1 Introduction

Modern statistical analysis frequently encounters *non-standard data*, which are data that are not given in classical statistical data formats such as nominal, ordinal, interval, or ratio scales, see, e.g., Stevens (1946). Examples include multivariate data combining spatial and ordinal components or (partial) preference orders, where we observe a set of orders on fixed items. Addressing such data often requires either (implicitly) imposing additional assumptions, such as a metric space, see, e.g., the discussion in Blocher et al. (2024), or transforming the data at the cost of losing information, such as discretizing continuous variables see, e.g., Foss et al. (2019); Yanqing Zhang et al. (2024). Moreover, these methods are generally limited to specific types of non-standard data and lack a unified framework for broader applicability, see Stumme et al. (2023).

This limitation highlights a significant research gap: The absence of a general, flexible framework that reflects the inherent structure of non-standard data while avoiding unwanted assumptions or information loss. This gap creates a fundamental dilemma in statistical analysis. On the one hand, applying standard statistical methods can distort the underlying data structure and with it the results and interpretations. On the other hand, accounting for the true structure of the data can make standard methods inapplicable. Resolving this dilemma requires a novel approach that balances the data structure with the practical requirements of statistical analysis.

This article addresses the dilemma by introducing a novel, nonparametric method – the *union-free generic depth (ufg-depth)* – offering a new perspective on the analysis of non-standard data. The ufg-depth unifies the treatment of diverse data types without imposing further, eventually not justified assumptions. Unlike classical methods that rely on directly using data values (as in classical statistical tests like Student's t-test), the ufg-depth is based on natural groupings of the data elements. These groupings serve as the foundation for defining a center-outward order for the data.

The definition of the ufg-depth is based on combining *formal concept analysis (FCA)* and *depth functions*. Formal concept analysis provides the necessary tools to address the challenges of analyzing non-standard data by detecting and representing relationships within the data using mathematical lattice theory, see Ganter and Wille (2012). It has been successfully applied in diverse fields such as knowledge discovery, see Poelmans et al. (2013), bioinformatics, see Roscoe et al. (2022), and choice theory, see Ignatov and Kwuida (2022). Depth functions, on the other hand, extend the notion of quantiles in $\mathbb{R}$ to higher dimensional normed vector spaces. They provide a measure that denotes the centrality and outlyingness of data relative to a data cloud or a given probability distribution. These functions are widely used in nonparametric statistics on $\mathbb{R}^d$, see Chebana and Ouarda (2011); Liu et al. (1999). The ufg-depth introduced in this article builds on these concepts and generalizes the simplicial depth developed by Liu (1990). The simplicial depth defines a centrality measure as the probability that a point lies within a randomly drawn simplex (e.g., a triangle). Generalizing this idea for non-standard data involves two key challenges: redefining the concept of "lying in" and defining an appropriate analogue of a simplex. Formal concept analysis provides the theoretical foundation to address both challenges, enabling the development of the ufg-depth as a robust method for analyzing non-standard data.

By combining these two concepts, this article introduces for non-standard data a robust

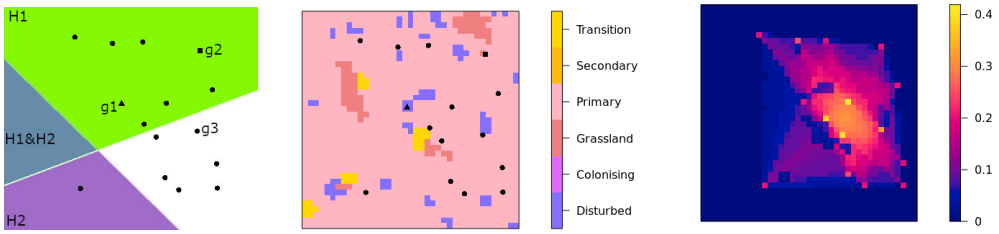

Figure 1: Excerpt of the spatial location of the GORILLAS data with (left) two closed half-spaces included in the plot, (middle) the vegetation component and (right) ufg-depth based on the spatial and vegetation component

center-outward order that considers broader spaces than solely classical statistic spaces and provides a unified framework to analyze these data. Moreover, this framework is very flexible and universal. For instance, we apply the ufg-depth on two real-word data problems – spatial-categorical-numerical data and hierarchical-nominal data – and analyze it generally using centrality notions derived in Blocher and Schollmeyer (2025). There the authors provide a general mapping structure for depth functions based on formal concept analysis. Moreover, they establish a systematic basis by adapting, among others, the desirable properties defined in Zuo and Serfling (2000a,b); Mosler and Mozharovskyi (2022).

This article is organized as follows: We first provide a detailed illustration of the conceptual strategy underlying the ufg-depth and the main definitions of formal concept analysis using two concrete examples. Next, we generally define the ufg-depth and analyze its properties, drawing on the framework established in Blocher and Schollmeyer (2025). Section 5 presents concrete applications of the method to real-world non-standard data. Finally, we conclude with a discussion of the ufg-depth's contributions and limitations. Supplementary materials include detailed introduction to formal concept analysis, proofs and further side notes.

2 Illustration of the Concepts behind the UFG-Depth

In this section, we describe the idea behind the definition of ufg-depth using a snippet of concrete data examples: The GORILLAS data set, see Funwi-Gabga and Mateu (2012), containing nesting sites of gorillas and data from the German General Social Survey (GGSS) concerning occupations, see GESIS - Leibniz-Institut für Sozialwissenschaften (2023). A detailed and complete analysis of both data sets can be found in Section 5. Moreover, we introduce the main concept of formal concept analysis needed in this article. For more details on formal concept analysis, see the supplementary or Ganter and Wille (2012).

Example 1. The GORILLAS data provide a point pattern consisting of a spatial component (gorilla nesting sites) and categorical, numerical observations (e.g. vegetation or elevation). For further details on the data see Section 5.1. In the following, we use an excerpt of 15 observations of the gorilla nesting sites to illustrate our approach, see Figure 1. We start by considering only the spatial component and describe the link to simplicial depth, see Liu (1990).

Formal concept analysis (FCA) serves as the foundation of our method, offering a powerful tool for uncovering relationships within data. It is based on the formalization

of a cross-table. The rows of the cross-table correspond to the ground space (in formal concept analysis called *objects* and denoted by G) and the columns represent attributes (in formal concept analysis also called *attributes* and denoted by M) that can either be true or false for an object. The number of attributes and objects can be infinite. A cross in the cross-table indicates that the attribute holds for the element/object. These crosses are formalized by an *incidence relation* $I \subseteq G \times M$. The triple $K = (G, M, I)$ is called *formal context*. We want to point out that an attribute can be either true or false for an object. In particular, no degree in between can be assigned. In most cases, however, the values of the data/ground space are not binary. The transformation of many-valued data into a set of binary attributes is called *(conceptual) scaling method* and has been studied extensively, for example, see (Ganter and Wille, 2012, Chapter 1.3.). For the GORILLAS data, the objects $G_{\mathbb{R}^2}$ equal $\mathbb{R}^2$. Now, we have to apply a scaling method that represents the data by a set of binary attributes. Therefore, we use the method developed in Blocher and Schollmeyer (2025) where the attributes $M_{\mathbb{R}^2}$ are all topologically closed half-spaces.[1] The incidence relation $I_{\mathbb{R}^2}$ represents whether the element lies in the half-space or not. We denote this formal context by $\mathbb{K}_{\mathbb{R}^2} = (G_{\mathbb{R}^2}, M_{\mathbb{R}^2}, I_{\mathbb{R}^2})$. Figure 2 (left) shows an excerpt of the infinite (w.r.t. $G_{\mathbb{R}^2}$ and $M_{\mathbb{R}^2}$) formal context using the indicated objects g_1, g_2 and g_3 and half-spaces H_1 and H_2 in Figure 1 (left).

In the next step, we use this formal context/cross-table to define the grouping procedure and resulting grouping system/groups. Let us take a subset of attributes and consider the maximal group of objects that are all valid for these attributes. This summarizes all objects that are in a certain relation (i.e. via the attribute subset). Thus, when taking all attribute subsets and the respective maximal object sets, we obtain a grouping system representing the relationship between the objects. In the case of the GORILLA excerpt, we take a subset of half-spaces as a subset of attributes, and the corresponding objects are all elements that lie in every half-space of the subset. Similarly, the relationship between attributes can be considered by starting with the object set. Note that combining these two operations provides an operator that also leads to the same grouping system as starting with the attribute sets. Formally, this is given by the maps $\Phi : 2^M \to 2^G, B \to B' := \{g \in G \mid \forall m \in B\colon gIm\}$ and $\Psi : 2^G \to 2^M, A \to A' := \{m \in M \mid \forall g \in A\colon gIm\}$.[2] The grouping procedure on the object set is then given by $\gamma := \Phi \circ \Psi$. We denote the resulting grouping system by $\mathcal{E} = \gamma(2^G)$ and call it *(set of) extents*. Now, we can exploit that the the extents $\gamma(2^G)$ and operator γ define a *closure system* and a *closure operator* and have a one-to-one correspondence. A closure operator is a function on the power set 2^G that is extensive (i.e. $A \subseteq \gamma(A)$ for all $A \subseteq G$), monotone (i.e. $\gamma(B) \subseteq \gamma(A)$ for all $B \subseteq A \subseteq G$) and idempotent (i.e. $\gamma(A) \subseteq \gamma(\gamma(A))$ for all $A \subseteq G$). A closure system is a subset $H \subseteq 2^G$ such that $G \in H$ and the intersection of elements in H is again an element of H. For $\mathbb{K}_{\mathbb{R}^2}$ the closure operator corresponds to the convex hull operator $\gamma_{\mathbb{R}^2}$ that maps a set onto the smallest closed convex set containing this set. The extents $\mathcal{E}_{\mathbb{R}^2}$ equal all convex sets.

Moreover, both, the closure system and operator, can be uniquely described by a *family of implications*. An implication is a statement of the form $A \to B$ with $A, B \subseteq G$ which

[1] In the following we always consider topologically closed half-spaces/convex sets. For simplicity, we will drop the term topological and closed from now on. Note that we will use also the term closed when referring to closed based on a closure operator.

[2] For simplicity we write $\Phi(g)$ instead of $\Phi(\{g\})$ for $g \in G$. The same applies to all operators on the power set.

	H_1	H_2	...	...	...	...
g_1	x					
g_2	x					
g_3					x	
...						

	tran.	sec.	prim.	grass.	colo.	dist.
g_1						x
g_2				x		
g_3			x			
...						

Figure 2: Scaling the spatial (left) and vegetation (right) component of the point pattern in Figure 1 (left).

claims that if A is a subset of a group/extent, then B must also be part of that group/extent, i.e. $\gamma(B) \subseteq \gamma(A)$. Here, we call A *premise* and B *conclusion* of the implication. Reverse, when we have a set of implications $\mathcal{I}$ of a set G, then we say that $D \subseteq G$ *respects an implication* $A \to B$ iff either $A \not\subseteq D$ or $A \subseteq D$ then $B \subseteq D$ also follows. With this, we obtain the extent set back by $\mathcal{E}_{\mathcal{I}_G} = \{D \subseteq G \mid D \text{ respects every implication in } \mathcal{I}_G\}$. The implications of $\mathbb{K}_{\mathbb{R}^2}$ are statements $A \to B$ where B lies in the every convex set that contains also A. More formally using the convex hull operator $\gamma_{\mathbb{R}^2}$, an implication $A \to B$ holds for the extent set given by $\mathbb{K}_{\mathbb{R}^2}$ iff $\gamma_{\mathbb{R}^2}(A) \supseteq \gamma_{\mathbb{R}^2}(B)$.

However, some of these implications are redundant. For example, let $B \subsetneq \gamma_{\mathbb{R}^2}(A)$. Then statement $A \to B$ is true but redundant since its information is already given by implication $A \to \gamma_{\mathbb{R}^2}(A)$. These semantic redundancy structures are summarized by Armstrong (1974) as inference axioms, see (Maier, 1983, p. 45): Let $A, B, C, D, A_1, A_2, B_1, B_2 \subseteq G$. Then we say that the axiom of *reflexivity* holds iff $A \to A$, the axiom of *augmentation* holds iff $A_1 \to B$ implies $A_1 \cup A_2 \to B$, the axiom of *additivity* holds iff $A \to B_1$ and $A \to B_2$ imply $A \to B_1 \cup B_2$, axiom of *projectivity* holds iff $A \to B_1 \cup B_2$ implies $A \to B_1$, axiom of *transitivity* holds iff $A \to B$ and $B \to C$ imply $A \to C$, and the axiom of *pseodotransitivity* holds iff $A \to B$ and $B \cup C \to D$ imply $A \cup C \to D$. Note, however, that when deleting implications that follow from the above semantic structures, one may delete too many implications and end up not representing the same closure system, see Section 3.2. Therefore we say that a reduced family of implications $\mathcal{I}_G$ is *complete* if the reduced family describes the same extent set as the unreduced one.

For the ufg-depth, we now consider the reduction of the family of implications based on reflexivity, augmentation, additivity and projectivity. With this, a complete reduction of the implications describing the spatial context is given by the set of all implications $A \to \gamma_{\mathbb{R}^2}(A)$ with $\#A = \{2, 3\}$. Based on this, we can now define the ufg-depth of $g \in \mathbb{R}^2$ as the probability that g lies in $\gamma_{\mathbb{R}^2}(C)$, where C is a randomly drawn line or triangle according to an (empirical) probability measure on $\mathbb{R}^2$. (For details on how we weight the randomly drawn triangles vs. lines, see Section 3.2.) Hence, if we assume that the measure is absolutely continuous to the Lebesgue measure, we obtain the well-known simplicial depth.

Example 2. We consider again the GORILLAS data of Example 1. Now we add the vegetation observations to the analysis. Figure 1 (left and middle) show the point pattern of Example 1 together with its vegetation component. For clarity, the round dots represent observations in *primary vegetation*, the triangle in *disturbed vegetation*, and the square in *grassland vegetation*. Now, the ground space is $G = \mathbb{R}^2 \times V$, where $V = \{\text{tran., sec., prim., grass., colo., dist.}\}$ consists of all possible vegetation outcomes, see Section 5.1 for details.

We proceed similarly to Example 1. First, we use a scaling method to define a formal context/cross-table. For the spatial component we use the half-spaces discussed in Example 1. The vegetation component is categorical and therefore we use the so-called *nominal scaling*, see (Ganter and Wille, 2012, p. 42) where the attributes are all possible vegetation outcomes V. The table in Figure 2 (right) represents the vegetation part of the objects g_1, g_2, g_3 denoted in Figure 1 (middle). Joining the two cross-tables of Figure 2 by the object set G gives us the formal context that represents the spatial and categorical component of each data element of $G = \mathbb{R}^2 \times V$ with attribute set $M_{\mathbb{R}^2 \times V} = M_{\mathbb{R}^2} \cup M_V$. The resulting groups/extents arise by considering all possible combinations of attributes and summarizing all data elements that apply to them. Thus, the extents are all sets $C \times \tilde{V}$, where $C \subseteq \mathbb{R}^2$ is a convex set and $\tilde{V} \in \binom{V}{1} \cup V$. We set $\binom{V}{1} = \{\{v_1\}, \ldots, \{v_k\}\}$ for $V = \{v_1, \ldots, v_k\}$. Note that due to the nominal scaling, $\tilde{V}$ either has cardinality one or is directly the entire set. This represents the dependencies between the groups, since the relation between two vegetation categories is the same as to any other vegetation, so all other vegetation categories are also included.

Analogously to Example 1, we utilize the fact that the extents define a closure system that can be uniquely described by a family of implications. Finally, we consider only a subset of all valid implications by deleting redundancies. With this, the ufg-depth of an element g is the proportion of non-redundant implications with positive empirical probability mass that imply g. The ufg-depth indicates how supportive/typical/central the observation g is with respect to all other observations. This is because an object g that lies in many non-redundant sets must have many attributes that are shared by the elements in those sets. Therefore, in the extreme case where there is an object that has all attributes, that object has maximum depth. The other extreme case, where an object has not many attributes that other objects have, denotes a small depth value. For example here, we only observed the vegetation disturbed once. Therefore, even though this observation is in the center of spatial component of the data cloud, it has a low depth value. The vegetation categories transition and grassland are never observed. So for these two categories we have that the depth is zero. The ufg-depth of an element $g \in \mathbb{R}^2 \times V$ is then given by Figure 1 (right).

Example 3. To highlight the wide variety of different data types that fall under the term non-standard data, we consider occupational data from the German General Social Survey (GGSS), see GESIS - Leibniz-Institut für Sozialwissenschaften (2023). These occupations are categorized using a hierarchy of different levels given by the International Standard Classification of Occupations (ISCO) of 2008[3]. To define a formal context representing the different occupational groups, the occupations are successively classified into categories and subcategories. First, on a basic level (Level 1), each data element is assigned to exactly one category (coded here with digits $1, 2, \ldots, 9, 0$, see (GESIS - Leibniz-Institut für Sozialwissenschaften, 2023, Appendix D)). For example, the level-1 categories of ISCO-08 are 1: *Managers*; 2: *Professionals*; etc. Each of these categories is then split on a finer level (Level 2) into further subcategories and again each element of a single Level-1 category is assigned to exactly one subcategory of Level 2. For example, the Level-1 Category 3: *Technicians and associate professionals* is further divided into the Level-2 categories

[3]see `https://ilostat.ilo.org/methods/concepts-and-definitions/classification-occupation/` (last accessed: 14.12.2024) for details

31: *Science and engineering associate professionals*; 32: *Health associate professionals*; etc. Then, again the Level 2 categories are divided into further subcategories, and so on. For the ISCO-08 classification scheme, we have 4 levels with different numbers of possible categories on each level (ranging from 1 to 10 categories). Now, we build a formal context for the representation of our data structure by introducing the following attributes: Every sequence $x_1x_2\dots x_k$ with $x_i \in \{1,\dots,9,0\}$ (and $k \in \{1,\dots,4\}$ for ISCO-08) describes the category on Level i. For each sequence, we introduce one attribute $x_1x_2\dots x_k$. An object g has this attribute if it belongs to the respective occupational category up to Level k. Note that for this conceptual scaling, in contrast to Example 1, there are usually different objects that have exactly the same attributes.

Remark 1. Finally, we want to emphasize that the closure system, the implications and later the ufg-depth, depends on the application of a reasonable scaling method. The closure system and the implications provide a tool for analyzing/discussing the relational structure in detail, but the starting point is the scaling method. In particular, all the underlying assumptions of the ground space structure are determined by the scaling method. For example, in Example 2 we have $G = \mathbb{R}^2 \times V$ as ground space. Therefore, if we have two observations in the same place with the same vegetation, we assume them to be duplicates of the same objects. Hence, the information about the two identical observations is only included in the empirical probability measure and not in the formal context itself. Another approach, not discussed here, is to consider each individual nesting point as an observation that cannot be a duplicate, but is another object in the ground space.

3 The Union-Free Generic Depth

In this section, we introduce the *union-free generic* depth function. It provides a centrality and outlyingness measure for data that cannot be embedded in the multidimensional real vector space. In particular, this depth function makes direct use of the relational structure provided by formal concept analysis. The definition of the union-free generic depth function is in the spirit of the simplicial depth function on $\mathbb{R}^d$, see Liu (1990). We transfer the idea of using simplices to the framework of formal concept analysis.

3.1 The Simplicial Depth from the Perspective of Formal Concept Analysis

Recall Example 1 where we discussed the formal context $\mathbb{K}_{\mathbb{R}^2} = (\mathbb{R}^2, M_{\mathbb{R}^2}, I_{\mathbb{R}^2})$ with $G = \mathbb{R}^2$ as data/objects and the set of half-spaces as attributes. In this section we have a look at the simplicial depth from the perspective of formal concept analysis. Therefore, let $\mathcal{I}_{\mathbb{R}^2,\mathrm{ufg}} = \{C \to \gamma_{\mathbb{R}^2}(C) \mid \ C \subseteq \mathbb{R}^2 \text{ vertices of a simplex and } 2 \leq \#C \leq 3\ \}$ be the reduced family of implications of $\mathcal{I}_{\mathbb{R}^2}$.

Using Carathéodory's theorem, see Eckhoff (1993), we can first show that this family of implications describes the convex sets. Moreover, we obtain that compared to $\mathcal{I}_{\mathbb{R}^2}$, $\mathcal{I}_{\mathbb{R}^2,\mathrm{ufg}}$ has deleted all implications that follow from the inference axioms of reflexivity, augmentation, additivity, and projectivity, see Lemma 2.2 in the supplementary for details. Note that the transitivity and pseodotransitivity axioms are not applied for the deletion. This is done because when restricting to the inference axioms of reflexivity, augmentation,

additivity and projectivity, the information about the *betweenness* of the data points is preserved directly, otherwise this information is given only indirectly. For example, let $A \subseteq \mathbb{R}^2$ with $\#A = 3$ and let $A_1, A_2 \subseteq A$ be a division of A such that $A_1 \cup A_2 = A$. Then $\gamma_{\mathbb{R}^2}(A_i), i \in \{1, 2\}$ is the line between the two elements of A_i, and using then transitivity and pseodotransitivity together, we get that $\gamma_{\mathbb{R}^2}(\gamma_{\mathbb{R}^2}(A_1) \cup \gamma_{\mathbb{R}^2}(A_2)) = \gamma_{\mathbb{R}^2}(A)$ the whole triangle.

Example 4. Note that all the considerations above for $\mathbb{R}^2$ (Example 1) can be easily adopted to general $\mathbb{R}^d$ with $d \in \mathbb{N}$. The formal context is defined analogously with half-spaces in $\mathbb{R}^d$ and as closure system/operator we get again the closed convex sets with the corresponding convex closure operator. Similar to Lemma 2.2 in the supplementary we obtain all implications with $2 \leq k \leq d + 1$ points defining a vertice of a simplex as premises of an ufg-implication.

With the above in mind, let us take a closer look at

$$D : \mathbb{R}^d \to \mathbb{R}, g \mapsto \sum_{i=2}^{d+1} P(g \in \gamma(\{X_1 \dots X_i\}) \mid X_1 \dots X_i \text{ define vertices of a proper simplex})$$

for a probability measure P on $\mathbb{R}^d$ and independent random variables $X_1, \dots, X_{d+1} \sim P$. This gives us the sum of the probabilities that g lies in a proper simplex of cardinality $2 \leq k \leq d + 1$. Assuming that the probability measure P is absolutely continuous with respect to the Lebesgue measure, we obtain that D exactly mimics the simplicial depth function.

3.2 Definition of the Union-Free Generic Depth

Now we take the next step and transfer the idea based on simplicial depth to general data represented by a formal context and the resulting closure system/operator. Let $\mathbb{K} = (G, M, I)$ be a formal context with corresponding closure system $\mathcal{E}_\mathbb{K}$ and closure operator $\gamma_\mathbb{K}$. In the style of the above section, we define a family of implications that is reduced based on the Armstrong rules of reflexivity, augmentation, additivity, and projectivity.

Definition 3.1. The *union-free generic family of implications* (*ufg-family of implications* for short) for a formal context $\mathbb{K}$ on a object set G is defined by

$$\mathcal{I}_{G,\text{ufg}} := \{A \to \gamma_G(A) \mid A \textit{ fulfills (C1) and (C2)}\}$$

with the conditions on A: (C1) $A \subsetneq \gamma_G(A)$ and (C2) for all families $(A_j)_{j \in J}$ with $A_j \subsetneq A$ for $j \in J$ we have that $\cup_{j \in J} \gamma_G(A_i) \neq \gamma_G(A)$. For an implication $A \to B \in \mathcal{I}_{G,\text{ufg}}$, we say that $A \in \mathcal{I}^{prem}_{G,\text{ufg}}$ is the *ufg-premise* and $B \in \mathcal{I}^{concl}_{G,\text{ufg}}$ the *ufg-conclusion*. For examples, see Section 3.1 and Section 5.

We call $\mathcal{I}_{G,\text{ufg}}$ generic following Bastide et al. (2000) where they called an implication with minimal premise and maximal conclusion to be generic. The minimality of the premise follows from Condition (C2). Since we set the conclusion to $\gamma_G(A)$ the maximality is also directly given. The term union-free describes the idea behind the Condition (C2) as it covers more then only generic.[4] Note that there can still exist non-redundant implications which follow semantically from other implications by use of transitivity and pseudotransitivity.

[4]Such families are also called proper and contracted in (Ganter and Wille, 2012, p. 82).

We want to point out that the ufg-family of implications does not necessarily result in a family of implications that describes the closure systems as it can reduce too much information. An example is the closure system $\mathcal{E}_{\mathbb{N}} = \{A \subseteq \mathbb{N} \mid \#A \text{ finite}\} \cup \mathbb{N}$ on $\mathbb{N}$. Here the family of all implications is given by $\mathcal{I}_{\mathbb{N}} = \{A \to B \mid \#A = \infty, A \subseteq \mathbb{N}, B \subseteq \mathbb{N}\} \cup \{A \to B \mid \#A < \infty, B \subseteq A\}$. So every implication does not satisfy Condition (C1) or (C2). Hence $\mathcal{I}_{\mathbb{N},\text{ufg}} = \emptyset$. Note that this can only happen if the underlying space is infinite, and even in infinite cases this limitation does not hold in general, as can be seen in the spatial case. However, when using the ufg-depth, this is another aspect that needs to be considered in the scaling method definition.

Now we transfer the idea of the simplicial depth and define the ufg-depth as weighted probability that an object/element lies in a randomly drawn ufg-conclusion. The definition is in line with the general mapping structure given in Blocher and Schollmeyer (2025). To simplify the notation, we set $\mathcal{I}^{prem,j}_{\mathbb{K},\text{ufg}}$ to be the set of all ufg-premises given by the formal context $\mathbb{K}$ of cardinality $j \in \mathbb{N}$ and define $f^j_g : 2^G \to \{0,1\}, A \mapsto 1_{\gamma(A)}(g) 1_{\mathcal{I}^{prem,j}_{\mathbb{K},\text{ufg}}}(A)$ with $\mathcal{F}^j = \{f^j_g \mid g \in G\}$.[5] Let $h^j : G \times \ldots \times G \to \{0,1\}, (g_1, \ldots, g_j) \mapsto 1_{\mathcal{I}^{prem,j}_{\mathbb{K},\text{ufg}}}(g_1, \ldots, g_j)$. Moreover, we define the functional U-statistics for every function i

$$U^j_{(g_1,\ldots,g_n)}[i] = \begin{cases} \binom{n}{j}^{-1} \sum\limits_{1 \le i_1 < \ldots < i_j \le n} i(g_{i_1}, \ldots, g_{i_j}), & j \le n \\ 0, & j > n \end{cases}.$$

.

Definition 3.2. Let G be a set. We set $\varkappa_G \subseteq \{\mathbb{K} \mid G \text{ is object set of } \mathbb{K}\}$ to be a set of formal contexts with object set G and $\mathcal{P}_G$ to be a family of probability measures on G such that for every $P \in \mathcal{P}_G$ every extent of every $\mathbb{K} \in \kappa_G$ is measurable. Moreover, we assume that for every $\mathbb{K} \in \kappa_G$ there exists a one-to-one correspondence between the ufg-family of implications and the formal context. Let the weights $C_j \in]0, \infty[$ be fix for all $j \in \mathbb{N}$.[6]

Then the *union-free generic depth (ufg-for short)* with $J_{P,\mathbb{K}} = \{j \subseteq \mathbb{N} \mid P((X_1, \ldots, X_j) \in \mathcal{I}_{\mathbb{K},\text{ufg}}) > 0\}$ with $X_1, \ldots, X_j \overset{i.i.d.}{\sim} P$ is given by

$$D : \begin{cases} G \times \kappa_G \times \mathcal{P}_G \to \mathbb{R}^d, \\ (g, \mathbb{K}, P) \mapsto \sum_{j \in J_{P,\mathbb{K}}} \frac{C_j}{\mathbb{E}[h^j]} \mathbb{E}[f^j_g] \end{cases}.$$

Where the expectation is based on the product measure of $P \in \mathcal{P}_G$. The object(s) with the highest ufg-depth value is(are) called *ufg-median*.

Remark 2. First, the ufg-premises being finite is not necessary, but as the sum only includes finite ufg-premises infinite ufg-premises are not taken into account. At a first glance the definition of $J_{P,\mathbb{K}}$ seems to be tricky. However, it is sufficient to know an upper bound for $\max J_{P,\mathbb{K}}$, since for an index that is not part of $J_{P,\mathbb{K}}$, that part of the sum is zero by default in the empirical version. To obtain such an upper bound one can utilize the structure of the formal context, see, e.g. Lemma 2.3. or Lemma 2.13. in the supplementary. Moreover, we want to point out that the definition of the ufg-family of implication in concrete settings

[5] For the consistency proof, we will later need the dual definitions $f^j_A : G \to \{0,1\}, g \mapsto 1_{\gamma(A)}(g) 1_{\mathcal{I}^{prem,j}_{\mathbb{K},\text{ufg}}}(A)$ and $\tilde{\mathcal{F}}^j = \{f^j_A \mid A \subseteq G\}$.

[6] The weights C_j can also be random depending on P, see Blocher et al. (2024).

can differ strongly in their complexity. Thus, in the definition of the formal context also the computation aspect should be taken into account. For details see Section 5.

Second, note that the weights C_j allow flexibility in the definition. For $C_j = 1$ for all $j \in J_{P,\mathbb{K}}$ we obtain the conditional probabilities.

Definition 3.3. Let $G, \varkappa_G$ and $\mathcal{P}_G$ as in Definition 3.2. Let $P \in \mathcal{P}_G$ and we set $J_{P,\mathbb{K}}$ as in Definition 3.2. Let $x_1, \ldots, x_n \overset{i.i.d.}{\sim} P$ for $n \in \mathbb{N}$. Again, we assume fixed weights $C_j \in]0, \infty[$ for every $j \in \mathbb{N}$. Then the *empirical ufg-depth* is given by

$$D^{(n)} : \begin{cases} G \times \kappa_G \to \mathbb{R}^d, \\ (g, \mathbb{K}) \mapsto \sum_{j \in J_{P,\mathbb{K}}} \frac{C_j}{U^j_{(x_1,\ldots,x_n)}[h^j]} U^j_{(x_1,\ldots,x_n)}[f^j_g] \end{cases} .$$

For a rigorous definition, from now on we set $r/0$ to zero for $r \in \mathbb{R}$.

From now on, $\varkappa_G$ and $\mathcal{P}_G$ are two families where every extent from $\mathbb{K} \in \varkappa_G$ is measurable for every $P \in \mathcal{P}_G$. Moreover, we omit the index in $D, \mathbb{K}, \mathcal{I}, M, I$ and γ from now on if the ground space is clear.

4 Structural Properties

In the previous sections, we explored the connection between ufg-depth and simplicial depth. Here, we aim to determine exactly how the ufg-depth is a measure of centrality. Therefore, we build on the *structural properties* given by Blocher and Schollmeyer (2025). These properties provide a systematic basis for discussing centrality and outlyingness for non-standard data represented via formal concept analysis. These structural properties address two aspects. First, the adaptation of existing desirable properties in $\mathbb{R}^d$, see Zuo and Serfling (2000a,b); Mosler and Mozharovskyi (2022). Some, such as quasiconcavity which relies on a notion of "lying in", can be easily transferred. For others, e.g. vanishing to infinity, this is not the case. Second, the structural properties cover the inherited centrality/outlyingness of the data structure itself. In this section, we analyze the ufg-depth function in terms of these structural properties. We use the examples above to provide the idea behind the structural properties. For overview, we underline the structural properties discussed in the theorem.

4.1 Representation Properties

The first two properties ensure that the depth functions preserve the structure imposed by the formal context on G. Concretely, this means that representing the data G by a different formal context, which results in the same closure system on G, should not affect the depth as long as the probability measure is preserved. In other words, if a different scaling method is used that represents the relationship between the data elements in the same way, then the relationship structure, and not the attributes and incidence relation used, should be crucial. The second part assumes that two objects having the same attributes, and therefore are not distinguishable from the perspective of formal concept analysis, need to have the same depth values.

Theorem 4.1. *Let $P, \tilde{P} \in \mathcal{P}_G$ be two probability measures on G and let $\mathbb{K}, \tilde{\mathbb{K}} \in \varkappa$ be two formal contexts on G.*
Invariance on the extents: Assume that there exists a bijective and bimeasureable function $i : G \to G$ such that the extents are preserved (i.e. E extent w.r.t. $\mathbb{K} \Leftrightarrow i(E)$ extent w.r.t. $\tilde{\mathbb{K}}$) and the probability as well (i.e. $P(E) = \tilde{P}(i(E))$). Then $D_G(g, \mathbb{K}, P) \leq D_G(\tilde{g}, \mathbb{K}, P) \Leftrightarrow \tilde{D}_G(i(g), \tilde{\mathbb{K}}, \tilde{P}) \leq \tilde{D}_G(i(\tilde{g}), \tilde{\mathbb{K}}, \tilde{P})$ is true for all $g, \tilde{g} \in G$.
Invariance on the attributes: Let $g_1, g_2 \in G$ with $\Psi_{\mathbb{K}}(g_1) = \Psi_{\mathbb{K}}(g_2)$, then $D(g_1, \mathbb{K}, P) = D(g_2, \mathbb{K}, P)$ holds.

4.2 Order-Preserving Properties

The order-preserving properties cover the idea of "maximality at the center", "monotonicity relative to the deepest point", and "quasiconcavity" properties in $\mathbb{R}^d$, see Mosler and Mozharovskyi (2022). At the same time, these properties also represent the structure of the ground space, such as an inherited centrality/outlyingness structure. In contrast to $\mathbb{R}^d$ where no element has a predetermined tendency to be more central than another element, this can appear for non-standard data. For example, consider the case of two objects g_1, g_2 with $\Psi(g_1) \supseteq \Psi(g_2)$. Hence, g_2 has all attributes that g_1 has and therefore lies in every closure set that contains also g_1. g_2 is, in some sense, more specific than g_1 and therefore the depth of g_2 should be as least as high as the depth of g_1. The property that formalizes this is called *isotonicity*. Note that in some cases center and outlying elements are then directly implied. When an element lies in every closure set it needs to have maximal depth. This property is called *maximality* property. The reverse is called *minimality* property and states that an object that lies only in the most general extent, i.e. the entire set, needs to have minimal. Both properties follow directly from the isotonicity by Theorem 2 of Blocher and Schollmeyer (2025).

Theorem 4.2. *Let $P \in \mathcal{P}_G$ and formal context $\mathbb{K} \in \varkappa$ with $g_1, g_2 \in G$ such that $\gamma_{\mathbb{K}}(\{g_1\}) \supseteq \gamma_{\mathbb{K}}(\{g_2\})$. Then the isotonicity property $D(g_1, \mathbb{K}, P) \leq D(g_2, \mathbb{K}, P)$ is true.*

The isontonicity property can be seen as a pre-property for the stricter *starshaped* and *quasiconcavity/contourclosed* properties. As the name implies, the starshaped property is inspired by the "monotone relative to the deepest point" property in $\mathbb{R}^d$, see Zuo and Serfling (2000a). It says that if we have a center (e.g., given by the maximality property), then any element g_2 implied by the center c and another element g_1 (i.e., $g_2 \in \gamma(c, g_1)$) has at least as high a depth as g_1. A depth function satisfies the quasiconcave property iff for every $\alpha \in \mathbb{R}$ the contour set $Cont_{D,\alpha} = \{g \in G \mid D(g, \mathbb{K}, P) \geq \alpha\}$ defines an extent set. Since the convex sets correspond to the extents, this is a direct translation of the quasiconcavity property in Mosler and Mozharovskyi (2022). Recall that for the formal context $\mathbb{K}_{\mathbb{R}^d}$ (see Example 2) the ufg-depth coincides with the simplicial depth in $\mathbb{R}^d$. With Zuo and Serfling (2000b) we immediately obtain that the ufg-depth is neither starshaped nor quasiconcave.

In many cases, however, one can easily define an adopted ufg-depth function that is starshaped or quasiconcave. We show in Section 4.4 that such an adaptation can lead to a quasiconcave depth function with as few ties as possible. One approach builds on order theory and we aim to obtain the smallest quasiconcave function that still lies above the original ufg-depth. This gives us $D^{qc}(\cdot, \mathbb{K}, P) : G \to \mathbb{R}, x \mapsto \sup\{\alpha \in \mathbb{R} \mid Cont_{D,\alpha} \to g\}$.

Theorem 4.3. *Let $D(\cdot, \mathbb{K}, P)$ be a depth based on formal concept analysis. Then D^{qc} is quasiconcave.*

Note that searching for the most similar quasiconcave function based on a loss function is another approach to get a a quasiconcave function, see the supplement. Also note that our focus here is on quasiconcave, but one can adapt these ideas to starshapedness.

4.3 (Empirical) Sequence Properties

The previous sections focused on how the structure of the formal context is represented in the data. In this section, we have a fixed formal context, but consider a sequence of (empirical) probability measures. These properties address issues such as duplication in a sample, how outlying objects influence the more central ones and consistency considerations.

First, let us assume we have a sample $(g_1, \ldots, g_n)$. In the first case, the *reflecting duplication* property, we assume that there are two objects g_i, g_ℓ in the sample that cannot be distinguished by the formal context (i.e. the same object is observed twice, or they have exactly the same attributes). Then the depth of g_i should be higher when considering the entire sample compared to the sample where the duplication is deleted. For the second property, we assume that there is an object g_i that is completely different from all other observed objects. This means that the only extent containing g_i and any subset of the sample is directly the entire set G. Then this object g_i should not affect the order of the remaining objects, in the sense that it does not matter whether it is in the sample or not. This property is called *stability of the order*.

Theorem 4.4. *Let $\mathbb{K} \in \varkappa$. Let $g_1, \ldots, g_n$ be a sample of G with $n \in \mathbb{N}$. We denote with $P^{(n)}$ the empirical probability measure given by $g_1, \ldots, g_n$ and by $P^{(n,-\ell)}$ the empirical probability measure based on $g_1, \ldots, g_{\ell-1}, g_{\ell+1}, \ldots, g_n$ with $\ell \in \{1, \ldots, n\}$.*
Respecting duplication: Let $i, \ell \in \{1, \ldots, n\}$ with $i \neq \ell$ and for every extent $E \in \mathcal{E}$ we have $g_\ell \in E$ iff $g_i \in E$. Moreover, assume that there exists $j \in J_{P,\mathbb{K}}$ and ufg-premises $A_1, A_2 \in 2^{\{g_1,\ldots,g_{\ell-1},g_{\ell+1},\ldots,g_n\}} \cap \mathcal{I}_{ufg}^{prem,j}$ with $g_i \in A_1$ and $g_i \notin \gamma(A_2)$. Then, we have $D_G(g_i, \mathbb{K}, P^{(n,-\ell)}) < D_G(g_i, \mathbb{K}, P^{(n)})$.
Stability of the order: Assume that the only extents E that contains g_ℓ for $\ell \in \{1, \ldots, n\}$ as well as any subset of $g_1, \ldots, g_{\ell-1}, g_{\ell+1}, \ldots$ is $E = G$. Then for $g, \tilde{g} \in \{g_1, \ldots, g_{\ell-1}, g_{\ell+1}, \ldots, g_n\}$ we have $D_G(g, \mathbb{K}, P^{(n)}) \leq D_G(\tilde{g}, \mathbb{K}, P^{(n)}) \Leftrightarrow \tilde{D}_G(g, \mathbb{K}, P^{(n,-\ell)}) \leq \tilde{D}_G(\tilde{g}, \mathbb{K}, P^{(n,-\ell)})$.

Finally, we discuss the consistency of the ufg-depth based on an i.i.d. sample. Let $(P^{(n)})_{n\in\mathbb{N}}$ be a sequence of empirical probability measures based on i.i.d. samples. We show that the ufg-depth is consistent when the set of all ufg-conclusions has a finite VC-dimension. The VC-dimension of a family of sets $\mathcal{C} \subseteq 2^G$ is the largest number such that there exists a set $\{g_1, \ldots, g_{vc}\} \subseteq G$, $vc \in \mathbb{N}$, with $\{C \cap \{g_1, \ldots, g_{vc}\} \mid C \in \mathcal{C}\} = 2^{\{g_1,\ldots,g_{vc}\}}$, see Dudley et al. (1991). In other words, the VC-dimension of $\mathcal{C}$ denotes the largest possible set that can be still shattered be $\mathcal{C}$.

Theorem 4.5. *Let $\mathbb{K} \in \varkappa, P \in \mathcal{P}$ and $J_{P,\mathbb{K}} \subseteq \mathbb{N}$ be the same as in Definition 4.2. of the main article. Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$. We assume that $\#J_{P,\mathbb{K}} < \infty$. Moreover, we assume that for every $j \in J_{P,\mathbb{K}}$ $\mathcal{I}_{\mathbb{K},ufg}^{concl,j}$ has finite VC-dimension. With this, we get the consistency property with $\sup_{g\in G} |D(g, \mathbb{K}, P^{(n)}) - D(g, \mathbb{K}, P)| \to 0$ almost surely for n to infinity. (We assume that this supremum is measurable.)*

4.4 Universality Properties

As discussed in Section 4.2, the ufg-depth D is generally not quasiconcave, but one can work instead with the quasiconcave version D^{qc} from Theorem 4.3. In this section we show that D^{qc} provides a depth function that is quasiconcave and, in a sense, as flexible as possible, e.g. having only ties that are actually needed for the quasiconcavity. This property is formalized by the *universality properties* introduced in Blocher and Schollmeyer (2025). The idea behind universality w.r.t. a property Q, here quasiconcavity, is to say that a depth function (here, D^{qc}) is as flexible as possible if it can have the same orderings of the depth values like that of another arbitrary depth function E with the same property Q, if it is only equipped with an appropriate probability measure P^*. If P^* is allowed to be chosen arbitrarily, then we speak about *weak freenness*. If P^* is only allowed to be chosen from a set of probability measures that are arbitrary close to each other, then we speak about *strong freenness*, which is of course a stronger property than weak freenness. For the mathematical details, we refer to Blocher and Schollmeyer (2025) and the supplementary. As it turns out, under some technical assumptions, the ufg-depth is approximately weakly free. Because of some technical subtleties in the assumptions and in the exact formulation of the statement we decided to move the corresponding theorem to the supplementary, see Theorem 2.9 and Remark 2, where also a short discussion about some cases in which the assumptions are fulfilled can be found. The following theorem now gives a concrete situation under which the ufg-depth is also strongly free. This is a main advantage compared to the generalized Tukey depth,[7] which is not strongly free, as shown in (Blocher and Schollmeyer, 2025, Theorem 10).

Theorem 4.6. *Let $\mathbb{K} = (G, M, I)$ be a formal context given by hierarchical-nominal data with $L \geq 2$ levels, $K \geq 3$ categories on each level, and the scaling method presented in Example 3 of Section 2 of the main article. We assume that for each object $g \in G$ there exists another object $\tilde{g} \in G$ with $g \neq \tilde{g}$ and $\Psi(\{g\}) = \Psi(\{\tilde{g}\})$.*

We set $C_1, C_2 > 0$ in the ufg-depth. Then the quasiconcave version D^{qc} of the ufg-depth is strongly free with respect to the property quasiconcavity. This means that for every $\varepsilon > 0$ there exists a family $\mathcal{P}^\varepsilon$ of probability measures with diameterless than or equal to ε such that for any other arbitrary quasiconcave depth function E and any arbitrary probability measure P there exists a measure $P^* \in \mathcal{P}^\varepsilon$ such that

$$\forall g, \tilde{g} \in G : E(g, \mathbb{K}, P) > E(\tilde{g}, \mathbb{K}, P) \Longrightarrow D^{qc}(g, \mathbb{K}, P^*) > D^{qc}(\tilde{g}, \mathbb{K}, P^*).$$

5 Examples

In this section we provide two application examples of the ufg-depth. First, we analyze the Gorillas data, see Example 2 and second, the occupational data from the German General Social Survey (GGSS), see Example 3.[8] We want to emphasize that analyzing

[7] The generalized Tukey depth is based on Schollmeyer (2017b,a) and was firstly formally introduced in Blocher et al. (2022). A description of the used basic concepts and an in-depth analysis of the properties of the generalized Tukey depth can be found in Blocher and Schollmeyer (2025). The generalized Turkey's depth T of an object g w.r.t. a formal context $\mathbb{K}$ and w.r.t. a probability measure P is defined as $T(g) := 1 - \sup\{P(E) \mid E \text{ extent of } \mathbb{K} : g \notin E\}$.

[8] Both analysis can be found on GitHub: `https://anonymous.4open.science/r/ufg_depth_application-0567/`. (last accessed: 14.12.2024)

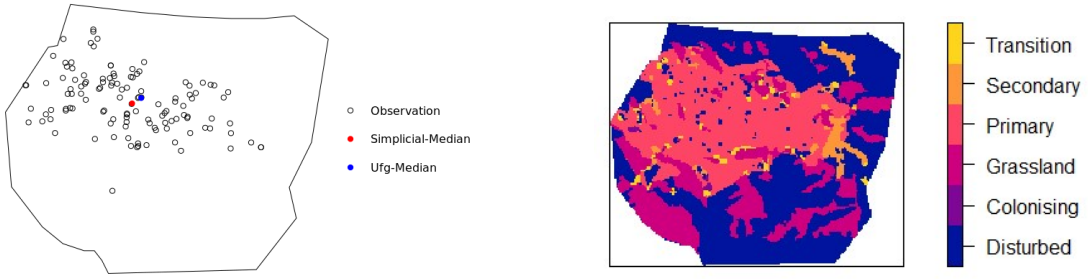

Figure 3: The gorilla nesting sites (left) and vegetation (right) of the Kagwene National Park (Cameroon) for the year 2006.

the ufg-depth for a concrete data type of interest can lead to a simplified definition of the ufg-depth and, in particular, improve the computation time by exploiting the further data structure. This has been done in Blocher et al. (2024), where the authors defined, analyzed and applied the ufg-depth on the special case of partial orders as ground space.

5.1 Mixed Categorical, Numeric and Spatial Data

Recall Example 2, where we used a snippet of the gorilla nesting sites data to motivate our approach. Now, we want to extend this example by adding a further covariate (elevation) and considering a larger sample. Besides, we outline how the situation of the ground space in Section 2 was simplified.

The data set is stored in the R–package GORILLAS and GORILLAS.EXTRA and both are provided by the R-packages SPATSTAT, see Baddeley and Turner (2005). The observations are a point pattern where each point represents one nesting site of the gorilla population at the Kagwene Gorilla Sanctuary in Cameroon. The nesting sites where observed from 2007 and 2009 and it consists of 647 observations. For more details we refer to Funwi-Gabga and Mateu (2012). In the following, we analyze the sample of the gorilla nesting sites observed in 2006. In total we have 121 points which are plotted in Figure 3 (left). Besides the spatial observation, we include the vegetation and elevation component, see Figure 3 (right) and Figure 4 (left). Mainly *primary* (76 points) and *disturbed* (24 points) vegetation category are associated to the observed points. The corresponding elevation values range from 1340 to 2053.

In the illustration in Example 2, we simplified the data situation for the sake of accessibility. There, the underlying ground space was assumed to be $\mathbb{R}^d \times V$ with $V = \{transition, secondary, primary, grassland, colonising, disturbed\}$. However, since we are only interested in the nesting sites of the gorillas within the Kagwene Gorilla Sanctuary in Cameroon, we now reduce the spatial set to $K \subseteq \mathbb{R}^2$ which represents the area of the national park. Moreover, at the same location, the ground space in Section 2 assumed that two different vegetation categories are possible. This assumption does not hold as the vegetation is a fixed covariate and unique to the location part. Hence, the ground space should be $\{(x, v) \mid x \in K$ with unique corresponding vegetation $v \in V\} \subsetneq \mathbb{R}^2 \times V$. Finally, we extent our analysis and add the elevation as a further observation value. With this, we

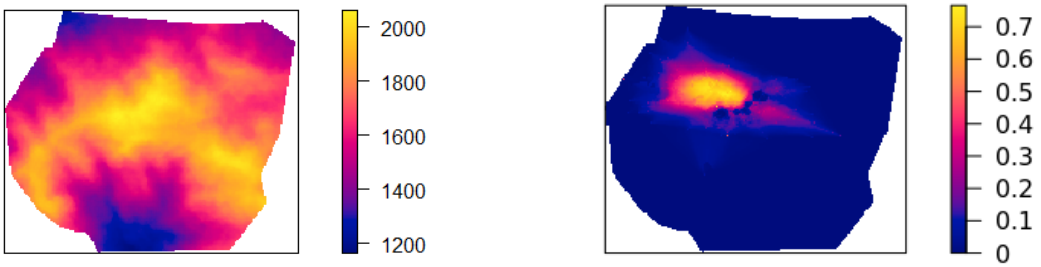

Figure 4: The elevation (left) of the Kagwene National Park (Cameroon) and the computed ufg-depth (right).

get as underlying ground space

$$G = \left\{(x, v, e)\middle|\ x \in K, x \text{ with unique corresponding vegetation } v \in V \text{ and elevation } e \in \mathbb{R}\ \right\}.$$

The formal context now results from scaling the spatial component as in Example 1, the categorical variable using nominal scaling, see Example 2, and for the ordinal component we use *interordinal scaling*, see (Ganter and Wille, 2012, p. 42). The interordinal scaling equals the spatial scaling and since half-spaces in $\mathbb{R}^1$ are one-side unbounded intervals, we get as attributes "$\leq x$" and "$\geq x$" for all $x \in \mathbb{R}$. Despite the changes to the ground space, the extents and implications are similar to those described in Example 2. Using the notation introduced in Section 2, we obtain as the extent set

$$\left\{C \times \tilde{V} \times [a, b] \mid\ C \subseteq \mathbb{R}^2 \text{ closed convex set}, \tilde{V} \in \tbinom{V}{1} \cup V, a \leq b\ \right\}. \tag{1}$$

The set of implications are all statements $A \to B$ with $A \subseteq G$ and $B \subseteq \gamma_{\mathbb{R}^2} \circ \pi_{\mathbb{R}^2}(A) \times \tilde{V} \times [\min\{\pi_{\mathbb{R}}(A)\}, \max\{\pi_{\mathbb{R}}(A)\}]$ with $\tilde{V} = \pi_V(A)$ if $\#\pi_V(A) = 1$ and $\tilde{V} = V$ else. So if an implication $A \to B$ is true. Then all elements in B must be inside the smallest convex hull containing the spatial part of A. Furthermore, all elements must lie between the minimum and maximum value of the elevation component in A, and finally, if A consists of only one vegetation class, then all elements in B are of the same category.

The next step is to consider the calculation of ufg-implications $\mathcal{I}_{\text{ufg}}$. Therefore, we first utilize that the formal context can be divided into three formal contexts: the spatial, the elevation and the vegetation part.

Lemma 5.1. *For the formal context $\mathbb{K}_G$ with extent set given by (1), we have for the ufg-family of implications*

$$\mathcal{I}_{ufg} \subseteq \left\{ A \to B \middle| \begin{array}{l} A \subseteq \mathbb{R}^2 \times V \times \mathbb{R} \text{ and } 2 \leq \#A \leq 4, \\ \pi_{\mathbb{R}}(B) = [\min\{\pi_{\mathbb{R}}(A)\}, \max\{\pi_{\mathbb{R}}(A)\}],\ \pi_{\mathbb{R}^2}(B) = \gamma_{\mathbb{R}^2} \circ \pi_{\mathbb{R}}(A), \\ \pi_V(B) \in \tbinom{V}{1} \cup V : \pi_V(B) = \pi_V(A) \text{ if } \#\pi_V(A) = 1, \pi_V(B) = V \text{ else} \end{array} \right\}.$$

Figure 4 (right) shows the calculated ufg depth for the observed point pattern in 2006. The ufg-median is unique and has a depth of 0.765.[9] It is an observed point and lies within the primary vegetation category at an elevation of 1805. The ufg-median is thus in the most

[9] All values are rounded to three decimal places.

frequently observed vegetation category and is also relatively close to the median of the numerical elevation component (52.1% observed elevation values are strictly below 1805). The spatial component is also relatively close to the median of the median computed by the simplicial depth, see Figure 3 (left), where only the spatial part is considered. Note that the observation with the highest simplicial depth has an elevation value of 1900 and is therefore further away from the center from the perspective of the elevation component. The minimum depth value is zero. In particular, the ufg-depth is always zero when the elevation corresponding to a location is strictly below (or above) the minimum (or maximum) of the observed elevation values. This is the reason why the area in the center of the image, from a purely spatial perspective, has a low or even zero ufg-depth value. It can also be seen that the ufg-depth reflects that the vegetation categories colonization (1 point), grassland (7 points), secondary (5 points) and transitional (6 points) are not often observed.

5.2 Hierarchical-Nominal Data

As a further example, we analyze the ufg-depth for occupational data as described in Example 3. We use data from the German General Social Survey (GGSS) of the year 2021, see GESIS - Leibniz-Institut für Sozialwissenschaften (2023). Additionally, we also compare the ufg-depth to three other *measures of central tendency*. Namely one approach that only uses the categories on the finest level, secondly, a *top down* approach that analyses the frequencies of occupations successively, going from coarser levels to finer levels, and thirdly, the median based on the generalized Tukey depth. This comparison aims to illustrate the fact that the ufg-depth approach is different from these other approaches in a substantial way which is to some extent surprising given the meager structure of hierarchical-nominally scaled data.[10]

For the specification of the hierarchical categories of occupation we use the International Standard Classification of Occupations (ISCO) 2008. It consists of occupational categories on 4 levels with up to 10 categories on each level, coded by digits $0-9$. We analyze the set of all 2700 respondents for which the ISCO-08 status is a available. The sample is not drawn i.i.d., the respondents in east Germany were over-sampled. We account for this by simply reweighting the obtained empirical measure accordingly. Figure 5 (left and right) depicts the distribution of the occupations by drawing histograms on all 4 levels of the hierarchy of occupations.[11] While the left figure goes from Level-1 Category 3 to Level-1 Category 4, the right picture zooms into Level-1 Category 3 and goes from Level-2 Category 32 to Level-2 Category 33. The vertical lines indicate different further measures of location (orange: occupation with the highest ufg-depth; green: category with highest frequency on the finest level; purple: median according to the top down approach (see below)).

The *ufg-median*, i.e., the occupation with the highest ufg-depth, is occupation 3221: *Nursing Associate Professionals* (indicated with the orange vertical line in the plots). The depth for these persons was 0.927. The smallest ufg-depth value has occupation 6210:

[10]For example, for hierarchical-nominal data, the formal extents are either nested or they have an empty intersection.

[11]Level 1: black; Level 2: blue; Level 3: yellow; Level 4: pink. Left: Level-1 Category 3 to Level-1 Category 4. Right: Zoom into Level-1 Category 3 (from Level-2 Category 32 to Level-2 Category 33). The height of the bars corresponds to the absolute observed frequencies (counts) within the corresponding categories on a log scale.

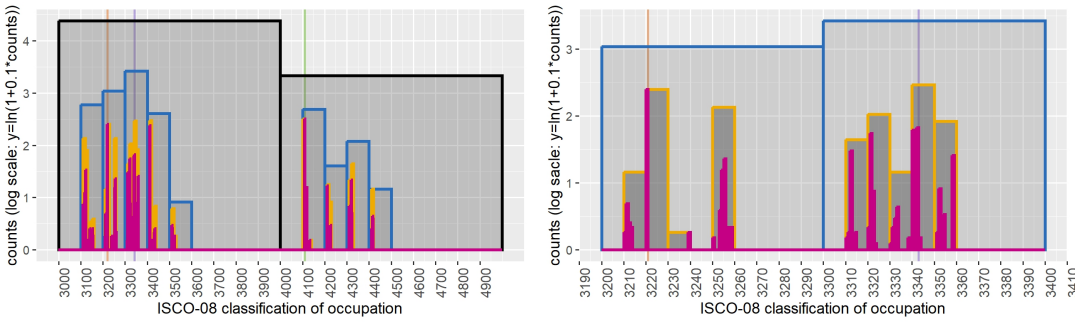

Figure 5: Histogram of the occupations on all 4 levels.

Forestry and related workers with a depth value of 0.824. The ufg-depth has all-together 285 unique depth values which induce 285 contour sets. These contour sets $Cont_{D,\alpha}$ induce three different attribute sets $\Psi(Cont_{D,\alpha})$ containing all attributes each object in the contour set has, namely these sets are $B_1 = \{$Level-4 Category 3221: *Nursing associate professionals*; plus all corresponding categories on the coarser Levels 1-3$\}$; $B_2 = \{$Level-1 Category 3: *Technicians and associate professionals*$\}$ and the $B_3 = \emptyset$. These three attribute sets induce the corresponding extents $\Phi(\Psi(Cont_{D,\alpha})) = \gamma(Con_{D,\alpha})$ that are (due to construction) exactly the contour sets of D^{qc}. Note that generally, for hierarchical-nominal data with L levels, a quasiconcave depth function, and in particular D^{qc}, can only have up to $L+1$ different depth values (c.f. the proof of Theorem 2.11 in the supplementary). Here, with 3 levels, the quasiconcave version D^{qc} is more flexible compared to e.g., the generalized Tukey depth in this data situation.[12]

We now compare the ufg-depth with other measures of central tendency for occupational data. First, note that the ufg-median here differs from the modus, i.e. the occupation (at the finest level) with the highest frequency. The modus is occupation 4110: *General office clerks* (indicated by the green vertical line). The modus only considers the categories at the most detailed level (Level 4). The structure at the more general Levels 1-3 is not taken into account at all. Opposed to this, the ufg-depth does take the other levels into account: First, note that the ufg-premises are exactly the one-element sets and the sets of two objects with different occupational ISCO-08 categories, see Theorem 3.10. in the supplementary. If we only consider the one-element ufg-premises, we end up with the frequencies at the finest levels. However, the ufg-depth approach also uses the two-element ufg-premises. Let $p = \{x, y\}$ with $x = x_1x_2 \dots x_{\ell-1}x_\ell \dots x_k$ and $y = x_1x_2 \dots x_{\ell-1}y_\ell \dots y_k$ be a two-element ufg-premise, where the first occurring classification difference is x_ℓ and y_ℓ. Then p contributes to the depth values of each object in $\gamma(p)$, which consists of all objects that share the first $\ell-1$ category assignments with x and y, i.e. are also categorized in the group $x_1x_2 \dots x_{\ell-1}$. For each object that can be distinguished from x and y based on the first $\ell-1$ categories, the ufg-premise p does not contribute to the ufg-depth of this object. Therefore, generally, the two-element ufg-premises contribute to the ufg-depth on all levels of the hierarchy. Let us now compare with a further construction of a median, which could be called the *top down* approach. A simple way of ordering the hierarchical categories is to look first at Level 1

[12]The generalized Tukey depth function has only two different depth values, which is typical, compare the discussion and the proof of the non-freenness of the generalized Tukey depth in Blocher et al. (2024).

and take the modal category at that level, i.e. the category with the highest frequency of occurrence (here Level 1 Category 3: *Technical and associate professionals*). Then, within this modal category, one could look at the subcategories at Level 2 and again take the modal subcategory at level 2 (here, Level 2 Category 33: *Business and administration associate professionals*), and so on. In our data set, this approach gives us the median occupation 3343: *Administrative and executive secretaries* (indicated by the purple vertical line). Like the ufg-depth approach, the top down approach uses all levels of the hierarchical structure. However, the mode at Level 1 predetermines the Level-1 category of the final median, and unlike the ufg-approach, if a data point does not fall into the modal Level 1 category, it can never become the median, even if – due to high frequencies – it is a clear median candidate from the perspective of all other levels. This property of predetermination of the Level 1 mode is also shared by the median according to the *generalized Tukey depth* (c.f., the proof of the failure of strong freenness of the generalized Tukey depth given in Blocher and Schollmeyer (2025), Theorem 10). In addition, the generalized Tukey depth has only two different depth values for this data set. Specifically, all occupations with Level-1 Category 3 have a generalized Tukey depth of 0.747, and all other occupations have a generalized Tukey depth of 0.0.710.

6 Conclusion

Providing statistical methods that take into account the underlying data structure is essential in statistics. The ufg-depth introduced here is a non-parametric and user-friendly method that uses the theory of formal concept analysis and data depth to define a statistical method for non-standard data. While this article presented and analyzed the ufg-depth and provided two descriptive examples showing the benefits of the ufg-depth, it also raised further research questions:

Statistical Inference: With the exception of the consistency property, the analysis of the ufg-depth and the examples focus on descriptive analysis. However, building on the consistency property, a further research question is how to define statistical inference tests. These tests can build on approaches provided by Li and Liu (2004) in $\mathbb{R}^d$.

Deeper analysis of the quasiconcave version of a depth function: In Section 4.2 we briefly touched on the topic of the quasiconcave version of a depth function and in Section 5.2 we showed that this does indeed provide a meaningful and non-trivial depth function. It is of interest to explore this topic in more detail, in particular with a closer look at $\mathbb{R}^d$ and the large variety of depth functions already defined for $\mathbb{R}^d$.

Other data sets: We applied the ufg-depth to two data types, the categorical-numerical-spatial data and the hierarchical-nominal data. In Blocher et al. (2024) the authors applied the ufg-depth to partial orders. These three data types are by no means all possible non-standard data. The investigation of further data types, and in particular of scaling methods that transform the data into a formal context, is a further interesting research area.

Further generalizations of depth functions in $\mathbb{R}^d$: So far the Tukey depth, see Tukey (1975), and the simplicial depth, see Liu (1990) are generalized to non-standard data. Also a discussion on the convex-hull-peeling depth, see Blocher et al. (2022), has been started. Similarly, this can be done with many other depth functions, such as the projection depth, see, e.g., Zuo and Serfling (2000b).

SUPPLEMENTARY MATERIAL

In the following, we provide the supplementary material and information to the main article *Union-Free Generic Depth for Non-Standard Data*. This includes a short introduction to formal concept analysis, a further discussion on the quasiconcave ufg-depth and all the proofs of the claims made in the main article. Unless otherwise stated, all references to equations, lemmas, etc. are to the supplementary material.

The repository corresponding to the main article can be found at `https://anonymous.4open.science/r/ufg_depth_application-0567/`(last accessed: 14.12.2024). There we also provide all the information about the reproducibility of the results in Section 6 of the main article.

7 Formal Concept Analysis

Formal concept analysis can be seen as applied lattice theory, which describes the relationship between data elements in a user-friendly and unified way. It is based on the formalization of a cross-table, see (Ganter and Wille, 2012, p. 17):

Definition 7.1. The triple $\mathbb{K} = (G, M, I)$ defines a *formal context* with G s set of *objects* and M a set of *attributes*. $I \subseteq G \times M$ states a binary relation between G and M.

In our case, the objects G correspond to the data described by the attributes M. Note that an object/data element can either have this attribute or not. While in some cases binary attributes, such as yes or no responses to a yes-no question, are naturally given, this is generally not the case. Therefore, we use so-called *scaling methods*, see (Ganter and Wille, 2012, Chapter 1.3.). These methods convert non-binary information about the data into attributes with a binary incidence relation. Examples can be found in Ganter and Wille (2012); Blocher and Schollmeyer (2025) and in the examples below. With the scaling method, we achieve that all types of data are presented through a formal context in a unified way.

Example 5. Recall Example 2 in the main article with $G = \mathbb{R}^2 \times V$ as ground space and attributes $M_{\mathbb{R}^2 \times V} = M_{\mathbb{R}^2} \cup M_V$. $I_{\mathbb{R}^2 \times V}$ now describes the incidence of both the spatial and the categorical component, where we say that the categorical attribute holds if the data element has that category. A snippet of this formal context is the joint (by the objects) tables of Figure 2 in the main article.

Especially this formalization of a cross-table is the basis to rigorously define the grouping procedure. Therefore, consider the following *derivation operators*, see (Ganter and Wille, 2012, p. 18):

$$\Psi : 2^G \to 2^M, A \to A' := \{m \in M \mid \forall g \in A \colon gIm\} \qquad \text{and}$$
$$\Phi : 2^M \to 2^G, B \to B' := \{g \in G \mid \forall m \in B \colon gIm\}.$$

Ψ maps a set of objects A to each attribute that each object in A has. Φ does the same, only with the roles of attribute set and object set reversed. In particular, the composition $\gamma_G := \Phi \circ \Psi$ now groups the objects based on the attributes in a maximal way. More precisely, the set $\gamma_G(A)$ composes all objects that share the same attributes given by $\Psi(A)$.

Definition 7.2. We call $\gamma_G(A)$ with $A \subseteq G$ an extent and $\Psi(A)$ an intent of $\mathbb{K}$. Additionally, we denote the set of all extents by $\mathcal{E}_G$.

The set of extents can be partially ordered using the subset relation. If $\gamma_G(A) \subseteq \gamma_G(B)$ for $A, B \subseteq G$, we can conclude that the objects in $\gamma_G(A)$ are more specific than those in $\gamma_G(B)$. This means that the attributes common to all objects in B are a subset of the attributes common to all objects in A, see (Ganter and Wille, 2012, Chapter 1.) for details. By examining this order on the extents, we can determine whether the relationship between the elements/objects is reasonable or not. We also get an idea of how fine the grouping is, i.e. if we are close to the power set.

In addition, the set of extents defines a closure system on G with the corresponding closure operator γ_G, which builds the bridge to lattice theory, see (Ganter and Wille, 2012, Chapter 0).

Definition 7.3. Let G be a set. Then $\gamma_G : 2^G \to 2^G$ is a *closure operator* on G if and only if γ_G is *extensive* (for all $A \subseteq G$: $A \subseteq \gamma_G(A)$), *monotone* (for all $A \subseteq B \subseteq G$ we have $\gamma_G(A) \subseteq \gamma_G(B)$) and *idempotent* (for all $A \subseteq G$, $\gamma_G(A) = \gamma_G(\gamma_G(A))$).

$\gamma_G(2^G)$ induces the corresponding *closure system*. Closure systems are families of sets which are closed under arbitrary intersections (let $(A_j)_{j\in J} \subseteq \gamma_G(2^G)$ then $\cap_{j\in J} A_j \in \gamma_G(2^G)$) and contain the entire set $G \in \gamma_G(2^G)$.[13]

Note that there exists a one-to-one correspondence between closure operators and closure systems/extents $\mathcal{E}_G$, see (Ganter and Wille, 2012, p. 8).

Example 6. Consider the formal context $\mathbb{K}_{\mathbb{R}^2\times V}$ defined in Example 5. Then we get as set of extents $\mathcal{E}_{\mathbb{R}^2\times V} = \left\{ C \times \tilde{V} \mid \ C \text{ topologically closed convex set } \wedge\ \tilde{V} \in \binom{V}{1} \cup V \right\}$. Note that due to nominal scaling, $\tilde{V}$ either has cardinality one or is directly the entire set. This follows from the fact that if two different categories are grouped together, then the relation between these two categories is the same as to any other category, so all other categories are also included in order not to state a relation between these two categories that does not exist.

As we saw in Example 6, the closure system/extent set contains the structure of the data and describes the dependencies between data elements. This becomes even clearer when we exploit the fact that every closure system can be described by a family of implications. In the context of the closure operator γ_G, we define implications as follows, see (Ganter and Wille, 2012, Chapter 2.3):[14]

Definition 7.4. Let G be a set. An *implication* is a tuple $(A_1, A_2) \in G \times G$. We say that A_1 implies A_2 and denote this by $A_1 \to A_2$. We call A_1 the *premise* and A_2 the *conclusion* of the implication $A_1 \to A_2$.

Let γ_G be a closure operator on G with a corresponding closure system $\mathcal{E}_G$. Then, the closure system defines a family of implications consisting of statements $A_1 \to A_2$ with $\gamma_G(A_1) \supseteq \gamma_G(A_2)$. The family of all implications provided by $\mathcal{E}_G$ is denoted by $\mathcal{I}_G$. For a given closure system $\mathcal{E}_G$ we say that an *implication* $A_1 \to A_2$ *holds* if and only if $\gamma_G(A_1) \supseteq \gamma_G(A_2)$.

[13] In the following, we use both the terms "extent set" and "closure system", depending on whether we want to emphasize that it is based on a formal context or that we exploit the mathematical structure.

[14] Note that in Ganter and Wille (2012) the authors discuss attribute implications. The results can be applied to object implications discussed here.

Example 7. Recall Example 5 and 6. For $\mathbb{K}_{\mathbb{R}^2\times V}$ we have as family of all implications

$$\mathcal{I}_{\mathbb{R}^2\times V} = \left\{ A \to B \,\middle|\, \begin{array}{l} A \subseteq \mathbb{R}^2 \times V \text{ and } \pi_{\mathbb{R}^2}(B) \subseteq \gamma_{\mathbb{R}^2} \circ \pi_{\mathbb{R}^2}(A) \text{ and} \\ \pi_V(B) \in \binom{V}{1} \cup V : \pi_V(A) = \pi_V(B) \text{ if } \Pi_V(A) = \#1, \pi_V(B) = V \text{ else} \end{array} \right\}$$

with $\pi_{\mathbb{R}^2} : \mathbb{R}^2 \times V \to \mathbb{R}^2$ being the projection onto $\mathbb{R}^2$. Similarly, we set π_V.

From the definition of a closure system the definition of the family of implications is straight forward. Reverse, one can obtain a closure system based on a family of implications as follows.

Definition 7.5. Let $\mathcal{I}_G$ be a family of implications. We say that $D \subseteq G$ *respects an implication* $A \to B$ if and only if either $A \not\subseteq D$ or $A \subseteq D$ then $B \subseteq D$ also follows. We set $\mathcal{E}_{\mathcal{I}_G} = \{D \subseteq G \mid D \text{ respects every implication in } \mathcal{I}_G\}$.

As this definition already suggests, there is a one-to-one correspondence between closure systems/operators and the set of all closed families of all implications:

Lemma 7.6. *Let $\mathcal{E}_G$ be a closure system and $\mathcal{I}_G$ the family of all implications that respect $\mathcal{E}_G$. Then $\mathcal{I}_G$ is unique and $\mathcal{E}_G = \mathcal{E}_{\mathcal{I}_G}$. In particular, this then states that $\mathcal{I}_G$ uniquely defines a closure system.*

Proof. The uniqueness follows directly. For the second part, assume in contradiction that $\mathcal{E}_G \neq \mathcal{E}_{\mathcal{I}_G}$. In the first case, we assume that $E \in \mathcal{E}_G \setminus \mathcal{E}_{\mathcal{I}_G}$. Since $E \notin \mathcal{E}_{\mathcal{I}_G}$, there exists an implication $A \to B \in \mathcal{I}_G$ with $A \subseteq E$, but $B \not\subseteq E$. But since $\mathcal{I}_G$ consists of all implications that hold for $\mathcal{E}_G$ this implies that $E \notin \mathcal{E}_G$. This contradicts the assumption.

For the reverse, assume that $E \in \mathcal{E}_{\mathcal{I}_G} \setminus \mathcal{E}_G$. This means that for all implications $A \to B \in \mathcal{I}_G$, if $A \subseteq E$, then $B \subseteq E$ is also true. Since $E \notin \mathcal{E}_G$ we get that $g \in \gamma(E) \setminus E$. However, this implies that $E \to g$ is an implication that holds in $\mathcal{E}_G$ and therefore should lie in $\mathcal{I}_G$. So $E \notin \mathcal{E}_{\mathcal{I}_G}$, which is a contradiction, and we obtain the claim. □

Before we continue, let us take a closer look at the set of implications $\mathcal{I}_G$. We can immediately see that some implications follow semantically from others. For example, if $A_1 \to A_2$ and $B \supseteq A_1$, then we get $B \to A_2$. So the implication $B \to A_2$ is somewhat redundant, since it follows from $A_1 \to A_2$. These semantic structures are summarized by Maier (1983) as inference axioms, see (Maier, 1983, p. 45): Let $A, B, C, D, A_1, A_2, B_1, B_2 \subseteq G$. Then we say that the axiom of *reflexivity* holds iff $A \to A$, the axiom of *augmentation* holds iff $A_1 \to B$ implies $A_1 \cup A_2 \to B$, the axiom of *additivity* holds iff $A \to B_1$ and $A \to B_2$ imply $A \to B_1 \cup B_2$, axiom of *projectivity* holds iff $A \to B_1 \cup B_2$ implies $A \to B_1$, axiom of *transitivity* holds iff $A \to B$ and $B \to C$ imply $A \to C$, and the axiom of *pseodotransitivity* holds iff $A \to B$ and $B \cup C \to D$ imply $A \cup C \to D$.

Armstrong proved, see Armstrong (1974), that the iterative repetition of these inference axioms on a set of implications (on a set G) leads to a family of implications that equals the set of all implications that hold for a closure system on G. Note, however, that when deleting implications that follow from others, one may delete too many implications and end up not representing the same closure system, see Section 3.2. of the main article. Therefore, we say that a family of implication $\mathcal{I}_G$ is complete iff every implication that holds for a closure system follows semantically from $\mathcal{I}_G$, see (Ganter and Wille, 2012, p. 81):

Definition 7.7. Let $\mathcal{E}_G$ be a closure system with corresponding closure operator γ_G and $\mathcal{I}_G$ a family of implications. Then $\mathcal{I}_G$ is *complete* w.r.t $\mathcal{E}_G$ if and only if $\mathcal{E}_G = \mathcal{E}_{\mathcal{I}_G}$.

Remark 3. Finally, we want to point out that everything, the closure system, the implications and later the ufg-depth, depends on the application of a reasonable scaling method. The closure system and the implications provide a tool for analyzing/discussing the relational structure in detail, but the starting point is the scaling method. In particular, all the underlying assumptions of the ground space structure are determined by the scaling method. For example, consider the GORILLAS example in Section 2 of the main article. There the ground space is $\mathbb{R}^2 \times V$. Therefore, if we have two observations in the same place with the same vegetation, we assume them to be duplication of the same objects. Another approach, not discussed here, is to consider each individual nesting point as an observation that cannot be a duplicate, but is another object in the ground space. In this way we can observe more than one object at the same place with the same vegetation. The main article sticks to the first perspective given in Example 1 and 2.

8 Claims and Proofs

In this section we present the proofs for the claims made in the main part that do not have a reference to the literature containing the proof. We divided the claims into the corresponding sections in the main article. Since we discuss further lemmas to show the claims, the enumeration of lemmas, theorems, etc. differs from that in the main article.

Claims and Proofs of Section 4 - The Union-Free Generic Depth

First of all, we consider a general observation for ufg-implications.

Lemma 8.1. *Let $\mathbb{K} = (G, M, I)$ be a formal context. Then we have $A \in \mathcal{I}_{ufg}^{prem}$ if and only if there exists $b \in \gamma(A)$ such that for all $a \in A$ and all $\tilde{a} \in A \setminus a$ exists $m \in \Psi(A \setminus a)$ with $(\tilde{a}, m) \in I$ and $(b, m), (a, m) \notin I$. In other words, A is an ufg-premise if and only if there exists an element in the conclusion where every element in A is needed.*

Proof. The claim that the second statement implies the first statement follows directly by the definition of the ufg-premise. For the reverse, assume that A is an ufg-premise. Then for $A_g = A \setminus \{g\}$ with $g \in A$ we have by Condition (C2) that $g \in \gamma(A) \setminus \cup_{g \in A} \gamma(A_g)$ which is exactly the second statement. □

In Section 4 of the main part of the article, we formalize that the triangles in $\mathbb{R}^2$ together with the convex closure operator define indeed the set of ufg-implications based on formal context $\mathbb{K}_{\mathbb{R}^2}$.

Lemma 8.2. *For $\mathbb{K}_{\mathbb{R}^2}$, the spatial formal context of Example 1 in Section 2, we have $\mathcal{E}_{\mathcal{I}_{\mathbb{R}^2,ufg}} = \mathcal{E}_{\mathcal{I}_{\mathbb{R}^2}}$. Moreover, $\mathcal{I}_{\mathbb{R}^2,ufg}$ is the reduced version of $\mathcal{I}_{\mathbb{R}^2}$ without the implications following from the Armstrong rules of reflexivity, augmentation, additivity, and projectivity.*

Proof. First, we prove $\mathcal{E}_{\mathcal{I}_{\mathbb{R}^2}} = \mathcal{E}_{\mathcal{I}_{\mathbb{R}^2,\text{ufg}}}$. Since $\mathcal{I}_{\mathbb{R}^2,\text{ufg}} \subseteq \mathcal{I}_{\mathbb{R}^2}$, we have that if $D \subsetneq \mathbb{R}^2$ respects all implications in $\mathcal{I}_{\mathbb{R}^2}$, then it also respects all implications in $\mathcal{I}_{\mathbb{R}^2,\text{ufg}}$. Therefore, $\mathcal{E}_{\mathcal{I}_{\mathbb{R}^2}} \subseteq \mathcal{E}_{\mathcal{I}_{\mathbb{R}^2,\text{ufg}}}$. For the subset relation, let $D \in \mathcal{E}_{\mathcal{I}_{\mathbb{R}^2,\text{ufg}}}$ and $A \to \gamma_{\mathbb{R}^2}(A)$ be an arbitrary implication

in $\mathcal{I}_{\mathbb{R}^2} \setminus (\mathcal{I}_{\mathbb{R}^2,\text{ufg}} \cup \{A \to A \mid \#A = 1\})$. By Carathéodory's theorem, see Eckhoff (1993), we get that for every $g \in \gamma_{\mathbb{R}^2}(A)$ there exist $a_1^g, a_2^g, a_3^g \in A$ such that $g \in \gamma_{\mathbb{R}^2}(\{a_1^g, a_2^g, a_3^g\})$. In particular, $\{a_1^g, a_2^g, a_3^g\} \to \gamma_{\mathbb{R}^2}(\{a_1^g, a_2^g, a_3^g\}) \in \mathcal{I}_{\mathbb{R}^2,\text{ufg}}$. Since D respects all implications in $\mathcal{I}_{\mathbb{R}^2,\text{ufg}}$, $\gamma(A) \supseteq \bigcup_{g\in\gamma_{\mathbb{R}^2}(A)}\{a_1^g, a_2^g, a_3^g\}$ and $\bigcup_{g\in\gamma(A)} \gamma_{\mathbb{R}}(\{a_1^g, a_2^g, a_3^g\}) = \gamma_{\mathbb{R}^2}(A)$ we get that D also respects $A \to \gamma_{\mathbb{R}^2}(A)$. So we have $\mathcal{E}_{\mathcal{I}_{\mathbb{R}^2}} = \mathcal{E}_{\mathcal{I}_{\mathbb{R}^2,\text{ufg}}}$.

Now we have to show that $\mathcal{I}_{\mathbb{R}^2,\text{ufg}}$ does not contain any further implications that follow from reflexivity, augmentation, additivity, and projectivity. Since the conclusion is set to $\gamma_{\mathbb{R}^2}(A)$ for some premise A, we get that there cannot be a proper superset of $\gamma_{\mathbb{R}^2}(A)$ such that the implication holds for the convex sets. A similar argument provides that it is reduced for the additivity rule. The reflexivity and augmentation follow from the fact that we consider non-degenerate simplices together with Carathéodory's Theorem. □

Moreover, we provide generally that the maximal cardinality of an ufg-premise is bounded by the VC-dimension of the extent set of the formal context.

Lemma 8.3. *Let $\mathbb{K}$ be a formal context that has a unique ufg-family of implications $\mathcal{I}_G$. Let vc be the VC dimension of the extent sets. Then* $\max\{\#A \mid A \in \mathcal{I}_{G,ufg}^{prem}\} \leq vc$.

Proof. This proof is a slight adaptation of the proof given in Blocher et al. (2024), Theorem 4. To prove $\max\{A \in \mathcal{I}_{G,\text{ufg}}^{prem}\} \leq vc$ take an arbitrary subset $Q = \{g_1, \ldots, g_k\}$ and ufg-premise of size $k > vc$. Then this subset is not shatterable because vc is the largest cardinality of a shatterable set. Thus, there exists a subset $R \subseteq Q$ that cannot be obtained as an intersection of Q and some $\gamma(A)$ with $A \subseteq G$. In particular, this holds for $R = A$. Thus, $R \neq \gamma(R) \cap Q$ and with the extensitivity of γ we get $R \subsetneq \gamma(R) \cap Q$. This means that there exists an object $\tilde{g}$ in $\gamma(R) \cap Q \backslash R$ for which the formal implication $R \to \{\tilde{g}\}$ holds. Thus, (because of the Armstrong rules, cf., (Armstrong, 1974, p. 581)) the object $\tilde{g}$ is redundant in the sense of $Q\backslash\{\tilde{g}\} \to Q$ and thus Q is not minimal with respect to γ. Therefore, Q is not an ufg-premise which completes the proof. □

Claims and Proofs of Section 5 - Structural Properties

Section 5 in the main article discusses the structural properties of depth functions using formal concept analysis given by Blocher and Schollmeyer (2025). Here, we provide the proofs to the claims done in the main article.

Theorem 8.4. *Let $P, \tilde{P} \in \mathcal{P}_G$ be two probability measures on G and let $\mathbb{K}, \tilde{\mathbb{K}} \in \varkappa$ be two formal contexts on G.*
Invariance on the extents: Assume that there exists a bijective and bimeasureable function $i : G \to G$ such that the extents are preserved (i.e. E extent w.r.t. $\mathbb{K} \Leftrightarrow i(E)$ extent w.r.t. $\tilde{\mathbb{K}}$) and the probability as well (i.e. $P(E) = \tilde{P}(i(E))$). Then $D_G(g, \mathbb{K}, P) \leq D_G(\tilde{g}, \mathbb{K}, P) \Leftrightarrow \tilde{D}_G(i(g), \tilde{\mathbb{K}}, \tilde{P}) \leq \tilde{D}_G(i(\tilde{g}), \tilde{\mathbb{K}}, \tilde{P})$ is true for all $g, \tilde{g} \in G$.
Invariance on the attributes: Let $g_1, g_2 \in G$ with $\Psi_{\mathbb{K}}(g_1) = \Psi_{\mathbb{K}}(g_2)$, then $D(g_1, \mathbb{K}, P) = D(g_2, \mathbb{K}, P)$ holds.

Proof. Observe that the ufg-depth is based on the extent set. Thus, if two formal contexts result in the same extent set and the probability measure is also preserved by a function i, then the ufg-depth does not change. For the invariance on the attributes we use that for every $E \in \mathcal{E}$ we have $g_1 \in E$ iff $g_2 \in E$. So g_1 is in an ufg-conclusion iff g_2 is in the ufg-conclusion and therefore the ufg-depths must be equal. □

Theorem 8.5. *Let $P \in \mathcal{P}_G$ and formal context $\mathbb{K} \in \varkappa$ with $g_1, g_2 \in G$ such that $\gamma_{\mathbb{K}}(\{g_1\}) \supseteq \gamma_{\mathbb{K}}(\{g_2\})$. Then the isotonicity property $D(g_1, \mathbb{K}, P) \leq D(g_2, \mathbb{K}, P)$ is true.*

Proof. This follows immediately from the fact that for every ufg-premise U with $U \to g_1$ we have $U \to g_2$. So the probability of ufg-conclusions containing g_1 is a smaller than of those containing g_2, which provides the claim. □

Theorem 8.6. *Let $D(\cdot, \mathbb{K}, P)$ be a depth based on formal concept analysis. Then D^{qc} is a quasiconcave function.*

Proof. We show that for every $\alpha \in \mathbb{R}$ $\gamma(Cont_{D^{qc},\alpha}) = Cont_{D^{qc},\alpha}$ is true. Assume in contradiction that there exists $\alpha \in \mathbb{R}$ such that $g \in \gamma(Cont_{D^{qc},\alpha}) \setminus Cont_{D^{qc},\alpha}$. Then $Cont_{D^{qc},\alpha} \to g$ is a valid implication.

Case 1: $Cont_{D^{qc},\alpha} = \emptyset$. Then we know that every subset $A \subseteq G$ implies g. Hence, for every α we get that $Cont_{D,\alpha} \to g$ is true and therefore $D^{qc}(g, \mathbb{K}, P)$ has a maximum depth value. So it can never contradict the quasiconcavity assumption by having a depth value that is too small.

Case 2: $Cont_{D^{qc},\alpha} \neq \emptyset$. Since $D^{qc}(g, \mathbb{K}, P) < \alpha$, we get that there exists $\varepsilon > 0$ such that for every $\alpha' > \alpha - \varepsilon$ it holds that $Cont_{D,\alpha'} \not\to g$ is true, otherwise the depth of $D^{qc}(g, \mathbb{K}, P)$ must be at least α'. However, the construction of the quasiconcave depth function gives us that for every $\alpha' < \alpha$ we have $Cont_{D,\alpha'} \to Cont_{D^{qc},\alpha}$. So by the inference axioms we know that $Cont_{D,\alpha'} \to g$ is also valid for every $\alpha - \varepsilon < \alpha' < \alpha$ which is a contradiction. □

Theorem 8.7. *Let $\mathbb{K} \in \varkappa$. Let $g_1, \ldots, g_n$ be a sample of G with $n \in \mathbb{N}$. We denote with $P^{(n)}$ the empirical probability measure given by $g_1, \ldots, g_n$ and by $P^{(n,-\ell)}$ the empirical probability measure based on $g_1, \ldots, g_{\ell-1}, g_{\ell+1}, \ldots, g_n$ with $\ell \in \{1, \ldots, n\}$.*
Respecting duplication: Let $i, \ell \in \{1, \ldots, n\}$ with $i \neq \ell$ and for every extent $E \in \mathcal{E}$ we have $g_\ell \in E$ iff $g_i \in E$. Moreover, assume that there exists $j \in J_{P,\mathbb{K}}$ and ufg-premises $A_1, A_2 \in 2^{\{g_1,\ldots,g_{\ell-1},g_{\ell+1},\ldots,g_n\}} \cap \mathcal{I}_{ufg}^{prem,j}$ with $g_i \in A_1$ and $g_i \notin \gamma(A_2)$. Then, we have $D_G(g_i, \mathbb{K}, P^{(n,-\ell)}) < D_G(g_i, \mathbb{K}, P^{(n)})$.
Stability of the order: Assume that the only extents E that contains g_ℓ for $\ell \in \{1, \ldots, n\}$ as well as any subset of $g_1, \ldots, g_{\ell-1}, g_{\ell+1}, \ldots$ is $E = G$. Then for $g, \tilde{g} \in \{g_1, \ldots, g_{\ell-1}, g_{\ell+1}, \ldots, g_n\}$ we have $D_G(g, \mathbb{K}, P^{(n)}) \leq D_G(\tilde{g}, \mathbb{K}, P^{(n)}) \Leftrightarrow \tilde{D}_G(g, \mathbb{K}, P^{(n,-\ell)}) \leq \tilde{D}_G(\tilde{g}, \mathbb{K}, P^{(n,-\ell)})$.

Proof. The assumption of the respecting duplication property implies that there are g_i, g_ℓ in the sample with $i \neq j$ and $\Psi(g_i) = \Psi(g_\ell)$. We assume that in the full sample there exists a further ufg-premise containing g_i than in the reduced sample. Since the proportion is not already one, this provides the claim.

Now, we show the stability of the order property. Let g_ℓ be an element of the sample that is completely different to the rest. Then for every ufg-premise $A \in 2^{\{g_1,\ldots,g_n\}} \cap \mathcal{I}_{ufg}^{prem}$ we have that either $g_\ell \in A$ and the entire sample lies in the conclusion or $g_\ell \notin A$. Thus, for every $j \in J_{P,\mathbb{K}}$ and every $g \in \{g_1, \ldots, g_{\ell-1}, g_{\ell+1}, \ldots, g_n\}$ the amount added in the proportion equals when including the observation g_ℓ.

□

Remark 4. Note that the assumptions on the existence of the two ufg-implications in the invariance on the extents property are indeed necessary. The first assumption, that there exists an implication $A_1 \to \gamma(A_1)$ with $g_i \in A$, is necessary because otherwise this object

g_i, and hence also object g_ℓ, has no effect on the ufg-depth. Note that this assumption is generally true, and in particular holds for all the examples discussed in the main article.

The second assumption, that there is an ufg-premise $A_2 \to \gamma(A_2)$ with $g_i \notin \gamma(A_2)$, ensures that the proportion does indeed increase. If there is no such ufg-implication, then the ufg-depth of g_i is already maximal and therefore cannot increase. (Note that this property has a strictly larger in its definition).

Theorem 8.8. *Let $\mathbb{K} \in \varkappa, P \in \mathcal{P}$ and $J_{P,\mathbb{K}} \subseteq \mathbb{N}$ be the same as in Definition 4.2. of the main article. Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$. We assume that $\#J_{P,\mathbb{K}} < \infty$. Moreover, we assume that for every $j \in J_{P,\mathbb{K}}$ $\mathcal{I}^{concl,j}_{\mathbb{K},ufg}$ has finite VC-dimension. With this, we get the consistency property with $\sup_{g\in G} |D(g, \mathbb{K}, P^{(n)}) - D(g, \mathbb{K}, P)| \to 0$ almost surely for n to infinity. (We assume that this supremum is measurable.)*

Proof. Recall the notation defined before Definition 4.2. of the main article. The proof can be divided into three parts. First, we proof that for all $j \in J_{P,\mathbb{K}}$: $\sup_{g\in G} |U^j_{(X_1,\ldots X_n)}[f^j_g] - \mathbb{E}[f^j_g]| \to 0$ almost surely. Second, we show $|\frac{C_j}{U^j_{(X_1,\ldots X_n)}[h^j]} - \frac{C_j}{\mathbb{E}[h^j]}| \to 0$ almost surely for every $j \in J_{P,\mathbb{K}}$. In the last part, we combine the first two to provide the claim.

Part 1: For $j \in J_{P,\mathbb{K}}$ consider the dual set $\tilde{\mathcal{F}}^j$. Since $\mathcal{I}^{concl,j}_{\mathbb{K},\text{ufg}}$ has a finite VC-dimension, the sub-graphs of $\tilde{\mathcal{F}}^j$ have also a finite VC-dimension. By Assouad (1983) we obtain that the sub-graphs of $\mathcal{F}^j$ have also finite VC-dimension. With Arcones and Giné (1993) we get the first part.

Part 2: Let $j \in J_{P,\mathbb{K}}$. We use Theorem 2.3. of Christofides (1992) to obtain $|U^j_{(X_1,\ldots X_n)}[h^j] - \mathbb{E}[h^j]| \to 0$ almost surely. In particular, this also implies that $U^j_{(X_1,\ldots X_n)}[h^j]$ is almost surely positive for n large enough (note that $\mathbb{E}[h^j] > 0$ for $j \in J_{P,\mathbb{K}}$). Hence, the function $x \to 1_{x>0}1/x$ is almost surely evaluated only at a positive argument, if n is large enough. Therefore, because this function is continuous for positive arguments we obtain Part 2.

Part 3: For $j \in J_{P,\mathbb{K}}$ we consider the following inequality, which uses that function h^j is independent of $g \in G$, a decomposition of the factors and the triangle inequality.

$$\begin{aligned}
&\sup_{g\in G}\left|\frac{C_j}{U^j_{(X_1,\ldots X_n)}[h^j]}U^j_{(X_1,\ldots X_n)}[f^j_g] - \frac{C_j}{\mathbb{E}[h^j]}\mathbb{E}[f^j_g]\right| \\
\leq &\sup_{g\in G}\left|U^j_{(X_1,\ldots X_n)}[f^j_g]\right|\left|\frac{C_j}{U^j_{(X_1,\ldots X_n)}[h^j]} - \frac{C_j}{\mathbb{E}[h^j]}\right| + \left|\frac{C_j}{\mathbb{E}[h^j]}\right|\sup_{g\in G}\left|U^j_{(X_1,\ldots X_n)}[f^j_g] - \mathbb{E}[f^j_g]\right|
\end{aligned}$$

Since the first two components of the multiplications can be bounded by above for every g and every (empirical) probability measure. We obtain with Part 1 and 2 and $J_{P,\mathbb{K}}$ being finite the claim.

□

Theorem 8.9. *Let $\mathbb{K} = (G, M, I)$ be a formal context and $J_{P,\mathbb{K}}$ be finite with $\sup\{j \mid P \in \mathcal{P}, j \in J_{P,\mathbb{K}}\} = K \in \mathbb{N}$. Additionally, we assume that $\mathbb{K}$ satisfies the following further conditions:*

(A1) For all $A \subseteq G$ and all $B \subseteq G \setminus \gamma(A)$ with B finite there exists an extent $S \subseteq G\setminus B$ such that for all $g \in A$ there exists an ufg-premise $U \subseteq S$ with $U \to \{g\}$.

(A2) There exists $L > 0$ such that $\frac{\max_{j \in J_{P,\mathbb{K}}} C_j/\mathbb{E}[h^j]}{\min_{j \in J_P} C_j/\mathbb{E}[h^j]} \leq L$ for every $P \in \mathcal{P}$.

Then D^{qc} is approximately weakly free w.r.t. quasiconcavity in the following sense: For every quasiconcave depth function E on $\mathbb{K}$, for every probability measure $P \in \mathcal{P}$ and every finite $\tilde{G} \subseteq G$ there exists a probability measure P^* on G (with finite support) such that for all $g, \tilde{g} \in \tilde{G}$ we have

$$E(g, \mathbb{K}, P) > E(\tilde{g}, \mathbb{K}, P) \Rightarrow (D_{|\tilde{G}})^{qc}(g, \mathbb{K}, P^*) > (D_{|\tilde{G}})^{qc}(\tilde{g}, \mathbb{K}, P^*). \tag{2}$$

Proof. The proof is divided into two parts. First we define the probability measure P^* and show that the depth values $D(g)$ of the objects in $\tilde{G}$ w.r.t. the original ufg-depth D satisfy (2). In the second step, we consider the depth values of the quasiconcave version D^{qc} of D and show that they also fulfill property (2).

Part 1: Let $e_{(1)}, \ldots, e_{(k)}$ be the increasingly ordered (unique) depth values of $E(\tilde{G}, \mathbb{K}, P)$ and let $G_{(i)} = \{g \in \tilde{G} \mid E(g, \mathbb{K}, P) = e_{(i)}\}$.

Now, we go step by step through the layers $G_{(i)}$ given by E. First we set $P^{(k+1)} = 0$ for all $g \in G$ and will modify it in the following process.

Step 1: We start with layer $G_{(k)}$ corresponding to the highest value $e_{(k)}$. We set $A_{(k)} = G_{(k)}$ and $B_{(k)} = \tilde{G} \setminus A_{(k)} \subseteq G \setminus A_{(k)}$. Note that $B_{(k)}$ is finite. Due to assumption (A1) there exists an extent $S_{(k)} \subseteq G \setminus B_{(k)}$ such that for all $g \in A_{(k)}$ there exists an ufg-premise $U_g \subseteq S_{(k)}$ with $U_g \to g$. Since $S_{(k)}$ is an extent, we know that no element of $B_{(k)}$ is implied by U_g. We set $U_{(k)} = \bigcup_{g \in A_{(k)}} U_g$, $\mathcal{U}_{(k)} = \{U_g \mid g \in A_{(k)}\}$.

Step 2: Now, we proceed with $G_{(k-1)}$. Similar to Step 1 we set $A_{(k-1)} = G_{(k)} \cup G_{(k-1)}$ and $B_{(k-1)} = \tilde{G} \setminus A_{(k-1)}$. Again, Assumption (A1) provides us a set $S_{(k-1)} \subseteq G \setminus B_{(k-1)}$ such that for every $g \in A_{(k-1)}$ exists an ufg-premise $U_g \subseteq S_{(k-1)}$ with $U_g \to g$ but U_g implies no element in $B_{(k-1)}$. Note that for every such U_g there exists at least one $u \in U_g$ with $u \notin U_{(k)}$, because otherwise we would have $U_g \to g$. We set $U_{(k-1)} = \bigcup_{g \in A_{(k-1)}} U_g$, $\mathcal{U}_{(k-1)} = \{U_g \mid g \in A_{(k-1)}\}$.

Step 3 to $k-1$. We proceed similar until we defined $\mathcal{U}_{(1)}$ and $A_{(1)}$.

Part 2: Now, we set $U = \bigcup_{i=1}^{k} U_{(i)}$, $\mathcal{U} = \bigcup_{i=1}^{k} \mathcal{U}_{(i)}$ and $c := L \cdot \#\mathcal{U} \cdot k$. Note that c does not depend on the probability measure P^* that we will now construct. Based on the above definitions, we can now define again successively the probability measure P^* which will have support U:

Step a: We define $p_{(k)}$ and $P^{(k)}$ corresponding to the contour set of E having the highest depth value $e_{(k)}$.

$$0 < p_{(k)} \leq \frac{1}{\#U \cdot k}$$

$$P^{(k)}(g) := \begin{cases} p_{(k)}, & g \in U_{(k)} \\ 0 & else \end{cases}$$

Step b: We define $p_{(k-1)}$ and $P^{(k-1)}$ corresponding to the contour set of E having the second highest depth value $e_{(k)}$.

$$0 < p_{(k-1)} < p_{(k)} \text{ such that } p_{(k)}^K > c \cdot p_{(k-1)}$$
$$P^{(k-1)}(g) := \begin{cases} p_{(k-1)}, & g \in U_{(k-1)} \\ 0, & else \end{cases}$$

Step c and on: Analogously to Step b we define $p_{(i)}$ and $P^{(i)}$ for $i \in \{1, \dots k-2\}$. Note that also for arbitrary i, j with $i < j$ we have $p_{(j)}^K > c \cdot p_{(i)}$.

Finally, we set $P(g) = \sum_{i=1}^{k} P^{(i)}(g)$ for all $g \in U$ and with this we obtain that

$$0 < \sum_{g \in U} P(g) \leq \sum_{g \in U} \left(\sum_{i=1}^{k} P^{(i)}(g) \right) \leq \sum_{g \in U} \left(\sum_{i=1}^{k} \frac{1}{\#U \cdot k} \right) \leq 1.$$

Thus, we get $0 \leq \varepsilon := 1 - \sum_{g \in U} P(g) \leq 1$ and with this we can now define the probability measure P^*:

$$P^*(g) = \begin{cases} P(g) + \frac{\varepsilon}{\#U_{(k)}}, & g \in U_{(k)} \\ P(g), & else \end{cases}.$$

With this, P^* defines a probability measure on G.

Let us take a look at the upper and lower bounds of the ufg-depth $D(g, \mathbb{K}, P^*)$ for $g \in \tilde{G}$. Let $i \in \{1, \dots, k\}$ and $g \in G_{(i)}$. We know that there exists at least one ufg-premise U in $\mathcal{U}_{(i)}$ with maximal cardinality K and with $U \to \{g\}$. Thus, we have as lower bound for $g \in G_{(i)}$

$$D(g, \mathbb{K}, P^*) \geq \frac{C_i}{\mathbb{E}\left[h^i\right]} p_{(i)}^K$$

For $i \in \{1, \dots, k-1\}$ and $g \in G_{(i)}$, we get as an upper bound:

$$D(g, \mathbb{K}, P^*) \leq \max_{j \in J} \frac{C_j}{\mathbb{E}[h^j]} \#\mathcal{U} \cdot k \cdot p_{(i)}$$

With this, for $i, \ell \in \{1, \dots, k\}$ with $i < \ell$ and $g \in G_{(i)}$ and $\tilde{g} \in G_{(\ell)}$ we immediately get

$$\begin{aligned} D(g, \mathbb{K}, P^*) &\leq \max_{j \in J} \frac{C_j}{\mathbb{E}[h^j]} \cdot \#\mathcal{U} \cdot k \cdot p_{(i)} \\ &\leq L \cdot \frac{C_\ell}{\mathbb{E}[h^\ell]} \cdot \#\mathcal{U} \cdot k \cdot p_{(i)} \\ &= \frac{C_\ell}{\mathbb{E}(h^\ell)} \cdot c \cdot p_{(i)} \\ &\leq \frac{C_\ell}{\mathbb{E}(h^\ell)} \cdot p_\ell^K \leq D(\tilde{g}, \mathbb{K}, P^*). \end{aligned}$$

Until now, we showed that for $g, \tilde{g} \in \tilde{G}$ we have $E(g, \mathbb{K}, P) > E(\tilde{g}, \mathbb{K}, P) \Longrightarrow D(g, \mathbb{K}, P^*) > D(\tilde{g}, \mathbb{K}, P^*)$. Now we show that the same holds for $\left(D_{|\tilde{G}}\right)^{qc}$: Let $i, j \in \{1, \ldots, k\}$ with $j < i$ and let $g_i \in G_i, g_j \in G_j$. Then we know that

$$D(g_i, \mathbb{K}, P^*) > D(g_j, \mathbb{K}, P^*).$$

Assume in contradiction that $(D_{|\tilde{G}})^{qc}(g_i, \mathbb{K}, P^*) \leq (D_{|\tilde{G}})^{qc}(g_j, \mathbb{K}, P^*)$. Then there exists a set $A \subseteq G_{(j+1)} \cup \ldots G_{(k)}$ such that the implication $A \to g_j$ with $D(g, \mathbb{K}, P^*) \geq D(g_i, \mathbb{K}, P^*)$ for all $g \in A$ holds. But this implication is in clear contradiction with the quasiconcavity of E, because $E(g, \mathbb{K}, P) > E(g_j, \mathbb{K}, P)$ for all $g \in A$. Therefore, in fact we have $E(g, \mathbb{K}, P) > E(\tilde{g}, \mathbb{K}, P) \Longrightarrow \left(D_{|\tilde{G}}\right)^{qc}(g, \mathbb{K}, P^*) > \left(D_{|\tilde{G}}\right)^{qc}(\tilde{g}, \mathbb{K}, P^*)$ for all $g, \tilde{g} \in \tilde{G}$.

□

Remark 5. In general, Assumptions $(A1)$ and $(A2)$ are very strong. Assumption $(A1)$ can be seen as a separability condition. It is satisfied, e.g., in the case of $\mathbb{R}^d$ together with the conceptual scaling discussed in Example 1 of the main article. Assumption $(A1)$ is also true if every one-element set is an ufg-premise, because in this case one can set $U_g := \{g\}$. For example, this condition hods for the case of hierarchical-nominal data if duplication is allowed.

Assumption $(A2)$ is generally difficult to satisfy. If G is finite, Assumption $(A2)$ cannot be satisfied at all. However, a slightly modified choice of the coefficients C_j as $C_j := \mathbb{E}[h^j]/(\varepsilon + \mathbb{E}[h^j]) \approx 1$ with some small fixed constant $\varepsilon > 0$ makes Assumption $(A2)$ satisfied. Beyond that, in the case of $\mathbb{R}^d$ with the scaling method in Example 1 of the main article and an absolutely continuous probability measure, we have that $\mathbb{E}[h^j]$ is one if for $j \in \{2, \ldots, d+1\}$ and zero otherwise. Of course, when we construct P^* as in the proof, we use a discrete probability measure. Note, however, that for $\mathbb{R}^d$ one can always use points that are affine independent for the construction of the ufg-premises U_g such that, independent of the probability values $p_{(i)}$, we always have $\mathbb{E}[h^j] = 1$ for $j \in \{2, \ldots, d+1\}$ and zero otherwise. Note also that it is sufficient to assume that the boundedness Assumption $(A2)$ for P holds over all probability measures implicitly used in the construction of P^* (which is of course a hard condition to keep track of).

In the following we consider the formal context $\mathbb{K} = (G, M, I)$ given by hierarchical-nominal data and the scaling method presented in Example 3 of Section 2 in the main article. Analogously we denote the attributes by $x_1 x_2 \ldots x_k$ with x_i describing the category on level i depending on the previous levels $1, \ldots, i-1$.

Lemma 8.10. *Let $\mathbb{K} = (G, M, I)$ be a formal context given by hierarchical-nominal data with $L \geq 2$ levels and at each level at least 3 categories. We use the scaling method presented in Example 3 of Section 2 of the main article. We assume that for each object $g \in G$ there exists another object $\tilde{g} \in G$ with $g \neq \tilde{g}$ and $\Psi(\{g\}) = \Psi(\{\tilde{g}\})$. Then the ufg-premises have cardinality one or two.*

Proof. First, we show that there are ufg-premises of cardinality one and two. Let $g \in G$. By assumption there exists an object $\tilde{g} \in G$ with $g \neq \tilde{g}$ and $\Psi(\{g\}) = \Psi(\{\tilde{g}\})$. So $g \to \gamma(\{g\}) \supseteq \{g, \tilde{g}\}$ is a ufg-implication. Since we have at least two levels, we know that there are two objects $\tilde{g} \in G$ with $\Psi(\{g\}) \neq \Psi(\{\tilde{g}\})$ such that they differ at least at Level L. Thus the set $\{g, \tilde{g}\}$ implies all objects that can also be sorted into the $x_1, \ldots x_k$ categories

with $k < L$. Thus $\{g, \tilde{g}\}$ is union-free, and since it obviously cannot be reduced without also reducing the conclusion, it is also generic.

Finally, we show that every set $\{g_1, \dots, g_n\}$ for $n \geq 3$ is not an ufg-premise. Let $\{g_1, \dots, g_n\} \to B$ be a valid implication given by $\mathbb{K}$. Then B is a subset of $\Phi \circ \Psi(\{g_1, \dots, g_n\})$. By constructing the attributes in Example 3 of Section 2 of the main article, we get that $\Psi(\{g_1, \dots, g_n\})$ are exactly the attributes describing the first k level categories on which all objects $g_1, \dots, g_n$ agree. These attributes can also be described by only two objects g_i, g_j with $i, g \in \{1, \dots, n\}$, namely by two objects g_i, g_j that agree up to level k, but that disagree on level $k+1$. Therefore, the implication $\{g_i, g_j\} \to B$ is also valid and thus, the implication $\{g_1, \dots, g_n\} \to B$, is not an ufg implication since the premise is not minimal. □

Theorem 8.11. *Let $\mathbb{K} = (G, M, I)$ be a formal context given by hierarchical-nominal data with $L \geq 2$ levels, $K \geq 3$ categories on each level, and the scaling method presented in Example 3 of Section 2 of the main article. We assume that for each object $g \in G$ there exists another object $\tilde{g} \in G$ with $g \neq \tilde{g}$ and $\Psi(\{g\}) = \Psi(\{\tilde{g}\})$.*

We set $C_1, C_2 > 0$ in the ufg-depth definition. Then the quasiconcave version D^{qc} of the ufg-depth is strongly free with respect to the property quasiconcavity. This means that for every $\varepsilon > 0$ there exists a family $\mathcal{P}^\varepsilon$ of probability measures with diameter[15] less than or equal to ε such that for any other arbitrary quasiconcave depth function E and any arbitrary probability measure P there exists a measure $P^* \in \mathcal{P}^\varepsilon$ such that

$$\forall g, \tilde{g} \in G : E(g, \mathbb{K}, P) > E(\tilde{g}, \mathbb{K}, P) \Longrightarrow D^{qc}(g, \mathbb{K}, P^*) > D^{qc}(\tilde{g}, \mathbb{K}, P^*).$$

Proof. First, we introduce some notation for simplicity. By assumption, we have on the finest category-level $N = K^L$ categories. Moreover, due to the scaling method, we can divide G into N subsets $G_1, \dots, G_N$, where each G_i corresponds to a set of objects with the same category on the finest level. This means that $G_i \cap G_j = \emptyset$ for different $i, j \in \{1, \dots, N\}$ and $G_1 \cup \dots \cup G_N = G$ is true. In particular, there exists an attribute $x_i = (x_i)_1 (x_i)_2 \dots (x_i)_N$ such that $G_i = \Phi(\{x_i\})$. So G_i is an extent and for all $g \in G_i$, $\gamma(\{g\}) = G_i$ is true.

Let $\varepsilon > 0$. We define

$$\mathcal{P}^\varepsilon := \left\{ P \text{ probability measure} \mid \forall g \in G : P(\{g\}) \in \left[\frac{1/N - \varepsilon/(2N)}{\#\gamma(\{g\})}, \frac{1/N + \varepsilon/(2N)}{\#\gamma(\{g\})} \right] \right\}.$$

Note that $\mathcal{P}^\varepsilon$ has a diameter smaller than or equal to ε. Let E be a quasiconcave depth function. This means that the contour sets are extents. In addition, the contour sets are nested due to their construction. For now on let $\Phi(\{y_1 y_2 \dots y_K\})$ be the objects with the largest depth value with respect to E. Then the contour sets of E are a subset of the extents $G \supseteq \Phi(\{y_1\}) \supseteq \Phi(\{y_1 y_2\}) \supseteq \dots \supseteq \Phi(\{y_1 y_2 \dots y_K\})$, where $\Phi(y_1 \dots y_K\}$ is the contour set of the objects with the highest depth. Note that $\Phi(\{y_1 y_2 \dots y_K\})$ is equal to one of the set G_i corresponding to a division on the finest level. W.l.o.g. we set $\Phi(\{y_1 y_2 \dots y_K\}) = G^* = G_N$.

Now, we construct $P^* \in \mathcal{P}$

$$P^*(\{g\}) \begin{cases} \frac{1/N + \varepsilon/(2N)}{\#G^*}, & \text{if } g \in G^* \\ \frac{1/N - \varepsilon/(2N \cdot (N-1))}{\#\gamma(g)}, & \text{else} \end{cases}.$$

[15] The diameter of a family $\mathcal{P}$ of probability measures on a measurable space (G, Σ) is defined as $diam(\mathcal{P}) := \sup_{A \in \Sigma; P, Q \in \mathcal{P}} |P(A) - Q(A)|$

In the following we show that $D(\cdot, \mathbb{K}, P^*)$ provides the same order as E. So we set $y_0 = \emptyset$ and show that for every $\ell \in \{0, \dots, L-1\}$ and every $g_1 \in \Phi(\{y_1 y_2 \dots y_i\}) \setminus \Phi(\{y_1 y_2 \dots y_{i+1}\})$ and $g_2 \in \Phi(\{y_1 y_2 \dots y_{i+1}\})$, $D(g_1, \mathbb{K}, P^*) < D(g_2, \mathbb{K}, P^*)$ is true. Let $\ell \in \{0, \dots, L-1\}$ be arbitrary. To obtain the depth function, we need to discuss the ufg implications. From Lemma 8.10, we know that the ufg-premises have either cardinality one or two.

Part 1 - ufg-premises of cardinality one: Each implication $g \to \gamma(\{g\})$ implies only those objects which have exactly the same attributes. By the definition of $G_1, \dots, G_N$ there exists $i \in \{1, \dots, N\}$ such that $g \in G_i$ and $\gamma(\{g\}) = G_i$. This gives us

$$\frac{C_1}{\mathbb{E}[h^1]}\mathbb{E}[f_g^1] = \frac{C_1}{\mathbb{E}[h^1]}\mathbb{E}[1_{\gamma(A)}(g) 1_{\mathbb{K}_{ufg}^{prem,1}}(A)] = \begin{cases} 1/N + \varepsilon/(2N), & \text{if} g \in G^* \\ 1/N - \varepsilon/(2N \cdot (N-1)), & \text{else} \end{cases}.$$

Part 2 - ufg-premises of cardinality two: Let G_i and G_j be such that their corresponding attributes differ for at least on the finest level L (see proof of Lemma 8.10). Then for every $g_i \in G_i$ and $g_j \in G_j$ we have that $\{g_i, g_j\}$ defines an ufg-premise. In particular, with this procedure we obtain all possible ufg-premises of cardinality two, see the proof of Lemma 8.10. We denote all these pairs by $\mathcal{G}^2$. Due to symmetry (on each level we have exactly the same number of categories) we obtain that the number of pairs $G_i, G_j \in \mathcal{G}^2$ such that $g \in \gamma(G_i \cup G_j)$ is the same for all $g \in G$. Let $g \in G$, then we get

$$\begin{aligned}
&\frac{C_2}{\mathbb{E}[h^2]}\mathbb{E}[f_g^2] = \frac{C_2}{\mathbb{E}[h^2]}\mathbb{E}[1_{\gamma(A)}(g) 1_{\mathbb{K}_{ufg}^{prem,2}}(A)] \\
&= \frac{C_2}{\mathbb{E}[h^2]}\Big[\mathbb{E}[1_{\gamma(A)}(g) 1_{\mathbb{K}_{ufg}^{prem,2}}(A) 1_{\{A \subseteq G \setminus G^*\}}] \\
&\qquad\qquad + \frac{C_2}{\mathbb{E}[h^2]}\mathbb{E}[1_{\gamma(A)}(g) 1_{\mathbb{K}_{ufg}^{prem,2}}(A) 1_{\{A \cap G^* \neq \emptyset\}}]\Big] \\
&= \frac{C_2}{\mathbb{E}[h^2]}\Big[\sum_{\substack{G_i, G_j \in \mathcal{G}^2 \text{ with} \\ G_i \neq G^* \neq G_j \text{ and} \\ g \in \gamma(G_i \cup G_j)}} (1/N - \varepsilon/(2N \cdot (N-1)))^2 \\
&\qquad\qquad + \sum_{\substack{G_i, G_j \in \mathcal{G}^2 \text{ with} \\ G_i = G^* \text{ or } G^* = G_j \text{ and} \\ g \in \gamma(G_i \cup G_j)}} (1/N - \varepsilon/(2N \cdot (N-1)))(1/N + \varepsilon/(2N))\Big].
\end{aligned}$$

If $g_1 \in G^*$ and $g_2 \in G$ the difference between the pairs is that the set of pairs G_i, G_j where at least one is equal to G^* is strictly larger for g_1 than for g_2. Hence, there are more pairs in the second part of the sum above. With this, we immediately get that $\frac{C_2}{\mathbb{E}[h^2]}\mathbb{E}[f_{g_1}^2] > \frac{C_2}{\mathbb{E}[h^2]}\mathbb{E}[f_{g_2}^2]$.

Now Part 1 and 2 together with the definition of the ufg-depth show that E and $D(\cdot, \mathbb{K}, P^*)$ give the same order of the objects G. So $D(\cdot, \mathbb{K}, P^*)$ is already quasiconcave and with $D(\cdot, \mathbb{K}, P^*) = D^{qc}(\cdot, \mathbb{K}, P^*)$ we prove the claim. □

Claims and Proofs of Section 6 - Examples

Section 6 of the main article discusses two concrete data examples: mixed spatial-categorical-numerical data and hierarchical-nominal data. Here we provide the proof of the claim that

simplifies the calculation of the ufg depth for mixed spatial-categorical-numerical data. Therefore, we consider the special case of joined formal contexts. Let us assume that we have two formal contexts on the same object set G but with two different attribute sets, $\mathbb{K}_1 = (G, A_1, I_1)$ and $\mathbb{K} = (G, A_2, I_2)$. Then consider the joined formal context $\mathbb{K} = (G, A_1 \cup A_2, I_1 \cup I_2)$. Analogously we denote the derivation and closure operators. Then for this joint formal context, we get:

Lemma 8.12. *Let $\mathbb{K}_1, \mathbb{K}_2$ and $\mathbb{K}_{1,2}$ together with the closure operator γ_1, γ_2 and $\gamma_{1,2}$ be defined as in the beginning of this section. Let G be the set of objects and $A \subseteq G$. Then $\gamma_{1,2}(A) = \gamma_1(A) \cap \gamma_2(A)$ is true.*

Proof. The proof follows from

$$\begin{aligned} a \in \gamma_{1,2}(A) &\Leftrightarrow \Psi_{1,2}(A) = \Psi_1(A) \dot{\cup} \Psi_2(A) \subseteq \Psi_{1,2}(a) \\ &\Leftrightarrow a \text{ has every attribute in } \Psi_1(A) \text{ and } a \text{ has every attribute in } \Psi_2(A) \\ &\Leftrightarrow a \in \gamma_1(A) \cap \gamma_2(A). \end{aligned}$$

□

Lemma 8.13. *Let $\mathbb{K}_1, \mathbb{K}_2$ and $\mathbb{K}_{1,2}$ together with the closure operator γ_1, γ_2 and $\gamma_{1,2}$ be defined as in the beginning of this section. Let $\max\{\#A \mid A \in \mathcal{I}_{1,ufg}\} = u_1$ and $\max\{\#A \mid A \in \mathcal{I}_{2,ufg}\} = u_2$, then every $A \subseteq G$ with $\#A > u_1 + u_2$ cannot be an ufg-premise of $\mathbb{K}$.*

Proof. This proof follows directly from Lemma 8.1. □

Lemma 8.14. *For the formal context $\mathbb{K}_G$ with extent set given by Equation (1) in Section 6.1. of the main article, we have for the ufg-family of implications*

$$\mathcal{I}_{ufg} \subseteq \left\{ A \to B \,\middle|\, \begin{array}{l} A \subseteq \mathbb{R}^2 \times V \times \mathbb{R} \text{ and } 2 \leq \#A \leq 4, \\ \pi_{\mathbb{R}}(B) = [\min\{\pi_{\mathbb{R}}(A)\}, \max\{\pi_{\mathbb{R}}(A)\}],\ \pi_{\mathbb{R}^2}(B) = \gamma_{\mathbb{R}^2} \circ \pi_{\mathbb{R}}(A), \\ \pi_V(B) \in \binom{V}{1} \cup V : \pi_V(B) = \pi_V(A) \text{ if } \#\pi_V(A) = 1, \pi_V(B) = V \text{ else} \end{array} \right\}.$$

Proof. Note that for $A \subseteq G$ we have $\gamma(A) = \gamma_{\mathbb{R}^2} \circ \pi_{\mathbb{R}^2}(A) \times \tilde{V} \times [\min\{\pi_{\mathbb{R}}(A)\}, \max\{\pi_{\mathbb{R}}(A)\}]$ with $\tilde{V} = \pi_V(A)$ if $\#\pi_V(A) = 1$ and $\tilde{V} = V$ else. Thus, we have to show that for every $A \subseteq G$ with $\#A = \mathbb{N} \setminus \{2, 3, 4\}$ is not an ufg-premise. For $g \in G$ we have $\gamma(\{g\}) = \{g\}$ which is a contradiction to (C1) in Definition 4.2. of the main article. For the upper bound, we first utilize that the formal context can be divided into three formal context $\mathbb{K}_{\text{spatial}}, \mathbb{K}_{elevation}$ and $\mathbb{K}_{vegetation}$. One can easily show that the ufg-premises of these formal contexts are bounded from above by $3, 2$ and 2. Hence, applying Lemma 8.13 provides us with an upper bound of 7. To show that the maximal cardinality is 4, we prove for cardinalities 5, 6 and 7 directly that they cannot be an ufg-premise.

So let $A \subseteq G$ with $\#A = 5$ and $g \in \gamma(A)$. Then there exists $A_1 \subseteq A$ with $\#A_1 = 3$, so that $\pi_{\mathbb{R}^2}(g) \in \pi_{\mathbb{R}^2} \circ \gamma(A_1)$. If $\pi_{\mathbb{R}}(g) \in \pi_{\mathbb{R}} \circ \gamma(A_1) = [\min\{\pi_{\mathbb{R}}(A)\}, \max\{\pi_{\mathbb{R}}(A)\}]$ let $\tilde{g}$ be another point in A with a different vegetation than that in A_1 (if it doesn't exist, just take an arbitrary one) and we set $A_g = A_1 \cup \tilde{g}$. Then $\#A_g = 4$ and $g \in \gamma(A_g)$ is true. If $\pi_{\mathbb{R}}(g) \notin \pi_{\mathbb{R}} \circ \gamma(A_1)$, then there exists $\overline{g} \in A$ such that $\pi_{\mathbb{R}}(g) \in \pi_{\mathbb{R}} \circ \gamma(A_1 \cup \overline{g})$. Now look at these four elements $A_1 \cup \overline{g}$, because of the geometry in $\mathbb{R}^2$ (i.e. there are only two cases, either $\pi_{\mathbb{R}^2}(\overline{g}) \not\in \pi_{\mathbb{R}^2} \circ \gamma(A_1)$ or $\pi_{\mathbb{R}^2}(\overline{g}) \in \pi_{\mathbb{R}^2} \circ \gamma(A_1)$), there exists a subset $A_2 \subsetneq A_1$ with

$\pi_{\mathbb{R}^2}(g) \in \pi_{\mathbb{R}^2} \circ \gamma(A_2 \cup \overline{g})$. Now we are back in uppercase, and using the uppercase argument, we can define A_g with $\#A_g = 4$ so that $g \in \gamma(A_g)$.

This can be done for every $g \in \gamma(A)$ and we obtain a division of $\gamma(A)$ by $\cup_{g \in A} \gamma(A_g) = \gamma(A)$. This is a contradiction to the union-free condition (C2). Hence, A with $\#A = 5$ cannot be an ufg-premise.

Similar one can show that $\#A = 6, 7$ cannot be an ufg-premise either which gives the claim. □

9 Quasiconcavity from the Perspective of Loss Functions

In this section, we shortly outline that a quasiconcave version of a depth function D^{qc} can be also defined to be the depth function that has minimal loss w.r.t. one specific loss function. In general we define:

Definition 9.1. Let $D(\cdot, \mathbb{K}, P)$ be a depth function on G with corresponding formal context $\mathbb{K}$ and probability measure P. Let L be a loss function on the function space and $\mathcal{Q}$ a subset of quasiconcave functions on G based on $\mathbb{K}$. We say that a depth function $\tilde{D}(\cdot, \mathbb{K}, P)$ is a *close quasiconcave version of D w.r.t. $\mathbb{K}, P, L$ and $\mathcal{Q}$* if and only if

$$D^{qc,L}(\cdot, \mathbb{K}, P) = \arg \min_{E(\cdot,\mathbb{K},P) \in \mathcal{Q}} \int_G L(E(\cdot, \mathbb{K}, P), D(\cdot, \mathbb{K}, P)) dP.$$

Note that this definition is only well-defined when the minimum is attained and that the $D^{qc,L}(\cdot, \mathbb{K}, P)$ is only unique except for a null set.

Let us now consider the special case of the following loss function $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R} \cup \{\infty\}, (x, y) \mapsto (+\infty)1_{x<y} + (x - y)1_{x \geq y}$. Thus, when looking at the order provided by a depth function, we force that the resulting quasiconcave depth function only reorders upwards and not downwards (except for null sets).

Theorem 9.2. *Let $D(\cdot, \mathbb{K}, P)$ be a depth based on formal concept analysis on a finite object set G and for every $g \in G$ we have $P(g) > 0$. Let $\mathcal{Q}$ be the set of all quasiconcave functions on G where the quasiconcavity is defined by $\mathbb{K}$. Then D^{qc} is a close quasiconcave version of D w.r.t. $\mathbb{K}, P, L$ and $\mathcal{Q}$.*

Proof. The proof that D^{qc} is quasiconcave follows from Theorem 4.3. We show that D^{qc} is a closed quasiconcave version of D w.r.t. $\mathbb{K}, P, L$ and $\mathcal{Q}$. Note that reordering any object in D^{qc} lower than in D already gives an infinite loss (since every object in G has a positive probability). So the depth function must be the smallest quasiconcave depth function that has D as a point-wise lower bound. This is exactly D^{qc}. □

References

Arcones, M. and E. Giné (1993). Limit theorems for U-processes. *The Annals of Probability 21*(3), 1494–1542.

Armstrong, W. (1974). Dependency structures of data base relationships. In *International Federation for Information Processing Congress*, Volume 74, pp. 580–583. North-Holland Publishing Company.

Assouad, P. (1983). Densité et dimension. *Annales de l'institut Fourier 33*(3), 233–282.

Baddeley, A. and R. Turner (2005). Spatstat : An r package for analyzing spatial point patterns. *Journal of Statistical Software 12*(6), 1–42.

Bastide, Y., N. Pasquier, R. Taouil, G. Stumme, and L. Lakhal (2000). Mining minimal non-redundant association rules using frequent closed itemsets. In J. Lloyd, V. Dahl, U. Furbach, M. Kerber, K.-K. Lau, C. Palamidessi, L. M. Pereira, Y. Sagiv, and P. J. Stuckey (Eds.), *International Conference on Computational Logic*, pp. 972–986. Springer.

Blocher, H. and G. Schollmeyer (2025). Data depth functions for non-standard data by use of formal concept analysis. *Journal of Mulitvariate Analysis 205*, 105372.

Blocher, H., G. Schollmeyer, and C. Jansen (2022). Statistical models for partial orders based on data depth and formal concept analysis. In D. Ciucci, I. Couso, J. Medina, D. Ślęzak, D. Petturiti, B. Bouchon-Meunier, and R. R. Yager (Eds.), *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pp. 17–30. Springer.

Blocher, H., G. Schollmeyer, M. Nalenz, and C. Jansen (2024). Comparing machine learning algorithms by union-free generic depth. *International Journal of Approximate Reasoning 169*, 109166.

Chebana, F. and T. B. M. J. Ouarda (2011). Depth-based multivariate descriptive statistics with hydrological applications. *Journal of Geophysical Research 116*(D10).

Christofides, T. (1992). A strong law of large numbers for u-statistics. *Journal of Statistical Planning and Inference 31*(2), 133–145.

Dudley, R. M., E. Gin, and J. Zinn (1991). Uniform and universal glivenko-cantelli classes. *Journal of Theoretical Probability 4*(3), 485–510.

Eckhoff, J. (1993). Chapter 2.1 - Helly, Radon, and Carathéodory type theorems. In P. Gruber and J. Wwillis (Eds.), *Handbook of Convex Geometry*. North-Holland Publishing Company.

Foss, A. H., M. Markatou, and B. Ray (2019). Distance metrics and clustering methods for mixed–type data. *International Statistical Review 87*(1), 80–109.

Funwi-Gabga, N. and J. Mateu (2012). Understanding the nesting spatial behaviour of gorillas in the kagwene sanctuary, cameroon. *Stochastic Environmental Research and Risk Assessment 26*(6), 793–811.

Ganter, B. and R. Wille (2012). *Formal Concept Analysis: Mathematical Foundations*. Springer.

GESIS - Leibniz-Institut für Sozialwissenschaften (2023). Allgemeine bevölkerungsumfrage der sozialwissenschaften allbus 2021. GESIS, Köln. ZA5280 Datenfile Version 2.0.1, https://doi.org/10.4232/1.14238.

Ignatov, D. I. and L. Kwuida (2022). On shapley value interpretability in concept-based learning with formal concept analysis. *Annals of Mathematics and Artificial Intelligence 90*(11-12), 1197–1222.

Li, J. and R. Y. Liu (2004). New nonparametric tests of multivariate locations and scales using data depth. *Statistical Science 19*(4), 686–696.

Liu, R. (1990). On a notion of data depth based on random simplices. *The Annals of Statistics 18*(1), 405–414.

Liu, R., J. Parelius, and K. Singh (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference, (with discussion and a rejoinder by liu and singh). *The Annals of Statistics 27*(3), 783–858.

Maier, D. (1983). *The Theory of Relational Databases*. Computer Science Press.

Mosler, K. and P. Mozharovskyi (2022). Choosing among notions of multivariate depth statistics. *Statistical Science 37*(3), 348–368.

Poelmans, J., D. I. Ignatov, S. O. Kuznetsov, and G. Dedene (2013). Formal concept analysis in knowledge processing: A survey on applications. *Expert Systems with Applications 40*(16), 6538–6560.

Roscoe, S., M. Khatri, A. Voshall, S. Batra, S. Kaur, and J. Deogun (2022). Formal concept analysis applications in bioinformatics. *ACM Computing Surveys 55*(8), 1–40.

Schollmeyer, G. (2017a). Application of lower quantiles for complete lattices to ranking data: Analyzing outlyingness of preference orderings. Technischer Report, LMU. last accessed: 14.12.2024.

Schollmeyer, G. (2017b). Lower quantiles for complete lattices. Technischer Report, LMU. last accessed: 14.12.2024.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science 103*(2684), 677–680.

Stumme, G., D. Dürrschnabel, and T. Hanika (2023). Towards ordinal data science: 39 pages. *Transactions on Graph Data and Knowledge 1*(1), 6:1–6:39.

Tukey, J. (1975). Mathematics and the picturing of data. In R. James (Ed.), *Proceedings of the International Congress of Mathematicians Vancouver*, pp. 523–531. Mathematics-Congresses.

Yanqing Zhang, Qi Xu, Niansheng Tang, and Annie Qu (2024). Differentially private data release for mixed-type data via latent factor models. *Journal of Machine Learning Research 25*(116), 1–37.

Zuo, Y. and R. Serfling (2000a). General notions of statistical depth function. *The Annals of Statistics 28*(2), 461–482.

Zuo, Y. and R. Serfling (2000b). Structural properties and convergence results for contours of sample statistical depth functions. *The Annals of Statistics 28*(2), 483–499.

Contribution 3

Hannah Blocher, Georg Schollmeyer, and Christoph Jansen (2022). "Statistical Models for Partial Orders based on Data Depth and Formal Concept Analysis". In: *Information Processing and Management of Uncertainty in Knowledge-based Systems*. Ed. by Davide Ciucci, Inés Couso, Jesús Medina, Dominik Ślęzak, Davide Petturiti, Bernadette Bouchon-Meunier, and Ronald Yager. Cham: Springer, 17–30

Statistical Models for Partial Orders Based on Data Depth and Formal Concept Analysis

Hannah Blocher(✉), Georg Schollmeyer, and Christoph Jansen

Department of Statistics, Ludwig-Maximilians-Universität Munich, Munich, Germany
{hannah.blocher,georg.schollmeyer,christoph.jansen}@stat.uni-muenchen.de

Abstract. In this paper, we develop statistical models for partial orders where the partially ordered character cannot be interpreted as stemming from the non-observation of data. After discussing some shortcomings of distance based models in this context, we introduce statistical models for partial orders based on the notion of data depth. Here we use the rich vocabulary of formal concept analysis to utilize the notion of data depth for the case of partial orders data. After giving a concise definition of unimodal distributions and unimodal statistical models of partial orders, we present an algorithm for efficiently sampling from unimodal models as well as from arbitrary models based on data depth.

Keywords: Partial orders · Partial rankings · Data depth · Formal concept analysis · Unimodality · Quasiconcavity

1 Introduction

Orders play a role in a broad range of scientific disciplines. In many of these disciplines like revealed preference theory, social choice theory, decision making under uncertainty, social-economics (Human Development Index, costumer preference rankings etc.) or statistics and machine learning, studying *partial* orders has attracted more and more researchers (see [5,10,18,19,21,28,33] and [13] for recent works in the respective discipline). As an example, one can consider pair comparison data sets as in [9] and [12]. Consequently, there are many approaches that can deal with partial orders. However, in most approaches known to the authors, the incompleteness of the involved orders is interpreted as stemming from missing data, see, e.g., [25,30]. In other words, an explicit missing mechanism is modeled or at least assumed. In contrast, in this paper we explicitly assume that the incompleteness of the order is not due to missing of data. Instead, we understand incomporabilities within observed partial orders as precise observation of a factual incomparability that actually exists.

The aim of this paper is to define statistical models on the set of partial orders that do not assume a missing mechanism and that can be easily specified by defining a location and a scale parameter. Beyond this methodological contribution we propose an efficient algorithm for sampling from such models. Our

D. Ciucci et al. (Eds.): IPMU 2022, CCIS 1602, pp. 17–30, 2022.
https://doi.org/10.1007/978-3-031-08974-9_2

statistical models are based on formal concept analysis (**FCA**) and the concept of data depth. Via the chosen representation of the data set using a formal context (which is a formalization of a cross table defined in FCA), the information that a pair of elements does not exist in a partial order is explicitly included. Furthermore, we embed the notion of depth function in the theory of FCA. Data depth functions are commonly used in robust and nonparametric statistics and can be viewed as a generalization of univariate quantiles. Applications range from outlier detection over nonparametric statistical tests and confidence intervals to robust regression, for a short overview see, e.g., [26] and the references therein. Generally, a depth function measures the outlyingness and centrality of an observation w.r.t. a data cloud or an underlying probability measure. While common depth functions are defined for data in $\mathbb{R}^d$, accompanied by a geometric intuition, within this paper, we aim at generalizing the concept of data depth to the abstract setting of data points that are objects of a formal context. Unlike many of the existing distance based models for partial order data (cf., e.g., [14,25,27,30]), we are therefore able to use data depth functions instead of a distance measure on the set of partial orders. While depth functions compute the depth value with respect to all other data points, distance measures compute only the distance between the data points of a partial order with respect to a predefined partial order representing the center, see e.g. [6]. Note further that many distance measures are based on the linear extensions of the partial orders involved and thus do not take into account the incomparable character, but mimic an underlying true linear order. Since we want to define a simple location-scale type statistical model, we also further define unimodality in the context of FCA and we consider in particular depth functions that satisfy this property. As far as the authors are aware, there is no other work that directly links FCA and data depth or uses FCA to define a model for the set of partial orders.

In Sect. 2 we begin with an overview of currently used distance measures and point out their explicit and implicit assumptions. In Sect. 3 we give a short introduction to FCA and define unimodality in this framework. Then, in Sect. 4 and 5 we define and discuss the formal context representing the set of partial orders, and, afterwards, propose some concrete depth functions. In Sect. 6, we introduce an algorithm for sampling from the proposed statistical models and finally we give a brief conclusion.

2 Motivation and Related Work

To illustrate how the currently used distance measures implicitly mimic the missing mechanism and other counter-intuitive structures, let us start by discussing the current approaches that use distance measures for (partial) orders. There are several proposals for adequately defining a meaningful distance concept between (partial) orders in the literature (cf, e.g., [6,11]) which can be used to establish distance based statistical models for partial orders. Throughout the paper let $\mathcal{X}$ be a finite ground space with $n \geq 3$ elements and let $\mathscr{P}$ denote the set of

all partial orders[1] (i.e., all reflexive, transitive and anti-symmetric binary relations) on $\mathcal{X}$. Two prominent distance measures for partial orders are discussed for example in [6]: The *nearest neighbour* and the *Hausdorff distance*. Both of these distances rely on the idea of first computing the set of all linear extensions of the considered partial orders and then, each in its own manner, generalizing the well-known *Kendall's* τ-distance (see [22]) for linear orders (i.e. counting pairs that are ranked oppositely by the considered orders). However, such an approach has the following counter-intuitive property: The nearest neighbour distance systematically assigns lower distance values if sparse partial orders are involved. The nearest neighbour distance is defined as

$$d_{NN}(P_1, P_2) := \min_{L_1 \in lext(P_1)} \min_{L_2 \in lext(P_2)} \tau(L_1, L_2)$$

for two orders P_1, P_2 where $lext(P)$ denotes the set of all linear extensions of a partial order P and τ denotes the Kendall's τ-distance for linear orders mentioned before. Then it is immediate from the definition that $d_{NN}(\tilde{P}_1, P_2) \leq d_{NN}(P_1, P_2)$ for arbitrary partial orders $\tilde{P}_1 \subseteq P_1$, since this implies $lext(P_1) \subseteq lext(\tilde{P}_1)$ and therefore the minimum is taken over a super-set of the original one. Most extremely, the minimal distance is attained whenever one of the considered partial orders is the trivial one consisting solely of the diagonal $D_{\mathcal{X}} := \{(x, x) : x \in \mathcal{X}\}$, whereas two partial orders differing only in few pairs receive non-minimal distance value. This seems to be a very counter-intuitive property of this generalized distance measure. An analogous line of argumentation applies when the nearest neighbour distance is replaced by the directed Hausdorff hemi-metric

$$m_H(P_1, P_2) := \max_{L_1 \in lext(P_1)} \min_{L_2 \in lext(P_2)} \tau(L_1, L_2).$$

Then, in a dual manner, $D_{\mathcal{X}}$ (if seen as the first argument in the Hausdorff hemi-metric) has always the maximal distance to other orders whereas a linear order L has always a smaller distance to other orders compared to any other partial order $P \subseteq L$. Similar arguments can be given for the usual symmetrized non-directed Hausdorff distance defined by

$$d_H(P_1, P_2) := max\{m_H(P_1, P_2), m_H(P_2, P_1)\}.$$

Alternatively, one could directly generalize Kendall's τ to partial orders without looking at linear extensions. This would result in one of the two expressions

$$\tau_s(P_1, P_2) := |\Delta(P_1, P_2)| = |(P_1 \cup P_2) \setminus (P_1 \cap P_2)| \quad \text{or}$$
$$\tau_a(P_1, P_2) := |\{(x, y) \mid x \neq y, (x, y) \in P_1, (y, x) \in P_2\}|,$$

both, in a way, generalizing the idea of counting pairs that are ranked oppositely by the considered *partial* orders. However, whereas τ_a has the same problem like the nearest neighbour distance, the expression τ_s would lead, as will be shown in

[1] In the sequel, we will also shortly say order instead of partial order.

Sect. 3, Example 1, to statistical models that are not completely quasiconcave, which means that it seems to be impossible to build a simple unimodal model with such a distance (cf., Definition 1). Furthermore, τ_s treats pairs which are in the relation and pairs being not in the relation in exact the same way, and one can ask if this is natural. As we will see later, our approach that uses a depth function treats pairs being in the relation or not seemingly differently. (Note that a partial order is transitive but not necessarily negatively transitive, so there is in fact some asymmetry between a pair being in the relation or not.) With these problems in mind, we propose statistical modelling of partial orders based on a depth function. The model idea is analogous to a distance based version

$$P(X = x) = C_\lambda \cdot \Gamma\left(\lambda \cdot d\left(\mu, x\right)\right),$$

where, C_λ is a normalizing constant, $d : \mathscr{P} \times \mathscr{P} \longrightarrow \mathbb{R}_{\geq 0}$ is a distance function, $\Gamma : \mathbb{R}_{\geq 0} \longrightarrow \mathbb{R}_{\geq 0}$ is a (weakly decreasing) decay function, $\mu \in \mathscr{P}$ is a location parameter and $\lambda \in \mathbb{R}_{>0}$ is a scale parameter. Now, instead of a distance function, in this paper we work with a depth function and a statistical model given by

$$P(X = x) = C_\lambda \cdot \Gamma\left(\lambda \cdot \left(1 - D^\mu\left(x\right)\right)\right) \tag{1}$$

where now D^μ is a depth function that is maximal at partial order μ. Since depth functions are usually only used for data in $\mathbb{R}^d$ we have to adapt the notion of data depth to partial order data, for which we use FCA.

3 Formal Concept Analysis, Data Depth and Unimodality

In this section we only touch a few aspects about the theory of formal concept analysis (FCA) and we refer the reader to [17] for more details. The basis of FCA is the definition of a **formal context** $\mathbb{K} = (G, M, I)$ which is a generalization and formalization of a cross table. Here, G is a set of objects, M a set of binary attributes and $I \subseteq G \times M$ a relation. We say that an object g has an attribute m if $(g, m) \in I$ is true. For example Cross Table 1 describes a formal context with $G = \{\mu, g, h, i\}$, $M = \{m_1, \ldots, m_6\}$ and the relation I is given by the crosses. By the use of the following **derivation operators**, we obtain a description of the relation between the object and attribute set:

$$\Psi : 2^G \to 2^M : A \mapsto \{m \in M \mid \forall g \in A\colon (g, m) \in I\}$$
$$\Phi : 2^M \to 2^G : B \mapsto \{g \in G \mid \forall m \in B\colon (g, m) \in I\}.$$

Here $\Psi(A)$ contains all the attributes that each object in A has, and $\Phi \circ \Psi(A) \subseteq G$ are all objects that have all attributes in $\Psi(A)$. The tuple $(\Phi \circ \Psi(A), \Psi(A))$ for $A \subseteq G$ is called a **formal concept**, $\Phi \circ \Psi(A)$ its **extent**, and $\Psi(A)$ its **intent**. The construction of the two derivation operators allows to determine the relation I when the set of all formal concepts is known. Note that $\Psi(A) = \Psi \circ \Phi \circ \Psi(A)$ holds, and thus each formal concept is uniquely described by its extent or intent. Moreover, the set of extents and the set of intents yield a closure system with

$\Phi \circ \Psi$ and $\Psi \circ \Phi$, respectively, the corresponding closure operator. Note that if $A \subseteq G$ lies in an extent E, then the closure operator $\Phi \circ \Psi$ ensures that every object having all attributes of $\Psi(A)$ is also an element of E. Thus, $A \subseteq E$ implies that $\Phi \circ \Psi(A) \subseteq E$. With this, we say that the pair $A, B \subseteq G$ is an **(object) implication** (we denote this by $A \to B$) if $\Phi \circ \Psi(A) \supseteq \Phi \circ \Psi(B)$ holds. Moreover, one can show that the set of all implications that follow from the extent set completely describes the extent set itself, see, e.g., [17, p. 80, Proposition 20]. Within this paper, we use formal implications between objects to model a notion of betweenness. For example $\{g, h\} \longrightarrow \{i\}$ can be interpreted as "object i lies between object g and object h" (or "object i lies in the space that is spanned by the objects g and h"), because object i has all attributes that are shared by both g and h. (Note that we do not restrict the premise of a formal implication to have exactly two objects.) For further discussion of a family of implications, see [2] and [17]. If non-binary attributes are considered, then they are converted into a set of binary attributes by using a so-called conceptual scaling method (see Sect. 4).

Our approach is to represent the set of partial orders by a formal context and, using the properties of a formal context, to define the notion of unimodality and depth function. By using the following properties that a function $f\colon G \to \mathbb{R}$ can satisfy on a formal context $\mathbb{K}$, we define the notion of unimodality.

Definition 1. *Let $\mathbb{K} = (G, M, I)$ be a formal context and let $f : H \longrightarrow \mathbb{R}$ with $H \subseteq G$ be a function. Then f is called*

i) ***isotone*** *if for all $g, h \in H$ we have $\{g\} \longrightarrow \{h\} \implies f(g) \leq f(h)$;*
ii) ***2-quasiconcave*** *if for arbitrary objects $g, h, i \in H$ we have $\{g, i\} \longrightarrow \{h\} \implies f(h) \geq \min\{f(g), f(i)\}$;*
iii) ***completely quasiconcave*** *if for every finite set of objects $\{g_1, \ldots, g_n\} \subseteq H$ we have $\{g_1, \ldots, g_{n-1}\} \longrightarrow \{g_n\} \implies f(g_n) \geq \min\{f(g_1), \ldots f(g_{n-1})\}$;*
iv) ***strongly quasiconcave*** *if for every finite set $\{g_1, \ldots, g_n\} \subseteq H$ of size $n \geq 2$ we have $\{g_1, \ldots, g_{n-1}\} \longrightarrow \{g_n\} \implies f(g_n) > \min\{f(g_1), \ldots f(g_{n-1})\}$;*
v) ***star-shaped*** *if there exists a center $c \in H$ such that for all $g \in H$ we have $\{c, g\} \longrightarrow \{h\} \implies f(h) \geq \min(f(c), f(g))$.*

Additionally, a probability measure P on a finite G is called ***unimodal (strictly unimodal)*** *if its probability function, restricted to its support $\{g \in G \mid P(\{g\}) > 0\}$, is completely quasiconcave (strongly quasiconcave).*

In general, depth functions measure outlyingness and centrality of an observation w.r.t. a data cloud or an underlying probability measure. We apply the concept of data depth to partial order data represented by a formal context and we denote it by $D\colon G \to \mathbb{R}_{\geq 0}$. Note that it depends on the formal context. Moreover, if we ensure that the depth function is completely quasiconcave (strongly quasiconcave), then the statistical model given in (1) is unimodal (strictly unimodal).

Our notion of quasiconcavity is an adaption of classical quasiconcavity which was already used (e.g., in [29]) for classical data depth for $\mathbb{R}^d$. In particular, here

we emphasize (complete) quasiconcavity because it most adequately renders the idea of an unimodal distribution of partial orders that would be induced by a statistical model that uses a quasiconcave depth function: Quasiconcavity would ensure that we have no point that is a local minimum of the probability function w.r.t. the notion of betweenness that is appropriate for a FCA view on partial orders. Another nice feature of complete quasiconcavity is the fact that this property is equivalent to the property that the upper level sets $D_\alpha := \{g \in G \mid D(g) \geq \alpha\}$ of the depth function D are extents. Thus, every upper level set can be nicely described by a formal concept which makes them descriptively accessible, especially the fact that they cannot only be exactly described by objects, but also by attributes, is very convincing.

Example 1. Let $\mathbb{K} = (G, M, I)$ be given by Cross Table 1. Then, the depth function D^μ with mode μ given by $D^\mu(g) := |\Psi(\mu) \cap \Psi(g)|$, together with the conceptual scaling of Sect. 4 can be shown to be exactly the depth-based formulation of a distance based approach with τ_s. It is 2-quasiconcave but in general not completely quasiconcave and therefore is not appropriate to define a unimodal distribution. Note that for arbitrary contexts, D^μ is generally not 2-quasiconcave. Note further that D^μ is at least star-shaped for arbitrary contexts. Furthermore, a generalization of Tukey's depth $\mathcal{T}$ (cf., [31]) and a localized version of Tukey's depth $\mathcal{T}^\mu$ with mode μ can be defined via

$$\mathcal{T}(g) := 1 - \max_{m \in M \setminus \Psi(\{g\})} \frac{|\Phi(\{m\})|}{|G|}; \quad \mathcal{T}^\mu(g) := 1 - \frac{\max\limits_{\substack{m \in M \setminus \Psi(\{g\}),\\ \mu I m}} |\Phi(\{m\})|}{|G|}, \quad (2)$$

respectively. (Here the empty maximum is defined as 0.) Both $\mathcal{T}$ and $\mathcal{T}^\mu$ are completely quasiconcave functions.

Table 1. Illustration of the difference between complete and 2-quasiconcavity.

	m_1	m_2	m_3	m_4	m_5	m_6
μ		x	x	x	x	x
g		x				
h	x				x	x
i	x		x	x		

4 Formal Context Defined by All Partial Orders

In our case the set G is exactly the set $\mathscr{P}$ of all partial orders on $\mathcal{X}$. Note that we regard a partial order not necessarily as a linear order together with

a missing mechanism. Therefore, as attributes we also include the property of being incomparable pairs and get

$$M := \underbrace{\{\text{“}x_i \leq x_j\text{”} \mid i,j = 1,\dots,n,\, i \neq j\}}_{=:M_\leq} \cup \underbrace{\{\text{“}x_i \not\leq x_j\text{”} \mid i,j = 1,\dots,n,\, i \neq j\}}_{:=M_{\not\leq}}.$$

Since we consider only reflexive relations the attributes “$x_i \leq x_i$” and “$x_i \not\leq x_i$” are redundant and therefore not included here. Note that each order g has $n(n-1)$ many attributes $B = \Psi(\{g\})$ which can be divided into the set $B_\leq \subseteq M_\leq$ and $B_{\not\leq} \subseteq M_{\not\leq}$. In particular, we have that either (x_i, x_j) lies in g or not and thus we can conclude $(g, \text{“}x_i \leq x_j\text{”}) \in I \Leftrightarrow (g, \text{“}x_i \not\leq x_j\text{”}) \notin I \;\&\; (g, \text{“}x_j \not\leq x_i\text{”}) \in I$. This means if a pair (x_i, x_j) exists then the attribute “$x_i \not\leq x_j$” cannot hold, but “$x_j \not\leq x_i$” must be true. The same is true for the reverse. Indeed, ensuring that a pair $(x_i, x_j), i \neq j$ is in an order g or not has a different strength of restriction, i.e., if we assume that $(x_i, x_j) \in g$, then $g^{-1} := \{(x_j, x_i) \mid (x_i, x_j) \in g\}$ satisfies the condition $(x_i, x_j) \notin g^{-1}$. Thus, the number of orders $\tilde{g}$ fulfilling the condition $(x_i, x_j) \notin \tilde{g}$ is larger than the number of orders g fulfilling $(x_i, x_j) \in g$. Additionally, because of symmetry these numbers are independent of the concrete pair (x_i, x_j).

First let us go one step back and consider the formal context given only by the attribute set $M_\leq$. In this case, for an isotone depth function D and two orders g, h with $g \subseteq h$ we have $D(g) \leq D(h)$. Thus, we would obtain again a depth concentration on linear orders. Furthermore, if the depth function is additional 2-quasiconcave and we consider the space of all partial orders, then at least half of all partial orders must have equal depth. More precisely, the depth must be minimal. To see this, let g be an order and let g^{-1} be the inverse order. We obtain that $\{g, g^{-1}\} \longrightarrow G$ and therefore either the depth of g or the depth of g^{-1} is minimal. Thus, because the map $g \mapsto g^{-1}$ is a bijection, half of all orders have minimal depth. Note that the stronger property of strong quasiconcavity cannot be fulfilled by any depth function: Assume we have four orders $g_1, \dots, g_4$ where all pairs of orders have no attribute $m \in M_\leq$ in common. Then $\{g_1, g_2\} \longrightarrow G$ and therefore $\min\{D(g_1), D(g_2)\} < D(g_i), i = 3, 4$. But since $\{g_3, g_4\} \longrightarrow G$ this is a contradiction to $\min\{D(g_3), D(g_4)\} < D(g_i), i = 1, 2$. Note that for $|\mathcal{X}| \geq 3$ there exist four linear orders fulfilling this property. Let us now return to the formal context given by the entire attribute set M.

Then, the same argument for the non-existence of a strongly quasiconcave depth function from above would still apply for the extended attribute set M. Beyond this, now the context defined here contains no two different orders g and h such that $\Psi(g) \subseteq \Psi(h)$. Thus, for an isotone depth, isotonicity alone does not imply that the depth value of one order is constrained by the depth value of any other order. In contrast, 2-quasiconcavity would still lead to some restriction on the depth function: Let g be an order such that the complement order (i.e. $g^c := (\mathcal{X} \times \mathcal{X} \setminus g) \cup D_\mathcal{X}$) is also an order. Then one of the orders must have minimal depth, since $\{g, g^c\} \longrightarrow G$. If we take G as the set of all partial orders, then examples of such orders are exactly the linear orders. If, in contrast, one had chosen G as the set of all quasiorders, then exactly all negatively transitive

orders g would have the property that also g^c is in G and therefore one of g or g^c would have minimal depth.

5 Specifying Unimodal Distributions of Partial Orders

In this section we want to analyze how one can specify a generic non-null model (i.e. a model that is different from a uniform distribution, see [27]) with the help of a depth function and Eq. (1) by specifying only two parameters, namely location and scale. More specifically, we discuss methods for generating unimodal distributions of partial orders based on three concrete depth functions. Firstly, we will discuss Tukey's depth defined by Eq. (2). Secondly, we define a generalization of the convex hull peeling depth. The classical convex hull peeling depth for $\mathbb{R}^d$-valued data was introduced in [4] and we will generalize it here for the case of data represented by a formal context. We will call this depth function peeling depth. It is sometimes said that the convex hull peeling depth has the disadvantage that it can only order the data points from outwards to inwards. In contrast, in our situation, we are able to directly specify a mode of the distribution and therefore we know beforehand, where 'the inwards', i.e., the mode, is exactly located. With this, we can in fact order the data points from inwards to outwards by starting from the mode and successively enclosing further layers. Thus, thirdly, we can define a new depth function that we call enclosing depth, here. The generalization of Tukey's depth for data values or probability distributions on arbitrary complete lattices or formal contexts was introduced in [31] and applied to the case of ranking data in [32] and in the context of social choice theory in [20]. The definition is given in Eq. (2). Before discussing all three data depths, we firstly define the remaining two:

Definition 2. *Define the* ***peeling depth*** $\mathcal{P}$ *by* $\mathcal{P}(extr(G)) := \frac{1}{|G|}$ *and*

$$\mathcal{P}\left(extr\left(G\backslash\mathcal{P}^{-1}\left(\left[0,\frac{i}{|G|}\right]\right)\right)\right) = \frac{i+1}{|G|}, \quad i = 1, 2, \ldots$$

Additionally, define the ***localized peeling depth*** $\mathcal{P}^\mu$ *w.r.t. mode* $\mu \in G$ *simply by adding a high enough amount of objects which have exactly the same attributes as* μ *to the original context* G*. The operator extr is here the extreme point operator which maps a set* A *to the set of all its extreme points.*[2] *Note that this definition is only well defined if the underlying context is meet-distributive.*[3] *Furthermore,*

[2] A point $g \in A$ is an extreme point of A if $A\backslash\{h \in G \mid \Psi(\{h\}) = \Psi(\{g\})\} \nrightarrow \{g\}$.

[3] A context is called meet-distributive, if every extent is generated by all extreme points of the extent. In our situation, the underlying context is not meet-distributive, but it is possible to replace the extreme point operator by another appropriate operator that maps a set A to a set $B \subseteq A$ that implies A and is minimal w.r.t. this property. Note that for such an operator the obtained depth function is generally not quasiconcave anymore. Another possibility would be the operator $\tilde{extr}(A) := extr(A) \cup A\backslash(\Phi \circ \Psi)(extr(A))$. This operator would lead to a completely quasiconcave depth function. Note further that this operator is generally not minimal which means that the number of depth layers is usually lower compared to a minimal operator.

we define the ***enclosing depth*** $\mathcal{E}^\mu$ *w.r.t. mode* μ *by* $\mathcal{E}^\mu\left((\Phi \circ \Psi)(\{\mu\})\right) = 1$ *and*

$$\mathcal{E}^\mu\left(encl\left((\mathcal{E}^\mu)^{-1}\left(\left[\frac{i}{|G|}, 1\right]\right)\right)\right) = \frac{i-1}{|G|}; \quad i = |G|, |G|-1, \ldots$$

Here, encl denotes an operator which we would like to call an enclosing operator. Concretely, we have in mind an operator $encl : H \longrightarrow 2^G$ *with* $H \subseteq 2^G$ *that for all* $A \in H$ *satisfies the three properties i):* $encl(A) \cap A = \emptyset$*, ii):* $encl(A) \longrightarrow A$ *and iii):* $(\Phi \circ \Psi)\left(encl\left(A\right)\right)$ *is minimal w.r.t. properties i) and ii).*

Now we discuss, how one can specify with the above depth functions a unimodal distribution of orders with a given mode and one scale parameter. The simplest distribution, which can be always defined in a finite setting, is the uniform distribution. To specify a distribution that is in some certain sense distributed around a given mode, one simple approach would be to first generate every partial order exactly one time (this would correspond to a uniform distribution) and then to simply add a big amount of partial orders that are identical to the mode. Then, based on the corresponding data depth that is obtained for this data set, one can define a distribution according to Eq. (1). (Note that generally the obtained distribution is different from a mixture of a uniform distribution and a distribution that equals the mode with probability one.) However, for Tukey's depth, due to reasons of symmetry one can show that the obtained distribution of orders would assign the mode one probability p and every other order that differs from the mode exactly one of two probability values q or r. More concretely, the localized Tukey's depth could then be written as

$$\mathcal{T}^\mu(g) = 1 - \max\left\{\max_{(p,q)\in\mu\setminus g} \alpha_{p,q}, \max_{(p,q)\in g\setminus\mu} \beta_{p,q}\right\} \text{ with } \alpha_{p,q}, \beta_{p,q} \in [0,1],$$

[4]where actually $\alpha_{p,q}$ and $\beta_{p,q}$ do not depend on p or q. This seems to be somehow unsatisfying. Of course, one can use Tukey's depth based on another (empirically or analytically) given distribution, but then, in the first place one is back at the "... major outstanding problem in ranking theory ..." and has to specify a "... suitable population of ranks in non-null cases..." ([23]). Alternatively, one can replace $\alpha_{p,q}$ and $\beta_{p,q}$ by other weights that depend on the pairs (p, q), actually fortunately without losing the quasiconcavity. For this weighted Tukey-type depth function one would have to specify only n^2 values instead of $2^{n^2/4}$ or more values (cf., [24]), which would be needed for a completely nonparametric approach. Because this can still be very demanding, we will later use an analysis of the enclosing depth to get a rough guidance for specifying the weights. For the peeling depth there seems to be not so much ties compared to Tukey's depth. However, it seems a little bit counter-intuitive to specify a distribution of orders that are distributed around a mode by not locally looking at a neighbourhood of the mode but instead by globally ordering the data points from outwards to inwards. Compared to other applications of data depth where one does not know

[4] This also shows an asymmetry between pairs that are in relation and pairs that are not in relation.

the location beforehand but where the problem is actually the estimation of the mode of the distribution, here we are in the comfortable situation that we can simply specify the mode. Therefore, in the sequel, we will focus on the enclosing depth (applied for the case $G := \mathscr{P}$). Also here, because of the high amount of symmetries there are many ways of defining an enclosing operator and corresponding depth layers. One way out of this would be to compute in a first step for every partial order the expected depth value under a stochastic choice of the layer that is built in every step. This is of course possible and also a simulation from such a model can be exactly done. However, the obtained depth function is not completely quasiconcave. Therefore, one can in a second step build the closure of every depth contour to obtain a completely quasiconcave depth function. For this, one has to analyze in detail, how the expected depth values of the first step exactly look like, which seems to be a very difficult problem. Therefore, we only analyze the situation for total orders and a totally ordered mode μ under a conceptual scaling of the partial orders that uses only $M_{\leq}$. With this analysis we are able to roughly oversee the situation for the enclosing depth and we will use the results to guide the specification of the weights within the modified Tukey's depth (see above) under a conceptual scaling that uses both $M_{\leq}$ and $M_{\not\leq}$: Let $(p, q) \in \mu$. Define $\Delta_\mu(p, q)$ simply as the "distance" between p and q w.r.t. the mode μ measured by the number of pairs between p and q w.r.t. the covering relation of μ. Furthermore, for $x \in \mathscr{P}$ define

$$s_\mu(x) := \max_{(p,q)\in\mu\setminus x} \Delta_\mu(p, q).$$

Then one can show that total orders x with a higher $s_\mu(x)$ have a lower depth value w.r.t. the enclosing depth $\mathscr{E}^\mu$. Thus, for a weighted version of Tukey's depth function one can weight pairs (p, q) with a higher $\Delta_\mu(p, q)$ correspondingly with a higher weight, e.g., via $\alpha_{p,q} \propto \Delta_\mu(p, q)$. (For pairs with the same value it would be natural to choose the same weight.) Now, the problem is to specify the corresponding weights for pairs $(p, q) \in x\setminus\mu$. Because we would like to think from the direction of the mode μ and not from the perspective of x, we do not want to simply change the roles of μ and x. The problem here is that it seems to be somehow difficult to order pairs (p, q) w.r.t. the mode μ that are not in relation w.r.t. μ. However, there are some possibilities to rank such pairs. The following definition is somehow inspired by the work in [15]: For $(p, q) \notin \mu$ one could define[5]

$$\Delta_\mu(p, q) := |\{r \in \mathcal{X} \mid p \wedge_\mu q \leq_\mu r \leq_\mu p \vee_\mu q\}| - 1.$$

This definition extends the original definition of Δ_μ and it can be used to specify the weights (e.g., via $\alpha_{p,q} \propto \Delta_\mu(p, q); \beta_{p,q} \propto \Delta_\mu(p, q)$) for the modified Tukey's depth that uses the whole attribute set M for the conceptual scaling.

[5] If the considered partial order does not build a complete lattice one could simply compute the Dedekind-MacNeille completion beforehand.

6 Simulation

By representing the set of partial orders as defined in Sect. 4 and applying one of the data depth functions of Sect. 5, we obtain a statistical model on the set of partial orders on $\mathcal{X}$ by Eq. (1). In this section we derive an algorithm to sample from such a statistical model. The algorithms is based on the acceptance-rejection method and the idea of the algorithm is based on [16]. For a small number of elements, we can directly compute all reflexive, transitive, and anti-symmetric orders. Thus, we can easily draw a sample from one of the above distributions. Since the runtime of the computation of all partial orders grows with the number of elements faster or equal to $2^{n^2/4}$ (see [24]), the direct computation is not feasible for larger n. Therefore, we provide an algorithm based on the following structure: First, we systematically draw a partial order and calculate the number of possible paths to obtain this partial order. Finally, we compute the acceptance probability such that we sample with probability of interest f. The algorithm uses that each partial order is a subset of at least one linear order. A linear order has $\frac{1}{2}(n-1)n$ many pairs of the form (x_i, x_j) with $i \neq j$ and, in particular, if we randomly delete some of these pairs, then, by computing the transitive hull, we obtain a partial order. To obtain step 1, we first take a uniform sample of a linear order and then randomly delete some pairs by a uniform variable on all subsets. Note that drawing a linear order is only a tool to obtain a partial order, and due to the definition of the acceptance probability, does not affect the probability of selecting a partial order. By computing all linear extensions, we can compute the probability that this partial order was sampled. Finally, we adjust the acceptance probability so that the sample ends up consisting of the probability function f we are interested in. More precisely, the probability that a given order g is computed in step 1 is:

$$P_{algo\_select}(g) = |lext(g)| \cdot 2^{|g|-|reduc(g)|} \cdot \left(n!2^{n(n-1)/2}\right)^{-1}$$

where $reduc(g)$ is the transitive reduction[6] of g and $n! \cdot 2^{n(n-1)/2}$ is the number of all paths to obtain a partial order by the procedure above. Since the number of pairs of each linear order is the same, the probability that the partial order g is sampled is identical for each linear order from the linear extension of g. Let f be the probability function from which we want to draw a sample, then the acceptance function is given by

$$acc(g) = f(g) \cdot \left(P_{algo\_select}(g) \cdot n!2^{n(n-1)/2}\right)^{-1}. \tag{3}$$

Lemma 1. *Algorithm 1 samples a partial order with probability function f on all partial orders of G.*

[6] The transitive closure of a relation is the smallest transitive relation containing it, and the transitive reduction is a minimal relation having the same transitive closure.

Algorithm 1: Sampling a partial order

Input: n: number of items;
f: probability function with the set of all partial orders as domain;
Result: partial order sampled according to the probability given by f.
repeat
  # sampling the order
  LIN_ORDER ← sample uniformly a linear order;
  DEL_PAIRS ← sample uniformly a subset of $\{1, \ldots, (n-1)n/2\}$;
  PARTIAL_ORDER ← uniformly delete DEL_PAIRS many pairs and compute the transitive closure;
  # compute the acceptance probability (thereby we have to compute the transitive reduction)
  ACCEPT_PROB ← computation of (3);
until *random]0,1]* $\leq$ ACCEPT_PROB;

The proof is analogous to the one given in [16]. Note we could use also a modified version of the acceptance function: $\tilde{acc} = c \cdot acc$ with constant $c \geq \max_g f(g)/P_{algo\_select}(g)$. This modified version must assure that for all partial orders g, $f(g) \leq c \cdot P_{algo\_select}(g)$ is true. Unfortunately, the computation of all linear extensions is # P-complete (see [7]). Note that for some subsets of all partial orders the running time of the computation of the linear extension is smaller, i.e., if we consider only the set of trees (see [3]). Additionally, to improve the runtime of the algorithm we generally could also use an approximation for the number of all linear extensions $|lext(g)|$, for which e.g. [8] gives approximation approaches.

7 Conclusion

In this paper, we developed statistical models for partial orders based on data depth and FCA. Therefore we embedded the terms data depth and unimodality into FCA. We think that with this approach, opposed to statistical models based on distances, we are in fact able to appropriately incorporate the incomparability of two items as well as the notion of unimodality of a statistical model for partial orders. In particular, we think that a notion of unimodality based on concepts of lattice theory is more appropriate compared to notions based on metrics or based on the embedding of partial orders into a linear space. In particular, the simulation approach allows to perform simulation studies in order to analyze the statistical behavior of certain statistical procedures for partial order data. In addition, this enables the construction of parametric bootstrap procedures to quantify the uncertainty inherent in any statistical procedure for partial order data. What is left open for further research is the question how to exactly specify the decay function and the weights within the approach that uses Tukey's depth. A further analysis of the newly developed enclosing depth, especially w.r.t. the question whether this depth function can be also applied if one does not know

the mode beforehand, is also of high interest. Additionally, the application of our approach to concrete data situations is another important line of further research.

Acknowledgments. Hannah Blocher and Georg Schollmeyer gratefully acknowledge the financial and general support of the LMU Mentoring program. Further, we thank the three anonymous reviewers for their constructive and insightful comments that helped to improve the paper.

References

1. Alvo, M., Cabilio, P.: Rank correlation methods for missing data. Canad. J. Stat. **23**(4), 345–358 (1995)
2. Armstrong, W.: Dependency structures of data base relationships. Int. Fed. Inf. Process. Congress **74**, 580–583 (1974)
3. Atkinson, M.: On computing the number of linear extensions of a tree. Order **7**, 23–25 (1990)
4. Barnett, V.: The ordering of multivariate data (with discussion) J. Roy. Stat. Soc. Ser. A **139**(3), 318–352 (1976)
5. Boutilier, C., Rosenschein, J.: Incomplete information and communication in voting. In: Moulin, H., Brandt, F., Conitzer, V., Endriss, U., Lang, Procaccia, A. (eds.) Handbook of Computational Social Choice, pp. 223–258, Cambridge University Press (2016)
6. Brandenburg, F.J., Gleißner, A., Hofmeier, A.: Comparing and aggregating partial orders with kendall tau distances. In: Rahman, M.S., Nakano, S. (eds.) WALCOM 2012. LNCS, vol. 7157, pp. 88–99. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-28076-4_11
7. Brightwell, G., Winkler, P.: Counting linear extensions is # P-complete. In: Proceedings 23rd ACM Symposium on the Theory of Computing, pp. 175–181 (1991)
8. Bubley, R., Dyer, M.: Faster random generation of linear extensions. Discret. Math. **201**(1–3), 81–88 (1999)
9. Collins-Thompson, K., Callan, J.: A language modeling approach to predicting reading difficulty. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL, pp. 193–200 (2004)
10. Comim, F.: Beyond the HDI? Assessing alternative measures of human development from a capability perspective. In: Background paper of the Human Development Report. UNDP Human Development Report (2016)
11. Critchlow, D.: Metric methods for analyzing partially ranked data. Lecture Notes in Statistics, 34. Springer (1985)
12. Dittrich, R., Hatzinger, R., Katzenbeisser, W.: Modelling the effect of subject-specific covariates in paired comparison studies with an application to university rankings. J. R. Stat. Soc. Ser. C **47**(4), 511–525 (1998)
13. Fahandar, M., Hüllermeier, E., Couso, I.: Statistical inference for incomplete ranking data: The case of rank-dependent coarsening. In: Proceedings of the 34th International Conference on Machine Learning, vol. 70, pp. 1078–1087 (2017)
14. Fligner, M., Verducci, J.: Distance based ranking models. J. Roy. Stat. Soc. B **48**(3), 359–369 (1986)

15. Gäbel-Hökenschnieder, T., Schmidt, S.: Generalized metrics and their relevance for FCA and closure operators. Concept Lattices and their Applications, pp. 175–186 (2016)
16. Ganter, B.: Random extents and random closure systems. Concept Lattices and their Applications, pp. 309–318 (2011)
17. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer Science & Business Media (2012)
18. Jansen, C., Blocher, H., Augustin, T., Schollmeyer, G.: Information efficient learning of complexly structured preferences: elicitation procedures and their application to decision making under uncertainty. Int. J. Approximate Reasoning **144**, 69–91 (2022)
19. Jansen, C., Schollmeyer, G., Augustin, T.: Concepts for decision making under severe uncertainty with partial ordinal and partial cardinal preferences. Int. J. Approximate Reasoning **98**, 112–131 (2018)
20. Jansen, C., Schollmeyer, G., Augustin, T.: A probabilistic evaluation framework for preference aggregation reflecting group homogeneity. Math. Soc. Sci. **96**, 49–62 (2018)
21. Jena, S., Lodi, A., Palmer, H., Sole, C.: A partially ranked choice model for large-scale data-driven assortment optimization. Informs J. Optim. **2**(4), 297–319 (2020)
22. Kendall, M.: A new measure of rank correlation. Biometrika **30**(1/2), 81–93 (1938)
23. Kendall, M.: Discussion on symposium on ranking methods. J. Roy. Stat. Soc. B **12**, 153–162 (1950)
24. Kleitman, D., Rothschild, B.: The number of finite topologies. Proc. Am. Math. Soc. **25**(2), 276–282 (1970)
25. Lebanon, G., Mao, Y.: Non-parametric modeling of partially ranked data. J. Mach. Learn. Res. **9**(10), 2401–2429 (2008)
26. Liu, R., Parelius, J., Singh, K.: Multivariate analysis by data depth: descriptive statistics, graphics and inference (with discussion and a rejoinder by Liu and Singh). Ann. Stat. **27**(3), 783–858 (1999)
27. Mallows, C.: Non-null ranking models. I. Biometrika **44**(1/2), 114–130 (1957)
28. Mangaraj, B., Aparajita, U.: Constructing a generalized model of the human development index. Socio-Econ. Plann. Sci. **70**, 100778 (2020)
29. Mosler, K.: Depth statistics. In: Becker, C., Fried, R., Kuhnt, S. (eds.) Robustness and Complex Data Structures, pp. 17–34. Springer, Heidelberg (2013)
30. Nakamura, K., Yano, K., Fumiyasu, K.: Learning partially ranked data based on graph regularization. arXiv:1902.10963 (2019)
31. Schollmeyer, G.: Lower quantiles for complete lattices. Technical Report 207. Department of Statistics. LMU Munich (2017)
32. Schollmeyer, G.: Application of lower quantiles for complete lattices to ranking data: Analyzing outlyingness of preference orderings. Technical Report 208. Department of Statistics. LMU Munich (2017)
33. Stewart, R.: Weak pseudo-rationalizability. Math. Soc. Sci. **104**, 23–28 (2020)

Contribution 4

Hannah Blocher, Georg Schollmeyer, Malte Nalenz, and Christoph Jansen (2024). "Comparing Machine Learning Algorithms by Union-free Generic Depth". In: *International Journal of Approximate Reasoning* 169, 109166. (Invited Paper for the ISIPTA 2023 Special Issue)

International Journal of Approximate Reasoning 169 (2024) 109166

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

journal homepage: www.elsevier.com/locate/ijar

Comparing machine learning algorithms by union-free generic depth

Hannah Blocher *, Georg Schollmeyer, Malte Nalenz, Christoph Jansen

Department of Statistics, LMU Munich, Ludwigstr. 33, Munich, 80539, Bavaria, Germany

ARTICLE INFO

Keywords:
Partial orders
Data depth
Benchmarking
Algorithm comparison
Outlier detection
Non-standard data

ABSTRACT

We propose a framework for descriptively analyzing sets of partial orders based on the concept of depth functions. Despite intensive studies in linear and metric spaces, there is very little discussion on depth functions for non-standard data types such as partial orders. We introduce an adaptation of the well-known simplicial depth to the set of all partial orders, the *union-free generic (ufg)* depth. Moreover, we utilize our ufg depth for a comparison of machine learning algorithms based on multidimensional performance measures. Concretely, we provide two examples of classifier comparisons on samples of standard benchmark data sets. Our results demonstrate promisingly the wide variety of different analysis approaches based on ufg methods. Furthermore, the examples outline that our approach differs substantially from existing benchmarking approaches, and thus adds a new perspective to the vivid debate on classifier comparison.[1]

1. Introduction and related literature

We begin with the general motivation for this paper and an overview of the contributions of our paper to the comparison of machine learning algorithms. We also provide references to related literature.

1.1. Motivation

Partial orders – and the systematic incomparabilities of objects encoded in them – occur naturally in a variety of problems in a wide range of scientific disciplines. Examples range from decision theory, where the agents under consideration might be unable to arrange the consequences of their actions into total orders (see, e.g., [52,35]) or have partial cardinal preferences (see, e.g., [31,29]), over social choice theory, where a fair aggregate order might only be possible by incorporating systematic incomparabilities (see, e.g., [43,32]), to finance, where risky assets do not always have to be comparable (see, e.g., [37,13]). Of course, many other relevant examples exist.

In the specific context of statistics and machine learning, the incompleteness of the considered orders often originates from the fact that the objects to be ordered are compared with respect to several criteria and/or on several instances *simultaneously*: only if there is unanimous dominance of one object over another, this order is included in the corresponding relation. Quite a number of

* Corresponding author.
E-mail address: hannah.blocher@stat.uni-muenchen.de (H. Blocher).

[1] **Open Science:** Reproducible implementation and data analysis are available at: https://github.com/hannahblo/Comparing_Algorithms_Using_UFG_Depth.

https://doi.org/10.1016/j.ijar.2024.109166
Received 8 December 2023; Received in revised form 14 February 2024; Accepted 4 March 2024

Available online 7 March 2024

research papers recently have been devoted to such comparison in the specific context of classification algorithms, either with respect to multiple quality metrics (e.g., [21,30]) or across multiple data sets (e.g., [18,4]) or with respect to genuinely multidimensional performance criteria like receiver operating characteristic (ROC) curves (e.g., [14]).

Another source of partial incomparability of classifiers is the case of classifiers that make only imprecise predictions, like for example the naive credal classifier (cf., [62]) or credal sum-product networks (cf., [39]). In this case, the imprecision in the predictions may take over to incomparabilities of the then possibly interval-valued performance measures.[2]

Within the application field of machine learning and statistics, one further aspect is of special importance: Since the instances generally depend on chance, the same is true for the partial orders considered. Consequently, instead of a single partial order, random variables must then be analyzed that map into the set of *all possible* partial orders on the set of objects under consideration. For example, in the aforementioned comparison of classification algorithms, the concrete order obtained depends on the random instantiation of the data set on which they are applied. In this paper, we are interested in exactly this situation: we discuss ways to *descriptively* analyze samples of such partial order-valued (or short: *poset-valued*[3]) random variables.

Of course, a descriptive analysis of samples of partial orders requires a completely different mathematical apparatus than the analysis of standard data. A suitable formal framework is by no means obvious.[4] Fortunately, it turns out that the concept of *(data) depth function* and the data representation given by *formal concept analysis* can be promisingly combined and applied to poset-valued random variables.[5] In this paper, we adapt the concept of depth to poset-valued random variables by introducing the *union-free generic (ufg) depth*. In general, depth functions define a notion of centrality and outlyingness of observations with respect to an entire data cloud or an underlying distribution, see e.g. [64,40]. So far, depth functions have mainly been applied to $\mathbb{R}^d$-valued random variables. Exceptions are e.g. the definition of depth functions on lattices in [49,48], on total orders, see [26], and a general discussion on depth functions based on formal concept analysis in [6]. Since we transfer this depth concept to poset-valued data, some classical descriptive statistics can naturally be adapted to this particular *non-standard data* type.

1.2. Comparing machine learning algorithms

Before turning to the main part of the paper, we first indicate which contributions (beyond benchmarking) our methodology can add to the general task of analyzing machine learning (ML) algorithms.

While the basic task of performance comparison is very common in machine learning (cf., [28] and the references therein), our methodological contribution deviates from the typical benchmark setting with regard to at least two points:

(I) First, we compare algorithms not with respect to one unidimensional criterion like, e.g., balanced accuracy, but instead look at a whole set of performance measures. We then judge one algorithm as at least as good as another one if it is not outperformed with respect to any of these performance measures. With this, for every data set, we get a partial order of algorithms and since we are not looking at only one, but a whole population/sample of data sets, we get a poset-valued random variable. Importantly, we do not hold the view that there is an underlying true (random) total order together with a coarsening mechanism that generates the (random) partial order. Such views are often termed *epistemic*, cf., [15]. Applications of this view in the context of partial order data can be found for example in [36,42]. In contrast, within the nomenclature of [15], we see our approach more in the spirit of the *ontic view* that is usually applied to set-valued data.[6] However, since in our case the random objects are partial orders, the term *ontic* seems to fit not perfectly. Instead, we understand poset-valued data as a special type of non-standard data.[7]

(II) Second, we are not interested in which algorithm is the best or most competitive. Instead, we are interested in how the relative performance of different algorithms is distributed over a population/sample of different data sets. Analyzing the distribution of performance relations is in our view a research question of its own statistical importance that may add further insights to other types of analyses like that of [30,33], which are of utmost importance when it comes to choosing between different machine learning algorithms.

These deviations can have very different motivations: Analyzing multidimensional criteria (of performance, here) is already motivated by the fact that different performance measures might be conceptually on an equal footing (at least, if one has no further concrete, e.g., decision-theoretic desiderata at hand). Therefore it appears natural to take more than one measure at the same time into account. Beyond this, there are far more possible motivations for dealing with multidimensional criteria: For example, if one accounts for distributional shifts within covariates in classification, then for different covariate distributions the class balance of the class labels will vary. This can naturally be captured by looking at different weightings of the true positives and the true negatives

[2] For example one could think of comparing classifiers with utility-discounted predictive accuracy, cf., [63, p. 1292] under the usage of a whole range $[\underline{a}, \overline{a}]$ for the coefficient of risk aversion.

[3] Note that in fact, we speak here about random variables that have posets (on a common underlying ground space) as outcomes. This should not be confused with random variables which have values in a partially ordered set.

[4] Many approaches proposed for analyzing poset-valued data rely on distance measures, see, e.g., [16,10]. For further discussion see [7], Section 2.

[5] The representation of posets via formal concept analysis can be expressed as a closure operator, see Definition 1. Therefore, formal concept analysis is not necessary to follow this paper. For more information and discussion of this representation, see [7].

[6] Concrete applications can be found e.g. in [46] for the ontic, and in [45] for the epistemic view. A case where the ontic and the epistemic view coincide is discussed in [50]. Beyond this, in the field of partial identification in the context of generalizing confidence intervals to confidence regions for the so-called identified set, the question about an ontic vs. an epistemic view is also implicitly asked (without reference to these terms) by asking if such a confidence set should cover the true parameter or the whole identified set with prespecified probability, cf., [54].

[7] Note that we do not want to generally rule out epistemic treatment, but this is not the focus of this paper. Such a treatment could use that every partial order can be described by the set of all its linear extensions. For further discussions, cf., [7].

within the construction of a classical performance measure.[8] Alternatively, one can take into account different discrimination thresholds for the classifiers simultaneously, which would correspond to looking at a whole region of the receiver operating characteristic.

Also, the motivation for the second point can be manifold: Generally, it seems somehow naive to search for one best algorithm per se. For example, the scope of application of an algorithm can vary very strongly and different algorithms could be best for different situations. Further, in certain situations different algorithms can be comparable in their performance with respect to one specific measure, but incomparable if one looks at different performance measures at the same time. Therefore, it can be of high interest, how the conditions between different algorithms change over the distribution of different data sets or application scenarios. For example, if in one very narrowly described data situation the performances of different algorithms vary extremely from case to case, but not so much across algorithms, then, at some point, it would become more or less hopeless to search for the best algorithm in the training phase, because one knows that in the prediction setting, the situation is too different to the training situation.

Another aspect is outlier detection: If one knows that in a large, maybe automatically generated benchmark suite there are data sets that have some bad data quality (for example if some covariates are meaningless because of some data formatting error, etc.), then it would be reasonable to try to exclude such outlying data sets from a benchmark analysis beforehand. Candidates of such outliers are then naturally data sets with a low depth value.

1.3. Outline of the paper

This paper is an extended version of [8] presented at the 13th International Symposium on Imprecise Probabilities: Theories and Applications (ISIPTA 2023) conference in Oviedo, Spain. Compared to the ISIPTA 2023 conference, this paper focuses more on the application. We added an entire section on comparison with other methods for evaluating poset-valued data and a sensitivity analysis with respect to the number of performance measures, see Section 7. We also extend the analysis and implementation of the ufg depth, see Sections 5 and 6.

Our paper is organized as follows: We begin with a brief practitioner's summary, focusing on the special case of classifier comparisons. In Section 3, we briefly discuss the required mathematical definitions and concepts. We give a formal definition of our depth function, the ufg depth, in Section 4 and discuss some of its properties in Section 5. While Section 6 prepares our application by providing implementation details, Section 7 is devoted to applying our framework to specific examples, namely the analysis of the goodness of classification algorithms on different data sets. First, we focus on comparing our benchmarking approach to further benchmarking concepts, see Section 7.1 and afterward, we give a detailed demonstration of the ufg method, see Section 7.2. Section 8 concludes by elaborating on some promising perspectives for future research.

2. Summary for practitioners

In this section we demonstrate how our general methodology can be used for the problem setting of selecting a reasonable classifier among a set of candidate classifiers within an application situation. Selection methods for classification algorithms are of immense importance for applications in a wide range of industries. Depending on the concrete branch, large amounts of money or even human lives can depend on it, e.g., treatment decision. As many practitioners have broad expertise in their specialty but are less familiar with statistical methodology, *reliable* guidance on classifier selection is crucial. To qualify as reliable, a selection method should at least meet the following standards:

(I) **Option to make no decision:** If the inherent uncertainty of the comparison situation is too large to identify a clear best classifier, the method should report and not obscure this fact.
(II) **Utilization of available information:** If the evidence is sufficient to identify a clear best classifier, the method should recognize this and not make overly cautious recommendations.
(III) **Intuitive accessibility:** The method should enable the user to develop an intuitive understanding of it (an exact understanding of the underlying mathematics is not necessary).

We present a framework for comparing classifiers that addresses requirements (I) to (III) and is based on two simple observations: *First*, the suitability of a classifier depends on the specific question, i.e., the specific data set being analyzed. Accordingly, the search should not (solely) be for the best classifier per se, but represent how *central/typical* a data set and the induced classifier ranking is. *Second*, the performance of certain classifiers might be incomparable already on one data set because, for example, multiple or vaguely defined performance measures are considered that conflict each other in the quality dimensions.

These two simple observations serve as a guiding principle of our construction: To compare data sets in terms of their centrality (think of the empirical *median* as most central one and the high and low *quantiles* as more outlying ones), we use the statistical theory of *data depth*: the deeper a data set lies (in some abstract space), the more central it is. To account for the complexity of the (potentially multidimensional or vaguely defined) concept of performance, we allow classifiers to be only *partially* ordered (already when considering only a single data set).

Our resulting framework does indeed manage to meet the above standards: By using partial orders rather than total ones, in the spirit of (I), a clear best classifier is identified only if the underlying evidence is strong enough and always accompanied by

[8] Note that usual performance measures are more or less simple transformations of the vector of the true positives and the true negatives (and the class balance).

information on the typicality of its superiority. In the spirit of (II), our framework avoids the incomparability of classifiers if the evidence is strong enough for a clear favorite (who is still accompanied by information on the typicality of its superiority). Finally, standard (III) is met as well: Since data depth is a generalization of the elementary statistical concept of a *quantile* (where, e.g., the deepest point corresponds to the median), every user is directly guided by natural intuition when applying the proposed concepts to real-world applications.

In a nutshell, our contributions can be summarized as follows: We offer practitioners the opportunity to reliably compare classification algorithms with regard to several performance measures and taking into account the typicality/centrality of the specific situation analyzed. Beyond the theoretical soundness of our framework, we provide well-documented implementations of our method, which are easily transferable to comparable applications. Last but not least, we propose a customized depth function for our application, the *ufg depth*, and enable the reliable comparison of classifiers by performing the following simple three-step scheme:

1. Compute the ufg depth function. We provide an R-script that can be easily adapted. The script can be found on GitHub, see https://github.com/hannahblo/Comparing_Algorithms_Using_UFG_Depth.
2. Use the partial order among classifiers corresponding to the highest depth value as a reference and the undominated algorithms as recommendations. We perform this analysis in Section 7.2 on Fig. 4.
3. Check how stable the dominance structure is by checking what the partial orders with the k highest depth values have in common. An example analysis using the intersection of the deepest posets can be found in the Section 7.2 discussing Fig. 5.

Finally, we want to emphasize that our concept can be used to compare any objects that arise from a sample of instances and where the comparison is based on multiple numeric measures. Thus, the concept introduced here is not limited to classification problems.

3. Preliminaries

Partial orders (posets) sort the elements of a set M, where we allow that two elements $y_1, y_2 \in M$ are incomparable. Formally stated: Let M be a fixed set. Then $p \subseteq M \times M$ defines a partial order (poset) on M if and only if p is *reflexive* (for each $y \in M$ holds $(y, y) \in p$), *antisymmetric* (if $(y_1, y_2), (y_2, y_1) \in p$ then $y_1 = y_2$ is true) and *transitive* (if $(y_1, y_2), (y_2, y_3) \in p$ then also $(y_1, y_3) \in p$). If p is also strongly connected (for all $y_1, y_2 \in M$ either $(y_1, y_2) \in p$ or $(y_2, y_1) \in p$), then p defines a *total/linear order*. The reverse, where the poset consists only of the reflexive part, is called *trivial order* p_Δ. For a fixed set M, various posets sort the set M. In this paper, we are interested in all posets that can occur on the set M with cardinality $\#M$ being finite. We denote the set of all posets on M by $\mathcal{P}_M$ (or $\mathcal{P}$ for short). Sometimes it can be useful to consider only the *transitive reduction* of a poset p, this means that for poset p we delete all pairs (y_1, y_2) which can be obtained by a transitive composition of other elements in p. We denote the transitive reduction of a poset p by $tr(p)$. Note that there exists a one-to-one correspondence between the transitive reduction of posets and the posets itself. In particular, this transitive reduction is often used to simplify the diagram used to represent the partial order. These diagrams are called *Hasse diagrams*. They consist of edges and knots where the knots represent the elements of M and the edges state the relation between the elements. More precisely, if $(a, b) \in p$ for a poset p, then there is a path from a to b that points strictly upwards, e.g., see Fig. 4. The reverse, where we add all pairs that follow from transitivity, is called *transitive hull*. We denote by $th(p)$ the transitive hull of an antisymmetric and reflexive relation p. We refer to [25] for further readings on partial orders. From now on, let $\mathcal{P}$ be all posets for a fixed set M. We denote the elements of M by y.

The concept for analyzing poset-valued observations presented in this paper is based on a *closure operator* on $\mathcal{P}$, see Section 4. In general, a closure operator $\gamma_\Omega : 2^\Omega \to 2^\Omega$ on set Ω is an operator which is *extensive* (for $A \subseteq \Omega$, $A \subseteq \gamma_\Omega(A)$ holds), *isotone*, (for $A, B \subseteq \Omega$ with $A \subseteq B$, $\gamma_\Omega(A) \subseteq \gamma_\Omega(B)$ is true) and *idempotent* (for $A \subseteq \Omega$, $\gamma_\Omega(A) = \gamma_\Omega(\gamma_\Omega(A))$ holds). The set $\gamma_\Omega(2^\Omega)$ is called the *closure system* and we say that $\gamma(B)$ for $B \subseteq \Omega$ is the *closure* of B. In what follows, we use that every closure operator (and therefore the closure system) can be uniquely described by an implicational system. Generally, an *implicational system* $\mathcal{I}$ on Ω is defined by a subset of $2^\Omega \times 2^\Omega$. The implicational system corresponding to the closure operator γ_Ω is defined by all pairs $(A, B) \in 2^\Omega \times 2^\Omega$ satisfying $\gamma_\Omega(A) \supseteq \gamma_\Omega(B)$. For short, we denote this by $A \to B$. For more details on closure operators and implicational systems, see [5].

We aim to define a centrality and outlyingness measure on the set of all posets $\mathcal{P}$ based on a fixed and finite set M. In general, functions that measure the centrality of a point with respect to an entire data cloud or an underlying distribution are called *(data) depth functions*. Depth functions on $\mathbb{R}^d$ have been studied intensively by, e.g., [64] and [41], and various notions of depth have been defined, such as Tukey's depth, see [57], and simplicial depth, see [38]. The idea behind the ufg depth introduced here is an adaptation of the *simplicial depth* on $\mathbb{R}^d$ to posets. The simplicial depth on $\mathbb{R}^d$ is based on the convex hull/closure operator that is defined as follows:

$$\gamma_{\mathbb{R}^d} : \begin{array}{l} 2^{\mathbb{R}^d} \to 2^{\mathbb{R}^d} \\ A \mapsto \left\{ x \in \mathbb{R}^d \,\middle|\, \begin{array}{l} x = \sum_{i=1}^k \lambda_i a_i \text{ with } a_i \in A, \\ \lambda_i \in [0,1], \sum_{i=1}^k \lambda_i = 1, k \in \mathbb{N} \end{array} \right\}. \end{array}$$

The simplicial depth considers those sets $\gamma(A)$ with $A \subseteq \mathbb{R}^d$ where the cardinality of A is $d+1$ and A forms a $(d+1)$ simplex (when no duplicates occur). Before going on, let us take a closer look at the set of all $(d+1)$ simplexes. First, the set of all $(d+1)$ simplexes is a proper subset of $2^{\mathbb{R}^d}$. By using Carathéodorys' Theorem on convex sets, see [20], we obtain that any set B of $d+1$ unique points is the smallest set such that no family of proper subsets $(A_i)_{i \in \{1,\dots,\ell\}}$ with $A_i \subsetneq B$ and $\bigcup_{i \in \{1,\dots,\ell\}} \gamma_{\mathbb{R}^d}(A_i) = \gamma_{\mathbb{R}^d}(B)$ exists. Moreover, the closure of every set B which consists of more than $d+1$ points can be divided in the sense that there exists a family of proper

subsets $(A_i)_{i\in\{1,\dots,\ell\}}$ with $A_i \subsetneq B$ such that $\bigcup_{i\in\{1,\dots,\ell\}} \gamma_{\mathbb{R}^d}(A_i) = \gamma_{\mathbb{R}^d}(B)$. Thus, the $(d+1)$ simplexes still completely characterize the corresponding closure system. Based on this set, the simplicial depth of a point $x \in \mathbb{R}^d$ is the probability that x lies in $\gamma(A)$ where A consists of $d+1$ many points that are randomly drawn from the underlying (empirical) distribution. For $\mathcal{M}$ being the set of all probability measures on $\mathbb{R}^d$ the simplicial depth is then given by

$$D: \begin{array}{l} \mathbb{R}^d \times \mathcal{M} \to [0,1], \\ (x,\nu) \mapsto \nu(x \in \gamma_{\mathbb{R}^d}\{X_1,\dots,X_{d+1}\}), \end{array}$$

where $X_1,\dots,X_{d+1} \overset{i.i.d.}{\sim} \nu$ with i.i.d. being *independent and identically distributed* for short. When we consider a sample $x_1,\dots,x_n \in \mathbb{R}^d$; $n \in \mathbb{N}$, we use the empirical probability measure instead of a probability measure ν. Thus, for a sample $x_1,\dots,x_n \in \mathbb{R}^d$ with empirical measure ν_n we obtain as empirical simplicial depth

$$D_n: \begin{array}{l} \mathbb{R}^d \to [0,1], \\ x \mapsto \binom{n}{d+1} \sum_{1\le i_1<\dots<i_{d+1}\le n} \mathbb{1}_{\gamma_{\mathbb{R}^d}\{x_{i_1},\dots,x_{i_{d+1}}\}}(x). \end{array}$$

Hence, if $x_1,\dots,x_n$ are affine independent, then the depth of a point x is the proportion of $(d+1)$ simplexes given by $x_1,\dots,x_n$ that contain x.

4. Union-free generic depth on posets

Now, we introduce the union-free generic (ufg) depth function for posets which is in the spirit of the simplicial depth function, see Section 3. To define the depth function, we start, similar to the simplicial depth, by defining a closure operator on $\mathcal{P}$. This closure operator maps a set of posets P onto the set of posets where each poset is a superset of the intersection of P and a subset of the union of P. In other words, any p lying in the closure of $\gamma(P)$ satisfies the following condition: First, every pair $(y_1,y_2) \in M \times M$ that lies in every poset in P is also contained in p, and second, for every pair (y_1,y_2) that lies in p, there exists at least one $\tilde{p} \in P$ such that $(y_1,y_2) \in \tilde{p}$. Note that while the intersection of posets defines a poset again, this does not hold for the union.

Definition 1. Let $\mathcal{P}$ be the set of posets on M. We define the mapping

$$\gamma: \begin{array}{l} 2^{\mathcal{P}} \to 2^{\mathcal{P}} \\ P \mapsto \left\{ p \in \mathcal{P} \mid \bigcap\limits_{\tilde{p}\in P} \tilde{p} \subseteq p \subseteq \bigcup\limits_{\tilde{p}\in P} \tilde{p} \right\}. \end{array}$$

Remark 1. This definition is based on the theory of formal concept analysis, see [25], and the formal context introduced in [7]. Using this formal context and the corresponding theory, we immediately obtain that γ defines a closure operator on $\mathcal{P}$ with associated closure system $\gamma(2^{\mathcal{P}})$.

Analogously to the definition of the simplicial depth, we now only consider a subset of $2^{\mathcal{P}}$. We denote this family by $\mathscr{S}$. $\mathscr{S}$ is a proper subset of $2^{\mathcal{P}}$, see Theorem 3 for details, which reduces $2^{\mathcal{P}}$ by redundant elements in the following sense: First, all subsets $P \subseteq \mathcal{P}$ with $\gamma(P) = P$ are trivial and therefore not included in $\mathscr{S}$. Second, if there exists a proper subset $\tilde{P} \subsetneq P$ with $\gamma(\tilde{P}) = \gamma(P)$, then P is also not in $\mathscr{S}$. These two properties can be generalized to arbitrary closure systems, and referring to [3], we call a set fulfilling these two properties *generic*. The final reduction is to delete also all sets P where P can be decomposed by a family of proper subsets $(A_i)_{i\in\{1,\dots,\ell\}}$ of P. Moreover, the union of $(\gamma(A_i))_{i\in\{1,\dots,\ell\}}$ needs to equal $\gamma(P)$. Note that due to isotonicity, the assumption $\cup_{i\in\{1,\dots,\ell\}}\gamma(A_i) \subseteq \gamma(P)$ is always true. We call sets respecting this third part *union-free*. Thus $\mathscr{S}$ consists of elements that are union-free and generic. The following definition summarizes these conditions.

Definition 2. Let $\mathcal{P}$ be the set of posets on the finite set M. We set

$$\mathscr{S} = \{ P \subseteq \mathcal{P} \mid \text{Condition } (C1) \text{ and } (C2) \text{ hold for } P \}$$

with Conditions (C1) and (C2) given by:

(C1) $P \subsetneq \gamma(P)$,
(C2) There does not exist a family $(A_i)_{i\in\{1,\dots,\ell\}}$ such that for all $i \in \{1,\dots,\ell\}$ $A_i \subsetneq P$ and $\bigcup_{i\in\{1,\dots,\ell\}} \gamma(A_i) = \gamma(P)$.[9]

We call $\mathscr{S}$ the family of union-free generic sets.

Example 1. As a concrete example, consider the set $\mathscr{S}$ based on all posets on $\{y_1,y_2,y_3\}$. Let p_1,p_2 and p_3 be posets given by the transitive hull of $\{(y_1,y_2)\}$, $\{(y_1,y_2),(y_1,y_3)\}$ and $\{(y_1,y_3),(y_2,y_3)\}$. One can show that $\gamma(\{p_1,p_3\}) = \gamma(\{p_1,p_2,p_3\})$. Thus, $\{p_1,p_2,p_3\}$

[9] In formal concept analysis this is sometimes called *proper*, see [25, p. 81].

contradicts Condition (C2). For a single poset p we can immediately prove that the closure contains only itself. Therefore, any set consisting of only one poset does not satisfy Condition (C1). In contrast, $\{p_1, p_3\}$ satisfies both Condition (C1) and Condition (C2), since it implies the trivial poset $p_\Delta := \{(y, y) \mid y \in M\}$. Thus, $\{p_1, p_3\}$ is an element of $\mathscr{S}$.

Now, we define the *union-free generic (ufg) depth* of a poset p to be the weighted probability that p lies in a randomly drawn element of $\mathscr{S}$.

Definition 3. Let $\mathcal{M}$ be the set of probability measures on $\mathcal{P}$ equipped with $2^{\mathcal{P}}$ as σ-field. The *union-free generic (ufg for short) depth on posets* is given by

$$D : \begin{array}{l} \mathcal{P} \times \mathcal{M} \to [0,1] \\ (p, \nu) \mapsto \begin{cases} 0, & \text{if for all } S \in \mathscr{S} : \ \prod_{\tilde{p} \in S} \nu(\{\tilde{p}\}) = 0 \\ c \sum_{S \in \mathscr{S} : \ p \in \gamma(S)} \prod_{\tilde{p} \in S} \nu(\{\tilde{p}\}), & \text{else} \end{cases} \end{array}$$

with $c = \left(\sum_{S \in \mathscr{S}} \prod_{\tilde{p} \in S} \nu_n(\{\tilde{p}\}) \right)^{-1}$.[10]

The two cases in Definition 3 are needed because the constant c is not defined in the first case. Note that if there exists an $S \in \mathscr{S}$ with $\prod_{\tilde{p} \in S} \nu(\tilde{p}) \neq 0$, then $D \not\equiv 0$. The case that $D \equiv 0$ only occurs in two specific situations that result from the structure of the probability mass, see the non-triviality property in Corollary 8 in Section 5 for details. Note that in contrast to the simplicial depth where only sets of cardinality $d+1$ are considered, the elements of $\mathscr{S}$ differ in their cardinality. Thus, different approaches on how to include the different cardinalities are possible, i.e., by weighting. In Definition 3 we use weights equal to one.

The empirical version of the ufg depth uses the empirical probability measure ν_n given by a sample of posets $\underline{p} = (p_1, \dots, p_n)$, $n \in \mathbb{N}$ instead of the probability measure ν in Definition 3. Thus, the empirical ufg depth of a poset p is therefore the normalized weighted sum of drawn sets $S \in \mathscr{S}$ which imply p.

Definition 4. Let ν_n be an empirical probability measure based on sample $\underline{p} = (p_1, \dots, p_n)$, $n \in \mathbb{N}$ (equipped with $2^{\mathcal{P}}$ as σ-field). The *empirical union-free generic (ufg) depth* is then given by

$$D_n : \begin{array}{l} \mathcal{P} \to [0,1] \\ p \mapsto \begin{cases} 0, & \text{if for all } S \in \mathscr{S} : \ \prod_{\tilde{p} \in S} \nu_n(\{\tilde{p}\}) = 0 \\ c_n \sum_{S \in \mathscr{S}, p \in \gamma(S)} \prod_{\tilde{p} \in S} \nu_n(\{\tilde{p}\}), & \text{else} \end{cases} \end{array}$$

with $c_n = \left(\sum_{S \in \mathscr{S}} \prod_{\tilde{p} \in S} \nu_n(\{\tilde{p}\}) \right)^{-1}$.

Note that when restricting $\mathscr{S}$ to the set $\mathscr{S}_{obs} := \{S \in \mathscr{S} \mid \text{all } p \in S \text{ are observed}\}$, this does not change the depth value. This holds since for other elements $S \in \mathscr{S}$, the empirical measure for at least one $p \in S$ is zero.

Example 2. Returning to Example 1, suppose that we observe (p_1, p_2, p_3). Then for the trivial poset p_Δ, the empirical ufg depth is $D_n(p_\Delta) = 1/2$. For the set p_4 given by the transitive hull of $\{(y_3, y_1)\}$, the value of the empirical ufg depth is zero. For p_{total} given by the transitive hull of $\{(y_1, y_3), (y_3, y_2)\}$, the empirical ufg depth value is again zero.

5. Properties of the UFG depth and $\mathscr{S}$

For a better understanding of the ufg depth, we now discuss some properties of D, D_n, and $\mathscr{S}$. The properties of D_n and D describe the mutual influence between the (empirical) measure and the ufg depth.

5.1. Properties of $\mathscr{S}$

We begin with introducing some properties of $\mathscr{S}$. Later, these properties are used to analyze the (empirical) ufg depth and to improve the computation. First, we want to introduce a second equivalent definition of Condition (C1) and Condition (C2). This says that $S \in \mathscr{S}$ if and only if there exists a poset q such that every poset $p \in S$ contributes to q. More precisely, for every $p \in S$ we have that either p has an edge (y_1, y_2) which only p and q share, or p is the only poset in S which does not have an edge (y_1, y_2) which also q does not have. Note that if we have ensured that every poset $p \in S$ contributes to a poset $q \in \gamma(S)$, then S satisfies Conditions (C1) and (C2). This will help later to prove the connectedness property, see Theorem 3, give an upper bound for $\#S$ with $S \in \mathscr{S}$, see Theorem 5, and to improve the implementation, see Section 6.

[10] Note that Condition (C1) and (C2) can be applied to the convex closure operator on $\mathbb{R}^d$, see Section 3, and we obtain an adapted $\mathscr{S}_{convex}$. Then, $\mathscr{S}_{convex}$ together with $\mathcal{M}_{convex}$ the set of measures that are absolute continuous to the Lebesgue measure, leads to the simplicial depth.

Lemma 1. *Let $S \subseteq \mathcal{P}$. Then $S \in \mathscr{S}$ if and only if there exists a poset $q \in \gamma(S) \setminus S$ such that for all $p \in S$, $q \notin \gamma(S \setminus \{p\})$.*

Proof. Let us first assume that $S \in \mathscr{S}$. Then, due to Condition (C1), we know that $\gamma(S) \setminus S$ is nonempty. Since $\left((S \setminus \{x\})_{x \in S}\right)$ is a family of proper subsets of S, we obtain by Condition (C2) that there must exist an element $q \in \gamma(S) \setminus S$ such that for all $x \in S$, $q \notin \gamma(S \setminus \{x\})$.

Conversely, suppose that there exists $q \in \gamma(S) \setminus S$ such that for all $x \in S$, $q \notin \gamma(S \setminus \{x\})$. Condition (C1) follows immediately. To prove Condition (C2), let $(A_i)_{i \in \{1,\dots,\ell\}}$ be a family of all sets with $A_i \subsetneq S$ for all $i \in \{1,\dots,\ell\}$. Then for every $i \in \{1,\dots,\ell\}$ there exists an $p_i \in S$ with $p_i \notin A_i$ (follows from $A_i \subsetneq S$). Since γ is isotone, we know that $\gamma(A_i) \subseteq \gamma(S \setminus \{p_i\})$. By assumption we get that $q \notin \gamma(S \setminus \{x_i\})$ and, thus, $q \notin \gamma(A_i)$. This argument holds for every $i \in \{1,\dots,\ell\}$ and we obtain $q \notin \bigcup_{i \in \{1,\dots,\ell\}} \gamma(A_i)$. With $(A_i)_{i \in \{1,\dots,\ell\}}$ arbitrarily chosen, the claim follows.

Definition 5. We call such a poset q given by Lemma 1 an *ufg element* w.r.t. S.

As we pointed out above, if q is an ufg element w.r.t. some set $S \in \mathscr{S}$, then each poset $p \in S$ has two different ways of contributing to q: Either it has an edge that only p and q share, or it has an edge that only p and q do not have. With this we immediately get a description of the ufg elements w.r.t. some set $S \in \mathscr{S}$. Therefore we define the following two sets, which sort the posets into two sets w.r.t. $S \in \mathscr{S}$. One consists of those posets that have an edge that all other posets in S do not have, and conversely the other set contains all posets that do not have an edge that all other posets in S have. Note that these two sets are not necessarily disjoint.

Definition 6. Let $S \subseteq \mathcal{P}$ and $p \in S$. We define

$$D_S^{p,\mathrm{edge}} = \{(a,b) \in M \times M \mid (a,b) \in p \text{ and } \forall \tilde{p} \in S \setminus \{p\} \text{ we have } (a,b) \notin \tilde{p}\},$$
$$D_S^{p,\cancel{\mathrm{edge}}} = \{(a,b) \in M \times M \mid (a,b) \notin p \text{ and } \forall \tilde{p} \in S \setminus \{p\} \text{ we have } (a,b) \in \tilde{p}\}.$$

$D_S^{p,\mathrm{edge}}$ consists of all edges (a,b) which the poset p has, but which each other poset $\tilde{p} \in S \setminus \{p\}$ does not have. Reverse, $D_S^{p,\cancel{\mathrm{edge}}}$ consist of all edges which the poset p does not have, but every other poset $\tilde{p} \in S \setminus \{p\}$ does have.

Using these definitions we obtain by Lemma 1 and its corresponding discussion a concrete definition of an ufg element w.r.t. S if $S \in \mathscr{S}$, see the next corollary.

Corollary 2. *Let $S \subseteq \mathcal{P}$. Then q is an ufg element w.r.t. S if and only if for all $p \in S$ we have that $q \cap D_S^{p,\mathrm{edge}} \neq \emptyset$ or $(M \times M) \setminus q \cap D_S^{p,\cancel{\mathrm{edge}}} \neq \emptyset$.*

Moreover, for every ufg element q w.r.t. S there exists an ufg element $\tilde{q} \subseteq q$ w.r.t. S given by

$$\tilde{q} = th\left(\left(\bigcap_{p \in S} p\right) \cup \left(\bigcup_{p \in \tilde{S}} (y_1^p, y_2^p)\right)\right), \tag{1}$$

with $\tilde{S} = \{p \in S \mid D_S^{p,\cancel{\mathrm{edge}}} = \emptyset\}$ and for all $p \in \tilde{S}$ we set (y_1^p, y_2^p) to be precisely one edge $(y_1^p, y_2^p) \in D_S^{p,\mathrm{edge}}$.

Proof. The first part follows directly from the discussion corresponding to Definition 5 and Lemma 1. Therefore, we focus on showing that for each ufg element q w.r.t. $S \in \mathscr{S}$ there exists a further ufg element $\tilde{q} \subseteq q$ that is given by Equation (1). For each $p \in S$ we know that $q \cap D_S^{p,\mathrm{edge}} \neq \emptyset$ or $(M \times M) \setminus q \cap D_S^{p,\cancel{\mathrm{edge}}} \neq \emptyset$ is true. In particular by the definition of $\tilde{S}$, we obtain that for all $\tilde{p} \in \tilde{S}$ we know $D_S^{\tilde{p},\mathrm{edge}} \neq \emptyset$. Thus, we get $(y_1^{\tilde{p}}, y_2^{\tilde{p}}) \in D_S^{\tilde{p},\mathrm{edge}}$ for all $\tilde{p} \in \tilde{S}$. We now use these edges $(y_1^{\tilde{p}}, y_2^{\tilde{p}})$ to define $\tilde{q}$ as in Equation (1). With $q \in \gamma(S)$ we know that $\cap_{p \in S} p \subseteq q$ needs to be true and therefore, we obtain $th(\tilde{q}) \subseteq th(q) = q \subseteq \cup_{p \in S} p$ which proves the claim.

Note that there are cases where q and $\tilde{q}$ of Corollary 2 are different. Recall Example 1. Then $\{p_2, p_3\} \in \mathscr{S}$ and $q = \{(y_1, y_2)\} \in \gamma(\{p_2, p_3\})$ is an ufg element with respect to $\{p_2, p_3\}$. Furthermore, $\tilde{q} = p_\Delta$ is also an ufg element w.r.t. $\{p_2, p_3\}$, which can be given by Equation (1). We have $\tilde{q} \subseteq q$.

After discussing the different definitions of sets in $\mathscr{S}$, let us now consider the claims made in Section 4. These claims are that sets of cardinality one cannot be union-free and generic, and that only special cases of sets of cardinality two are union-free and generic. These two claims are proved in the following theorem. Furthermore, this theorem shows that the sets in $\mathscr{S}$ are connected in the sense that for every $S \in \mathscr{S}$ with $\#S = m \geq 3$, there exists $S_m \in \mathscr{S}$ such that $S_m \subsetneq S$ and $\#S_m = m - 1$.

Theorem 3. *The family of sets $\mathscr{S}$ given in Section 4 fulfills the following properties.*

1. *For every $p \in \mathcal{P}$, $\{p\} \notin \mathscr{S}$.*
2. *Let $\{p_1, p_2\} = S \in 2^{\mathcal{P}}$. Then $S \notin \mathscr{S}$ if and only if the transitive reductions $tr(p_1)$ and $tr(p_2)$ differ only on one edge (y_i, y_j) which is contained in either $tr(p_1)$ or $tr(p_2)$. This means that either $\#\{(tr(p_1) \setminus tr(p_2))\} = 1$ or $\#\{(tr(p_2) \setminus tr(p_1))\} = 1$ holds.*
3. *$\mathscr{S}$ is connected in the sense that for every set $S \in \mathscr{S}$ of size $k \geq 3$ there exists a subset $S_m \subsetneq S$ of size $k - 1$ that is in $\mathscr{S}$ too.*

Proof. Claim 1. follows directly from Condition (C1) of Definition 1 as $\gamma(\{p\}) = \{p\}$ for every $p \in \mathcal{P}$.

Now, we show the second claim. Let us first assume that $\{p_1, p_2\} = S \notin \mathscr{S}$. Using Part 1. we get that Condition (C1) is not fulfilled. Hence, there exists no $p \in \mathcal{P}$ such that $p \in \gamma(S) \setminus \{p_1, p_2\}$. Thus, the intersection must be either p_1 or p_2, (otherwise $p_1 \cap p_2 \in \gamma(S) \setminus S$). W.l.o.g., let $p_1 = p_1 \cap p_2$. Then p_2 must be a superset of p_1 where there is no poset lying between p_1 and p_2. Therefore, $\#\{tr(p_2) \setminus tr(p_1)\} = 1$ is true. Conversely, assume that $S \in \mathscr{S}$ and that p_1 is a superset of p_2. With this and Condition (C1) we obtain that $\gamma(S) = \{p \in \mathcal{P} \mid p_2 \subseteq p \subseteq p_1\} \subsetneq \{p_1, p_2\}$. Further assume that $\#\{tr(p_1) \setminus tr(p_2)\} = 1$ holds. However, with this, we get $\gamma(S) = S$ is true since no $p \in \mathcal{P}$ can lie between p_1 and p_2. This is a contradiction which proves the claim.

Finally, we want to show the third part. Let $S \in \mathscr{S}$ with $\#S \geq 3$. We show that there exists $S_m \in \mathscr{S}$ with $S_m \subsetneq S$ and $\#S \setminus S_m = 1$. Let us distinguish the following two cases:

Case 1: There exists $\tilde{p} \in S$ such that $D_S^{p,\not{edge}} \neq \emptyset$ for all $p \in S \setminus \tilde{p}$. Then we set $q_m = \cap_{p \in S \setminus \tilde{p}} p$. Since for all $p \in S \setminus \{\tilde{p}\}$ we have $(M \times M) \setminus q_m \cap D_S^{p,\not{edge}} \neq \emptyset$, we can follow by Corollary 2 that q_m is an ufg element w.r.t. $S \setminus \{\tilde{p}\}$. Hence, $S_m = S \setminus \{\tilde{p}\} \in \mathscr{S}$.

Case 2: There exist $p_1, p_2 \in S$ such that $D_S^{p_i,\not{edge}} = \emptyset$ for $i \in \{1,2\}$. Before proceeding let us discuss the following claim:

Claim:* Let $S \in S$ with $\#S \geq 3$. Moreover, assume that there exist $\tilde{p_1}, \tilde{p_2} \in S$ such that $D_S^{\tilde{p}_i,\not{edge}} = \emptyset$ for $i \in \{1,2\}$. By Corollary 2 there exists an ufg element $\tilde{q}$ which is given by Equation (1). Analogously to this corollary, we set $\tilde{S} = \{p \in S \mid D_S^{p,\not{edge}} = \emptyset\} \supseteq \{\tilde{p}_1, \tilde{p}_2\}$. Then, we claim that there exists $p_0 \in \tilde{S}$ and $(a,b) \in \tilde{q} \cap D_S^{p_0,edge}$ such that there are no $y_1, \ldots, y_n \in M$ with $n \geq 3$ and $(a, y_1)(y_1, y_2), \ldots, (y_n, b) \in tr(\tilde{q})$. In other words, (a,b) cannot be given by a transitive chain and must be contained in the transitive reduction of $\tilde{q}$.

Assume in contradiction that Claim* is not true. Then for every $p_1 \in \tilde{S}$ and every element $(a,b) \in \tilde{q} \cap D_S^{p_1,edge}$ there must exists a sequence $y_1, \ldots, y_n \in M$ with $n \geq 3$ such that $(a, y_1)(y_1, y_2), \ldots, (y_n, b) \in tr(\tilde{q})$.

Let $p_1 \in \tilde{S}$ and $(a,b) \in \tilde{q} \cap D_S^{p_1,edge}$. By assuming the contradiction, there exists $a = y_0^1, y_1^1, \ldots, y_{n+1}^1 = b \in M$ such that $(y_0^1, y_1^1)(y_1^1, y_2^1) \ldots, (y_n^1, y_{n+1}^1) \in tr(\tilde{q})$. Since $\tilde{q}$ is in the style of Equation (1), the transitive reduction consists of elements that either lie in every poset or which lie in precisely one poset of S. Observe that every element (y_{i-1}, y_i) for $i \in \{1, \ldots, n+1\}$ cannot lie in the intersection of all posets since then we have $(a,b) \in \cap_{p \in S} p$ which is a contradiction to $(a,b) \in \tilde{q} \cap D_S^{p_1,edge}$. Hence, there exists $i \in \{1, \ldots, n\}$ and $p_2 \in \tilde{S}$ such that $(y_{i-1}^1, y_i^1) \in \tilde{q} \cap D_S^{p_2,edge}$. By assumption, we can again find a sequence of pairs in $tr(\tilde{q})$ such that (y_{i-1}^1, y_i^1) stems from transitivity. Note that (a,b) cannot be an edge of the sequence to obtain (y_{i-1}^1, y_i^1) by transitivity, because then this leads to $\tilde{q}$ containing a cycle and not being antisymmetric. Again, this sequence contains at least one pair $(y_{j-1}^2, y_j^2) \in \tilde{q} \cap D_S^{p_2,edge}$ and since we assume the contradiction of Claim* this pair can be obtained by a sequence representing the transitivity assumption. But now (a,b) and (y_{i-1}^1, y_i^1) cannot be used in the sequence to get (y_{j-1}^2, y_j^2). Since M is finite this leads to a contradiction as the process needs to stop at some point. This proves Claim*.

Let us go back to Case 2. We set $\tilde{S} = \{p \in S \mid D_S^{p,\not{edge}} = \emptyset\}$. Since $S \in \mathscr{S}$, there exists an ufg element $\tilde{q}$ w.r.t. S which is given by Equation (1). Now, we provide a procedure to give a modified version of $\tilde{q}$ (which we denote by q_m) which is then an ufg element w.r.t. a subset $S_m \subseteq S$ with $\#S \setminus S_m = 1$.

1. Step 1: By Claim* there exists $p \in \tilde{S}$ and $(a,b) \in \tilde{q} \cap D_S^{p,edge}$ which cannot be divided by a transitive sequence. Hence $\tilde{q} \setminus (a,b)$ is again a poset.
2. Step 2: If $\{(a,b)\} = \tilde{q} \cap D_S^{p,edge}$, then we set $\tilde{q}_m = (\tilde{q} \setminus \{(a,b)\})$ and we are finished ($\tilde{q}_m$ is then the ufg element w.r.t. $S_m = S \setminus \{p\}$ and we can apply Lemma 1). Else, we modify $\tilde{q}$ to $\tilde{q}^{next} = \tilde{q} \setminus (a,b)$. Note that $\tilde{q}^{next}$ is again an ufg element w.r.t. S which is in the style of Equation (1). This follows from the fact that there exists a further element $(\tilde{a}, \tilde{b}) \in \tilde{q} \cap D_S^{p,edge}$ with $(a,b) \neq (\tilde{a}, \tilde{b})$. Therefore, we can again apply Claim* on $\tilde{q}^{next}$ and go back to Step 1. This procedure will end after a finite number of steps since M is finite. Hence, we get to a point where for one $p \in \tilde{S}$ and the modified $\tilde{q}^{next}$ we have $1 = \#(\tilde{q}^{next} \cap D_S^{p,edge})$ and can apply the first part of Step 2.

Theorem 3 Part 1 gives us directly a lower bound for the cardinality of all sets $S \in \mathscr{S}$. For the upper bound, we use a complexity measure of $\mathscr{S}$, the Vapnik-Chervonenkis dimension (VC dimension for short), see [59]. The VC dimension of a family of sets $\mathscr{C}$ is the largest cardinality of a set A, such that A can still be shattered into the power set of A by $\mathscr{C}$.[11] With this, we obtain an upper bound for all $S \in \mathscr{S}$ is given by $\#S \leq vc$, with vc the VC dimension of the closure system $\gamma(2^{\mathcal{P}})$. Note that in our case of posets, the VC dimension is small compared to the number of all posets.

Theorem 4. *For all $S \in \mathscr{S}$, as defined in Section 4, $\#S \geq 2$ and $\#S \leq vc$ is true, where vc is the VC dimension of the set $\gamma(2^{\mathcal{P}})$.*

Proof. Let $S \in \mathscr{S}$. The proof for $\#S \geq 2$ follows immediately from Theorem 3.
To prove $\#S \leq vc$ take an arbitrary subset $Q = \{p_1, \ldots, p_k\} \in \mathscr{S}$ of size $k > vc$. Then this subset is not shatterable because vc is the largest cardinality of a shatterabel set. Thus there exists a subset $R \subseteq Q$ that cannot be obtained as an intersection of Q and some

[11] To be more precise: The intersection between a set A and a family of sets $\mathscr{C}$ is defined by $A \cap C = \{A \cap C \mid C \in \mathscr{C}\}$. We say that a set A can be shattered (by $\mathscr{C}$) if $\#(A \cap \mathscr{C}) = 2^{\#A}$ holds. The VC dimension of $\mathscr{C}$ is now defined as $vc = \max\{\#A \mid (A \cap \mathscr{C}) = 2^A\}$.

$\gamma(A)$ with $A \subseteq \mathcal{P}$. In particular, this holds for $R = A$. Thus, $R \neq \gamma(R) \cap Q$ and with the extensitivity of γ we get $R \subsetneq \gamma(R) \cap Q$. This means that there exists an order p_i in $\gamma(R) \cap Q \setminus R$ for which the formal implication $R \to \{p_i\}$ holds. Thus, (because of the Armstrong rules, cf., [1, p. 581]) the order p_i is redundant in the sense of $Q \setminus \{p_i\} \to Q$ and thus Q is not minimal with respect to γ. Therefore, Q is not in $\mathscr{S}$ which completes the proof.

Remark 2. In concrete applications, one usually does not observe all possible posets. Therefore, many summands in the definition of the empirical ufg depth (see Definition 4) are zero. Thus, one can restrict the analysis to $\mathscr{S}_{obs} := \{S \in \mathscr{S} \mid \text{ all } p \in S \text{ are observed}\}$. Then a similar argumentation shows that $\#S \leq vc_{obs}$ where vc_{obs} is the VC dimension of $\gamma(2^{\mathcal{P}_{obs}}) \cap \mathcal{P}_{obs}$ and $\mathcal{P}_{obs}$ is the set of all observed posets.

Beyond the bound for the size of a set $S \in \mathscr{S}$ that is given by the VC dimension and that is generally valid, we can additionally give another bound for the special case of poset data that is sharper than the VC bound in certain cases.

Theorem 5. *Let $m := \#M \geq 3$. Then, for every $S \in \mathscr{S}$ we have that $\#S \leq m(m-1)/2$.*

Proof. Let $M = \{x_1, x_2, \dots, x_m\}$. Let $S = \{p_1, \dots, p_\ell\} \in \mathscr{S}$ and let q be an ufg element w.r.t. S. Then, by Corollary 2, for all $p \in S$ we have that $q \cap D_S^{p,\text{edge}} \neq \emptyset$ or $(M \times M) \setminus q \cap D_S^{p,\text{edge}} \neq \emptyset$. This means that for every poset in S there is an edge $(x_i, x_j) \in p$ with $i, j \in \{1, \dots, m\}$ and $i \neq j$ such that p is needed to either

(i) contribute in q by providing an edge $(x_i, x_j) \in q \cap p$ that no other poset in S has, or
(ii) to contribute in q by not having an edge $(x_i, x_j) \notin q$ and $(x_i, x_i) \notin p$ that every other poset has.

This means that p has a unique existence-characteristic w.r.t. edge (x_i, x_j) and S, while all other posets in S agree to have the opposite existence-characteristic w.r.t. (x_i, x_j) and S. That is, if $(x_i, x_j) \in p$, for all $\tilde{p} \in S \setminus \{p\}$ we have $(x_i, x_j) \notin \tilde{p}$ and vice versa for $(x_i, x_j) \notin p$.
Since by Corollary 2 each poset must somehow contribute uniquely to q, we get that each edge can be used by only one single poset in S. This follows from the fact that $\#S \geq 3$ and all other posets $\tilde{p} \in S \setminus \{p\}$ need to agree on the opposite existence-characteristic w.r.t. (x_i, x_j). With this we obtain that $\#S \leq m(m-1)$ (since the reflexive part holds by default for every poset).
We continue with the above poset $p \in S$ and the corresponding edge (x_i, x_j) that uniquely contributes to q by p. We show that the inverse edge (x_j, x_i) cannot be a contributing element for any $\tilde{p} \in S \setminus \{p\}$. To show this, we split the proof into two cases:
Case 1: $(x_i, x_j) \in q \cap D_S^{p,\text{edge}}$. This means that all posets $\tilde{p} \in S \setminus \{p\}$ agree on $(x_i, x_j) \notin \tilde{p}$. We get that any poset $\tilde{p} \in S \setminus \{p\}$ cannot use $(x_j, x_i) \in \tilde{p}$ to uniquely contribute to q, since this edge never occurs in q (q is a poset and therefore antisymmetric). Moreover, since $(x_j, x_i) \notin p$ is true, $\tilde{p}$ cannot contribute uniquely to q by $(x_j, x_i) \notin \tilde{p}$. Since $\#S \geq 3$, there cannot be another poset $\tilde{p}$ that uses the inverted edge as a unique contribution to q.
Case 2: $(x_i, x_j) \in (M \times M) \setminus q \cap D_S^{p,\text{edge}}$. With this we get that for all posets $\tilde{p} \in S \setminus \{p\}$ we have $(x_i, x_j) \in \tilde{p}$. Due to antisymmetry we immediately get that $(x_j, x_i) \notin \tilde{p}$ for all $\tilde{p} \in S \setminus \{p\}$. With $\#S \geq 3$ we have that (x_j, x_i) cannot be used to uniquely contribute to q by a poset $\tilde{p}$.
With these two cases we have that only half of all possible edges can be used and therefore we have shown that $\#S \leq m(m-1)/2$.

Remark 3. The bound of Theorem 5 is tight in the sense that there exists a set $S \in \mathscr{S}$ with cardinality $m(m-1)/2$, namely the set

$$S := \{\{(x_i, x_j)\} \mid i, j \in \{1, \dots, m\}, i < j\}.$$

We conclude with a technical observation that we need to analyze the properties of the (empirical) ufg depth. The next lemma states that the set $\mathscr{S}$ can be rewritten.

Lemma 6. *For $p \in \mathcal{P}$ we get*

$$\{S \in \mathscr{S} \mid p \in \gamma(S)\} \tag{2}$$

$$= \bigcap_{(y_i, y_j) \in p} \{S \in \mathscr{S} \mid \exists p \in S : (y_i, y_j) \in p\} \cap \tag{3}$$

$$\bigcap_{(y_i, y_j) \notin p} \{S \in \mathscr{S} \mid \exists p \in S : (y_i, y_j) \notin p\}. \tag{4}$$

Proof. Let $p \in \mathcal{P}$. The proof is divided into two parts.
Part 1: We prove $\subseteq$. Let S be an element of (2). Since $p \in \gamma(S)$, we have $p \subseteq \cup_{\tilde{p} \in S} \tilde{p}$. So for every $(y_i, y_j) \in p$ there is a $\tilde{p} \in S$ such that $(y_i, y_j) \in \tilde{p}$. Therefore, S is an element of the intersection of (3). Also from $p \in \gamma(S)$ we get $\cap_{\tilde{p} \in S} \tilde{p} \subseteq p$ and thus we know that for every $(y_i, y_j) \notin p$ there exists a $\tilde{p} \in S$ such that $(y_i, y_j) \notin \tilde{p}$. Thus, S is an element of the intersection given by (4). This proves Part 1.

Part 2: We prove $\supseteq$. Let $S \in \mathscr{S}$ be an element of the right-hand side of the equation. We show that $p \in \gamma(S)$. Let S be in the intersection given by (3). Then we know that for every $(y_1, y_2) \in p$ there exists an $\tilde{p} \in S$ such that $(y_1, y_2) \in \tilde{p}$. Thus $p \subseteq \cup_{\tilde{p} \in S} \tilde{p}$. The second part of the intersection given by (4) analogously yields that $\cap_{\tilde{p} \in S} \tilde{p} \subseteq p$. Hence $p \in \gamma(S)$ and the second part is proven. The claim follows from Part 1 and Part 2.

5.2. Properties of the (empirical) UFG depth

In this subsection, we introduce properties of the (empirical) probability measure D_n and D. The following statements focus on D_n. Those properties that use only the empirical measure and not the concrete sample values can be transferred to D.

The first observation is that the ufg depth considers the orders as a whole, not just pairwise comparisons. More precisely, the ufg depth cannot be represented as a function of the sum-statistics

$$\left(w_{(a,b)} := \#\{i \in \{1, \dots, n\} \mid (a,b) \in p_i\}\right)_{(a,b) \in M \times M}$$

of the pairwise comparisons, see Theorem 7. Note that many classical approaches rely only on the sum-statistics. For example, within the Bradly-Terry-Luce model (cf., [9, p. 325]) or the Mallows Φ model (cf., [22, p. 360]), the likelihood function that is maximized depends only on the data through the sum-statistics.

Theorem 7. *D_n cannot be represented as a function of the sum-statistics $w_{(a,b)}$.*

Proof. We simply give two data sets $D = (p_1, p_2, p_3)$ and $\tilde{D} = (\tilde{p}_1, \tilde{p}_2, \tilde{p}_3)$ on the basic set $M = \{y_1, y_2, y_3\}$ with the same sum-statistics but different associated depth functions: Let p_1, p_2 and p_3 be given as the transitive reflexive closures of $\{(y_1, y_2)\}$; $\{(y_1, y_2), (y_1, y_3)\}$ and $\{(y_2, y_3), (y_1, y_3)\}$, respectively. Let $\tilde{p}_1$, $\tilde{p}_2$ and $\tilde{p}_3$ be the transitive reflexive closure of $\{(y_1, y_2)\}$; $\{(y_1, y_3)\}$ and $\{(y_1, y_2), (y_2, y_3)\}$, respectively. Then both data sets have the same sum-statistics $w_{(y_1,y_2)} = w_{(y_1,y_3)} = 2$; $w_{(y_1,y_3)} = 1$ and $w_{(y_i,y_j)} = 0$ for all other $y_i \neq y_j$. But the ufg depth of $p_1 = \tilde{p}_1$ is $1/2$ w.r.t. the first data set but $7/10$ w.r.t. the second data set. The corresponding code can be found at the link mentioned in Footnote 1.

Remarkably, this concretization by Theorem 7 formalizes precisely the analogy to the ontic notion of non-standard data mentioned at the beginning: Computing the depth of a partial order cannot be broken down via simple sum-statistics, but requires the partial order as a holistic entity. This is due to the fact that the involved set operations within the closure operator γ rely on the partial orders as a whole.

In Section 4, we defined the ufg depth in terms of two cases. If there exists at least one element $S \in \mathscr{S}$ such that every $p \in S$ has a positive empirical measure, then $D_n \not\equiv 0$. In Corollary 8 we specify this non-triviality property. We claim that $D_n \equiv 0$ occurs only when either the entire (empirical) probability mass lies on one poset or when the (empirical) probability mass is on two posets where the transitive reduction differs only in one pair, see Theorem 3.

Corollary 8. *$D(p) = 0$ for every $p \in \mathcal{P}$ if and only if the measure ν has either the entire positive probability mass on a single poset p or on exactly two posets p_1 and p_2 where the transitive reduction differs only in a pair (y_1, y_2), i.e., either $\#\{tr(p_1) \setminus tr(p_2)\} = 1$ or $\#\{tr(p_2) \setminus tr(p_1)\} = 1$.*

Proof. Note that $D(p) = 0$ for every $p \in \mathcal{P}$ is true if for all $S \in \mathscr{S}$, $\prod_{\tilde{p} \in S} \nu(\tilde{p}) = 0$. Theorem 3 Part 1. and 2. provide the cases when this holds which proves immediately the claim.

The converse follows analogously from Theorem 3.

The next observation relates to how the sampled posets affect the ufg depth value. Therefore, let us recall Example 1 and Example 2. From the structure of the sample, we can immediately see that p_Δ has a nonzero depth and that p_{total} must have a depth of zero. Now, let us take a closure look at how the structure of the sample effects D_n. Therefore, let $\underline{p} = (p_1, \dots, p_n)$ be a sample from $\mathcal{P}$. Let $(y_1, y_2) \in M \times M$ such that for all $i \in \{1, \dots, n\}$, $(y_1, y_2) \notin p_i$. Then for every $p \in \mathcal{P}$ with $(y_1, y_2) \in p$, we get $D_n(p) = 0$. This means that if a pair does not occur in any poset of the sample, then every poset that contains this pair needs to have zero empirical depth. Reverse, when looking at non-pairs, a similar statement is true. Let $p \in \mathcal{P}$ with $(y_1, y_2) \notin p$ but for all $i \in \{1, \dots, n\}$, $(y_1, y_2) \in p_i$ holds. Then, $D_n(p) = 0$. This follows from Corollary 9. This discussion is based on whether a pair has been observed or not. Now, we are interested in how duplicates influence the value of the empirical ufg depth D_n. This is immediately apparent by using the empirical measure ν_n. Thus, each element in $\mathscr{S}$ is weighted by the number of duplicates in the sample $\{p_1, \dots, p_n\}$.

Conversely to the question of how the sample effects the values of D_n, in some cases, structure in the sample can be inferred by the ufg depth values. In Example 1 and Example 2, knowing only the values of the depth function gives us some insight into the observed posets. For example, we know that there must be at least one pair (y_i, y_j) that is an element of p_{total}, but which is not given by any observed poset. Moreover, the fact that p_Δ has nonzero depth implies that there exists no pair (y_i, y_j) that every observed poset has. More precisely, the depth value of the trivial poset, which consists only of the reflexive part, as well as the values of the total orders, can provide further information about the sample. Therefore, let p_Δ be the trivial poset, and p_{total} be a total order. This implication of the outliers on the sample property is discussed by Corollary 9.

Corollary 9. *Let $(p_1,\dots,p_n)$ be a sample of $\mathcal{P}$. Let ν_n be the empirical probability measure induced by $(p_1,\dots,p_n)$. Furthermore, let ν_n be such a probability measure that $D_n \not\equiv 0$. Then for D_n, the following statements are true.*

1. *Assume that for all $p_i \in \{p_1,\dots,p_n\}$, $(y_1,y_2) \in p_i$ is true. Then for every poset $p \in \mathcal{P}$ with $(y_1,y_2) \notin p$, $D_n(p) = 0$ follows.*
2. *Assume that for all $p_i \in \{p_1,\dots,p_n\}$, $(y_1,y_2) \notin p_i$ holds. Then for every poset $p \in \mathcal{P}$ with $(y_1,y_2) \in p$, $D_n(p) = 0$ is true.*
3. *Let p_Δ be the poset consisting only of the reflexive part. $D_n(p_\Delta) = 0$ if and only if there exists a pair (y_1,y_2) such that for all $p_i \in \{p_1,\dots,p_n\}$, $(y_1,y_2) \in p_i$.*
4. *Let $p_{total} \in \mathcal{P}$ be a total order. $D_n(p_{total}) = 0$ if and only if there exists a pair $(y_1,y_2) \notin p_{total}$ such that for all $p_i \in \{p_1,\dots,p_n\}$, $(y_1,y_2) \in p_i$ is true.*

Proof. First, note that for $S \in \mathcal{S}$, where there exists an $\tilde{p} \in S$ such that $\nu_n(\tilde{p}) = 0$, S contributes nothing to D_n. So one can replace $\mathcal{S}$ in the definition of D_n by $\mathcal{S}_{obs}$, see Remark 2. The reduced set $\mathcal{S}_{obs}$ is used to show the claims.

The proof of Claim 1., 2., 3. and 4. are analogous. Hence, here we provide only the proof of Claim 1. Let $(y_1,y_2) \in M \times M$ such that for all $i \in \{1,\dots,n\}$ $(y_1,y_2) \in p_i$ and let $p \in \mathcal{P}$ such that $(y_1,y_2) \notin p$. Let $S \in \mathcal{S}_{obs}$ and take a closer look at Equation-part (3) of Lemma 6. Since $(y_1,y_2) \notin p$, S cannot be an element of the intersection of (3). Thus, $\{S \in \mathcal{S}_{obs} \mid p \in \gamma(S)\}$ is empty and with the comment above we get that $D_n(p) = 0$.

The last properties have summarized how the structure of a sample is reflected in the ufg depth and vice versa. Finally, we take a look at the consistency of the empirical ufg depth D_n. This means that D_n converges uniformly to D almost surely under the assumption of observing i.i.d. samples, see Theorem 10.

Theorem 10. *Let for all $n \in \mathbb{N}$ the sample $(p_1,\dots,p_n)$ be i.i.d. according to distribution $\nu \in \mathcal{M}$. Then the corresponding empirical depth function D_n almost surely uniformly to $D(\cdot,\nu)$ for n going to infinity.*

Proof. Due to the i.i.d. assumption and the law of large numbers, we know that for every $p \in \mathcal{P}$, $\left\|\nu_n(p) - \nu(p)\right\| \overset{n\to\infty}{\to} 0$ almost surely (a.s). Since $\#\mathcal{P}$ is finite, we get that ν_n also converges a.s. uniformly to ν. Finally, we use that D_n and D are both the same finite composition of ν_n and ν, respectively, and we obtain $\sup_{p\in\mathcal{P}} \left\|D_n(p) - D(p)\right\| \overset{n\to\infty}{\to} 0$ almost surely.

6. Implementation

In this section, we discuss the difficulties and corresponding solution approaches in computing D_n. Therefore, let $\underline{p} = (p_1,\dots,p_n)$ be a sample of posets. The naive approach is to just check all subsets of $\{p_1,\dots,p_n\}$ whether they are in $\mathcal{S}_{obs} := \{S \in \mathcal{S} \mid \text{all } p \in S \text{ are observed}\}$ is very time-consuming, especially since the subsets that are elements of $\mathcal{S}_{obs}$ can be very sparse in $2^{\{p_1,\dots,p_n\}}$. Moreover, it is difficult to test whether a subset is an element of $\mathcal{S}_{obs}$ or if it is not an element by taking Conditions (C1) and (C2) as a basis. Finally, even if we are able to compute the entire family $\mathcal{S}_{obs}$, computing all possible poset p on a set M to obtain some insights on D_n is for even small M with $\#M \geq 8$ very computation intensive and for larger $\#M$ it is currently not possible.[12] We address all these issues in the next paragraphs.

Let us start with computing $\mathcal{S}_{obs}$. First, we can use the lower and upper bound on the cardinality of $S \in \mathcal{S}_{obs}$, see Theorem 4 and Remark 2. Here we use the binary linear programming formulation described in [51, p.33f] to compute the VC dimension. Further, we use the connectedness of the elements $S \in \mathcal{S}$, see Theorem 3 Part 3. With this, we do not have to go through every subset that lies between the lower and upper bounds, but can stop the search earlier.

Still, we need to check whether a subset $P \subseteq \{p_1 \dots, p_n\}$ is an element of $\mathcal{S}_{obs}$. Therefore, we use Corollary 2 together with Definition 5 and 6. Hence, it is sufficient to test if an ufg element w.r.t. P that is in the style of Equation (1) exists. Based on this, we start with computing $D_P^{p,\text{edge}}$, $D_P^{p,\not\text{edge}}$ and $\tilde{S} = \{p \in P \mid D_P^{p,\not\text{edge}} = \emptyset\}$. With this, we can compute all $\tilde{q}$ that is in the style of Equation (1). If one $th(\tilde{q}) \subseteq \cup_{p\in P} p$ is true, then $P \in \mathcal{S}_{obs}$ since $\tilde{q}$ is an ufg element w.r.t. P by Lemma 1. Note that the reverse that $\cap_{p\in P} p \subseteq th(\tilde{q})$ follows directly from the definition given by $\tilde{q}$ via Equation (1).

Now we achieved to compute the whole family $\mathcal{S}_{obs}$. The next issue we approach is that computing all possible posets p on M can be difficult. For small $\#M$, say up to 6 or 7, this is computationally tractable. In this situation, we exploit the fact that every partial order can be represented as the intersection of its linear extensions. Furthermore, we use that the set of all partial orders, together with the relation $M \times M = \{(x,y) \mid x,y \in M\}$, which consists of every element of $M \times M$, defines a closure system on $M \times M$. This allows us to use efficient algorithms from formal concept analysis.[13]

[12] Even computing the number of possible posets is not feasible for $\#M > 18$, see https://oeis.org/A001035 (Accessed 10.11.2023).

[13] For a deeper understanding of this computational approach, knowledge of formal concept analysis is required, see [25]. This implementation is based on a so-called formal context (O,A,I), where each object $o \in O$ is a linear order L (on M) and each attribute $a \in A$ is a pair $(x,y) \in M \times M$ and $(L,(x,y)) \in I \iff (x,y) \in L$. Then, since every partial order is an intersection of a set of linear orders (more concretely, the set of all its linear extensions), the intents of this context are exactly all partial orders on the set M (plus the relation $M \times M$). Therefore, we can compute the set of all partial orders by computing the intents of this formal context instead, e.g. with the next closure algorithm, see [24].

Having explained how we compute all possible partial orders when it is tractable, we now discuss the case where computing all possible posets is no longer feasible. In this case, we use a binary linear program that gives us maximum and minimum depth values, as well as representative posets for each of these values. The binary linear program is defined as follows

$$\sum_{S\in\mathscr{S}_{obs}} \frac{w_S}{c_n}\cdot x_S + \sum_{(a,b)\in M\times M} 0\cdot x_{(a,b)} \longrightarrow \max \tag{5}$$

$$\begin{aligned} \text{subject to } & A_{poset}\, x \geq b_{poset} \\ & A_{intersect}\, x \geq b_{intersect} \\ & A_{union}\, x \geq b_{union} \\ & x_i \in \{0,1\} \quad \text{for } i \in \mathscr{S}_{obs} \cup (M\times M) \end{aligned}$$

where $\frac{w_S}{c_n}$ for $S \in \mathscr{S}_{obs}$ are the weights given by duplicates and the cardinality of S together with the normalization constant, see Definition 4. The first $\#\mathscr{S}_{obs}$ elements of x describe the union-free generic set, and the next $\#M \cdot \#M$ elements guarantee that we only discuss posets. First, we need to ensure that only relations that define a poset are discussed. This is represented in Program (5) by the constraint $A_{poset}\, x \geq b_{poset}$. Therefore matrix A_{poset} together with b_{poset} represents only the reflexive, transitive, and antisymmetric constraints on $x_{(a,b)}$ with $(a,b) \in M \times M$. So in this part, the matrix entries of the first $\#\mathscr{S}_{obs}$ columns are all zero. Second, we have to include in the binary linear program that a set $S \in \mathscr{S}_{obs}$ can only be included in the maximum if the considered poset lies in the intersection of all posets of S, see $A_{intersect}\, x \geq b_{intersect}$. Each row of this constraint corresponds to one ufg set $S \in \mathscr{S}$ and imposes a structure on $x_{(a,b)}$ with $(a,b) \in M \times M$ when this set S is included in the maximization of the objective function. Similarly, we ensure with $A_{union}\, x \geq b_{union}$ that a set $S \in \mathscr{S}_{obs}$ is only counted if the poset considered is a subset of the union of all posets in S. Now, computing the maximum leads to the highest depth value and a poset that has that highest depth value. However, if we are interested in the smallest depth value, this binary linear program does not work, because we have not yet enforced that a set $S \in \mathscr{S}_{obs}$ which has the currently analyzed poset in its closure must be represented in the calculation of the objective function. Therefore, we need to include further constraints so that when a poset is used to minimize the objective, every set $S \in \mathscr{S}_{obs}$ that has that poset in its closure is counted in the objective sum. The code with detailed comments can be found on GitHub, see Footnote 1.

All in all, we improved the computation compared to the naive approach by using the knowledge provided in Section 5. By the two bounds on the cardinality of the sets in $\mathscr{S}_{obs}$ we get an idea of the worst and best case of the computation time. By further using Lemma 1 with Corollary 2 together the connectedness property, see Theorem 3 Part 3, we could decrease the computation time. Although we currently cannot calculate the exact amount of this reduction in general as this depends on the complexity of the data set used. Note that the upper bound using the VC dimension is not fixed, but depends on the structure of the data set. However, the bound given by Theorem 5 holds in general, but can be quite loose in some data situations. Finally, by stating a binary linear program, we achieved that even if not all posets are computable, we get the highest and lowest ufg depth values with representative posets. Appendix A summarizes the computation time and complexity of the two examples in the next section.

7. Application on classifier comparison

In the above sections, we have discussed theoretically how the union-free generic depth captures the structure of the partial orders and how it can be computed. Now, we take the next step and apply our ufg depth to poset-valued data, where each poset arises from the comparison of machine learning algorithms based on multiple performance measures on data sets. Concretely, let us consider k algorithms which are evaluated on the basis of ℓ performance measures on n different data sets. Thus, for each data set and each algorithm, we have ℓ performance measures. This allows us to compare all algorithms based on a single data set. We say that algorithm i is better than/outperforms algorithm j if and only if there exists at least one performance measure such that algorithm i is strictly better than algorithm j and for all other performance measures algorithm i does not perform worse than algorithm j. If two performance measures contradict each other in the sense that for one measure algorithm i is strictly preferred and for another measure algorithm j is strictly preferred, we say that the two algorithms are incomparable. This gives us a poset for each data set, which describes the performance of the algorithms based on that data set. Thus, we observe in total n posets.[14] In what follows, we focus on benchmarking classification algorithms.

We provide two examples for comparing machine learning algorithms based on ufg depth. Both examples use openly available repositories containing data sets with binary classification problems. For each data set in the repository, multidimensional performance measures exist, and in this paper, we use these measures to obtain the corresponding posets. The aim of the first example is to highlight the difference between our approach and other approaches analyzing poset-valued data. The second example demonstrates how the ufg method can be used to gain insight into the performance of different algorithms.

[14] It may occur that two algorithms are equal on all performance measures. In this situation, the two algorithms are indifferent based on the performance measures used. Thus, the described procedure does not lead to a partial order, but only to a preorder. Note that in this situation it is not appropriate to say that there is no dominance structure between these two algorithms, and therefore to include this as one of the incomparability parts in the poset. This confuses the distinction between incomparability and indifference. Here we assume that there are only incomparabilities and no indifference. In the following, we restrict our analysis to algorithms that are substantially different and therefore do not produce the same performance measures.

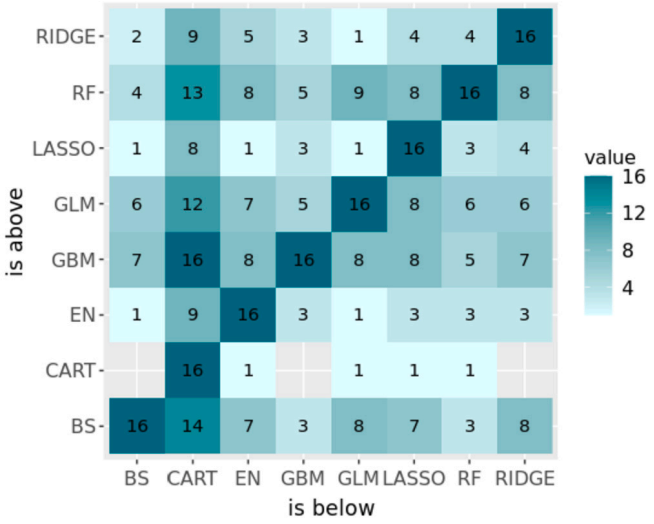

Fig. 1. UCI: Heatmap representing the sum-statistics, see Section 5. (For interpretation of the colors in the figure, the reader is referred to the web version of this article.)

7.1. UCI repository: comparison to related methods

In this section, we illustrate the difference between our approach and other existing approaches to analyze poset-valued data. We focus on an extension of the Bradley-Terry model to ties, see [9,17], and an approach based on generalized stochastic dominance, see [30]. Beyond these two approaches, there are others, such as the Plackett-Luce model, see [44,2], which are not discussed here. We use the data sets provided by the openly available UCI repository, see [19], to obtain the poset-valued data set.

7.1.1. Data set

We take 16 data sets from the UCI machine learning repository, see [19]. They all focus on classifier comparison and vary greatly in size, dimensionality, and class imbalance. The following poset-valued data describing the performance of classifiers is based on the performance evaluation and analysis in [30].

We are interested in the following supervised learning methods: *Boosted Decision Stumps* (BS), *Decision Tree* (CART), *Elastic net penalized logistic regression* (EN), *Gradient Boosting* (GBM), *Generalised Linear Model* (GLM), *L1 penalized logistic regression* (LASSO), *Random Forest* (RF) and *L2 penalized logistic regression* (RIDGE). Their performance is compared using *predictive accuracy, area under the curve* and *Brier score*. These performance measures were calculated by [30]. We refer to [30] (Section 6.1 and Appendix A.2) for more details on the data set selection, implementation, and evaluation of the performance measures.

By using the procedure given at the beginning of Section 7, we obtain one poset describing the performance structure of the eight classifiers for each data set. Thus, we observe 16 posets. These posets have no duplicates. Fig. 1 shows the heatmap of the pairwise comparison of the classifiers based on all three performance measures. We can observe that all posets agree that CART performs worse (in all dimensions) than GBM. We can also see that CART never dominates BS and RIDGE. However, only 14 (respectively nine for the comparison with RIDGE) posets state that BS (respectively RIDGE) is better than CART. For all other posets, CART and BS (or RIDGE) are incomparable in the sense that two performance measures give an opposite evaluation of their performance. Note that the diagonal of Fig. 1 needs to be 16 as all posets are reflexive.

Looking only at the pairwise comparisons, there exists no classifier that clearly dominates all other classifiers. At first glance, it seems that CART has the lowest performance.

7.1.2. Discussion on related methods

The ufg depth approach presented in this paper provides a depth measure for all possible partial orders given by the items *BS, CART, EN, GBM, GLM, LASSO, RF,* and *RIDGE*. The depth function is therefore a description of the distribution of all possible posets based on the observed posets. Applying the ufg depth to the posets obtained from the data sets provided by the UCI repository, we get that the maximum depth value of the observed and all possible posets is 0.32. The poset corresponding to the maximum depth value is given by the left side of Fig. 2. The minimum depth value is zero since any poset that states that CART dominates GBM or RIDGE has to have a depth value of zero according to Theorem 9.

In order to compare our approach with other existing approaches, we need to select a representative poset. Both the extended Bradley-Terry model, see [17], and the generalized stochastic dominance approach, see [30] provide (partial) performance order structures that have the most evidence from their respective point of view. For comparison, we choose the poset that corresponds to the highest depth value, as it contains the structure that is most supported by the observed posets. In our case, this is the poset given by Fig. 2 (left).

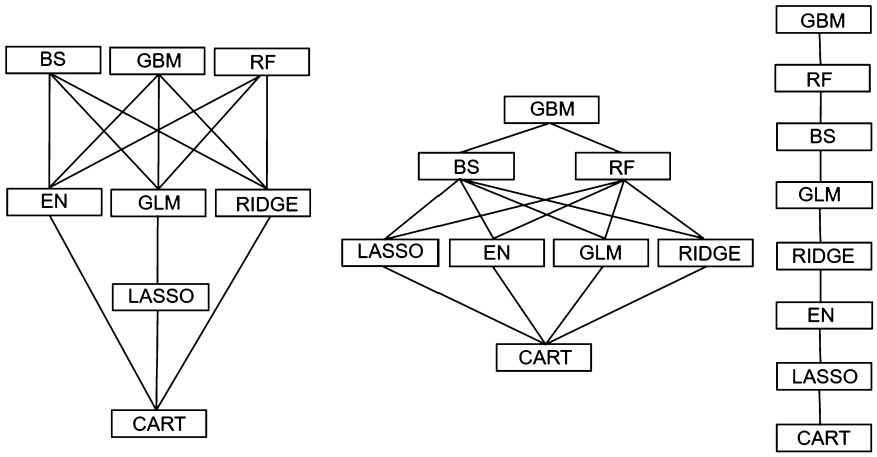

Fig. 2. UCI: Poset having maximal depth value based on all possible posets (left), the poset given by the generalized stochastic dominance approach (middle), see [30](Figure 5 upper graph), and the ranking given by the extended Bradley-Terry model (right). E.g. for all three posets we have that CART is dominated by all other classifiers.

The original Bradley-Terry model, see [9], was defined for total orders only. It is based on pairwise comparisons and assumes that these pairwise comparisons are independent. There are several approaches to extending the Bradley-Terry model, e.g. [17,47]. We focus on the extension of the Bradley-Terry model developed by [17]. This model was originally developed to analyze pairwise comparison data, where the participants can indicate a preference as well as have no preference. Sinclair [53] showed that this approach can be rewritten as a generalized linear model using the Poisson distribution with log link. Here, ties are considered to occur when there is no preference between two items. This extended Bradley-Terry model provides us with so-called worth parameters and a discrimination parameter that includes the no preferences of all comparisons. In our situation, the worth parameters model the latent performance structure of a classifier compared to another classifier. More precisely, if π_i denotes the worth of classifier i (with $\sum_i \pi_i = 1$) and ν the discrimination parameter, then the model assumes that the probability that classifier i is preferred over classifier j is given by $\pi_i/(\pi_i + \pi_j + \nu\sqrt{\pi_i \pi_j})$. Therefore, the classifier with the highest worth parameter is assumed to dominate all other classifiers in a pairwise comparison. Since the worth parameters are values in $[0, 1]$, it is unlikely that two classifiers will have exactly the same worth, and thus in most cases a total order results.

Note that Davidson developed this extension of the Bradley-Terry method with the aim of including Luce's choice axiom, see [12](Chapter 7). This means that the extended Bradley-Terry model assumes that the choice of one item over another is not influenced by other items. Furthermore, for two fixed classifiers i and j, it is assumed for the estimation of the worth parameters that all observations discussing the preference of these two classifiers are independent of each other. Thus, pairwise comparisons are again assumed to be independent. However, this seems questionable because, for a fixed data set, the pairwise comparisons between the performance of classifiers cannot be guaranteed to be independent. In other words, comparing the performance of classifiers i and j is often related to comparing the performance of classifiers i and k on a fixed data set. This suggests that the extended Bradley-Terry model may not be appropriate in this situation of comparing classifiers. In general, it is questionable whether the pairwise comparisons here can be modeled by a stochastic approach for a fixed data set. In contrast, in the ufg approach, the poset-valued observation and the underlying set cannot be reduced to pairwise comparisons. Instead, the posets are considered as a whole observation. More specifically, it is necessary to reflect that the relationship between the pairwise comparisons defines a poset.

Overall, the objective of the extended Bradley-Terry model is to estimate the true underlying worth parameters and thus obtain a total order of the items. In order to allow a proper comparison with our ufg method, we, therefore, need to select a single poset that represents our ufg depth approach. Since the ufg depth gives us a description of the distribution on all possible posets, we choose the poset corresponding to the highest ufg depth value.

Let us now compare the extended Bradley-Terry model and the ufg method using the data set discussed in the subsection above. The estimation of the extended Bradley-Terry model is based on pairwise comparisons and therefore on the data shown in Fig. 1. The estimated worth parameters are now sorted and we obtain the order shown in Fig. 2 (right). The entire estimated extended Bradley-Terry model can be found in Appendix B. Comparing our approach, see Fig. 2 (left), with the extended Bradley-Terry method, we see that the poset corresponding to the extended Bradley-Terry model is a linear extension of the poset representing the highest ufg depth value. Note that for a poset corresponding to the second highest ufg depth value, the result of the extended Bradley-Terry method is no longer a linear extension. For this poset with the second highest ufg depth, the performance of GBM is below that of RF. This analysis shows that the extended Bradley-Terry method provides a stronger structure than our ufg approach. As most of the observed posets are also not a total order, this seems to support our approach as it does not impose a total order structure. On the other hand, when we need to make a decision on exactly one classifier, the extended Bradley-Terry model gives us a classifier that dominates all, rather than just a selection, even if the evidence is weak.

Finally, we want to use the stochastic dominance approach of [30] to further clarify the difference of our approach to others. The mentioned paper analyzes, similar to the present one, the performance of classifiers with respect to several criteria on several data sets simultaneously. However, both the method and the objective differ fundamentally from those of the present paper: While we

include the entire distribution of the underlying poset-valued random variables in our analysis, the authors in [30] compose their representative ordering on the classifiers using a binary generalized stochastic dominance (GSD) relation.

The rough idea of this GSD-based approach is to first embed the range of the multivariate performance measure in a special type of ordered set, a so-called preference system, which then allows for also formalizing the entire information originating from the cardinal dimensions of the performance measure. A classifier is then judged at least as good as a competitor (similar to classic stochastic dominance) if its expected utility is at least as high with respect to every utility function representing (both the ordinal and the cardinal parts of) the preference system.[15] Opposed to GSD, our methodology is thus particularly applicable when it is not clear how the edges within each observed partial order were obtained. Second, while the objective of the present paper is to investigate descriptively how central and outlying partial orders are over the set of classifiers under investigation (this leads to a description of the distribution on all posets), [30] rather investigates if dominance relations hold between classifiers over a sample/population of data sets. This is particularly noticeable in the fact that – while our analysis takes place at the purely descriptive level – the authors there go further and test the descriptive GSD order obtained edge by edge for statistical significance. Despite the clear differences in content between the methods, it is interesting to work out those aspects that can nevertheless be compared with each other. As mentioned above, our method also offers the possibility of extracting a particularly representative partial order of the classifiers under consideration, namely one of the orders with maximum depth. For the analyzed data, this deepest order is unique and shown on the left side of Fig. 2. If we compare this order with the descriptive GSD order from [30] (middle of Fig. 2), we notice that both Hasse diagrams are very similar at first glimpse. The main difference in this case is that the analysis in [30] identifies a clear best classifier (GBM), while the deepest poset obtained by our method has three undominated elements (GBM, RF, and BS). The stronger structure of the GSD ordering can presumably be explained by the fact that it also exploits the cardinal information encoded in the individual quality metrics, whereas we perform a purely ordinal analysis.

Note that generally, one can not make definite statements. There are extreme situations where one method gives a very dense order as a result whereas the other method gives a very sparse poset: If the single performance measures are highly anti-correlated, then the method presented in this paper might lead to very sparse poset data and thus also to a very sparse order as the deepest data point. In contrast, the method of GSD only relies on the marginal distribution of the single performance vectors of the single classifiers over the data sets and not on the correlation structure between these vectors for different classifiers. Therefore, this method can still lead to very dense orderings as a result. On the other hand, if one has a strongly correlated structure of the performance measures and if additionally, the obtained orders are very similar on the majority of data sets, then the data depth method from here would give a very dense partial order. If at the same time, a small minority of measures is very different (i.e., outlying), then the method of [30] might lead to a very sparse result because it uses the cardinal scale of the performance measures.

7.2. OpenML repository: demonstration of UFG method

Now, we analyze data sets given by the OpenML Repository, see [58], to have a detailed discussion on the performance order of different classifiers. With this, we demonstrate the richness and great variety of descriptive analysis options that are possible using our ufg method.

7.2.1. Data set

To showcase the application of the ufg depth on machine learning algorithms we use openly available data from the OpenML benchmarking suite [58]. OpenML shares data sets and corresponding evaluations of classifiers based on different performance measures.

In our comparison we are interested in the performance of the following supervised learning methods: *Random Forests* (RF, implemented in the R-package `ranger` [61]), *Decision Tree* (CART, implemented via the `rpart` library [55]), *Logistic regression* (LR), *L1-penalized logistic regression* (LASSO, implemented through the `glmnet` library [23]), and *k-nearest neighbors* (KNN, through the `kknn` library [27]). For all methods, the choice of parameter settings depends on the goal of the user who uploaded the experiment and its results. Since different users may have different goals, this analysis does not necessarily extend to general statements about the performance of hyperparameter-tuned versions of the algorithms, where the goal is to increase their performance. The algorithms were chosen as a selection of widely used supervised learning methods that perform reasonably without much tuning, in contrast to methods such as neural networks or boosting, which require considerable tuning to perform well.

From the available data sets for which results for all the above algorithms are available in the OpenML database, we limit our analysis to binary classification data sets with more than 450 and less than 10000 observations, leading us to a total of 80 data sets for comparison. The data sets come from a variety of domains and strongly vary in their class balance as well as their overall difficulty. Included in our multidimensional criteria comparison are the measures *area under the curve*, *F-score*, *predictive accuracy*, and *Brier score*. These performance measures capture different aspects of performance, especially in the case of unbalanced data sets. Fig. 3 (right) shows the computed correlation between the performance measures.

The construction of the poset-valued data is analogous to the procedure described at the beginning of Section 7 and in Section 7.1. Since we consider here 80 data sets this leads to 80 observed posets. When including all four performance measures in the construction of the posets, we obtain that 58 of the 80 posets are unique. The sum-statistics, see Section 5.2, which count for each pair the number of occurrences along the 80 posets, can be seen in Fig. 3 (left). It shows that RF is very often above all other methods. So if

[15] For a more detailed discussion of generalized stochastic dominance, see also [34].

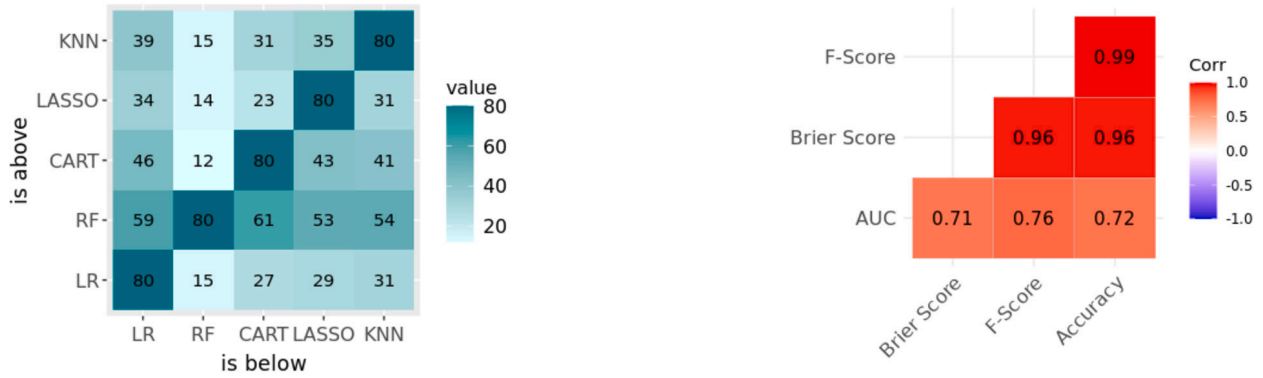

Fig. 3. OpenML: Heatmap representing the sum-statistics, see Section 5, based on all four performance measures (left). Compressed correlation matrix between the calculated performance measures (right). *AUC* is *area under the curve* for short. (For interpretation of the colors in the figures, the reader is referred to the web version of this article.)

one only looks at the sum-statistics RF is clearly the strongest method. The other methods are more balanced with respect to each other. Note that due to reflexivity the diagonal is always 80.

Finally, we want to highlight a structural observation about the order dimension of posets, which applies to all posets ordering five items. The order dimension of a finite poset is defined as the minimum number of linear extensions of this poset whose intersection is equal to this poset. In the case of posets based on five items, the order dimension is always less than or equal to two.[16] This implies that each poset can be constructed by an intersection of at most two total orders. Thus, theoretically, two different performance measures could be sufficient to obtain all possible posets. (This is indeed possible, consider for example two measures where one gives more weight to true positives and the other gives more weight to true negatives. Then any combination of two linear orders can be obtained.) In contrast, for the data sets given by the UCI repository with eight items, two performance measures are not sufficient to obtain every possible poset on eight items. For example, the crown S_4^0, see [56], is a poset on eight items with order dimension four. Therefore, in the case of Section 7.1, at least four performance measures are required. Note that the above observation should not be taken as advice on how to choose the (number of) performance measures. One aspect is that, given a set of performance measures A, it is questionable whether there are two performance measures that represent exactly the same order as that obtained by the set of performance measures A. This issue of reducing the number of performance measures based on the order dimension is discussed in the next section. In any case, the decision on which and how many measures to include in an analysis should always be based on substance considerations.

7.2.2. Analysis using UFG depth

In this subsection, we give a detailed illustration of how the ufg depth gives an insight into the performance structure of classifiers. This is done using the data sets, classifiers, and performance evaluations of the OpenML Repository described in the section above. This analysis consists of two parts: First, we discuss the poset-valued data obtained by using all four performance measures together. Here, we focus on aspects like what the deepest posets have in common as well as considerations on dispersion. In the second part, we tackle the question of how strong the influence of the choice of the performance measures on the resulting poset valued data as well as the depth function is.

Let us start with discussing the analysis based on the poset valued data using all four performance measures for their construction. This gives us a set of 80 posets corresponding to the heatmap in Fig. 3 (left). Here, 58 of the posets are unique. Based on these 80 observed posets we compute the empirical ufg depth. Evaluating this empirical ufg depth over the entire set of possible posets $\mathcal{P}$, we find that each of the 4231 posets has a unique depth value. The most central poset based on all possible posets with maximum depth value is a total order and can be seen in Fig. 4 (left). This poset is also observed and has a depth value of 0.34. Note that the poset with the highest depth value also has the most duplicates, meaning it is the most common pattern given by the posets obtained by the data sets. As described in Section 1.2, we are interested in the distribution of the observed posets. Nevertheless, we can consider the poset with the highest depth value as the poset whose structure is the most common one. Or, in other words, this poset is the one that is most supported based on all observations.

Fig. 5 describes which edges the posets with the $k \in \mathbb{N}$ highest depth values have in common. On the left-hand side, we focus on the observed posets and, the right-hand side is based on all possible posets. Note that while the underlying space on the right-hand side is larger, we include duplicates on the left-hand side. Thus, on the right-hand side, we can see that the deepest poset has seven duplicates in the data set. First, observe that the structure based on restricting to the observed posets or using all possible posets is very similar, i.e. the dominance structure eliminated differs only slightly between all or only the observed posets. For example, one can see that the dominance of RF over all classifiers based on all four performance measures holds for the 35 observed posets with the highest depth values and based on all possible posets for the 67 poset with the highest depth values. In particular, any other

[16] We established this fact by simply computing for each poset on 5 items its order dimension, see Footnote 1 for the code.

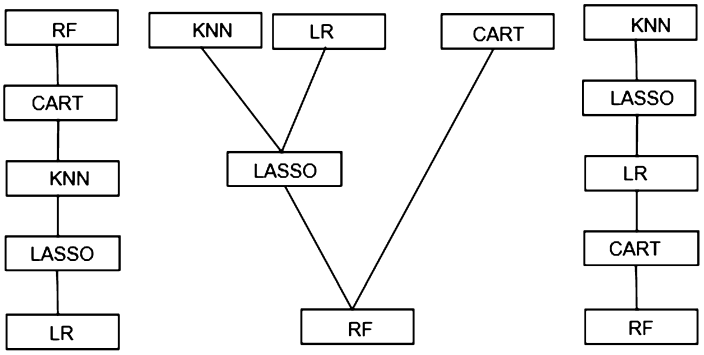

Fig. 4. OpenML based on all four performance measures: Poset with maximal depth based on all possible posets is plotted on the left. The poset with minimal ufg depth restricted to the observed one can be seen in the middle. The poset on the right denotes the poset with minimal depth value based on all possible posets.

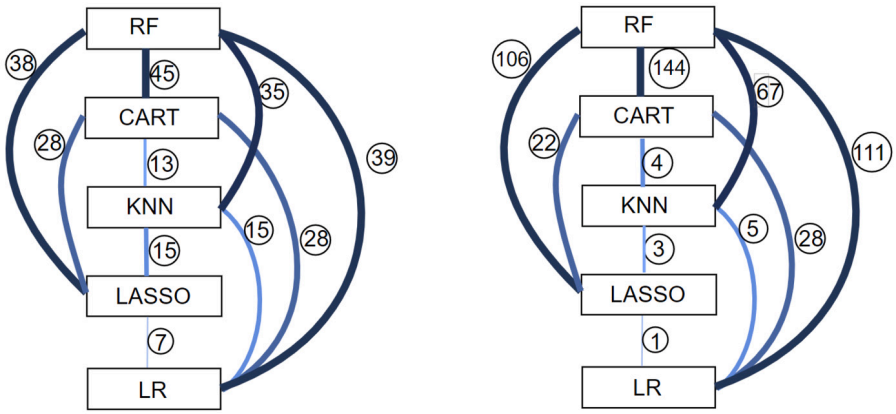

Fig. 5. OpenML based on all four performance measures: Represents what the (observed) posets with the k highest depth values have in common. On the left-hand side, we restrict the analysis to the observed posets and on the right-hand side, we focus on all possible posets. Compare with Fig. 4, where the poset with the highest depth value is plotted. Here each edge number $k \in \mathbb{N}$ indicates that the k deepest posets all contain this relation, but this is not true for the $k+1$ deepest poset.

classifier dominance (like CART outperforms KNN according to all performance measures), does not hold for so many posets with the highest depth values. Note that the observed posets with the highest 46 depth values have nothing more in common. For all possible posets, this is true for the 145 posets with the highest depth value.

Conversely, it is of interest to see what non-edges the posets have in common. Since the poset with the highest depth value is a total order, this is immediately apparent in Fig. 5. The posets with $k \in \mathbb{N}$ highest depth values have those non-edges in common, which are given by the inverse ordered poset of highest depth value intersecting with the inversely already deleted ones. For example, the observed posets with the nine highest depth values have in common that the RF is not dominated by CART, CART not by KNN, and KNN not by LASSO, but they do not agree on LASSO being not dominated by LR since the posets with the observed 8th highest depth value do not agree on this.

Unlike the posets with the highest depth values, the posets with low depth values do not have much in common. The posets (both observed and not observed) only agree on RF being dominated by another classifier. After that, no structure holds. All of these posets can be seen as outliers, or in other words, the corresponding data sets produce a performance structure on the classifiers which differ from the structure given by other data sets. The observed poset with the smallest depth value, which is 0.05, is plotted in the middle of Fig. 4. The poset on the right-hand side of Fig. 4 shows the poset corresponding to the smallest depth value (which is 0.01) based on all possible posets.

Finally, we want to give a notion of dispersion of the depth function. Therefore, we compute the depth function for every poset $p \in \mathcal{P}$ and compute the proportion of posets that lie in $\alpha \in [0,1]$ deepest observed depth values. For $\alpha = 0.25, 0.5$ and 0.75 we get $0.02, 0.10$ and 0.26. Thus, the empirical ufg depth seems to be clustered on small parts of the set $\mathcal{P}$. Note that this impression is a little vague, one computes the proportions of posets compared to the set of all posets. In certain situations, such as for the data sets provided by the UCI repository, not all posts can be obtained because the order dimension is four, but only two performance measures are used. Therefore it may be more natural to compute the proportions in comparison to the number of posets that could be obtained at all. However, as can be seen from the discussion of the order dimension from above (Section 7.2), for the case of the OpenML data set, we are not in such a situation. Finally, it may be still more adequate to compare values for the dispersion only directly between different data sets with the same or similar parameters like the number of items.

The above analysis discusses posets that represent the performance strength of classifiers using all four performance measures. The discussion of the order dimension shows that, in principle, two measures might be sufficient to define all possible posets representing

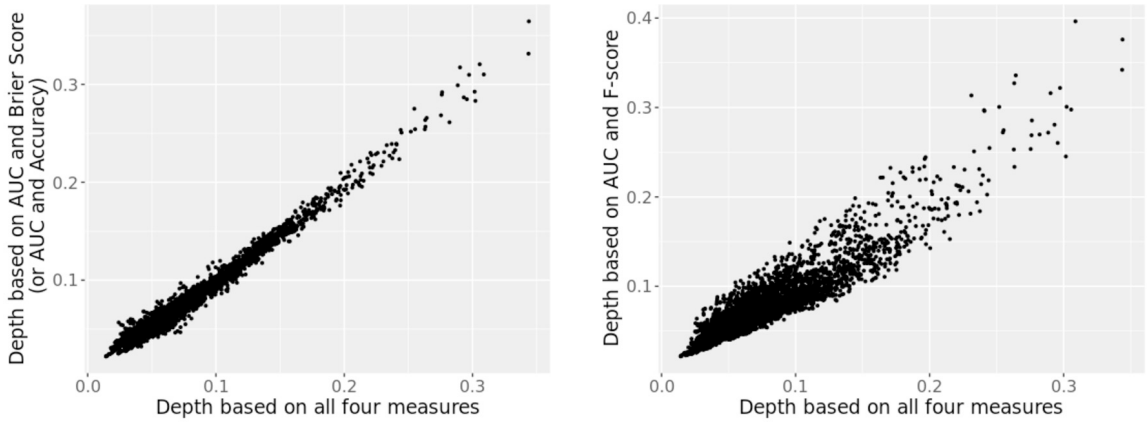

Fig. 6. OpenML: Each point in the scatter plot describes the depth of one poset based on the observed posets given by two different performance sets but based on the same data sets. The scatter plot on the right upper corner compares the depth of the posets based on the observed ones using *area under the curve* and *F score* with the observed posets using all four performance measures.

the performance strength of the five classifiers. However, if the two measures both aim to represent the performance strength of the algorithms, it is questionable whether such two measures exist. For example, in the analysis done above, all four performance measures are highly correlated, see Fig. 3 (right). It is also reasonable that the two measures that give all possible posets are derived from functions that have nothing to do with the performance of classifiers. Nevertheless, we want to discuss a heuristic approach to check whether a subset of size two of the performance measures discussed above is sufficient to obtain the same posets or at least a similar depth function. This heuristic approach is based on the idea that two performance measures with a low correlation tend to measure strength differently. Thus, using two performance measures with low correlation may give us similar posets and depth function structure as using all four performance measures at once. Therefore, we start with two performance measures that have a low correlation. To evaluate this heuristic approach, we also discuss the posets and resulting depth functions obtained by two performance measures with higher correlation.

Based on Fig. 3 (right), *area under the curve* and *Brier score* have a low correlation, as well as *area under the curve* and *predictive accuracy*. It is noteworthy that these two sets of performance measures give exactly the same observed posets. Therefore, the resulting depth functions are also the same. Compared to the posets obtained by using all four performance measures simultaneously, we see that four data sets produce different posets. Fig. 6 (left) compares the depth of all possible posets based on the posets obtained by all four performance measures (x-axis) and based on the posets obtained by the *area under the curve* and *Brier score* measures (y-axis). Each point in the scatter plot represents a poset and the axes denote the depth values. We observe that the depth values for all possible posets seem not to differ strongly between the two construction processes using *area under the curve* and *Brier score* or all four performance measures. If we look at the order of the posets using the two depth functions, we can see that the orders of the posets based on the depth functions are different after all. For a single depth function, we can order the posets according to their respective depth values. Applying this to both depth functions, we find that the maximum rank shift is 1679. More specifically, the poset with the maximal rank shift has the 71st smallest depth value based on all four performance measures, and using the posets obtained by *area under the curve* and *Brier score* only it has the 1750th smallest depth value. In total, 50% of the posets have a rank shift of at least 205 when comparing these two sets of performance measures. Note that this difference in the ranking is within the posets which are not observed. When restricting our analysis to the 76 posets which are observed by both sets of performance measures, then the shift is at most 6 ranks.

Now let us take the next step and look at the posets based on *area under the curve* and *F score*. Recalling Fig. 3 (right), we see that these two measures still have a low correlation of 0.76 compared to all other computed correlations. Nevertheless, there are now 19 data sets that give different poset structures using either *area under the curve* and *F score* or all four performance measures together. This results in a greater difference in the depth values, see Fig. 6 (right). This is also reflected in the ranking of the posets based on the depth values. Now 50% of the posets have a rank difference of at least 305 and the highest rank difference is 2260. Interestingly, if we look at the depth functions given by posets obtained by *F score* and *predicitive accuracy*, we have a smaller shift of the rank structure compared to the depth using all four performance measures. Here the shift of 50% of the posets is at least 226 and at most 1802. This is particularly interesting as *F score* and *predicitive accuracy* have the highest correlation with 0.99, see Fig. 3 (right).

Comparing all possible subsets of size two of the above performance measures, we see that using the two performance measures with the lowest correlation does indeed lead to posets and a depth function that are most similar to those obtained by all four performance measures. However, we have seen that the rank structure of the resulting depth functions is still different from that when all four performance measures are considered simultaneously. Furthermore, using *area under the curve* and *F score*, which also have a comparatively low correlation, leads to different posets and different depth functions. This does not support our heuristic approach. In particular, using the two measures with the highest correlation leads to posets and a depth function that are more similar to those using all four measures than using *area under the curve* and *F score*. Overall, this analysis highlights the point that even if two measures are theoretically sufficient to obtain all possible posets, we cannot simply pick two performance measures

and expect that this will lead to all posets being obtained by more than two performance measures. This supports our idea of using multiple performance measures simultaneously. Especially as the order dimension increases with a growing number of classifiers.

8. Conclusion

In this paper, we have shown how samples of poset-valued random variables can be analyzed (descriptively) by utilizing a generalized concept of data depth. For this purpose, we first introduced an adaptation of the simplicial depth, the so-called ufg depth, and studied some of its properties. Finally, we illustrated our framework with two examples of comparing classifiers using multiple performance measures simultaneously. In the process, we demonstrated how our approach differs from other methods used to analyze poset-valued data and highlighted the various ways in which poset-valued data can be analyzed based on the ufg depth. There are several promising avenues for future research, that include (but are not limited to):

Other ML Problems: Here, we focused on the comparison of classifiers. For example, the performance of different optimization algorithms could be also of interest. In this situation, one can take advantage of the fact that it is possible to tie back to the optimization problems that produced the performance structure. Thus, instead of analyzing the poset with the highest/lowest depth value, one can analyze the optimization problems that produce these posets.

Other Performance Criteria: Instead of using a set of unidimensional performance criteria, the analysis of classifiers with respect to other criteria could be an interesting modification. For example, one could use ROC curves, which can be easily incorporated into our order-based approach.

Inference: A first step towards inference for poset-valued random variables is already made by the consistency property in Section 5. Natural next tie-in points are provided by regression and statistical testing. Together with the results for modeling in [7], a complete statistical analysis framework for poset-valued random variables would then be achieved.

Statistical Uncertainty: Performance measures are only estimates of the true out-of-sample performance. Currently, we are not quantifying this underlying statistical uncertainty. Beyond the general need for uncertainty quantification, different classifiers (despite cross-validation, see e.g. [60,11]) actually differ very much in the dispersion and bias of their performance estimates. This is due to differences in the complexity of the classifiers (mainly because some classifiers have many hyperparameters and others do not). In principle, such heterogeneous statistical uncertainty can be incorporated into our approach by constructing confidence intervals and comparing the resulting intervals with interval orders.

Other Types of Non-Standard Data: Our analysis framework is by no means limited to poset-valued random variables. Since the ufg depth is based on a closure operator, all non-standard data types for which a meaningful closure operator exists can be analyzed with it. As seen in [7] such closure operators are easily obtained by formal concept analysis, thus, there exists a natural generalization of the ufg depth for non-standard data.

Fundings

Hannah Blocher and Georg Schollmeyer are financially and generally supported by the Mentoring Program of the Ludwig-Maximilians-Universität München, Munich. Hannah Blocher's dissertation project is supported by a scholarship of the Evangelisches Studienwerk Villigst e.V. grant 851516.

CRediT authorship contribution statement

Hannah Blocher: Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Formal analysis, Data curation, Conceptualization. **Georg Schollmeyer:** Writing – original draft, Validation, Software, Methodology, Formal analysis, Conceptualization. **Malte Nalenz:** Writing – original draft, Data curation. **Christoph Jansen:** Writing – original draft, Methodology, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

In footnote 1 of the article, we provide the link to the code/data.

Acknowledgements

We sincerely thank all four anonymous reviewers of the ISIPTA conference 2023 in Oviedo, Spain for valuable comments that helped to improve the paper. Moreover, we want to thank all participants at the ISIPTA'23 conference in Oviedo for all very helpful discussions. We are also grateful to the two reviewers for their review of this extended version. Hannah Blocher and Georg Schollmeyer gratefully acknowledge the financial and general support of the Mentoring Program of the Ludwig-Maximilians-Universität München, Munich. Hannah Blocher sincerely thanks Evangelisches Studienwerk Villigst e.V. grant 851516 for funding and supporting.

Appendix A. Computation time and complexity

Here, we state the computation time and complexity of the upper analyzed ufg depth functions.

A.1. UCI repository

We consider 16 different posets $p_1, \dots, p_{16}$, ordering eight elements. By the definition of empirical ufg depth, see Definition 4, we get that we only need to consider those union-free generic sets S which are a subset of $\{p_1, \dots, p_{16}\}$. Applying Theorem 3 Part 1. we obtain that it is sufficient to check all subsets of $\{p_1, \dots, p_{16}\}$ with size greater than two and test if they are union-free and generic. So we checked 65519 subsets and got that 4010 of them are union-free and generic. Testing all $S \subseteq \{p_1, \dots, p_{16}\}$ to see if they are union-free and generic and calculating the depth of the observed 16 posets required about 2 minutes. Since computing all possible posets for eight items is not feasible in a reasonable time, we used the binary linear programming approach described in Section 6. Defining the binary linear program required about 6 seconds. Calculating the highest depth value along with the corresponding poset required about 15 seconds. Calculation of the 20 highest depth values together with the corresponding posets required about 1 minute.

A.2. OpenML repository

Here, we considered in each computation 80 posets which order five items. Hereby, we considered different performance measures and therefore obtained different poset sets. In what follows, we break down the computation time and complexity based on all different observed poset sets.

Part 1: Using all four performance measures

Here we see that 58 of the 80 posets $\{p_1, \dots, p_{80}\}$ are unique (w.l.o.g. the first 58 posets are unique). We can use Theorem 3 to get that for all $S \in \mathscr{S}_{obs}$ we have $2 \leq \#S \leq 8$. Still, checking every single subset $S \in \{p_1, \dots, p_{58}\}$ with size between 2 and 8 is very time consuming, as these are 2262985652 sets (ignoring duplicated posets). Therefore we use the connectedness property, see Theorem 3. With this, computing all union-free generic sets based on the unique posets $\{p_1, \dots, p_{58}\}$ took a total of about 5 hours. We obtained 159382 union-free generic sets. Based on this, together with a weighting according to the duplicates, the computation time of the ufg depth for all possible posets is about 37 minutes.

Part 2: Using area under the curve and Brier score (or area under the curve and predictive accuracy)

Here we see that 57 of the 80 posets $\{p_1, \dots, p_{80}\}$ are unique (w.l.o.g. the first 57 posets are unique). We can use Theorem 3 to get that for all $S \in \mathscr{S}_{obs}$ we have $2 \leq \#S \leq 8$. Still, checking every single subset $S \in \{p_1, \dots, p_{57}\}$ with size between 2 and 8 is very time consuming, as these are 1957698535 sets (ignoring duplicated posets). Therefore we use the connectedness property, see Theorem 3. With this, computing all union-free generic sets based on the unique posets $\{p_1, \dots, p_{57}\}$ took a total of about 3 hours. We obtained 140118 union-free generic sets. Based on this, together with a weighting according to the duplicates, the computation time of the ufg depth for all possible posets is about 9 minutes.

Part 3: Using area under the curve and F-score

Here we see that 52 of the 80 posets $\{p_1, \dots, p_{80}\}$ are unique (w.l.o.g. the first 52 posets are unique). We can use Theorem 3 to get that for all $S \in \mathscr{S}_{obs}$ we have $2 \leq \#S \leq 7$. Still, checking every single subset $S \in \{p_1, \dots, p_{52}\}$ with size between 2 and 7 is very time consuming, as these are 157036191 sets (ignoring duplicated posets). Therefore we use the connectedness property, see Theorem 3. With this, computing all union-free generic sets of $\{p_1, \dots, p_{52}\}$ took a total of about 36 minutes. We obtained 69641 union-free generic sets. Based on this, together with a weighting according to the duplicates, the computation time of the ufg depth for all possible posets is about 5 minutes.

Part 4: Using predictive accuracy and F-score

Here we see that 56 of the 80 posets $(p_1, \dots, p_{80})$ are unique (w.l.o.g. the first 56 posets are unique). We can use Theorem 3 to get that for all $S \in \mathscr{S}_{obs}$ we have $2 \leq \#S \leq 8$. Still, checking every single subset $S \in \{p_1, \dots, p_{56}\}$ with size between 2 and 8 is very time consuming, as these are 1689096277 sets (ignoring duplicated posets). Therefore we use the connectedness property, see Theorem 3. With this, computing all union-free generic sets of $\{p_1, \dots, p_{56}\}$ took a total of about 2 hours. We obtained 120145 union-free generic sets. Based on this, together with a weighting according to the duplicates, the computation time of the ufg depth for all possible posets is about 7 minutes.

Appendix B. Estimated extended Bradley Terry model on the posets derived by the UCI repository

In this section, we briefly present the evaluation of the data sets provided by the UCI repository using the extended Bradley-Terry model, see [9,17]. It is based on the discussion and introduction in Section 7.1.

Sinclair [53] showed that the extended Bradley-Terry model can be rewritten as a generalized linear model. More precisely, for two different classifiers i and j, let $m_{i>j}$ be the number of comparisons that classifier i is preferred over classifier j. The parameters π_i and π_j are the worth parameters of the respective class (note that over all classifiers we assume that $\sum_\ell \pi_\ell = 1$), and ν is the discrimination parameter, which indicates the tendency of an answer to be no preference. The extended Bradley-Terry model is then a generalized linear model with Poisson distribution and log link. The linear predictor is given by

$$\log(m_{i>j}) = \mu_{ij} + \frac{1}{2}\log(\pi_i) - \frac{1}{2}\log(\pi_j), \quad \text{and}$$
$$\log(m_{i\sim j}) = \mu_{ij} + \log(\nu)$$

with $\mu_{ij} = \ln m - \ln\left(\sqrt{\pi_i/\pi_j} + \sqrt{\pi_j/\pi_i}\right)$ where m is the total number of pairwise comparisons.

We applied this to the posets provided by the UCI repository (this code is written in R):

```
> result_Uci_glm <- glm(cum_sum ~ -1 + mu + undecided +
+                       BS + CART + EN + GBM + GLM +
+                       LASSO + RF +RIDGE,
+                       family = 'poisson',
+                       data = design_mat)
                        # loglink is default
> summary(result_Uci_glm)

Call:
glm(formula = cum_sum ~ -1 + mu + undecided + BS + CART +
    EN + GBM + GLM + LASSO + RF + RIDGE,
    family = "poisson", data = design_mat)

Coefficients: (1 not defined because of singularities)
              Estimate Std. Error z value Pr(>|z|)
muBS_CART      0.73676    0.30950   2.380 0.017292 *
muBS_EN        1.41555    0.26296   5.383 7.32e-08 ***
muBS_GBM       1.52960    0.25459   6.008 1.88e-09 ***
muBS_GLM       1.52983    0.25457   6.010 1.86e-09 ***
muBS_LASSO     1.39271    0.26458   5.264 1.41e-07 ***
muBS_RF        1.53411    0.25424   6.034 1.60e-09 ***
muBS_RIDGE     1.44596    0.26078   5.545 2.94e-08 ***
muCART_EN      1.17764    0.28372   4.151 3.32e-05 ***
muCART_GBM     0.63859    0.31463   2.030 0.042392 *
muCART_GLM     0.82902    0.30456   2.722 0.006489 **
muCART_LASSO   1.21233    0.28135   4.309 1.64e-05 ***
muCART_RF      0.72483    0.31013   2.337 0.019432 *
muCART_RIDGE   1.12363    0.28728   3.911 9.18e-05 ***
muEN_GBM       1.36676    0.26644   5.130 2.90e-07 ***
muEN_GLM       1.45551    0.26007   5.597 2.19e-08 ***
muEN_LASSO     1.53309    0.25432   6.028 1.66e-09 ***
muEN_RF        1.40995    0.26336   5.354 8.62e-08 ***
muEN_RIDGE     1.53178    0.25441   6.021 1.74e-09 ***
muGBM_GLM      1.51639    0.25562   5.932 2.99e-09 ***
muGBM_LASSO    1.34024    0.26826   4.996 5.85e-07 ***
muGBM_RF       1.53066    0.25451   6.014 1.81e-09 ***
muGBM_RIDGE    1.40282    0.26392   5.315 1.06e-07 ***
muGLM_LASSO    1.43646    0.26145   5.494 3.93e-08 ***
muGLM_RF       1.52866    0.25466   6.003 1.94e-09 ***
muGLM_RIDGE    1.48011    0.25827   5.731 1.00e-08 ***
muLASSO_RF     1.38664    0.26500   5.233 1.67e-07 ***
muLASSO_RIDGE  1.52747    0.25475   5.996 2.02e-09 ***
muRF_RIDGE     1.44107    0.26114   5.518 3.42e-08 ***
undecided      0.37166    0.10989   3.382 0.000720 ***
BS             0.55687    0.17803   3.128 0.001760 **
CART          -1.24203    0.21482  -5.782 7.40e-09 ***
EN            -0.09093    0.17423  -0.522 0.601721
GBM            0.68258    0.18096   3.772 0.000162 ***
GLM            0.43440    0.17586   2.470 0.013505 *
LASSO         -0.15226    0.17485  -0.871 0.383885
RF             0.57238    0.17835   3.209 0.001331 **
RIDGE              NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
                '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 975.153  on 84  degrees of freedom
Residual deviance:  61.312  on 48  degrees of freedom
```

```
AIC: 404.11

Number of Fisher Scoring iterations: 5
```

Note that RIDGE is the reference level. Thus, the estimate of the worth parameter is the logarithm of zero. The upper result now gives us the estimated probability that classifier i is preferred over classifier j. For example, consider the estimated probability that CART will be outperformed by GBM. Note that this is the dominance structure that exists for the largest number $k \in \{1, \dots, 16\}$ in $\bigcap_{p \in \{p_{(1)}, \dots, p_{(k)}\}} p$, where $p_{(1)}, \dots, p_{(16)}$ are the observed posets, ordered in descending order of their ufg depth value. We have that this preference has a probability of 0.81. The calculation is as follows: The estimated worth parameters are given by $\pi_{\text{GBM}} = \exp(2 \cdot 0.68258) \approx 3.92$ and $\pi_{\text{CART}} = \exp(2 \cdot (-1.24203)) \approx 0.08$. With the estimated discrimination parameter $\nu = \exp(0.37166) \approx 1.45$ we get the estimated probability by $\pi_{\text{GBM}} / (\pi_{\text{GBM}} + \pi_{\text{CART}} + \nu\sqrt{\pi_{\text{GBM}}\pi_{\text{CART}}}) \approx 0.81$. The estimated probability that there is no preference between GBM and CART is $\nu\sqrt{\pi_{\text{GBM}}\pi_{\text{CART}}} / (\pi_{\text{GBM}} + \pi_{\text{CART}} + \nu\sqrt{\pi_{\text{GBM}}\pi_{\text{CART}}}) \approx 0.17$.

References

[1] W. Armstrong, Dependency structures of data base relationships, in: International Federation for Information Processing Congress 74, 1974, pp. 580–583.
[2] R. Baker, P. Scarf, Modifying Bradley–Terry and other ranking models to allow ties, IMA J. Manag. Math. 32 (2021) 451–463.
[3] Y. Bastide, N. Pasquier, R. Taouil, G. Stumme, L. Lakhal, Mining minimal non-redundant association rules using frequent closed itemsets, in: J. Lloyd, V. Dahl, U. Furbach, M. Kerber, K. Lau, C. Palamidessi, L. Pereira, Y. Sagiv, P. Stuckey (Eds.), Computational Logic — CL 2000, Springer, 2000, pp. 972–986.
[4] A. Benavoli, G. Corani, F. Mangili, Should we really use post-hoc tests based on mean-ranks?, J. Mach. Learn. Res. 17 (2016) 152–161.
[5] K. Bertet, C. Demko, J. Viaud, C. Guérin, Lattices, closures systems and implication bases: a survey of structural aspects and algorithms, Theor. Comput. Sci. 743 (2018) 93–109.
[6] H. Blocher, G. Schollmeyer, Data depth functions for non-standard data by use of formal concept analysis, https://www.foundstat.statistik.uni-muenchen.de/personen/mitglieder/blocher/blocheretal_properties23.pdf. (Accessed 21 November 2023), 2023.
[7] H. Blocher, G. Schollmeyer, C. Jansen, Statistical models for partial orders based on data depth and formal concept analysis, in: D. Ciucci, I. Couso, J. Medina, D. Slezak, D. Petturiti, B. Bouchon-Meunier, R. Yager (Eds.), Information Processing and Management of Uncertainty in Knowledge-Based Systems, Springer, 2022, pp. 17–30.
[8] H. Blocher, G. Schollmeyer, C. Jansen, M. Nalenz, Depth functions for partial orders with a descriptive analysis of machine learning algorithms, in: E. Miranda, I. Montes, E. Quaeghebeur, B. Vantaggi (Eds.), Proceedings of the Thirteenth International Symposium on Imprecise Probability: Theories and Applications, in: Proceedings of Machine Learning Research, 2023, pp. 59–71.
[9] R. Bradley, M. Terry, Rank analysis of incomplete block designs: I. The method of paired comparisons, Biometrika 39 (1952) 324–345.
[10] F. Brandenburg, A. Gleißner, A. Hofmeier, Comparing and aggregating partial orders with Kendall tau distances, in: S. Rahman, S. Nakano (Eds.), WALCOM: Algorithms and Computation 2012, in: Lecture Notes in Computer Science, 2012, pp. 88–99.
[11] G. Cawley, N. Talbot, On over-fitting in model selection and subsequent selection bias in performance evaluation, J. Mach. Learn. Res. 11 (2010) 2079–2107.
[12] C. Chambers, F. Echenique, Stochastic choice, in: Revealed Preference Theory, in: Econometric Society Monographs, Cambridge University Press, 2016, pp. 95–113.
[13] C. Chang, J. Jiménez-Martín, E. Maasoumi, T. Pérez-Amaral, A stochastic dominance approach to financial risk management strategies, J. Econom. 187 (2015) 472–485.
[14] L. Chang, Partial order relations for classification comparisons, Can. J. Stat. 48 (2020) 152–166.
[15] I. Couso, D. Dubois, Statistical reasoning with set-valued information: ontic vs. epistemic views, Int. J. Approx. Reason. 55 (2014) 1502–1518.
[16] D. Critchlow, Metric Methods for Analyzing Partially Ranked Data, Lecture Notes in Statistics, vol. 34, Springer, 1985.
[17] R. Davidson, On extending the Bradley-Terry model to accommodate ties in paired comparison experiments, J. Am. Stat. Assoc. 65 (1970) 317–328.
[18] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.
[19] D. Dua, C. Graff, Uci machine learning repository, http://archive.ics.uci.edu/ml. (Accessed 10 September 2023), 2017.
[20] J. Eckhoff, Chapter 2.1 - Helly, Radon, and Carathéodory type theorems, in: P. Gruber, J. Wwillis (Eds.), Handbook of Convex Geometry, North-Holland, Amsterdam, 1993, pp. 389–448.
[21] M. Eugster, T. Hothorn, F. Leisch, Domain-based benchmark experiments: exploratory and inferential analysis, Austrian J. Stat. 41 (2012) 5–26.
[22] M. Fligner, J. Verducci, Distance based ranking models, J. R. Stat. Soc., Ser. B, Methodol. 48 (1986) 359–369.
[23] J. Friedman, T. Hastie, R. Tibshirani, B. Narasimhan, K. Tay, N. Simon, J. Qian, Package glmnet, CRAN R Repository, 2021.
[24] B. Ganter, Two basic algorithms in concept analysis, in: Formal Concept Analysis: 8th International Conference, ICFCA 2010, Agadir, Morocco, March 15-18, 2010. Proceedings 8, Springer, 2010, pp. 312–340.
[25] B. Ganter, R. Wille, Formal Concept Analysis: Mathematical Foundations, Springer, 2012.
[26] M. Goibert, S. Clémençon, E. Irurozki, P. Mozharovskyi, Statistical depth functions for ranking distributions: definitions, statistical learning and applications, arXiv:2201.08105. (Accessed 13 November 2023), 2022.
[27] K. Hechenbichler, K. Schliep, Weighted k-nearest-neighbor techniques and ordinal classification, Technical Report, LMU, 2004, http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-1769-9. (Accessed 28 November 2023).
[28] T. Hothorn, F. Leisch, A. Zeileis, K. Hornik, The design and analysis of benchmark experiments, J. Comput. Graph. Stat. 14 (2005) 675–699.
[29] C. Jansen, H. Blocher, T. Augustin, G. Schollmeyer, Information efficient learning of complexly structured preferences: elicitation procedures and their application to decision making under uncertainty, Int. J. Approx. Reason. 144 (2022) 69–91.
[30] C. Jansen, M. Nalenz, G. Schollmeyer, T. Augustin, Statistical comparisons of classifiers by generalized stochastic dominance, J. Mach. Learn. Res. 24 (2023) 1–37.
[31] C. Jansen, G. Schollmeyer, T. Augustin, Concepts for decision making under severe uncertainty with partial ordinal and partial cardinal preferences, Int. J. Approx. Reason. 98 (2018) 112–131.
[32] C. Jansen, G. Schollmeyer, T. Augustin, A probabilistic evaluation framework for preference aggregation reflecting group homogeneity, Math. Soc. Sci. 96 (2018) 49–62.
[33] C. Jansen, G. Schollmeyer, T. Augustin, Multi-target decision making under conditions of severe uncertainty, in: V. Torra, Y. Narukawa (Eds.), Modeling Decisions for Artificial Intelligence, Springer, 2023, pp. 45–57.
[34] C. Jansen, G. Schollmeyer, H. Blocher, J. Rodemann, T. Augustin, Robust statistical comparison of random variables with locally varying scale of measurement, in: R.J. Evans, I. Shpitser (Eds.), Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence, in: Proceedings of Machine Learning Research, 2023, pp. 941–952.
[35] D. Kikuti, F. Cozman, R. Filho, Sequential decision making with partially ordered preferences, Artif. Intell. 175 (2011) 1346–1365.
[36] G. Lebanon, Y. Mao, Non-parametric modeling of partially ranked data, J. Mach. Learn. Res. 9 (2008) 2401–2429.

[37] H. Levy, A. Levy, Ordering uncertain options under inflation: a note, J. Finance 39 (1984) 1223–1229.
[38] R. Liu, On a notion of data depth based on random simplices, Ann. Stat. 18 (1990) 405–414.
[39] D. Mauá, F. Cozman, D. Conaty, C. Campos, Credal sum-product networks, in: A. Antonucci, G. Corani, I. Couso, S. Destercke (Eds.), Proceedings of the Tenth International Symposium on Imprecise Probability: Theories and Applications, in: Proceedings of Machine Learning Research, 2017, pp. 205–216.
[40] K. Mosler, Multivariate Dispersion, Central Regions, and Depth: The Lift Zonoid Approach, Springer, 2002.
[41] K. Mosler, P. Mozharovskyi, Choosing among notions of multivariate depth statistics, Stat. Sci. 37 (2022) 348–368.
[42] K. Nakamura, K. Yano, F. Komaki, Learning partially ranked data based on graph regularization, arXiv:1902.10963. (Accessed 28 November 2023), 2019.
[43] M. Pini, F. Rossi, K. Venable, T. Walsh, Incompleteness and incomparability in preference aggregation: complexity results, Artif. Intell. 175 (2011) 1272–1289.
[44] R. Plackett, The analysis of permutations, J. R. Stat. Soc., Ser. C, Appl. Stat. 24 (1975) 193–202.
[45] J. Plass, T. Augustin, M. Cattaneo, G. Schollmeyer, Statistical modelling under epistemic data imprecision: some results on estimating multinomial distributions and logistic regression for coarse categorical data, in: T. Augustin, S. Doria, E. Miranda, E. Quaeghebeur (Eds.), Proceedings of the Ninth International Symposium on Imprecise Probability: Theories and Applications, Aracne, 2015, pp. 247–256.
[46] J. Plass, P. Fink, N. Schöning, T. Augustin, Statistical modelling in surveys without neglecting the undecided: multinomial logistic regression models and imprecise classification trees under ontic data imprecision, in: T. Augustin, S. Doria, E. Miranda, E. Quaeghebeur (Eds.), Proceedings of the Ninth International Symposium on Imprecise Probability: Theories and Applications, Aracne, 2015, pp. 257–266.
[47] P.V. Rao, L. Kupper, Ties in paired-comparison experiments: a generalization of the bradley-terry model, J. Am. Stat. Assoc. 62 (1967) 194–204.
[48] G. Schollmeyer, Application of lower quantiles for complete lattices to ranking data: Analyzing outlyingness of preference orderings, Technical Report, LMU, 2017, http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-40452-9. (Accessed 28 November 2023).
[49] G. Schollmeyer, Lower quantiles for complete lattices, Technical Report, LMU, 2017, http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-40448-7. (Accessed 28 November 2023).
[50] G. Schollmeyer, A short note on the equivalence of the ontic and the epistemic view on data imprecision for the case of stochastic dominance for interval-valued data, in: J. De Bock, C. de Campos, G. de Cooman, E. Quaeghebeur, G. Wheeler (Eds.), Proceedings of the Eleventh International Symposium on Imprecise Probabilities: Theories and Applications, in: Proceedings of Machine Learning Research, 2019, pp. 330–337.
[51] G. Schollmeyer, C. Jansen, T. Augustin, Detecting stochastic dominance for poset-valued random variables as an example of linear programming on closure systems, Technical Report, LMU, 2017, http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-40416-0. (Accessed 28 November 2023).
[52] T. Seidenfeld, J. Kadane, M. Schervish, A representation of partially ordered preferences, Ann. Stat. 23 (1995) 2168–2217.
[53] C.D. Sinclair, Glim for preference, in: R. Gilchrist (Ed.), GLIM 82: Proceedings of the International Conference on Generalised Linear Models, Springer, 1982, pp. 164–178.
[54] J. Stoye, Statistical inference for interval identified parameters, in: T. Augustin, F. Coolen, S. Moral, M. Troffaes (Eds.), Proceedings of the Sixth International Symposium on Imprecise Probabilities: Theories and Applications, Aracne, 2009, pp. 395–404.
[55] T. Therneau, B. Atkinson, B. Ripley, Package rpart, http://cran.ma.ic.ac.uk/web/packages/rpart/rpart.pdf. (Accessed 15 February 2023), 2015.
[56] W. Trotter, Dimension of the crown skn, Discrete Math. 8 (1974) 85–103.
[57] J. Tukey, Mathematics and the picturing of data, in: R. James (Ed.), Proceedings of the International Congress of Mathematicians Vancouver, Mathematics-Congresses, Vancouver, 1975, pp. 523–531.
[58] J. Vanschoren, J. van Rijn, B. Bischl, L. Torgo, Openml: networked science in machine learning, SIGKDD Explor. 15 (2013) 49–60.
[59] V. Vapnik, A. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities, in: V. Vovk, H. Papadopoulos, A. Gammerman (Eds.), Measures of Complexity: Festschrift for Alexey Chervonenkis, Springer, 2015, pp. 11–30.
[60] S. Varma, R. Simon, Bias in error estimation when using cross-validation for model selection, BMC Bioinform. 7 (2006) 1–8.
[61] M. Wright, A. Ziegler, ranger: a fast implementation of random forests for high dimensional data in C++ and R, J. Stat. Softw. 77 (2017) 1–17.
[62] M. Zaffalon, The naive credal classifier, J. Stat. Plan. Inference 105 (2002) 5–21.
[63] M. Zaffalon, G. Corani, D. Mauá, Evaluating credal classifiers by utility-discounted predictive accuracy, Int. J. Approx. Reason. 53 (2012) 1282–1301.
[64] Y. Zuo, R. Serfling, General notions of statistical depth function, Ann. Stat. 28 (2000) 461–482.

Contribution 5

Julian Rodemann and Hannah Blocher (2024). "Partial Rankings of Optimizers". In: *The Second Tiny Papers Track at ICLR 2024, Tiny Papers @ ICLR 2024*. Ed. by Tom Burns and Krystal Maughan. Vienna: OpenReview.net

Published as a Tiny Paper at ICLR 2024

PARTIAL RANKINGS OF OPTIMIZERS

Julian Rodemann*, **Hannah Blocher***
Department of Statistics, Ludwig-Maximilians-Universität München, Munich, Germany
{julian.rodemann,hannah.blocher}@stat.uni-muenchen.de

ABSTRACT

We introduce a framework for benchmarking optimizers according to multiple criteria over a collection of test functions. Based on a recently introduced union-free generic depth function for partial orders/rankings, it fully exploits the ordinal information and allows for incomparability. Our method describes the distribution of all partial orders/rankings, avoiding the notorious shortcomings of aggregation. This permits to identify test functions that produce central or outlying rankings of optimizers and to assess the quality of benchmarking suites.[1]

1 INTRODUCTION

Despite its importance for machine learning research, there is no broad agreement on how to compare optimization algorithms on benchmark suites with regard to multiple criteria, see Hansen et al. (2022) for instance. This is particularly relevant for multi-objective optimization, which has diverse applications ranging from reinforcement learning (Basaklar et al., 2023; Zhu et al., 2023) to representation learning (Gu et al., 2023), neural architecture search (Lu et al., 2019) and large language models (Zhou et al., 2023). But such comparisons also arise when single-objective optimizers are evaluated with respect to several metrics, see Sivaprasad et al. (2020); Mattson et al. (2020); Dahl et al. (2023). A popular example is the duality of fixed-budget (performance) and fixed-target (speed) evaluation of deep learning optimizers, see e.g. Dewancker et al. (2016) or results in section 3.

We propose a novel framework for comparing optimizers with respect to multiple criteria over a benchmarking suite of test functions. It is motivated by two observations from benchmarking practice. Firstly, in many cases, ranking optimizers is the overall aim of benchmarking, which then renders the metric information a mere means to an end. Secondly, multiple criteria give rise to incomparability of optimizers ("better with respect to one metric, worse with respect to another one") – a fact that classical aggregation methods aiming at *total* orders fail to represent (Dewancker et al., 2016). In general, it has been proven that it is impossible to aggregate a set of total orders to a single total order while assuming natural conditions on the aggregation, see appendix F. In contrast, our framework is based on *partially* ordering optimizers according to their performance on a single test function, thus deliberately allowing for incomparability. When considering a benchmarking suite of test functions, we obtain a set of such partially ordered sets (posets) describing the performance of optimizers. To evaluate these posets, we use the concept of depth functions which provide a notion of centrality and outlyingness, see Zuo & Serfling (2000) and the adaptation to poset-valued data developed by Blocher et al. (2023) which is called *union-free generic (ufg) depth*. This gives us a description of the distribution and not just an aggregation. Thus, by applying the ufg depth to the posets given by a benchmark suite, we can identify those test functions that give a central/well-supported performance ordering of the optimizers and those test functions that return outliers, see appendices C and D. This paves the way for analyzing the diversity of problems covered by benchmarking suites.

2 METHOD

Depth functions describe an empirical distribution by indicating how central/outlying each point is relative to an entire data cloud or underlying distribution. Our framework relies on the union-free generic (ufg) depth presented in Blocher et al. (2023). This is an adaptation of the simplicial depth

*Authors contributed equally to this work. Total order was enforced by fair coin flip.

[1] **Code:** https://github.com/hannahblo/Posets_Optimizers

function, which denotes the probability that a point $x \in \mathbb{R}^d$ (with $d \in \mathbb{N}$) lies in a randomly drawn $d+1$ simplex (,i.e., in the output of the convex hull/closure operator, see appendix A). For adaptation to posets, let $\mathcal{P}$ be the set of all possible posets on a finite set M. The ufg depth is based on the closure operator $\gamma\colon 2^{\mathcal{P}} \to 2^{\mathcal{P}}, P \mapsto \{p \in \mathcal{P} \mid \cap_{\tilde{p}\in P}\, \tilde{p} \subseteq p \subseteq \cup_{\tilde{p}\in P}\, \tilde{p}\}$. Analogous to the simplicial depth, where only the edges of the $d+1$ simplices are used, the ufg depth considers only those subsets of $2^{\mathcal{P}}$ that are non-trivial, minimal, and not decomposable without loss of information based on the closure operator γ. More specifically, set $\mathscr{S} = \{P \subseteq \mathcal{P} \mid$ condition $(C1)$ and $(C2)$ hold$\}$ with (C1): $P \subsetneq \gamma(P)$, and (C2): there exists no family $(A_i)_{i\in\{1,\dots,\ell\}}$ such that for all $i \in \{1,\dots,\ell\}$ $A_i \subsetneq P$ and $\cup_{i\in\{1,\dots,\ell\}}\gamma(A_i) = \gamma(P)$. Thus, the ufg depth of a poset $p \in \mathcal{P}$ is the proportion of sets $S \in \mathscr{S}$ that contain (i.e. $p \in \gamma(S)$) the poset of interest. More details can be found in appendix B. Consider now the task of comparing d optimizers with respect to c criteria for each of n test functions. Based on a fixed test function, we say that optimizer i is better than optimizer j iff optimizer i is better for at least one criterion and not worse for all others. Thus, if there are two criteria that contradict each other (one criterion states that optimizer i performs better than optimizer j and the other criterion states the opposite), then we say that these two optimizers are incomparable.[2] Hence, when we consider n test functions, we obtain n posets describing the performance of d optimizers based on c criteria. Now, we apply the ufg depth on these posets. Then, the poset with the highest ufg depth value contains the structure that is most supported by the performance order given by all the test functions, while the poset with the lowest ufg depth gives a structure that can be seen as an outlier. Another interpretation is that a test function that produces the poset with a high (low) ufg depth value is a typical (an atypical) problem compared to the other test functions.

3 RESULTS

We illustrate our framework on `DeepOBS`, a benchmark suite for deep learning optimizers (Schneider et al., 2019). We closely mimic the setup in Schneider et al. (2019, section 4) and compare vanilla stochastic gradient descent (SGD), adam, and momentum as baselines on 8 test functions, which arise from training various models on different data sets, e.g., a Long Short-Term Memory network (LSTM) (Hochreiter & Schmidhuber, 1997) for character-level language modeling on Leo Tolstoi's *War and Peace*, see Schneider et al. (2019, Appendix A) for details. We benchmark SGD, adam, and momentum with respect to performance (minimal test loss achieved in a fixed time budget) and speed (time required to achieve a given test loss).[3] We obtain 8 posets describing the order of the optimizers. We observe 5 unique and 3 duplicated posets. From the 8 observed posets, figure 1 shows the poset corresponding to maximal and minimal ufg depth value. Here, an ascending chain of edges between any two optimizers means that the optimizer below is outperformed by the one above. The poset on the left has the highest ufg depth value with 0.65 and therefore the structure which is most supported by the observed benchmarking results. This poset is the duplicated one. In contrast, the poset on the right with the lowest ufg depth value of 0.29 can be seen as outlying. This means that the underlying problem (LSTM on *War and Peace*) produces an order structure that is atypical compared to the other 7 orders given by the test functions. Indeed, it appears surprising for vanilla SGD to outperform adam and momentum, the latter two being enhancements of SGD. We reckon the reason could be that the other 7 test functions in `DeepOBS` are all based on more modern network architectures than LSTM. Besides delivering benchmarking results, our framework also informs the benchmarking suite *designer* about the distribution of orderings produced by her suite. Depending on its overall aim, the designer might use this information to remove or add test functions, giving rise to a more targeted curation of benchmarking suites. For instance, the detection of LSTM on *War and Peace* as an outlier in `DeepOBS` could lead to the removal of the latter in case the designer wants the suite to test optimizers on modern network architectures, and vice versa. See appendix E for more details on how our framework can inform benchmarking suite designers.

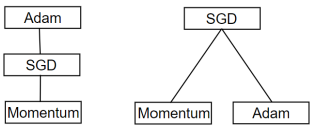

Figure 1: Orderings of optimizers corresponding to highest (0.65, left) and lowest (0.29, right) ufg depth.

[2]Equality of all criteria values does not occur in the following illustration and is discussed in appendix C.

[3]Further results for 11 optimizers on the Black-box Optimization Benchmarking (BBOB) suite (Hansen et al., 2010) comprising 24 test functions as well as for 7 multi-objective evolutionary algorithms (Wu et al., 2023) benchmarked on 13 test functions can be found in appendix C and D, respectively.

4 URM STATEMENT

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2024 Tiny Papers Track.

5 ACKNOWLEDGEMENTS

We sincerely thank Olaf Mersmann for sharing the BBOB results data with us. Another heartfelt thanks goes to Sebastian Fischer and Christoph Jansen for comments on earlier versions of this manuscript. We are also grateful to the three anonymous reviewers for their review of the first version of this manuscript. JR acknowledges support by the Bavarian Academy of Sciences (BAS) through the Bavarian Institute for Digital Transformation (bidt) and by the Federal Statistical Office of Germany within the co-operation project "Machine Learning in Official Statistics". HB sincerely thanks Evangelisches Studienwerk Villigst e.V. for funding and supporting her doctoral studies.

Both authors acknowledge support by the LMU mentoring program.

REFERENCES

Kenneth J Arrow. A difficulty in the concept of social welfare. *Journal of political economy*, 58(4): 328–346, 1950.

Anne Auger and Nikolaus Hansen. Performance evaluation of an advanced local search evolutionary algorithm. In *IEEE Congress on Evolutionary Computation*. IEEE, 2005.

Michael Bacharach. Group decisions in the face of differences of opinion. *Management Science*, 22 (2):182–191, 1975.

Toygun Basaklar, Suat Gumussoy, and Umit Ogras. PD-MORL: Preference-driven multi-objective reinforcement learning algorithm. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Hannah Blocher, Georg Schollmeyer, Christoph Jansen, and Malte Nalenz. Depth functions for partial orders with a descriptive analysis of machine learning algorithms. In *Proceedings of the Thirteenth International Symposium on Imprecise Probability: Theories and Applications (ISIPTA)*, volume 215, pp. 59–71. PMLR, 2023.

Jean Charles de Borda. Mémoire sur les élections au scrutin. *Histoire de l'Académie Royale des Sciences*, 12, 1781.

Marquis de Condorcet. Essai sur l'application de l'analyse a la probabilite des decisions rendues a la pluralite des voix, 1785. Paris.

George E Dahl, Frank Schneider, Zachary Nado, Naman Agarwal, Chandramouli Shama Sastry, Philipp Hennig, Sourabh Medapati, Runa Eschenhagen, Priya Kasimbeg, Daniel Suo, et al. Benchmarking neural network training algorithms. *arXiv preprint arXiv:2306.07179*, 2023.

Ian Dewancker, Michael McCourt, Scott Clark, Patrick Hayes, Alexandra Johnson, and George Ke. A strategy for ranking optimization methods using multiple criteria. In *Workshop on Automatic Machine Learning*, pp. 11–20. PMLR, 2016.

Cynthia Dwork, Ravi Kumar, Moni Naor, and Dandapani Sivakumar. Rank aggregation methods for the web. In *Proceedings of the 10th international conference on World Wide Web*, pp. 613–622, 2001.

Jürgen Eckhoff. Chapter 2.1 - Helly, Radon, and Carathéodory type theorems. In *Handbook of Convex Geometry*, pp. 389–448. North-Holland, Amsterdam, 1993.

Simon French and David Rios Insua. *Statistical Decision Theory: Kendall's Library of Statistics 9*. Wiley, 2010.

Irène Gijbels and Stanislav Nagy. On a general definition of depth for functional data. *Statistical Science*, 32(4):630–639, 2017.

Alex Gu, Songtao Lu, Parikshit Ram, and Lily Weng. Min-max multi-objective bilevel optimization with applications in robust machine learning. In *International Conference on Learning Representations (ICLR)*, 2023.

Nikolaus Hansen, Anne Auger, Raymond Ros, Steffen Finck, and Petr Pošík. Comparing results of 31 algorithms from the black-box optimization benchmarking bbob-2009. In *Proceedings of the 12th annual conference companion on Genetic and evolutionary computation*, pp. 1689–1696, 2010.

Nikolaus Hansen, Anne Auger, Raymond Ros, Olaf Mersmann, Tea Tušar, and Dimo Brockhoff. COCO: A platform for comparing continuous optimizers in a black-box setting. *Optimization Methods and Software*, 36:114–144, 2021. URL `http://numbbo.github.io/coco/`. (accessed: 08.12.2023).

Nikolaus Hansen, Anne Auger, Dimo Brockhoff, and Tea Tušar. Anytime performance assessment in blackbox optimization benchmarking. *IEEE Transactions on Evolutionary Computation*, 26 (6):1293–1305, 2022.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Christoph Jansen, Georg Schollmeyer, and Thomas Augustin. A probabilistic evaluation framework for preference aggregation reflecting group homogeneity. *Mathematical Social Sciences*, (96): 49–62, 2018a.

Christoph Jansen, Georg Schollmeyer, and Thomas Augustin. Concepts for decision making under severe uncertainty with partial ordinal and partial cardinal preferences. *International Journal of Approximate Reasoning*, 98:112–131, 2018b.

Christoph Jansen, Malte Nalenz, Georg Schollmeyer, and Thomas Augustin. Statistical comparisons of classifiers by generalized stochastic dominance. *Journal of Machine Learning Research*, 24 (231):1–37, 2023a.

Christoph Jansen, Georg Schollmeyer, Hannah Blocher, Julian Rodemann, and Thomas Augustin. Robust statistical comparison of random variables with locally varying scale of measurement. In Robin J. Evans and Ilya Shpitser (eds.), *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 941–952. PMLR, 31 Jul–04 Aug 2023b.

John G. Kemeny and J Laurie Snell. Preference ranking: An axiomatic approach. mathematical. *Mathematical Models in Social Science.*, pp. 9–23, 1962.

Daniel Kleitman and Bruce Rothschild. The number of finite topologies. *Proceedings of the American Mathematical Society*, 25(2):276, 1970.

Regina Liu. On a notion of data depth based on random simplices. *The Annals of Statistics*, 18: 405–414, 1990.

Zhichao Lu, Ian Whalen, Vishnu Boddeti, Yashesh Dhebar, Kalyanmoy Deb, Erik Goodman, and Wolfgang Banzhaf. Nsga-net: neural architecture search using multi-objective genetic algorithm. In *Proceedings of the genetic and evolutionary computation conference*, pp. 419–427, 2019.

Peter Mattson, Christine Cheng, Gregory Diamos, Cody Coleman, Paulius Micikevicius, David Patterson, Hanlin Tang, Gu-Yeon Wei, Peter Bailis, Victor Bittorf, David Brooks, Dehao Chen, Debo Dutta, Udit Gupta, Kim Hazelwood, Andy Hock, Xinyuan Huang, Daniel Kang, David Kanter, Naveen Kumar, Jeffery Liao, Deepak Narayanan, Tayo Oguntebi, Gennady Pekhimenko, Lillian Pentecost, Vijay Janapa Reddi, Taylor Robie, Tom St John, Carole-Jean Wu, Lingjie Xu, Cliff Young, and Matei Zaharia. MLPerf training benchmark. In *Proceedings of Machine Learning and Systems*, volume 2, pp. 336–349, 2020.

Olaf Mersmann, Heike Trautmann, Boris Naujoks, and Claus Weihs. Benchmarking evolutionary multiobjective optimization algorithms. In *IEEE Congress on Evolutionary Computation*, pp. 1–8. IEEE, 2010.

Olaf Mersmann, Mike Preuss, Heike Trautmann, Bernd Bischl, and Claus Weihs. Analyzing the bbob results by means of benchmarking concepts. *Evolutionary Computation*, 23(1):161–185, 2015.

Frank Schneider, Lukas Balles, and Philipp Hennig. Deepobs: A deep learning optimizer benchmark suite. In *International Conference on Learning Representations (ICLR)*, 2019. URL `https://deepobs.github.io/`. (accessed: 08.12.2023).

Prabhu Teja Sivaprasad, Florian Mai, Thijs Vogels, Martin Jaggi, and François Fleuret. Optimizer benchmarking needs to account for hyperparameter tuning. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119, pp. 9036–9045. PMLR, 2020.

Fei Wu, Wanliang Wang, Jiacheng Chen, and Zheng Wang. A dynamic multi-objective optimization method based on classification strategies. *Scientific Reports*, 13(1):15221, 2023.

Mihalis Yannakakis. The complexity of the partial order dimension problem. *SIAM Journal on Algebraic Discrete Methods*, 3(3):351–358, 1982.

Zhanhui Zhou, Jie Liu, Chao Yang, Jing Shao, Yu Liu, Xiangyu Yue, Wanli Ouyang, and Yu Qiao. Beyond one-preference-for-all: Multi-objective direct preference optimization. *arXiv preprint arXiv:2310.03708*, 2023.

Baiting Zhu, Meihua Dang, and Aditya Grover. Scaling pareto-efficient decision making via offline multi-objective RL. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Yijun Zuo and Robert Serfling. General notions of statistical depth function. *Annals of statistics*, pp. 461–482, 2000.

A PRELIMINARY DEFINITIONS

In this section, we state all general definitions necessary to this paper.

Partial orders (posets) are based on a fixed set M, which orders the elements of M. A partial order p is therefore a subset of $M \times M$ which is reflexive (for all $y \in M$ we have $(y, y) \in p$), antisymmetric (if $(y_1, y_2) \in p$ with $y_1 \neq y_2$ then $(y_2, y_1) \notin p$) and transitive (if $(y_1, y_2), (y_2, y_3) \in p$ then $(y_1, y_3) \in p$). A poset that is strongly connected (for all $y_1, y_2 \in M$ either $(y_1, y_2) \in p$ or $(y_2, y_1) \in p$ holds) is called a *total/linear order*. Note that a subset of $M \times M$ that is only reflexive and transitive is called a *preorder*.

A *closure operator* on a set Ω is a function $\gamma_\Omega : 2^\Omega \to 2^\Omega$ which is extensive (for $A \subseteq \Omega$ we have $A \subseteq \gamma_\Omega(A)$), idempotent (for $A \subseteq \Omega$ we have $\gamma_\Omega(A) = \gamma_\Omega(\gamma_\Omega(A))$), and increasing (for $A, B \subseteq \Omega$ with $A \subseteq B$ we have $\gamma_\Omega(A) \subseteq \gamma_\Omega(B)$). The *convex hull/closure operator* is defined on $\mathbb{R}^d$ with $d \in \mathbb{N}$. This operator maps each subset $A \subseteq \mathbb{R}^d$ to the smallest convex set that contains A. Note that it defines indeed a closure operator on $\mathbb{R}^d$.

B DEFINITION OF THE UNION-FREE GENERIC DEPTH FUNCTION ON POSETS AND SOME COMPUTATIONAL ASPECTS

In the following, we define the union-free generic (ufg) depth on the set of partial orders (posets), see appendix A and Blocher et al. (2023). Therefore, let M be a finite set and $\mathcal{P}$ the set of all possible posets on M.

Depth functions can be regarded as a generalization of the univariate median and quantiles to multi-dimensional spaces. They provide a natural ordering of data points from center to outwards based on an observed data cloud or underlying distribution. There have been several different depth functions defined, like the simplicial depth on $\mathbb{R}^d$, see Liu (1990), or depths on functional data, see Gijbels & Nagy (2017).

The union-free generic (ufg) depth function developed by Blocher et al. (2023) is based on the simplicial depth on $\mathbb{R}^d$, see Liu (1990). As described in section 2, the empirical simplicial depth uses the $d + 1$ simplices with observed edges and computes the proportion of simplices that contain the point of interest. Thus, it is based on the $d + 1$ simplices, where all edges are observed points $\{x_1, \ldots, x_n\} \in \mathbb{R}^d$ with $n \in \mathbb{N}$. Here, "contain" means that the point of interest is inside the simplex. Hence, it lies in the output of the convex hull/closure operator, see appendix A. In particular, from the perspective of the convex hull operator, the edges of $d + 1$ simplices are special subsets of $\mathbb{R}^d$. More precisely, these subsets are non-trivial in the sense that the output of the convex hull operator is a proper superset. They are minimal in that we cannot delete one point without changing the output. Moreover, they cannot be divided without loss of information. This means that dividing these edges into proper subsets and looking at the union of the output of the convex hull operator applied to these subsets yields a different set than applying it directly to the entire set, see Carathéodorys theorem on convex sets, see Eckhoff (1993). In Blocher et al. (2023) the authors describe this connection in more detail and exploit it in the definition of ufg depth.

Since the ufg depth is an adaptation of the simplicial depth, Blocher et al. (2023) starts by defining a closure operator, see appendix A, on $\mathcal{P}$:

$$\gamma \colon 2^{\mathcal{P}} \to 2^{\mathcal{P}}, \quad P \mapsto \left\{ p \in \mathcal{P} \mid \bigcap_{\tilde{p} \in P} \tilde{p} \subseteq p \subseteq \bigcup_{\tilde{p} \in P} \tilde{p} \right\}.$$

The next step is to adapt the $d + 1$ simplices to this new framework. Thus, the non-triviality and the union-freeness are defined more general by Blocher et al. (2023), and one obtains the following two conditions for $P \in \mathcal{P}$:

(C1) $P \subsetneq \gamma(P)$,

(C2) there does not exist a family $(A_i)_{i \in \{1,\ldots,\ell\}}$ such that for all $i \in \{1, \ldots, \ell\}$ $A_i \subsetneq P$ and $\bigcup_{i \in \{1,\ldots,\ell\}} \gamma(A_i) = \gamma(P)$.

The authors showed that by applying these two conditions to $\mathbb{R}^d$ together with the convex hull operator, one ends up with the $d+1$ simplices. Thus, Blocher et al. (2023) set

$$\mathscr{S} = \{P \subseteq \mathcal{P} \mid \text{ Condition } (C1) \text{ and } (C2) \text{ hold}\}$$

and say that it consists of union-free generic sets.

Example B.1. *We consider a subset of the observed posets given by the benchmarking suite* `DeepOBS` *by Schneider et al. (2019), see section 3. These are*

$$p_1 = \{(\textit{SGD}, \textit{Momentum})\},$$
$$p_2 = \{(\textit{SGD}, \textit{Adam})\}, \text{ and}$$
$$p_3 = \{(\textit{Momentum}, \textit{SGD}), (\textit{Momentum}, \textit{Adam}), (\textit{SGD}, \textit{Adam})\}.$$

Note that p_3 is the poset on the left of the figure 1. Then, since $\gamma(\{p_1, p_2, p_3\}) = \{p_1, p_2, p_3\}$ is trivial and gives no more information, we get $\{p_1, p_2, p_3\} \notin \mathscr{S}$. (Note that it is not trivial when it contains a poset that is not observed.) As $p^* = \{(\textit{SGD}, \textit{Momentum}), (\textit{SGD}, \textit{Adam})\} \in \gamma(\{p_1, p_2\})$ and no proper subset of $\{p_1, p_2\}$ contains the poset p^* in the output of its closure operator, we get that $\{p_2, p_3\} \in \mathscr{S}$.

Now, analogous to simplicial depth, the union-free generic depth function of a partial order p is the weighted proportion of sets $S \in \mathscr{S}$ containing p in its closure. The weight comes from the positive probability of observing the same poset more than once

Definition B.1. *Let $p_1, \ldots, p_n \in \mathcal{P}$ be a sample with corresponding empirical probability measure ν_n (equipped with the power set as σ-field). Then, the (empirical) union-free generic (ufg) depth is given by*

$$D_n \colon \begin{array}{l} \mathcal{P} \to [0,1] \\ p \mapsto \begin{cases} 0, & \text{if for all } S \in \mathscr{S} \colon \prod_{\tilde{p} \in S} \nu_n(\tilde{p}) = 0 \\ c_n \sum_{S \in \mathscr{S}, p \in \gamma(S)} \prod_{\tilde{p} \in S} \nu_n(\tilde{p}), & \text{else} \end{cases} \end{array}$$

with $c_n = \left(\sum_{S \in \mathscr{S}} \prod_{\tilde{p} \in S} \nu_n(\tilde{p}) \right)^{-1}$, see Blocher et al. (2023).

Note that since $\nu_n(p) = 0$ if $p \in \mathcal{P}$ is not observed, we can restrict the set $\mathscr{S}$ to $\mathscr{S}_{obs} = \{S \in \mathscr{S} \mid S \subseteq \{p_1, \ldots, p_n\}\}$ consisting only of the observed posets.

Example B.2. *Recall example B.1, assume that only p_1, p_2 and p_3 were observed (without duplicates) and let ν_3 be the corresponding empirical probability measure. By the remark after the definition of the ufg depth, we get that it is sufficient to consider only all union-free and generic sets that can be obtained by restricting to the observed posets. Thus, we obtain that $\mathscr{S}_{obs} = \{\{p_1, p_2\}, \{p_2, p_3\}, \{p_1, p_3\}\}$. With this, we get that the ufg depth value of all three observed posets is equal to $2/3$.*

In Blocher et al. (2023), the authors also introduce a population version. In addition, they showed some properties that the union-free generic depth function satisfies. This consists of properties such as how duplicates affect the result, what kind of posets have a depth of zero (e.g., when the poset has a preference structure that does not appear in any observed poset), and showing that the ufg depth approach cannot be reduced to a function based only on the pairwise comparisons. Furthermore, the authors prove that the ufg depth is consistent and therefore converges to the population version of the ufg depth under the independent and identically distributed (i.i.d.) assumption.

Finally, we want to discuss some computational aspects. The space of all possible observable posets grows with the number of elements $\#M = n$. Solving the exact number of posets for a set M with $n = \#M$ is NP-hard, a lower bound is given by $2^{n^2/4}$, see Kleitman & Rothschild (1970). Although the ufg depth function is defined over all possible posets, computing the depth of all possible posets can become computationally intensive. This is only needed if we are interested in inference, see appendix E. Another computationally intensive aspect is testing whether a subset satisfies conditions (C1) and (C2). In Blocher et al. (2023) the authors explain in detail how this can be reduced. However, the larger $n = \#M$ is, the longer the computation takes. This can be seen in the runtime of our three examples: `DeepOBS` (approximate runtime: 1 second), Multi-Objective Evolutionary Algorithms (approximate runtime: 10 seconds), and BBOB (approximate runtime: 7 hours). Note that not only is the $\#M$ different, but also the number of test functions is larger in the BBOB suite.

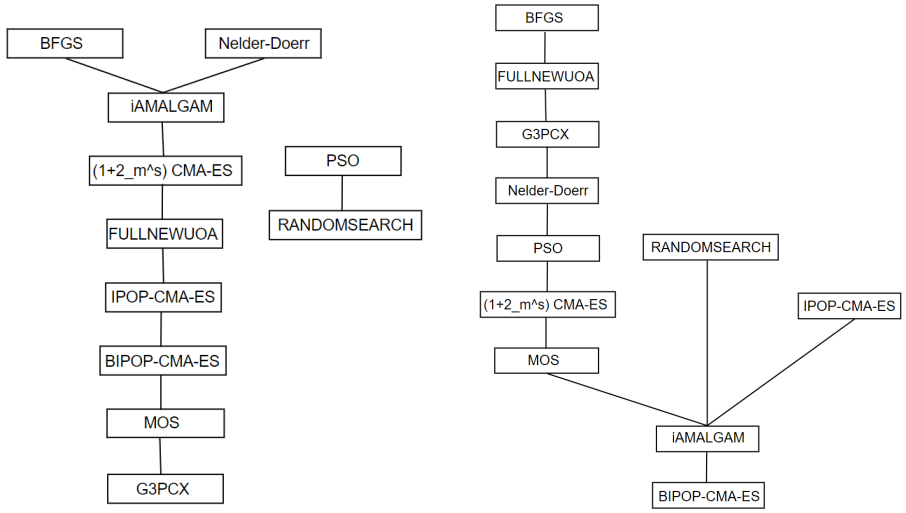

Figure 2: BBOB suite based on dimension 2: Poset corresponding to the maximal (0.21, left) and minimal (0.11, right) ufg depth value.

C Results on BBOB Suite

Many researchers rely on the Black-box Optimization Benchmarking (BBOB) suite of test functions for comparing new optimizers against existing ones. We focus on results and evaluations from the BBOB 2009 and 2010 workshops on 24 noiseless functions in continuous domain in a black-box optimization scenario, see Hansen et al. (2010). The aim of the test function selection was to provide difficult and commonly faced optimization problems. These test functions are scalable with dimensions 2, 3, 5, 10, 20, and 40, see Mersmann et al. (2015). To evaluate the performance we use the calculations of Mersmann et al. (2015). We focus on comparing the 11 algorithms chosen in Mersmann et al. (2015, Section 3.4) as representatives of 11 algorithm groups. These are RANDOMSEARCH, POS, G3PCX, MOS, BIPOP-CMA-ES, IPOP-CMA-ES, FULLNEWUOA, (1+2_mˆs) CMA-ES, iAMALGAM, BFGS and NELDER-DOERR. We also decided to limit our analysis to dimension 2. As performance measures, we use the expected runtime for precision level (distance to optimum) 0.001, see Auger & Hansen (2005), and the number of precision levels achieved overall.

Before discussing the ufg depth based on the posets given by test functions in the BBOB suite, we take a closer look at the construction of the partial orders. For each test function, we say that optimizer i outperforms optimizer j iff at least one criterion states that optimizer i is better and all other criteria say that optimizer i is not worse. If two criteria contradict each other in the sense that one criterion says optimizer i is better and another criterion states the opposite, then these two optimizers are incomparable. More precisely, we say that based on this set of criteria, the optimizers cannot be compared. However, this construction does not cover the possibility that two optimizers are equal in all criteria. In this situation, the optimizers cannot be distinguished on the basis of the criteria used. So they are indifferent. Note that this is actually a distinction from incomparability, where the individual criteria do not agree on the performance order. One can say that in the case of indifference, optimizer i is preferred over optimizer j, and vice versa, optimizer j is preferred over optimizer i. This precisely defines only the weaker structure of a preorder, and not a partial order (antisymmetry is missing, see definitions in appendix A). Thus the first naive approach, to include this indifference also as incomparability to the posets, does not do justice to the different interpretive backgrounds.[4]

The problem of indifferent optimizers arises with the 11 optimizers given by Mersmann et al. (2015) and the noiseless test functions in the BBOB suite with dimension 2. For test functions 1, 18, and 24,

[4]Generalizing the ufg depth to the set of preorders could be future work.

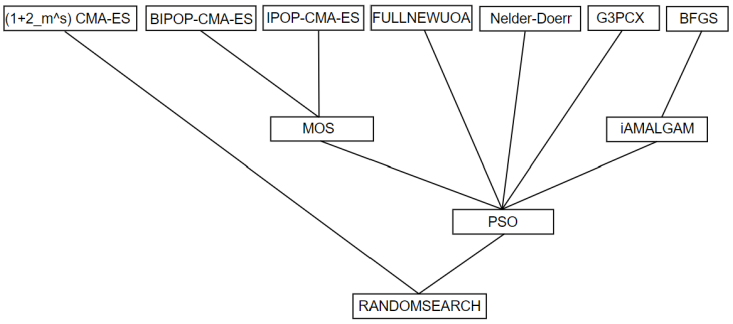

Figure 3: BBOB suite based on all dimensions: Three posets with the highest depth value have this dominance order in common, i.e., RANDOMSEARCH being dominated by all other optimizers is true for the three most central posets.

we have that at least two of the optimizers BFGS, FULLNEWUOA, RANDOMSEARCH, IPOP-CMA-ES, and G3PCX have equal criteria values. Therefore, we restrict our analysis of the ufg depth function to the posets given by the test functions without 1, 18, and 24.[5] Thus, we have 21 test functions resulting in 21 posets describing the performance order of the 11 optimizers where 20 are unique. Applying the ufg depth on these posets, we obtain a maximum depth value of 0.21 and a minimum depth value of 0.11. Since, in general, the ufg depth maps to $[0, 1]$, this indicates that there is no poset that is strongly supported because it lies in many observed union-free generic sets. Thus, each poset contains a dominance or non-dominance structure that is not supported by observed poset subsets. This result may be due to the large variety of different optimization problems reflected by the test functions. Nevertheless, a discussion still can give some insight into the performance order structure of the optimizers and the corresponding test functions.

The maximum ufg depth value is attained for the poset in figure 2 (left). This is a duplicated one and the corresponding test functions are 10 and 11. Note that the outperformance of NELDER-DOERR and BFGS over all other optimizers except (PSO, RANDMOMSEARCH, and FULLNEWUOA) exists within the posets with the 8 highest ufg depth values. The observation of the dominance of NELDER-DOERR for the test functions restricted to 2 dimensions is consistent with Hansen et al. (2010). Besides the poset in figure 2 (left) being the one where the dominance order (edges between the optimizers) is most supported, it is also the poset where the incomparability or non-dominance (non-edges) between two optimizers has the most typical order. Interestingly, we observe that the 9 most central posets all agree that RANDOMSEARCH is not dominated by any optimizer other than PSO. Note that this does not mean that RANDOMSEARCH dominates any other optimizer, only that for all corresponding test functions and for each optimizer there is at least one criterion where RANDOMSEARCH is better than the other optimizer. The poset in figure 2 (right) has the minimum ufg depth value of 0.11. Thus, the corresponding test function 22 produces a performance order of the optimizer that is more outlying compared to all other performance orders produced. However, since all ufg depth values are quite low, it is questionable whether this can really be considered an outlier compared to all others, or whether all posets differ from each other to some extent.

Finally, we want to compare our analysis with Mersmann et al. (2015, p. 176). In that article, the authors analyzed the optimizers based on all six dimensions. Here, the aim was to obtain an overall total order of the optimizers. This was done with Borda consensus ranking, see appendix F. Since the above analysis reflects the performance based only on dimension 2, we further evaluated the ufg depth function and posets, where for each of the test functions we use the expected running time of each dimension as a performance measure. We observe that the three posets corresponding to the three highest ufg depth values differ strongly and only agree on the dominance order in figure 3. The main takeaway is that the three posets that are most central according to ufg depth only agree on random search being dominated by all other optimizers, while optimizer PSO is dominated by all expect random search. Optimizers MOS and iAMALGAM both dominate PSO and random

[5]If we accept the abuse of incomparability and add indifference as well as incomparability and run the analysis on all 24 test functions, we get a similar result.

search, but are incomparable w.r.t each other and are dominated by all remaining optimizers, which are in turn incomparable among each other. Since the dominance order falls apart quickly, it is questionable whether an overall order as suggested in Mersmann et al. (2015, p. 176) is reasonable.

D Results on Multi-Objective Evolutionary Algorithms

We apply our framework on recently published benchmarking results for dynamic multi-objective evolutionary algorithms (Wu et al., 2023). The proposed algorithm by Wu et al. (2023) called DVC is compared against six state-of-the-art multi-objective evolutionary algorithms on 13 test functions with respect to the mean inverted generational distance (MIGD)[6] at four different phases (4 criteria).[7] This results in 13 (unique) posets describing the relation between the optimizers' 4 performance criteria on each of the 13 test functions. Figure 4 illustrates the posets with the highest and lowest depth value, respectively. The two depicted posets correspond to those 2 functions from the 13 considered test functions that entail the most typical relation of the optimizers' performances and the most outlying (atypical) one, respectively.

It becomes evident that the proposed evolutionary algorithm DVC is superior to all but two optimizers (namely, HPPCM and DSSP) on the most typical test function. On the most atypical one, it only dominates PPS and is incomparable to all others. Notably, this very simple illustration of our framework is enough to detect an oversimplified interpretation by Wu et al. (2023, page 10): "From an overall perspective, the DVC algorithm performs well in all phases, although it is slightly inferior to HPPCM in the FDA3 problem and the dMOP3 problem (2 of the 13 test functions, the authors), but performs better in all other test problems". Figure 4 (left graph) reveals that this statement is at least misleading. The DVC algorithm is in fact outperformed with respect to at least one criterion by DSSP. Otherwise, the two would not appear as incomparable in figure 4. A brief glimpse at the tables reporting the criteria values for all test functions and all optimizers in Wu et al. (2023, tables 2-5) confirms this. In fact, it can be seen that DVC is outperformed by several more optimizers with respect to at least one criterion.

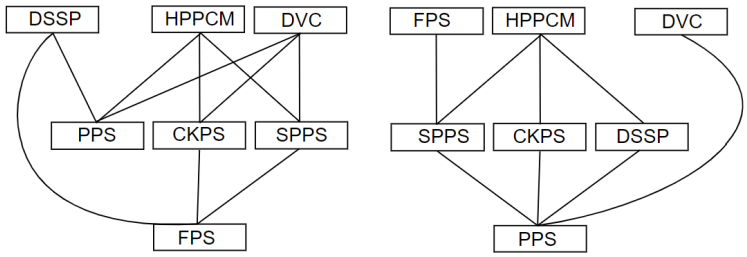

Figure 4: Multi-objective evolutionary algorithms: Orderings of optimizers corresponding to highest (0.39, left) and lowest (0.17, right) ufg depth.

E Outlook: Design and Curation of Benchmarking Suites

When designing and curating benchmarking suites, the question arises quite naturally as to what extent benchmarking results can be compared to each other. It might be of particular interest for benchmarking suite designers to gauge the diversity (dispersion) of the observed partial orderings.

We first note that the answer to this question depends on the definition of a benchmarking suite. While there is broad consensus that the latter entails the specification of test functions and evaluation criteria, it is not clear whether some tested optimizers are an integral part of a benchmarking

[6]The MIGD is based on the minimum sum of distances between solutions belonging to the actual Pareto front and the solutions generated by the algorithm.

[7]Notably, Wu et al. (2023) also benchmark with respect to the mean hypervolume difference at four different stages. For ease of demonstration, we abstain from including these additional criteria. Note that benchmarking according to 8 criteria would increase the number of incomparabilities.

suite, functioning as baselines for future benchmarking. The investigated `BBOB` benchmarking suite (Hansen et al., 2010; 2021), see section 2, does not come – to the best of our knowledge – with a pre-specified set of baseline optimizers. The `deepOBS` benchmarking suite (Schneider et al., 2019), however, comprises three deep learning optimizers as baselines (SGD, momentun, adam).

If there are no specified baselines, our descriptive analysis of the benchmarking results says little if anything about the benchmarking suite *as such*, because they are subject to the actually used optimizers in the benchmark. Moreover, the space of possible observable posets changes with the number of optimizers. More specifically, the cardinality of the possible observable posets grows with lower bound $2^{n^2/4}$ with n the number of items/optimizers, see Kleitman & Rothschild (1970).[8] Moreover, the different cardinality of possibly observable posets strongly affects the *a priori* probability of observing similar posets, e.g., when we have a discrete uniform distribution on posets with 3 optimizers versus on posets with 11 optimizers, it is more likely to observe 3 duplicated posets out of 8 observations for 3 optimizers than for 11 optimizers. However, if the benchmarked optimizers are considered a property of the benchmarking suite, our descriptive analysis is more meaningful: The depth values describe all partial orderings produced by the benchmarking suite. Neither more test functions ("observations") nor more optimizers ("features/items") are observable for a so-defined benchmarking suite. Statistically speaking, this translates to a complete survey, where the sample equals the population. From the perspective of a benchmarking suite designer, the ufg depth function values (and their dispersion, measured e.g., by their range) are thus a sensible tool to assess the diversity of the partial orders produced by the suite. They might also be used to compare their diversity to the results of other benchmarking suites, keeping in mind that they refer to different populations.

However, a benchmarking practitioner typically uses the benchmarking suite to compare a newly proposed optimizer with existing ones. She is thus more interested in *inferential* statements such as "Which test functions produce more likely a typical partial ordering of the old and new optimizers combined, and which produce an atypical order?" Practically speaking, she wants to get rid of test functions without changing the result. Since the new optimizer is unobserved with no prior knowledge, this question cannot be answered (by any ranking method). We can only restrict ourselves to some heuristics about how the distribution on the test functions changes for different optimizers. For example, we could assume that the test functions that produce typical or atypical poset structure on the benchmarking suite behave similar if we add a further optimizer from the same class of optimizers (e.g., evolutionary algorithms) to the benchmarking suite. This can make future experiments more time and energy efficient, while presumably not changing their results.

A second question that concerns a benchmarking practitioner is: "On another test function, what is the most likely poset structure to be observed?" At this stage we cannot answer such questions because our analysis is descriptive only, see also section 3. Future work will focus on statistical inference from posets. However, we want to point out that statistical inference will be difficult because the test functions, and thus the posets, can hardly be seen as independent and identically distributed over the space of all possible test functions/posets.

F Related Work

In this section, we provide some general background on benchmarking optimizers with respect to multiple criteria on a suite of test functions and review related work. Particular attention is paid to the notorious shortcomings of aggregating benchmarking results and how the presented approach avoids them. Before turning to concurring benchmark analysis methods in more detail, we start by a general description of the problem.

Comparing optimizers with respect to a single continuous criterion produces a complete ranking, i.e., a total order (see definition in appendix A). Aggregating several such orders arising from multiple criteria into one unique total order is a long-standing problem in social choice theory dating back to Borda (1781) and Condorcet (1785), still being subject of vivid discussion in economics. In the situation of this paper, the goal is to find a procedure/function that maps every possible set of total orders on any finite set M to a total order. "Arrow's Impossibility Theorem", see Arrow (1950), now

[8] Of course, the number of criteria used can restrict the number of possible observable posets as well, see order dimensions Yannakakis (1982).

states that this is not possible under natural desirable properties. These properties on the procedure, and especially on the resulting aggregated total order, contain the following three axioms among others. For all three axioms, let $M \supsetneq \{a, b\}$ be a finite set with $\{p_1, \ldots, p_n\}$ ($n \geq 2$) being a set of total orders.

1. If every ranking p_i for $i \leq n$ prefers a over b, then the aggregated ranking shall prefer a over b.
2. The aggregated preference of a over b shall not depend on changes of single total orders p_i for $i \leq n$ w.r.t. other items.
3. The aggregation shall take into account all single rankings instead of following one predetermined ranking ("non-dictatorship").

Then "Arrow's impossibility theorem" states that for each procedure there exists a set of total rankings on M such that the resulting aggregated total order does not fulfill all desired properties. For more details, see (French & Insua, 2010, chapter 7). Note that the second axiom implies the absence of cardinal information on the preferences. Cardinal information can be seen as some kind of strength between the comparisons, e.g., the difference of the estimated performance of one single measure between a and b is only 0.0001 instead of 1000. By allowing for such information (which is available in the typical case of continuous criteria) positive results for the aggregation problem are reachable, see Bacharach (1975) for instance. However, one may also argue that due to incommensurability (e.g., different scalings, or variation of test functions' difficulties) of different continuous criteria such a cardinal information is not given. Using only ordinal information, there exist many methods/procedures to obtain one single aggregated total order, e.g., Borda ranking Dwork et al. (2001); Borda (1781) or Kemeny-Snell method, see Kemeny & Snell (1962). Further approaches can be found in Jansen et al. (2018a). Note that these methods all have the problem of "Arrow's impossibility theorem".

Going back to the benchmarking approach with multiple test functions and multiple performance criteria, the question of aggregating total orders arises twice. First, when for each individual test function we have a set of total orders (given by the individual criteria values), where the aggregation needs to represent the performance structure on that test function. Second, when an overall ranking of all performance orders of the test function is desired. Current approaches handle this in different ways.

Despite the restriction of "Arrow's impossibility theorem", established benchmarking schemes usually proceed by aiming at an aggregated overall total order, see (Mersmann et al., 2010, section 5) or (Dewancker et al., 2016, section 2). Mersmann et al. (2010; 2015) take into account the problem of the first aggregation on single test functions by integrating the observed rankings as partial orders. Nevertheless, the overall goal was to find one single total ranking describing the performance of the optimizers, which is achieved by consensus ranking. They critically discuss the downsides of such an aggregation, see (Mersmann et al., 2010, section IV) for instance: "The choice of consensus method is crucial and at the same time highly subjective. There is no right way to choose and therefore no one best algorithm. Our choice of algorithm will always depend on our choice of consensus method." Similarly, the method of Dewancker et al. (2016) results in an overall total order that does justice to the partial order character of each individual test function. There, the authors propose to derive final scores for each optimizer from the individual partial orders through a voting mechanism. In contrast, Jansen et al. (2023a) returns an overall partial order. Here the authors work with so-called preference systems, which are an extension of the partial order to include the cardinal information, see also Jansen et al. (2018b; 2023b). Note that Jansen et al. (2023a) actually present a method to compare classifiers with respect to different quality criteria rather than optimizers, but their approach is principled and can be easily applied to optimizers.

Like Dewancker et al. (2016), we consider partially ordered optimizers arising from multiple criteria for a single test function. The main difference to all the approaches above is that our goal is not to obtain a single partial order that is an aggregation, but rather to describe the entire distribution of possible partial orders representing the performance of the optimizers. This is done by giving each partial order a measure of how central and outlying it is. Most importantly, we can also observe structures that are outlying/atypical and therefore find test functions that produce atypical performance structures. In this way, we take into account the fact that, from our point of view, a single performance structure is not sufficient to describe the overall performance. Nevertheless, the par-

tial order with the highest depth value is the most central partial order and can be seen as the most typical performance structure. With this, one can still compare our result with other benchmarking approaches. Appendix C gives an example of the method comparison.

Contribution 6

Christoph Jansen, Georg Schollmeyer, Hannah Blocher, Julian Rodemann, and Thomas Augustin (2023). "Robust Statistical Comparison of Random Variables with Locally Varying Scale of Measurement". In: *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*. Ed. by Robin Evans and Ilya Shpitser. Pittsburgh: PMLR, 941–952

Robust Statistical Comparison of Random Variables with Locally Varying Scale of Measurement

Christoph Jansen[1] **Georg Schollmeyer**[1] **Hannah Blocher**[1] **Julian Rodemann**[1] **Thomas Augustin**[1]

[1]Department of Statistics, Ludwig-Maximilians-Universität, Munich, Bavaria, Germany

Abstract

Spaces with locally varying scale of measurement, like multidimensional structures with differently scaled dimensions, are pretty common in statistics and machine learning. Nevertheless, it is still understood as an open question how to exploit the entire information encoded in them properly. We address this problem by considering an order based on (sets of) expectations of random variables mapping into such non-standard spaces. This order contains stochastic dominance and expectation order as extreme cases when no, or respectively perfect, cardinal structure is given. We derive a (regularized) statistical test for our proposed generalized stochastic dominance (GSD) order, operationalize it by linear optimization, and robustify it by imprecise probability models. Our findings are illustrated with data from multidimensional poverty measurement, finance, and medicine.

1 INTRODUCTION

Numerous challenges in statistics and machine learning can – at least theoretically – be broken down to comparing random variables $X, Y : \Omega \to A$ mapping between measurable spaces $(\Omega, \mathcal{S}_1)$ and $(A, \mathcal{S}_2)$. Consequently, much attention has been paid to find and apply well-founded *stochastic orderings* enabling such comparison. Examples range from classifier comparisons (e.g., Demsar [2006], Corani et al. [2017], or Blocher et al. [2023]) over ranking risky assets (e.g., Chang et al. [2015]) to deriving optimal (generalized) Neyman-Pearson tests (e.g., [Augustin et al., 2014b, §7.4]).

In the traditional case where the context allows to specify both a probability π on $\mathcal{S}_1$, and a *cardinal* scale $u : A \to \mathbb{R}$ representing the structure on A, a common order $\succsim_{E(u)}$ on $\left\{X \in A^\Omega : u \circ X \in \mathcal{L}^1(\Omega, \mathcal{S}_1, \pi)\right\}$ is obtained by setting $(X, Y) \in \succsim_{E(u)}$ if and only if

$$\mathbb{E}_\pi(u \circ X) = \int_\Omega u \circ X d\pi \geq \int_\Omega u \circ Y d\pi = \mathbb{E}_\pi(u \circ Y). \quad (1)$$

Here, random variables are ranked according to the expectations of their numerical equivalents induced by the scale u. We take the following perspective: This order $\succsim_{E(u)}$ would be the desired order if we were confronted with a problem under pure *aleatoric* uncertainty where an (objective) probability measure π and a cardinal scale u *were* available.[1]

Our paper addresses all situations where, in addition, *epistemic uncertainty* has to be taken into account. Then, such single π and u (and consequently the expectations in (1)) are not available, rendering a comparison by $\succsim_{E(u)}$ impossible. This non-availability corresponds to two facets (e.g. Hüllermeier and Waegeman [2021]) of epistemic uncertainty: Referring to π, *approximation* uncertainty arises since – as common in statistics – only samples of the considered variables are available.[2] Concerning u, on the other hand, *model* uncertainty is assumed to occur from weakly structured order information, making a non-singleton *set* $\mathcal{U}$ of candidate scales compatible with the structure on A.

Naturally, such situations can be approached in two steps: Focusing– in the first step – on model uncertainty, and thus assuming π still to be known, the order $\succsim_{E(u)}$ can be weakened to a *preorder* $\succsim_{(\mathcal{U},\pi)}$ on

$$\left\{X \in A^\Omega : u \circ X \in \mathcal{L}^1(\Omega, \mathcal{S}_1, \pi)\ \forall u \in \mathcal{U}\right\}$$

by setting $(X, Y) \in \succsim_{(\mathcal{U},\pi)}$ if and only if Inequality (1) holds for all candidate scales $u \in \mathcal{U}$. Depending on the concrete choice of the set $\mathcal{U}$, the relation $\succsim_{(\mathcal{U},\pi)}$ has some

[1]The term "aleatoric uncertainty" seems adequate only when π refers to a stochastic phenomenon. However, π might as well represent subjective beliefs which can be formalized by a probability measure such as, e.g., in the Bayesian school of thought.

[2]In Section 6 we go beyond approximation uncertainty and consider robustification by a candidate set of probabilities.

Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023), PMLR 216:941–952.

prominent special cases: If A is equipped with a preorder, and $\mathcal{U}$ is the set of all functions that are bounded and isotone w.r.t. this preorder, then $\succsim_{(\mathcal{U},\pi)}$ is (essentially) equivalent to (first-order) stochastic dominance. In contrast, if $(A, \mathcal{S}_2) = (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ and $\mathcal{U}$ consists of all bounded and *concave* functions, then $\succsim_{(\mathcal{U},\pi)}$ (essentially) corresponds to second-order stochastic dominance.

If – in a second step – information about π comes only from samples from the distributions of X and Y, then, instead of the order $\succsim_{(\mathcal{U},\pi)}$, one has to rely on the corresponding empirical version. Then, a statistical test is needed to control the probability of wrong conclusions from the data.

Motivation of our work: The main goal of the present work is to provide scientists from different fields of application with an inference methodology for the robust analysis of systematic distributional differences within a population. On the one hand, it is important to go beyond a simple comparison of location measures, similar to the case of classical stochastic dominance. On the other hand, we want to take into account the fact that classical (first-order) stochastic dominance systematically ignores potentially available metric information. We achieve this by a generalized stochastic dominance ordering (GSD), which is based on the flexible concept of preference systems. Specifically, we propose a nonparametric permutation test for subgroup comparison that robustifies (therefore further weakening the already parsimonious assumptions) towards the often-criticized assumption of exactly representative sampling.

Our contribution: We consider generalized stochastic dominance (GSD) that ensures exploiting the entire information encoded in data with locally varying scale of measurement. For that purpose, we (primarily) focus, technically speaking, on that specific class of preorders $\succsim_{(\mathcal{U},\pi)}$ where $\mathcal{U}$ is the set of representations of a *preference system* (cf. Sections 2 to 4). Then, using linear optimization, we derive a corresponding (regularized) test (cf. Section 5) and robustify it relying on imprecise probabilities (cf. Section 6). Particularly, our framework allows handling multidimensional structures with differently scaled dimensions in an information-efficient way (cf. Section 7). We illustrate this with data from multidimensional poverty measurement, finance, and medicine (cf. Section 8 and Supp. D) and conclude with a brief discussion (cf. Section 9). The proofs of Propositions 1 to 8, and Corollary 1 can be found in the supplementary material (cf., Supp. A). Our code is available under: `https://github.com/hannahblo/Robust_GSD_Tests`

Related work: Work on tests and/or checking algorithms for stochastic dominance (SD) outside preference systems includes McFadden [1989], Mosler and Scarsini [1991], Mosler [1995], Barrett and Donald [2003], Schollmeyer et al. [2017], Range and Østerdal [2019], Chetverikov et al. [2021]. Optimization under SD constraints was recently considered by, e.g., Dai et al. [2023]. Preference systems and related structures are discussed in a decision theoretic context in Pivato [2013] and Jansen et al. [2018, 2022a]. A test for GSD in the special case of a preference system arising from multiple quality metrics in classifier comparison is discussed in Jansen et al. [2022b].

Neighborhood models that are used to robustify tests are studied in e.g., Destercke et al. [2022], Augustin and Schollmeyer [2021], Montes et al. [2020]. Among others, Maua and de Campos [2021], Cabanas et al. [2020], Maua and Cozman [2020] study credal networks as robustifications of Bayesian networks, and, e.g., Utkin and Konstantinov [2022], Rodemann and Augustin [2022], Carranza and Destercke [2021], Utkin [2020], Abellan et al. [2018] have proposed robustifications and extensions of other machine learning procedures like forests or discriminant analyses by imprecise probabilities.

Accounting for both approximation uncertainty and model uncertainty is in line with recent deliberations in uncertainty quantification (e.g., Malinin and Gales [2018], Hüllermeier and Waegeman [2021], Bengs et al. [2022], Hüllermeier et al. [2022]).

2 BACKGROUND & PRELIMINARIES

We will consider *binary relations* at several points, relying on the following notation and terminology: A binary relation R on a set $M \neq \emptyset$ is a subset of the Cartesian product of M with itself, i.e. $R \subseteq M \times M$. R is called *reflexive*, if $(a, a) \in R$, *transitive*, if $(a, b), (b, c) \in R \Rightarrow (a, c) \in R$, *antisymmetric*, if $(a, b), (b, a) \in R \Rightarrow a = b$, *complete*, if $(a, b) \in R$ or $(b, a) \in R$ (or both) for arbitrary elements $a, b, c \in M$. A *preference relation* is a binary relation that is complete and transitive; a *preorder* is a binary relation that is reflexive and transitive; a *linear order* is a preference relation that is antisymmetric; a *partial order* is a preorder that is antisymmetric. If R is a preorder, we denote by $P_R \subseteq M \times M$ its *strict part* and by $I_R \subseteq M \times M$ its *indifference part*, defined by $(a, b) \in P_R \Leftrightarrow (a, b) \in R \wedge (b, a) \notin R$, and $(a, b) \in I_R \Leftrightarrow (a, b) \in R \wedge (b, a) \in R$.

This leads us to the central ordering structure under consideration in the present paper, namely *preference systems*. These formalize the idea of spaces with locally varying scale of measurement and were introduced in Jansen et al. [2018].[3]

Definition 1 *Let $A \neq \emptyset$ be a set, $R_1 \subseteq A \times A$ a preorder on A, and $R_2 \subseteq R_1 \times R_1$ a preorder on R_1. The triplet $\mathcal{A} = [A, R_1, R_2]$ is then called a **preference system** on A. We call $\mathcal{A}$ **bounded**, if there exist $a_*, a^* \in A$ such that $(a^*, a) \in R_1$, and $(a, a_*) \in R_1$ for all $a \in A$, and $(a^*, a_*) \in P_{R_1}$. Moreover, the preference system $\mathcal{A}' = [A', R_1', R_2']$ is called*

[3] For a study on representation results of the related concept of *incomplete difference preorders* see, e.g., Pivato [2013].

subsystem *of $\mathcal{A}$ if $A' \subseteq A$, $R_1' \subseteq R_1$, and $R_2' \subseteq R_2$. In this case, we call $\mathcal{A}$ a* ***supersystem*** *of $\mathcal{A}'$.*

The concrete definition of a preference system now also makes it possible to concretize the idea of a space with *locally varying scale of measurement*: While the relation R_1 formalizes the available ordinal information, i.e. information about the arrangement of the elements of A, the relation R_2 describes the cardinal part of the information in the sense that pairs standing in relation are ordered with respect to the intensity of the relation. Thus, intuitively speaking, the set A is locally almost cardinally ordered on subsets where R_1 and R_2 are very dense, while on subsets where R_2 is sparse or even empty, locally at most an ordinal scale of measurement can be assumed. A natural example is multi-dimensional structures with differently scaled dimensions, such as those that appear in the poverty analysis application discussed in Section 8: While variables like education can be assumed to have only ordinal scale of measurement, a variable like income is rather metrically scaled.

To ensure that R_1 and R_2 are compatible, we use a consistency criterion for preference systems relying on the idea that both relations should be simultaneously representable.

Definition 2 *The preference system $\mathcal{A} = [A, R_1, R_2]$ is* ***consistent*** *if there exists a* ***representation*** *$u : A \to \mathbb{R}$ such that for all $a, b, c, d \in A$ we have:*

i) If we have that $(a, b) \in R_1$, then it holds that $u(a) \geq u(b)$, where equality holds if and only if $(a, b) \in I_{R_1}$.

ii) If we have that $((a, b), (c, d)) \in R_2$, then it holds that $u(a) - u(b) \geq u(c) - u(d)$, where equality holds if and only if $((a, b), (c, d)) \in I_{R_2}$.

The set of all representations of $\mathcal{A}$ is denoted by $\mathcal{U}_\mathcal{A}$.

Especially when regularizing our test statistic in Section 5, normalized versions of the set $\mathcal{U}_\mathcal{A}$ play a crucial role.

Definition 3 *Let $\mathcal{A} = [A, R_1, R_2]$ be a consistent and bounded preference system with a_*, a^* as before. Then*

$$\mathcal{N}_\mathcal{A} := \left\{ u \in \mathcal{U}_\mathcal{A} : u(a_*) = 0 \ \wedge \ u(a^*) = 1 \right\}$$

is called the ***normalized representation set*** *of $\mathcal{A}$. Further, for $\delta \in [0, 1)$, we denote by $\mathcal{N}_\mathcal{A}^\delta$ the set of all $u \in \mathcal{N}_\mathcal{A}$ with*

$$u(a) - u(b) \geq \delta \quad \wedge \quad u(c) - u(d) - u(e) + u(f) \geq \delta$$

*for all $(a, b) \in P_{R_1}$ and for all $((c, d), (e, f)) \in P_{R_2}$. We call $\mathcal{A}$ δ-**consistent** if $\mathcal{N}_\mathcal{A}^\delta \neq \emptyset$.*

We conclude the section with an immediate observation of the connection between consistency and 0-consistency.

Proposition 1 *Let $\mathcal{A} = [A, R_1, R_2]$ be a bounded preference system. Then $\mathcal{A}$ is consistent if and only if it is 0-consistent.*

3 REGULARIZATION

We now discuss some thoughts on regularization in preference systems. Since our later considerations primarily concern statistical testing, regularization then aims at making the test statistic more sensitive, i.e., to increase discriminative power. In contrast to the usually advocated Thikonov-type regularization, here we think in terms of Ivanov-type regularization that constraints the space of functions (in our case $\mathcal{N}_\mathcal{A}$) over which later our optimization is done (cf., Section 5.1 where our test statistic is introduced as an infimum type test statistic). Beyond the different more or less equivalent ways of representing regularization in a Thikonov-, Ivanov- or in a Morozov-type style (cf. Oneto et al. [2016]), here additionally, two different ways of implementing regularization are conceivable: On the one hand, an *order-theoretic* regularization could be carried out by extending the considered preference system by additional comparable pairs (or pairs of pairs) to a consistent super system. On the other hand, a *parameter-driven* regularization could be performed to reduce the set of representations of the preference system. Both ways are schematically compared in Figure 1.

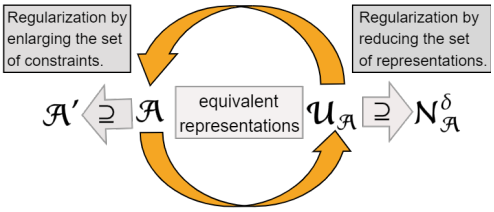

Figure 1: Two ways for regularizing a preference system.

Both approaches have their own strengths and weaknesses: In the case of order-theoretic regularization, the influence of the regularization on the content-related question can be controlled very precisely. However, this comes at the price that the concrete mathematical influence of the regularization can only be characterized with difficulty. The situation tends to be reversed in the case of parameter-driven regularization: Here, it is straightforward – by choosing larger and larger parameter values – to control the mathematical influence of the regularization. However, an interpretation of the regularization in the context of the content-related question is less direct than in the first case. Nevertheless, a possible interpretation in a decision-theoretic context is given in Jansen et al. [2018] by establishing a connection to Luce's *just noticeable differences* [Luce, 1956]. In this paper, we focus on parameter-driven regularization since, for regularization of the test statistic used later, the interpretation of the parameter is of secondary importance.

4 GENERALIZED DOMINANCE

As indicated at the outset, we now turn to a stochastic order between random variables with values in a preference system. This order rigorously generalizes stochastic dominance in the sense that it optimally exploits also the partial cardinal information encoded in these spaces. Therefore, it is neither limited to a purely ordinal analysis as first-order stochastic dominance nor requires perfect cardinal information as second-order stochastic dominance. Consequently, in cases without any cardinal information, i.e., where R_2 is the trivial preorder, the considered order reduces back to the first-order stochastic dominance.

We start by introducing some additional notation: For π a probability measure on $(\Omega, \mathcal{S}_1)$ and $\mathcal{A}$ a consistent preference system, we define by $\mathcal{F}_{(\mathcal{A},\pi)}$ the set

$$\Big\{X \in A^\Omega : u \circ X \in \mathcal{L}^1(\Omega, \mathcal{S}_1, \pi)\ \forall u \in \mathcal{U}_\mathcal{A}\Big\}.$$

We then can define the following preorder on $\mathcal{F}_{(\mathcal{A},\pi)}$.

Definition 4 *Let $\mathcal{A} = [A, R_1, R_2]$ be consistent. For $X, Y \in \mathcal{F}_{(\mathcal{A},\pi)}$, we say Y is $(\mathcal{A}, \pi)$-**dominated** by X if*

$$\mathbb{E}_\pi(u \circ X) \geq \mathbb{E}_\pi(u \circ Y)$$

*for all $u \in \mathcal{U}_\mathcal{A}$. The induced relation is denoted by $R_{(\mathcal{A},\pi)}$ and called **generalized stochastic dominance (GSD)**.*

We have the following immediate simplification if the underlying preference system $\mathcal{A}$ is additionally bounded.

Proposition 2 *If $\mathcal{A}$ is consistent and bounded with a_*, a^* as before, then $(X, Y) \in R_{(\mathcal{A},\pi)}$ iff*

$$\forall u \in \mathcal{N}_\mathcal{A} : \mathbb{E}_\pi(u \circ X) \geq \mathbb{E}_\pi(u \circ Y). \tag{2}$$

5 TESTING FOR DOMINANCE

Throughout this section, let $\mathcal{A} = [A, R_1, R_2]$ be *consistent* and *bounded* with $a_*, a^* \in A$ as in Definition 1.

We now turn to the statistical version of our investigation, where we do not know the underlying probability π but *i.i.d.* samples $\mathbf{X} = (X_1, \ldots, X_n)$ and $\mathbf{Y} = (Y_1, \ldots, Y_m)$ of X and Y are available. The fundamental question now is when we can, with a certain error probability, conclude from this information that $X, Y \in \mathcal{F}_{(\mathcal{A},\pi)}$ are in relation with respect to the GSD-relation $R_{(\mathcal{A},\pi)}$. Constructing a corresponding test, we first need to be clear about appropriate statistical hypotheses. Ideally, we would be interested in the following pair of hypotheses:

$$H_0^{id} : (X, Y) \notin R_{(\mathcal{A},\pi)} \quad \textbf{vs.} \quad H_1^{id} : (X, Y) \in R_{(\mathcal{A},\pi)} \tag{3}$$

In the pair (H_0^{id}, H_1^{id}) of hypotheses – as intended in a statistical test – the question actually of interest would be formulated as the alternative hypothesis. Then, the probability of falsely assuming it to be true could be controlled by the significance level. Unfortunately, similar to the situation of classical stochastic dominance as described, e.g., in Barrett and Donald [2003] and further investigated in Shaked and Shanthikumar [2007], or generally in the context of bioequivalence testing (e.g., Brown et al. [1997]), the hypothesis H_0^{id} seems to be too broad for a meaningful analysis, in the sense that the most conservative scenario under H_0^{id} is not clearly specifiable.[4] For this reason, we choose a pair of alternatives that deviates slightly from the actual question of interest and afterwards try to make the deviation from the actual pair of hypotheses of interest assessable by testing with the variables in reversed roles. The modified pair of hypotheses looks as follows:

$$H_0 : (Y, X) \in R_{(\mathcal{A},\pi)} \quad \textbf{vs.} \quad H_1 : (Y, X) \notin R_{(\mathcal{A},\pi)} \tag{4}$$

The advantage of the pair (H_0, H_1) is that a worst-case analysis of the distribution of a suitable test statistic under H_0 is possible: The test statistic would have to be analyzed under the most conservative case within H_0, namely $\pi_X = \pi_Y$, with π_X and π_Y the image measures of X and Y under π. The drawback to the pair (H_0, H_1) is that in the case of rejection of H_0 we can only control the erroneous conclusion on $(Y, X) \notin R_{(\mathcal{A},\pi)}$ (and not the one actually of interest on $(X, Y) \in R_{(\mathcal{A},\pi)}$) in its probability by the significance level. To mitigate this effect, we can test with the pair (H_0, H_1) of hypotheses additionally with X and Y in reversed roles.

5.1 THE CHOICE OF THE TEST STATISTIC

For defining an adequate test statistic, we first note that – due to the boundedness of $\mathcal{A}$ and Proposition 2 – it holds $(X, Y) \in R_{(\mathcal{A},\pi)}$ if and only if

$$D(X, Y) := \inf_{u \in \mathcal{N}_\mathcal{A}} (\mathbb{E}_\pi(u \circ X) - \mathbb{E}_\pi(u \circ Y)) \geq 0, \tag{5}$$

i.e., if the infimal expectation difference with respect to the available information is at least zero. Thus, a straightforward test statistic is the empirical version of $D(X, Y)$, i.e.,

$$d_{\mathbf{X},\mathbf{Y}} : \Omega \to \mathbb{R}$$

$$\omega \mapsto \inf_{u \in \mathcal{N}_{\mathcal{A}_\omega}} \sum_{z \in (\mathbf{XY})_\omega} u(z) \cdot (\hat{\pi}_X^\omega(\{z\}) - \hat{\pi}_Y^\omega(\{z\}))$$

with, for $\omega \in \Omega$ fixed, $\hat{\pi}_X^\omega(\cdot) := \frac{1}{n}|\{i : X_i(\omega) \in \cdot\}|$ and $\hat{\pi}_Y^\omega(\cdot) := \frac{1}{m}|\{i : Y_i(\omega) \in \cdot\}|$ the observed empirical image measures of X and Y,

$$(\mathbf{XY})_\omega = \{X_i(\omega) : i \leq n\} \cup \{Y_i(\omega) : i \leq m\} \cup \{a_*, a^*\},$$

and $\mathcal{A}_\omega$ the subsystem of $\mathcal{A}$ restricted to $(\mathbf{XY})_\omega$. If $d_{\mathbf{X},\mathbf{Y}}(\omega_0) \geq 0$ holds for some $\omega_0 \in \Omega$, we say there is

[4] The problem is due to the fact that the relation $R_{(\mathcal{A},\pi)}$ is a *partial* order. Compare also [Schollmeyer et al., 2017, p. 24-25].

in-sample GSD of X over Y in the sample induced by ω_0. If the underlying space $\mathcal{A}$ is not too complex[5] (under *i.i.d.* within every subgroup) this test statistic converges to the true value of $D(X,Y)$ and is therefore an adequate test statistic for our test.

As a further test statistic, we consider a regularized version $d^{\varepsilon}_{\mathbf{X},\mathbf{Y}}$ of $d_{\mathbf{X},\mathbf{Y}}$: The infimum in the definition of $d_{\mathbf{X},\mathbf{Y}}$ is now only computed among $[0,1]$-normalized representations of $\mathcal{A}_\omega$ that distinguish between strictly related alternatives over some prespecified threshold value. In this way, the regularized test statistic is also sensitive for distinguishing situations under dominance regarding their *extent* of dominance: While in-sample GSD (essentially) implies $d_{\mathbf{X},\mathbf{Y}}(\omega_0) = 0$, it often holds $d^{\varepsilon}_{\mathbf{X},\mathbf{Y}}(\omega_0) > 0$. Thus, for V, W with $(\mathbf{VW})_{\omega_0} = (\mathbf{XY})_{\omega_0}$ it might be that $d_{\mathbf{V},\mathbf{W}}(\omega_0) = 0$ and $d^{\varepsilon}_{\mathbf{X},\mathbf{Y}}(\omega_0) > d^{\varepsilon}_{\mathbf{V},\mathbf{W}}(\omega_0) > 0$ and, hence, that under regularization X (empirically) dominates Y more strongly than V dominates W.[6]

Formally, the regularized test statistic looks as follows:

$$d^{\varepsilon}_{\mathbf{X},\mathbf{Y}} : \Omega \to \mathbb{R}$$

$$\omega \mapsto \inf_{u \in \mathcal{N}_{\mathcal{A}_\omega}^{\delta_\varepsilon(\omega)}} \sum_{z \in (\mathbf{XY})_\omega} u(z) \cdot (\hat{\pi}^{\omega}_X(\{z\}) - \hat{\pi}^{\omega}_Y(\{z\}))$$

with $\varepsilon \in [0,1]$ and $\delta_\varepsilon(\omega) := \varepsilon \cdot \sup\{\xi : \mathcal{N}^{\xi}_{\mathcal{A}_\omega} \neq \emptyset\}$. Observe that $d_{\mathbf{X},\mathbf{Y}} = d^{0}_{\mathbf{X},\mathbf{Y}}$, i.e., the unregularized test statistic equals the regularized one if $\varepsilon = 0$.

5.2 A PERMUTATION-BASED TEST

As the distribution of $d_{\mathbf{X},\mathbf{Y}}$ and $d^{\varepsilon}_{\mathbf{X},\mathbf{Y}}$ can not be straightforwardly analyzed, we utilize that under the above i.i.d.-assumption a permutation-based test (see, e.g., Pratt and Gibbons [2012]) can be performed. For this, we assume we made observations of the i.i.d. variables, i.e., we observed

$$\mathbf{x} := (x_1, \ldots, x_n) := (X_1(\omega_0), \ldots, X_n(\omega_0)) \quad (6)$$

$$\mathbf{y} := (y_1, \ldots, y_m) := (Y_1(\omega_0), \ldots, Y_m(\omega_0)) \quad (7)$$

for some $\omega_0 \in \Omega$. The resampling scheme for analyzing the distributions of $d_{\mathbf{X},\mathbf{Y}}$ and $d^{\varepsilon}_{\mathbf{X},\mathbf{Y}}$, respectively, can then be described by the following steps:

[5] A concrete sufficient condition for consistency of $d_{\mathbf{X},\mathbf{Y}}$ is a finite VC dimension of the class of all indicator functions of the form $\{a \mid u(a) \geq c\}$ with $u \in \mathcal{N}_{\mathcal{A}}$. This property is usually given, for example if we have finitely many dimensions which have itself a finite VC dimension. Therefore, especially in our applications of Section 8 consistency is guaranteed.

[6] As an example, in the situation of a preference system guaranteeing a totally ordered space (i.e., R_2 is the trivial preorder, R_1 is complete) where the laws of the random variables build a location family $\{f(\cdot + c) \mid c \in \mathbb{R}\}$, the regularized statistic (with appropriately chosen δ) will capture the difference $\Delta = c - \tilde{c}$ between two populations distributed according to $f(\cdot + c)$ and $f(\cdot + \tilde{c})$, respectively, whereas the non-regularized test will not.

Step 1: Take the pooled data sample:

$$\mathbf{w} := (w_1, \ldots, w_{n+m}) := (x_1, \ldots, x_n, y_1, \ldots, y_m)$$

Step 2: Take all $k := \binom{n+m}{n}$ index sets $I \subseteq \{1, \ldots, n+m\}$ of size n. Evaluate $d_{\mathbf{X},\mathbf{Y}}$ resp. $d^{\varepsilon}_{\mathbf{X},\mathbf{Y}}$ for $(w_i)_{i \in I}$ and $(w_i)_{i \in \{1,\ldots,n+m\} \setminus I}$ instead of $\mathbf{x}$ and $\mathbf{y}$. Denote the evaluations by d_I resp. d^{ε}_I.

Step 3: Sort all d_I resp. d^{ε}_I in increasing order to get $d_{(1)}, \ldots, d_{(k)}$ resp. $d^{\varepsilon}_{(1)}, \ldots, d^{\varepsilon}_{(k)}$.

Step 4: Reject H_0 if $d_{\mathbf{X},\mathbf{Y}}(\omega_0)$ resp. $d^{\varepsilon}_{\mathbf{X},\mathbf{Y}}(\omega_0)$ is greater than $d_{(\ell)}$ resp. $d^{\varepsilon}_{(\ell)}$, with $\ell := \lceil (1-\alpha) \cdot k \rceil$ and α the significance level.

Note that, for large $\binom{n+m}{n}$, we can approximate the above resampling scheme by computing d_I resp. d^{ε}_I only for a large number N of randomly drawn I. Moreover, note that only the *i.i.d.* assumption is needed for the above test to be valid. (Precisely, it would already be enough to assume *exchangeable* observations of both variables.)

5.3 COMPUTATION OF $\mathbf{d_{X,Y}}$ AND $\mathbf{d^{\varepsilon}_{X,Y}}$

We show how the test statistics $d_{\mathbf{X},\mathbf{Y}}$ and $d^{\varepsilon}_{\mathbf{X},\mathbf{Y}}$ can be computed in concrete cases. For that, we consider samples $\mathbf{x}$ and $\mathbf{y}$ of the form (6) and (7), and we assume w.l.o.g. that

$$(\mathbf{XY})_{\omega_0} = \{z_1 = a_*, z_2 = a^*, z_3, \ldots, z_s\}$$

Further, we denote by $C(\mathbf{x},\mathbf{y})$ the set of all vectors $(v_1, \ldots, v_s, \xi) \in [0,1]^{s+1}$ such that $v_1 = 0$ and $v_2 = 1$ and for which it holds that

- $v_i = v_j$ if $(z_i, z_j) \in I_{R_1}$,
- $v_i - v_j \geq \xi$ if $(z_i, z_j) \in P_{R_1}$,
- $v_k - v_l = v_r - v_t$ if $((z_k, z_l), (z_r, z_t)) \in I_{R_2}$ and
- $v_k - v_l - v_r + v_t \geq \xi$ if $((z_k, z_l), (z_r, z_t)) \in P_{R_2}$.

Moreover, for $\xi_0 \in [0,1]$ fixed, we define $C_{\xi_0}(\mathbf{x},\mathbf{y})$ as $\{(v_1, \ldots, v_s) \in [0,1]^s : (v_1, \ldots, v_s, \xi_0) \in C(\mathbf{x},\mathbf{y})\}$, i.e., the set of all sample weights that respect the observed preference system and distinguish the strict part of its relations above a threshold of ξ_0. Both $C(\mathbf{x},\mathbf{y})$ and $C_{\xi_0}(\mathbf{x},\mathbf{y})$ are described by finitely many linear inequalities on $(v_1, \ldots, v_s, \xi)$ resp. $(v_1, \ldots, v_s)$. This allows to formulate Propositions 3 and 4. The first one demonstrates how to compute the maximum regularization threshold, whereas the second one captures the computation of $d_{\mathbf{X},\mathbf{Y}}$ and $d^{\varepsilon}_{\mathbf{X},\mathbf{Y}}$.

Proposition 3 *For samples* $\mathbf{x}$ *and* $\mathbf{y}$ *of the form (6) and (7) and* $\varepsilon \in [0,1]$*, we consider the linear program (LP)*

$$\xi \longrightarrow \max_{(v_1,\ldots,v_s,\xi)} \quad (8)$$

with constraints $(v_1, \ldots, v_s, \xi) \in C(\mathbf{x},\mathbf{y})$*. Denote by* ξ^* *its optimal value. It then holds* $\delta_\varepsilon(\omega_0) = \varepsilon \cdot \xi^*$.

Proposition 4 *For samples* $\mathbf{x}$ *and* $\mathbf{y}$ *of the form (6) and (7) and* $\varepsilon \in [0,1]$*, we consider the following LP*

$$\sum_{\ell=1}^{s} v_\ell \cdot \left(\frac{|\{i:x_i=z_\ell\}|}{n} - \frac{|\{i:y_i=z_\ell\}|}{m} \right) \longrightarrow \min_{(v_1,\ldots,v_s)} \tag{9}$$

with $(v_1,\ldots,v_s) \in C_{\varepsilon\xi^*}(\mathbf{x},\mathbf{y})$*, where* ξ^* *is the optimal value of (8). Denote by* $opt_\varepsilon(\mathbf{x},\mathbf{y})$ *its optimal value. Then:*

i) $opt_\varepsilon(\mathbf{x},\mathbf{y}) = d^\varepsilon_{\mathbf{X},\mathbf{Y}}(\omega_0)$.

ii) It holds in-sample GSD of X *over* Y *iff* $opt_0(\mathbf{x},\mathbf{y}) \geq 0$.

6 ROBUSTIFIED TESTING USING IP

Our test for GSD relies on i.i.d. samples of the populations of actual interest. It thus can be based directly on the observed empirical distributions. We now show how *imprecise probabilities (IP)* and *credal sets* (e.g., Walley [1991], Augustin et al. [2014a]) can be used to robustify our test towards deviations of its assumptions. Credal sets – and generally imprecise probabilities – form a consequent generalization of classical probability theory, which also accounts for partial probabilistic knowledge. Indeed, there are various reasons why the i.i.d. assumption can be violated, ranging from unobserved heterogeneity to dependencies arising from data collection. The latter reason is particularly prevalent in surveys, where the survey mode (e.g., phone, web, in-person) often results in unequal, and even outcome-dependent, chances of the units to be sampled. Although methods exist to tackle this, such as reweighting schemes or random routing, most of them come with flaws of their own kind. For example, Bauer [2014, 2016] shows that random routing may be substantially biased, leading to informatively distorted selection probabilities, hence non i.i.d. data.

6.1 THE ROBUSTIFIED TESTING FRAMEWORK

The rough idea of our robustification is to not analyze the test statistic based on $\hat{\pi}_X$ and $\hat{\pi}_Y$ alone, but use neighbourhood models or, more generally, *credal sets* $\mathcal{M}_X \ni \hat{\pi}_X$ and $\mathcal{M}_Y \ni \hat{\pi}_Y$ of candidate probability measures instead. Credal sets – introduced in Levi [1974] – model partial probabilistic information by the set of all non-contradictory probabilities and have gained popularity in machine learning (e.g., Corani and Zaffalon [2008], Lienen and Hüllermeier [2021], Shaker and Hüllermeier [2021], Jansen et al. [2022c], Rodemann et al. [2023], see also the corresponding literature referenced as related work in Section 1).

The concrete idea behind our robustification is that we allow our samples to be (potentially) biased. We assume that these biased samples are similar to the true ones in the sense that the associated true empirical laws are contained in the credal sets $\mathcal{M}_X$ and $\mathcal{M}_Y$ around the biased empirical laws, respectively. We start by only assuming both $\mathcal{M}_X$ and $\mathcal{M}_Y$ to be (random) convex polyhedra with extreme points collected in the finite sets $\mathcal{E}(\mathcal{M}_X)$ and $\mathcal{E}(\mathcal{M}_Y)$.

Now, we again want to test H_0 from Eq. (4), however, under the difficulty that the samples are biased. In the spirit of the concept of *cautious data completion* (see, e.g., [Augustin et al., 2014b, p. 181] or also Schollmeyer [2019] for the connections with stochastic dominance), one actually would adapt the resampling scheme discussed before by performing the test under all pairs of laws in the corresponding credal sets $\mathcal{M}_X$ and $\mathcal{M}_Y$. The null hypothesis H_0 from Eq. (4) would then be rejected whenever it is rejected for all such pairs. Since this adapted resampling scheme is computationally cumbersome, we instead look at the corresponding *lower envelopes* $\underline{d}_{\mathbf{X},\mathbf{Y}} : \Omega \to \mathbb{R}$ and $\underline{d}^\varepsilon_{\mathbf{X},\mathbf{Y}} : \Omega \to \mathbb{R}$, respectively, given by

$$\omega \quad \mapsto \quad \inf_{(\pi_1,\pi_2,u)\in\mathcal{D}} \sum_{z\in(\mathbf{XY})_\omega} u(z)\cdot(\pi_1(\{z\}) - \pi_2(\{z\}))$$

$$\omega \quad \mapsto \quad \inf_{(\pi_1,\pi_2,u)\in\mathcal{D}_\varepsilon} \sum_{z\in(\mathbf{XY})_\omega} u(z)\cdot(\pi_1(\{z\}) - \pi_2(\{z\}))$$

with $\mathcal{D} = \mathcal{M}^\omega_X \times \mathcal{M}^\omega_Y \times \mathcal{N}_{\mathcal{A}_\omega}$, $\mathcal{D}_\varepsilon = \mathcal{M}^\omega_X \times \mathcal{M}^\omega_Y \times \mathcal{N}^{\delta_\varepsilon(\omega)}_{\mathcal{A}_\omega}$ and $\mathcal{M}^\omega_X$ resp. $\mathcal{M}^\omega_Y$ the empirical credal sets given $\omega \in \Omega$. We compare these lower envelopes with the distribution (in the resamples) of the corresponding upper envelopes, $\overline{d}_{\mathbf{X},\mathbf{Y}}$ and $\overline{d}^\varepsilon_{\mathbf{X},\mathbf{Y}}$, that are obtained by replacing the part of $\inf$ concerning $\mathcal{M}^\omega_X \times \mathcal{M}^\omega_Y$ with the respective $\sup$ in the above definitions. This gives a conservative yet valid statistical test.

6.2 COMPUTATION OF $\underline{\mathbf{d}}_{\mathbf{X},\mathbf{Y}}$ AND $\underline{\mathbf{d}}^\varepsilon_{\mathbf{X},\mathbf{Y}}$

We now give an algorithm for the robustified test statistics.

Proposition 5 *For* $\mathbf{x}$ *and* $\mathbf{y}$ *of form (6) and (7),* $\varepsilon \in [0,1]$*, and* $(\pi_1,\pi_2) \in \mathcal{E}(\mathcal{M}^{\omega_0}_X) \times \mathcal{E}(\mathcal{M}^{\omega_0}_Y)$*, consider the LP*

$$\sum_{\ell=1}^{s} v_\ell \cdot (\pi_1(\{z\}) - \pi_2(\{z\})) \longrightarrow \min_{(v_1,\ldots,v_s)} \tag{10}$$

with $(v_1,\ldots,v_s) \in C_{\varepsilon\xi^*}(\mathbf{x},\mathbf{y})$ *and* ξ^* *the optimum of (8). Call* $opt_\varepsilon(\mathbf{x},\mathbf{y},\pi_1,\pi_2)$ *its optimum and* $\underline{opt}_\varepsilon(\mathbf{x},\mathbf{y})$ *the minimal optimum over* $(\pi_1,\pi_2) \in \mathcal{E}(\mathcal{M}^{\omega_0}_X) \times \mathcal{E}(\mathcal{M}^{\omega_0}_Y)$*. Then:*

i) $\underline{opt}_\varepsilon(\mathbf{x},\mathbf{y}) = \underline{d}^\varepsilon_{\mathbf{X},\mathbf{Y}}(\omega_0)$.

ii) There is in-sample GSD of X *over* Y *for any* π *with* $\hat{\pi}^{\omega_0}_X \in \mathcal{M}^{\omega_0}_X$ *and* $\hat{\pi}^{\omega_0}_Y \in \mathcal{M}^{\omega_0}_Y$ *if* $\underline{opt}_0(\mathbf{x},\mathbf{y}) \geq 0$.

Proposition 5 requires to solve $|\mathcal{E}(\mathcal{M}^{\omega_0}_X)| \cdot |\mathcal{E}(\mathcal{M}^{\omega_0}_Y)|$ linear programs. Depending on the concrete neighbourhood models, this is obviously limited: The number of programs is simply too large. A common strategy in such a case is to additionally assume 2-monotonicity of the considered credal sets, since this allows us – at least for R_1 complete

– to give closed formulas for the upper and lower expectations. Unfortunately, this is not so simple in the case of a partially ordered R_1: since the representation via the Choquet integral (e.g., Denneberg [1994]) depends on the order of elements of A, an optimum over all linear extensions of R_1 is needed to determine the most extreme Choquet integrals. In the worst case, this would lead to optimizing a non-convex function and thus hardly simplify the original problem (see Timonin [2012]).

Another strategy is restricting to credal sets with moderately many extreme points. We now consider one such possibility, namely the the γ- *contamination model* (or *linear-vacuous model*, see, e.g., [Walley, 1991, p. 147]). Here, we assume that for $\omega \in \Omega$, $\gamma \in [0,1]$, and $Z \in \{X, Y\}$ fixed, we have

$$\mathcal{M}_Z^\omega = \Big\{\pi : \pi \geq (1-\gamma)\cdot \hat{\pi}_Z^\omega\Big\}, \tag{11}$$

where $\geq$ is understood event-wise. For γ-contamination models there are exactly as many extreme points as there are observed distinct data points, concretely given by

$$\mathcal{E}(\mathcal{M}_Z^\omega) = \Big\{\gamma\delta_z + (1-\gamma)\hat{\pi}_Z^\omega : \exists j \text{ s.t. } Z_j(\omega) = z\Big\}, \tag{12}$$

where δ_z denotes the Dirac-measure in z (see again Walley [1991, p. 147]). Proposition 6 states that if the credal sets are both γ-contamination models, then a least favorable pair of extreme points can a priori be specified. The test statistics thus can be computed by solving one linear program.

Proposition 6 *Consider again the situation of Proposition 5, where additionally $\mathcal{M}_X^{\omega_0}$ and $\mathcal{M}_Y^{\omega_0}$ are of the form (11) with extreme points as in (12). It then holds:*

$$\underline{opt}_\varepsilon(\mathbf{x}, \mathbf{y}) = opt_\varepsilon(\mathbf{x}, \mathbf{y}, \pi_*, \pi^*), \textit{where}$$

$$\pi_* = \gamma\delta_{a_*} + (1-\gamma)\hat{\pi}_X^{\omega_0} \text{ and } \pi^* = \gamma\delta_{a^*} + (1-\gamma)\hat{\pi}_Y^{\omega_0}.$$

7 MULTIDIMENSIONAL SPACES WITH DIFFERENTLY SCALED DIMENSIONS

We now turn to a special case that is very common in applied research: multidimensional spaces whose dimensions may be of different scale of measurement.[7] While traditional empirical research and policy support (e.g., European Commission [2023]) summarizes such situations by indices/indicators that suffer eo ipso from "the subjectivity of choices associated with them" ([UNECE, 2019, p. 11]), the embedding into the framework considered here allows a faithful representation of the entire underlying information.

Concretely, we address $r \in \mathbb{N}$ dimensional spaces for which we assume – w.l.o.g. – that the first $0 \leq z \leq r$ dimensions are of cardinal scale (implying that differences of elements may be interpreted as such), while the remaining ones are purely ordinal (implying differences to be meaningless apart from the sign). Specifically, we consider (bounded subsystems of) the preference system[8]

$$\textit{pref}(\mathbb{R}^r) = [\mathbb{R}^r, R_1^*, R_2^*] \tag{13}$$

where

$$R_1^* = \Big\{(x,y) : x_j \geq y_j \ \forall j \leq r\Big\}, \text{ and}$$

$$R_2^* = \left\{((x,y),(x',y')) : \begin{array}{l} x_j - y_j \geq x'_j - y'_j \ \ \forall j \leq z \\ x_j \geq x'_j \geq y'_j \geq y_j \ \forall j > z \end{array}\right\}.$$

While R_1^* can be interpreted as a simple component-wise dominance relation, R_2^* deserves some more explanation: One pair of consequences is preferred to another one if it is ensured in the ordinal dimensions that the exchange associated with the first pair is not a deterioration to the exchange associated with the second pair and, in addition, there is component-wise dominance of the differences of the cardinal dimensions. The following proposition lists some important results for a more precise characterization of the GSD-relation on multidimensional structures.

Proposition 7 *Let π be a probability measure on $(\Omega, \mathcal{S}_1)$, and $X = (\Delta_1, \ldots, \Delta_r), Y = (\Lambda_1, \ldots, \Lambda_r) \in \mathcal{F}_{(\textit{pref}(\mathbb{R}^r),\pi)}$. Then, the following holds:*

i) $\textit{pref}(\mathbb{R}^r)$ is consistent.

ii) If $z = 0$, then $R_{(\textit{pref}(\mathbb{R}^r),\pi)}$ equals (first-order) stochastic dominance w.r.t. π and R_1^* (short: FSD(R_1^*, π)).

iii) If $(X,Y) \in R_{(\textit{pref}(\mathbb{R}^r),\pi)}$ and $\Delta_j, \Lambda_j \in \mathcal{L}^1(\Omega, \mathcal{S}_1, \pi)$ for all $j = 1, \ldots, r$, then

I. $\mathbb{E}_\pi(\Delta_j) \geq \mathbb{E}_\pi(\Lambda_j)$ for all $j = 1, \ldots, r$, and

II. $(\Delta_j, \Lambda_j) \in$FSD$(\geq, \pi)$ for all $j = z+1, \ldots, r$.

Additionally, if all components of X are jointly independent and all components of Y are jointly independent, properties I. and II. imply $(X,Y) \in R_{(\textit{pref}(\mathbb{R}^r),\pi)}$.

Part iii) of Proposition 7 is complete in the sense that the addition actually holds only under stochastic independence.

Remark 1 *The addition to iii) does not generally hold. A counterexample is $z = 1$, $r = 2$, $\Omega = \{\omega_1, \ldots, \omega_4\}$, and π the uniform distribution over Ω. Then, for $\Delta_1(\omega) = 1,1,2,2$, $\Delta_2(\omega) = 1,1,2,2$, $\Lambda_1(\omega) = 1,1,2,2$, and $\Lambda_2(\omega) = 1,2,1,2$ for $\omega = \omega_1, \ldots, \omega_4$, it holds that $\mathbb{E}_\pi(\Delta_1) = \mathbb{E}_\pi(\Lambda_1)$. In fact, the first components are equivalent with respect to first order stochastic dominance. The same holds for the second components. However, the whole*

[7] For recent applications of such special preference systems to classifier comparison or multi-target decision making see Jansen et al. [2022b], Jansen and Augustin [2022] and Jansen et al. [2023].

[8] One easily verifies that R_1^* and R_2^* are preorders.

vectors are incomparable with respect to first order stochastic dominance, since there is no corresponding mass transport from higher values to lower (or equal) values possible. Additionally, for $u(x,y) := x \cdot y$, we have that $u \in \mathcal{U}_{\mathsf{pref}(\mathbb{R}^r)}$, $\mathbb{E}_\pi(u \circ \Delta) = 10/4$, whereas $\mathbb{E}_\pi(u \circ \Lambda) = 9/4$. Thus, Δ and Λ can not be equivalent with respect to GSD.

As an immediate consequence of Proposition 7, we have the following corollary for bounded subsystems of $\mathsf{pref}(\mathbb{R}^r)$.

Corollary 1 *If $\mathcal{C} = [C, R_1^c, R_2^c]$ is a bounded subsystem of $\mathsf{pref}(\mathbb{R}^r)$ and $X, Y \in \mathcal{F}_{(\mathcal{C},\pi)}$, then $\mathcal{C}$ is 0-consistent and ii) and iii) from Prop. 7 hold, if we replace $R_{(\mathsf{pref}(\mathbb{R}^r),\pi)}$ by $R_{(\mathcal{C},\pi)}$, FSD(R_1^*, π) by FSD(R_1^c, π), and $(X,Y) \in R_{(\mathsf{pref}(\mathbb{R}^r),\pi)}$ by $\forall u \in \mathcal{N}_\mathcal{C} : \mathbb{E}_\pi(u \circ X) \geq \mathbb{E}_\pi(u \circ Y)$.*

Finally, we give a characterization of the set of all representations of $\mathsf{pref}(\mathbb{R}^r)$ if only one dimension is cardinal.

Proposition 8 *Let $z = 1$ and denote by $\mathcal{U}_{sep}$ the set of all $u : \mathbb{R}^r \to \mathbb{R}$ such that, for $(x_2, \ldots, x_r) \in \mathbb{R}^{r-1}$ fixed, the function $u(\cdot, x_2, \ldots, x_r)$ is strictly increasing and (affine) linear and such that, for $x_1 \in \mathbb{R}$ fixed, the function $u(x_1, \cdot, \ldots, \cdot)$ is strictly isotone w.r.t. the the componentwise partial order on $\mathbb{R}^{r-1}$. Then $\mathcal{U}_{sep} = \mathcal{U}_{\mathsf{pref}(\mathbb{R}^r)}$.*

8 APPLICATIONS

We now apply our framework on three examples: dermatological symptoms, credit approval data, and multidimensional poverty measurement. Results from the former two applications are presented in Supp. D, while Section 8.2 discusses results from poverty analysis. Before that, some details on the concrete implementation are given.

8.1 IMPLEMENTATION

To compute the test statistics for sample size s, we use a LP with constraints given by $C(x,y)$ (Section 5.3). The computation of the test statistics and the maximum regularization strength ξ^*, see Proposition 4 and 3, are LPs based on this same constraint matrix. The robustified statistics under γ-contamination are shifted versions of the original ones. Here, we utilize the linear connection between $\underline{d}^\varepsilon_{\mathbf{X},\mathbf{Y}}(\omega_0)$ and $d^\varepsilon_{\mathbf{X},\mathbf{Y}}(\omega_0)$, $\overline{d}^\varepsilon_I$ and d^ε_I, respectively, for fixed ϵ (see Supp. C).

Although one only needs to compute the constraint matrix once, the worst-case complexity of the computation is $\mathcal{O}(s^4)$. In the implementation, we focused on the case of two ordinal variables and only one numerical variable, using the preference system (13). We exploit the fact that sorting the data set allows some comparisons to be skipped immediately by considering only the ordinal components. In particular, if the ordinal variables have a small number of categories compared to the sample size s, this can lead to a large proportion of comparisons being skipped. In the most cases, this reduces the computational cost of computing the constraint matrix compared to a naive implementation. Of course, in the worst case, the computation time cannot be drastically reduced in this way. For further details on the implementation, see Supp. B.

8.2 EXAMPLE: POVERTY ANALYSIS

At least since the capability approach by Sen [1985], there is mostly consensus that poverty has more facets than income or wealth. It is perceived as multidimensional concept, involving variables that are often ordinally scaled, e.g., level of education. One common task in poverty analysis is to compare subgroups like men and women. Stochastic dominance is a popular way of comparing such subpopulations, see e.g. Garcia-Gomez et al. [2019]. Excitingly, our approach allows us to extend this to multidimensional poverty measurement with any kind (of scales) of dimensions.

Figure 2: Distributions of d^ε_I with $\varepsilon \in \{0, 0.25, 0.5, 0.75, 1\}$ obtained from $N = 1000$ resamples of ALLBUS data. Black stripes show exact positions of d^ε_I values. Vertical black line marks median. Red line shows value of the respective observed test statistics $d^\varepsilon_{\mathbf{X},\mathbf{Y}}(\omega)$.

In the following, we will use data from the German General Social Survey (ALLBUS) GESIS [2018] that accounts for three dimensions of poverty: income (numeric), health (ordinal, 6 levels) and education (ordinal, 8 levels), see also Breyer and Danner [2015]. We are using the 2014 edition and focus on a subsample with $n = m = 100$ men and women each. We are interested in the hypothesis that women are dominated by men with respect to GSD – differently put, that women are poorer than men regarding any compatible utility representation of income, health and education.

As discussed in Section 5, we test the hypotheses (4), where

X resp. Y correspond to the subpopulation of men resp. women. We deploy our test with varying regularization strength ε. Figure 2 displays the distribution of the test statistics obtained trough $N = 1000$ resamples (cf. Section 5.3). It becomes evident that our proposed regularization serves its purpose: As ε increases, the distribution of tests statistics becomes both more centered and closer to zero. Moreover, we reject for higher shares of the test statistics, see the position of $d^{\varepsilon}_{\mathbf{X},\mathbf{Y}}(\omega)$ (red line) compared to d^{ε}_{I} (black stripes). For $\varepsilon \in \{0.5, 0.75, 1\}$ we reject for the common significance level of $\alpha \approx 0.05$.

As touched upon in Section 8.1, the robustified versions of the test statistic under the linear-vacuous model are shifted versions of the regular test statistics, i.e., they do not have to be computed explicitly. Exploiting this fact, we visualize the share of regularized test statistics for which we do not reject the null hypothesis (black stripes right of red line in Figure 2), depending on the contamination parameter γ of the underlying linear-vacuous model, see Figure 3 (and Supp. C for details on computing the shares). It should be mentioned that these shares correspond to p-values telling at which significance levels α the test would be marginally rejected. Generally, it becomes apparent that even for small

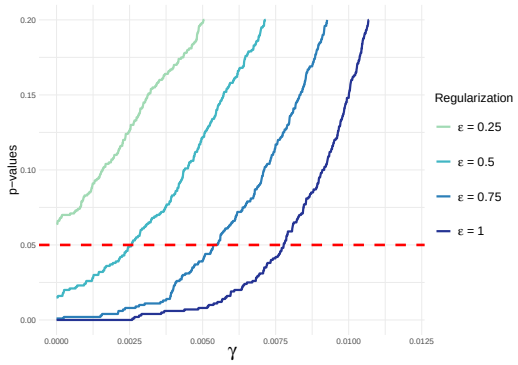

Figure 3: P-values as function of the contamination γ (see Supp. C) for tests with different regularization strength ε. Dotted red line marks significance level $\alpha = 0.05$.

values of γ the test statistics can be severely corrupted. If we allow more than 1% ($\gamma > 0.01$) of the data (2 observations) to be redistributed in any manner, the shares of rejections drop drastically. Therefore, ignoring an (even very tiny) contamination γ of the underlying distributions leads to a seriously inflated type I error. Remarkably, our regularization hedges against this to some extent: Given a significance level $\alpha = 0.05$, the fully regularized version (i.e., $\varepsilon = 1$) of our robustified test (cf., Section 6) comes to the same decision for γ up to 0.075. As explained in Section 5, rejecting H_0 does not necessarily mean that women are dominated by men; they could also be incomparable. However, our tests with reversed variables give no evidence of incomparability: all their observed p-values are above 0.95.

Further Applications: We also analyzed a dermatology data set that contains variables on symptoms of the eryhemato-squamous disease, see Demiroz et al. [1998] accessed via Dua and Graff [2017], as well as the German credit data set that consists of variables on credit applicants, see Dua and Graff [2017]. In case of the credit data, we reject the hypothesis that high-risk applicants are dominated by low-risk applicants w.r.t. GSD for a common significance level of $\alpha \approx 0.05$. In the first application we are interested in the hypothesis that patients without a family history of the disease are dominated by patients without a family history with respect to GSD. We reject again for $\alpha \approx 0.05$. However, the p-values are much higher than in the other two applications. For detailed results as well as more information on the data sets, we refer to the supplement.

9 CONCLUDING REMARKS

Summary: We have further explored a generalized stochastic dominance (GSD) order among random variables with locally varying scale of measurement. We focused on four aspects: First, the investigation of (regularized) statistical tests for GSD when only samples of the variables are available. Second, robustifications of these tests w.r.t. their underlying assumptions using ideas from imprecise probabilities. Third, a detailed investigation of our ordering for preference systems arising from multidimensional structures with differently scaled dimensions. Finally, applications to examples from poverty measurement, finance, and medicine.

Limitations and future research: Two particular limitations offer promising opportunities for future research.

Extending robust testing to belief function: In Section 6, we have focused – for computational complexity – to linear-vacuous models. However, the idea of identifying least favorable extreme points seems to generalize to any credals sets induced by belief functions in the sense of Shafer [1976].

Improving computational complexity: The LPs for checking in-sample GSD become computer intensive for larger amounts of data. Although complexity reduces for the special case of preference systems discussed in Section 7 (cf. Section 8.1), Proposition 8 suggests that a further drastic reduction can be expected for only one cardinal dimension.

Acknowledgements

We thank all reviewers for constructive comments that helped to improve the presentation of the paper. JR and TA gratefully acknowledge support by the Federal Statistical Office of Germany within the project "Machine Learning in Official Statistics". HB, JR, and GS gratefully acknowledge the financial and general support of the LMU Mentoring program. HB sincerely thanks Evangelisches Studienwerk e.V. for funding and supporting her PhD project.

References

J. Abellan, C. Mantas, J. Castellano, and S. Moral-Garcia. Increasing diversity in random forest learning algorithm via imprecise probabilities. *Expert Syst Appl*, 97:228–243, 2018.

T. Augustin and G. Schollmeyer. Comment: On focusing, soft and strong revision of Choquet capacities and their role in statistics. *Stat Sci*, 36(2):205–209, 2021.

T. Augustin, F. Coolen, G. de Cooman, and M. Troffaes, editors. *Introduction to Imprecise Probabilities*. Wiley, 2014a.

T. Augustin, G. Walter, and F. Coolen. Statistical inference. In T. Augustin, F. Coolen, G. de Cooman, and M. Troffaes, editors, *Introduction to Imprecise Probabilities*, pages 135–189. Wiley, 2014b.

G. Barrett and S. Donald. Consistent tests for stochastic dominance. *Econometrica*, 71(1):71–104, 2003.

J. Bauer. Selection errors of random route samples. *Sociol Method Res*, 43(3):519–544, 2014.

J. Bauer. Biases in random route surveys. *Journal of Survey Statistics and Methodology*, 4(2):263–287, 2016.

V. Bengs, E. Hüllermeier, and W. Waegeman. Pitfalls of epistemic uncertainty quantification through loss minimisation. In *Advances in Neural Information Processing Systems*, 2022.

H. Blocher, G. Schollmeyer, C. Jansen, and M. Nalenz. Depth functions for partial orders with a descriptive analysis of machine learning algorithms. In *International Symposium on Imprecise Probabilities: Theories and Applications*, 2023. PMLR (to appear).

B. Breyer and D. Danner. Skala zur Erfassung des Lebenssinns (ALLBUS). In *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS) (GESIS – Leibniz-Institut für Sozialwissenschaften)*, volume 10, 2015.

L. Brown, J. Hwang, and A. Munk. An unbiased test for the bioequivalence problem. *Ann Stat*, 25(6):2345 – 2367, 1997.

R. Cabanas, A. Antonucci, D. Huber, and M. Zaffalon. CREDICI: A Java library for causal inference by credal networks. In M. Jaeger and T. Nielsen, editors, *International Conference on Probabilistic Graphical Models*, volume 138 of *PMLR*, pages 597–600, 2020.

Y. Carranza and S. Destercke. Imprecise Gaussian discriminant classification. *Pattern Recogn*, 112:107739, 2021.

C. Chang, J. Jimenez-Martin, E. Maasoumi, and T. Perez-Amaral. A stochastic dominance approach to financial risk management strategies. *J Econometrics*, 187:472–485, 2015.

D. Chetverikov, D. Wilhelm, and D. Kim. An adaptive test of stochastic monotonicity. *Econometric Theory*, 37(3): 495–536, 2021.

G. Corani and M. Zaffalon. Learning reliable classifiers from small or incomplete data sets: The naive credal classifier 2. *J Mach Learn Res*, 9(4), 2008.

G. Corani, A. Benavoli, J. Demsar, F. Mangili, and M. Zaffalon. Statistical comparison of classifiers through Bayesian hierarchical modelling. *Mach Learn*, 106(11): 1817–1837, 2017.

H. Dai, Y. Xue, N. He, Y. Wang, N. Li, D. Schuurmans, and B. Dai. Learning to optimize for stochastic dominance constraints. In F. Ruiz, J. Dy, and J. van de Meent, editors, *Artificial Intelligence and Statistics*, volume 206 of *PMLR*, pages 8991–9009, 2023.

G. Demiroz, H. Govenir, and N. Ilter. Learning differential diagnosis of Eryhemato-Squamous diseases using voting feature intervals. *Artif Intell Med*, 13:147–165, 1998.

J. Demsar. Statistical comparisons of classifiers over multiple data sets. *J Mach Learn Res*, 7:1–30, 2006.

D. Denneberg. *Non-additive Measure and Integral*. Kluwer Academic Publishers, 1994.

S. Destercke, I. Montes, and E. Miranda. Processing distortion models: A comparative study. *Int J Approx Reason*, 145:91–120, 2022.

D. Dua and C. Graff. UCI machine learning repository, 2017. `http://archive.ics.uci.edu/ml`.

European Commission. Knowledge service: Competence centre on composite indicators and scoreboards, 2023. URL `https://knowledge4policy.ec.europa.eu/composite-indicators_en`. (Febr. 16, 2023).

C. Garcia-Gomez, A. Perez, and M. Prieto-Alaiz. A review of stochastic dominance methods for poverty analysis. *J Econ Surv*, 33(5):1437–1462, 2019.

GESIS. Allgemeine Bevölkerungsumfrage der Sozialwissenschaften ALLBUS 2014. GESIS Datenarchiv, Köln. ZA5240 Datenfile Version 2.2.0, https://doi.org/10.4232/1.13141, 2018.

E. Hüllermeier and W. Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Mach Learn*, 110(3):457–506, 2021.

E. Hüllermeier, S. Destercke, and M. Shaker. Quantification of credal uncertainty in machine learning: A critical analysis and empirical comparison. In J. Cussens and K. Zhang, editors, *Uncertainty in Artificial Intelligence*, volume 180 of *PMLR*, pages 548–557, 2022.

C. Jansen and T. Augustin. Decision making with state-dependent preference systems. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 729–742. Springer, 2022.

C. Jansen, G. Schollmeyer, and T. Augustin. Concepts for decision making under severe uncertainty with partial ordinal and partial cardinal preferences. *Int J Approx Reason*, 98:112–131, 2018.

C. Jansen, H. Blocher, T. Augustin, and G. Schollmeyer. Information efficient learning of complexly structured preferences: Elicitation procedures and their application to decision making under uncertainty. *Int J Approx Reason*, 144:69–91, 2022a.

C. Jansen, M. Nalenz, G. Schollmeyer, and T. Augustin. Statistical comparisons of classifiers by generalized stochastic dominance, 2022b. URL `https://arxiv.org/abs/2209.01857`. arXiv preprint.

C. Jansen, G. Schollmeyer, and T. Augustin. Quantifying degrees of E-admissibility in decision making with imprecise probabilities. In T. Augustin, F. Cozman, and G. Wheeler, editors, *Reflections on the Foundations of Probability and Statistics: Essays in Honor of Teddy Seidenfeld*, pages 319–346. Springer, 2022c.

C. Jansen, G. Schollmeyer, and T. Augustin. Multi-target decision making under conditions of severe uncertainty. In V. Torra and Y. Narukawa, editors, *Modeling Decisions for Artificial Intelligence*, pages 45–57. Springer, 2023.

I. Levi. On indeterminate probabilities. *The Journal of Philosophy*, 71:391–418, 1974.

J. Lienen and E. Hüllermeier. Credal self-supervised learning. *Advances in Neural Information Processing Systems*, 34:14370–14382, 2021.

R. Luce. Semiorders and a theory of utility discrimination. *Econometrica*, 24:178–191, 1956.

A. Malinin and M. Gales. Predictive uncertainty estimation via prior networks. *Advances in Neural Information Processing Systems*, 31, 2018.

D. Maua and F. Cozman. Thirty years of credal networks: Specification, algorithms and complexity. *Int J Approx Reason*, 126:133–157, 2020.

D. Maua and C. de Campos. Editorial to: Special issue on robustness in probabilistic graphical models. *Int J Approx Reason*, 137:113, 2021.

D. McFadden. Testing for stochastic dominance. In T. Fomby and T. Seo, editors, *Studies in the Economics of Uncertainty*, pages 113–134. Springer, 1989.

I. Montes, E. Miranda, and S. Destercke. Unifying neighbourhood and distortion models: Part II – new models and synthesis. *Int J Gen Syst*, 49:636–674, 2020.

K. Mosler. Testing whether two distributions are stochastically ordered or not. In H. Rinne, B. Rüger, and H. Strecker, editors, *Grundlagen der Statistik und ihre Anwendungen: Festschrift für Kurt Weichselberger*, pages 149–155. Physica-Verlag, 1995.

K. Mosler and M. Scarsini. Some theory of stochastic dominance. *Lecture Notes-Monograph Series*, 19:261–284, 1991.

L. Oneto, S. Ridella, and D. Anguita. Tikhonov, Ivanov and Morozov regularization for support vector machine learning. *Mach Learn*, 103:103–136, 2016.

M. Pivato. Multiutility representations for incomplete difference preorders. *Math Sco Sci*, 66:196–220, 2013.

J. Pratt and J. Gibbons. *Concepts of Nonparametric Theory*. Springer, 2012.

T. Range and L. Østerdal. First-order dominance: stronger characterization and a bivariate checking algorithm. *Math Program*, 173:193—219, 2019.

J. Rodemann and T. Augustin. Accounting for Gaussian process imprecision in Bayesian optimization. In *International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making (IUKM)*, pages 92–104. Springer, 2022.

J. Rodemann, C. Jansen, G. Schollmeyer, and T. Augustin. In all likelihoods: Robust selection of pseudo-labeled data. In *International Symposium on Imprecise Probabilities: Theories and Applications*, 2023. PMLR (to appear).

G. Schollmeyer. A short note on the equivalence of the ontic and the epistemic view on data imprecision for the case of stochastic dominance for interval-valued data. In *International Symposium on Imprecise Probabilities: Theories and Applications*, pages 330–337. PMLR, 2019.

G. Schollmeyer, C. Jansen, and T. Augustin. Detecting stochastic dominance for poset-valued random variables as an example of linear programming on closure systems, 2017. URL `https://epub.ub.uni-muenchen.de/40416/13/TR_209.pdf`. Technical Report 209, Department of Statistics, LMU Munich.

A. Sen. *Commodities and Capabilities*. Elsevier, 1985.

G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

M. Shaked and G. Shanthikumar. *Stochastic orders*. Springer, 2007.

M. Shaker and E. Hüllermeier. Ensemble-based uncertainty quantification: Bayesian versus credal inference, 2021. URL https://arxiv.org/abs/2107.10384. arXiv preprint.

M. Timonin. Maximization of the Choquet integral over a convex set and its application to resource allocation problems. *Ann Oper Res*, 196:543–579, 2012.

UNECE. Guidelines on producing leading, composite and sentiment indicators, 2019. URL https://unece.org/DAM/stats/publications/2019/ECECESSTAT20192.pdf. (Febr. 16, 2023).

L. Utkin. An imprecise deep forest for classification. *Expert Syst Appl*, 141:112978, 2020.

L. Utkin and A. Konstantinov. Attention-based random forest and contamination model. *Neural Networks*, 154: 346–359, 2022.

P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

Robust Statistical Comparison of Random Variables with Locally Varying Scale of Measurement (Supplementary Material)

Christoph Jansen[1] **Georg Schollmeyer**[1] **Hannah Blocher**[1] **Julian Rodemann**[1] **Thomas Augustin**[1]

[1]Department of Statistics, Ludwig-Maximilians-Universität, Munich, Bavaria, Germany

In the following, we give supplementary information and material to the main paper. This includes all mathematical proofs of the propositions and corollaries established in the main paper (Part A), further details on the implementation and reproducibility (Part B), further calculations for the robustified test statistics (Part C), and further analyses of the applications in the main paper (Part D). If not explicitly stated otherwise, from now on, all references to equations, propositions, etc. refer to the main part of the paper.

A PROOFS OF THE RESULTS IN THE MAIN PAPER

A.1 PROOFS FOR PROPOSITION 1 AND 2: BOUNDED PREFERENCE SYSTEMS

We start by proving Propositions 1 resp. 2 from Sections 2 resp. 4 that state that checking consistency resp. GSD simplifies if the underlying preference system is bounded.

Proposition 1 *Let $\mathcal{A} = [A, R_1, R_2]$ be a bounded preference system. Then $\mathcal{A}$ is consistent iff it is 0-consistent.*

Proof. If $\mathcal{A}$ is 0-consistent, then it is obviously also consistent, since every normalized representation is in particular a representation. For the other direction, assume $\mathcal{A}$ to be consistent. Choose $u \in \mathcal{U}_{\mathcal{A}}$ arbitrarily and denote by a_*, a^* the R_1-minimal resp. R_1-maximal elements satisfying $(a^*, a_*) \in P_{R_1}$. From the latter we know that $u(a^*) > u(a_*)$. Thus, the function

$$\tilde{u} : A \to [0,1] \ , \ a \mapsto \frac{u(a) - u(a_*)}{u(a^*) - u(a_*)}$$

is well-defined. Moreover, one easily verifies that $\tilde{u} \in \mathcal{U}_{\mathcal{A}}$, and $u(a_*) = 0$, and $u(a^*) = 1$. Thus, we can conclude that $\tilde{u} \in \mathcal{N}_{\mathcal{A}}$, which – by definition – implies 0-consistency. □

Proposition 2 *If $\mathcal{A}$ is consistent and bounded with a_*, a^* as before, then $(X, Y) \in R_{(\mathcal{A},\pi)}$ iff*

$$\forall u \in \mathcal{N}_{\mathcal{A}} : \mathbb{E}_\pi(u \circ X) \geq \mathbb{E}_\pi(u \circ Y).$$

Proof. The direction $\Rightarrow$ follows trivially by observing $\mathcal{N}_{\mathcal{A}} \subseteq \mathcal{U}_{\mathcal{A}}$. For the direction $\Leftarrow$, assume that it holds $\forall u \in \mathcal{N}_{\mathcal{A}} : \mathbb{E}_\pi(u \circ X) \geq \mathbb{E}_\pi(u \circ Y)$. Choose $u \in \mathcal{U}_{\mathcal{A}}$ arbitrarily. With the same argument as given in the proof of Proposition 1, we know that then $\tilde{u} \in \mathcal{N}_{\mathcal{A}}$, where $\tilde{u}$ is defined as in the proof of Proposition 1. Since $\tilde{u}$ is a positive (affine) linear transformation of u, we know that $\mathbb{E}_\pi(u \circ X) \geq \mathbb{E}_\pi(u \circ Y)$ if and only if $\mathbb{E}_\pi(\tilde{u} \circ X) \geq \mathbb{E}_\pi(\tilde{u} \circ Y)$. Since the latter is true by assumption (utilizing $\tilde{u} \in \mathcal{N}_{\mathcal{A}}$), the first also is true. As u was chosen arbitrarily, this completes the proof. □

A.2 PROOFS OF PROPOSITIONS 3 AND 4: COMPUTATIONS FOR THE PERMUTATION TEST

We now give proofs for Propositions 3 resp. 4 from Section 5.3 that concern the computation of the maximum regularization strength resp. the computation of the (regularized) test statistic for the permutation-test.

Proposition 3 *For samples* $\mathbf{x}$ *and* $\mathbf{y}$ *of the form (6) and (7) and* $\varepsilon \in [0,1]$*, we consider the linear program*

$$\xi \longrightarrow \max_{(v_1,\ldots,v_s,\xi)}$$

with constraints $(v_1,\ldots,v_s,\xi) \in C(\mathbf{x},\mathbf{y})$*. Denote by* ξ^* *its optimal value. It then holds* $\delta_\varepsilon(\omega_0) = \varepsilon \cdot \xi^*$*.*

Proof. The Proposition follows from standard results on linear optimization and the fact that $C(\mathbf{x},\mathbf{y})$ is compact. Set $I := \{\ell : (z_\ell, a^*) \in I_{R_1}\}$ and define the vector $\underline{v} := (0,1,v_3,\ldots,v_s,0) \in [0,1]^{s+1}$ by $v_\ell = 1$ if $\ell \in I$ and $v_\ell = 0$ otherwise. One then easily verifies that $\underline{v}$ is an admissible solution to the above linear program. Since $C(\mathbf{x},\mathbf{y})$ is compact, this implies the existence of an optimal solution. Denote thus by $\underline{v}^* := (0,1,v_3^*,\ldots,v_s^*,\xi^*)$ an arbitrary optimal solution. We have to show that

$$\xi^* = \sup\{\xi : \mathcal{N}^{\xi}_{\mathcal{A}_{\omega_0}} \neq \emptyset\} =: c.$$

Assume, for contradiction, the above equality does not hold. We distinguish two cases:

Case 1: $\xi^* < c$. Then, one easily verifies that for any function $u \in \mathcal{N}^{c}_{\mathcal{A}_{\omega_0}}$ the vector $(u(z_1),\ldots,u(z_s),c)$ defines an admissible solution to the above linear program with an objective value of c. This constradicts the optimality of $\underline{v}^*$.

Case 2: $\xi^* > c$. Then, setting $u : (\mathbf{XY})_{\omega_0} \to [0,1]$ with $u(z_\ell) := v_\ell^*$ defines an element of $\mathcal{N}^{\xi^*}_{\mathcal{A}_{\omega_0}}$, contradicting that c is the largest number for which $\mathcal{A}_{\omega_0}$ is c-consistent.

Thus, we have that $c = \xi^*$, implying $\delta_\varepsilon(\omega_0) = \varepsilon \cdot \xi^*$. □

Proposition 4 *For samples* $\mathbf{x}$ *and* $\mathbf{y}$ *of the form (6) and (7) and* $\varepsilon \in [0,1]$*, we consider the following linear program*

$$\sum_{\ell=1}^{s} v_\ell \cdot \left(\tfrac{|\{i:x_i=z_\ell\}|}{n} - \tfrac{|\{i:y_i=z_\ell\}|}{m}\right) \longrightarrow \min_{(v_1,\ldots,v_s)}$$

with constraints $(v_1,\ldots,v_s) \in C_{\varepsilon\xi^*}(\mathbf{x},\mathbf{y})$*, where* ξ^* *denotes the optimal value of (8). Denote by* $opt_\varepsilon(\mathbf{x},\mathbf{y})$ *its optimal value. It then holds:*

i) $opt_\varepsilon(\mathbf{x},\mathbf{y}) = d^{\varepsilon}_{\mathbf{X},\mathbf{Y}}(\omega_0)$*.*
ii) There is in-sample GSD of X *over* Y *if and only if* $opt_0(\mathbf{x},\mathbf{y}) \geq 0$*.*

Proof. i) By definition and Proposition 2 , we know that $\mathcal{N}^{\xi^*}_{\mathcal{A}_{\omega_0}} \neq \emptyset$. As these sets are nested with decreasing ξ-value and we have $\varepsilon\xi^* \leq \xi^*$, this implies that also $\mathcal{N}^{\varepsilon\xi^*}_{\mathcal{A}_{\omega_0}} \neq \emptyset$. Hence, we can choose $u \in \mathcal{N}^{\varepsilon\xi^*}_{\mathcal{A}_{\omega_0}}$. One then easily verifies that the vector $(u(z_1),\ldots,u(z_s))$ defines an admissible solution to the above linear program. Since $C_{\varepsilon\xi^*}(\mathbf{x},\mathbf{y})$ is compact, this implies the existence of an optimal solution. Thus, denote by $\underline{v}^* := (v_1^*,\ldots,v_s^*)$ an arbitrary such optimal solution. If we then define $u : (\mathbf{XY})_{\omega_0} \to [0,1]$ with $u(z_\ell) := v_\ell^*$, then one easily verifies that $u \in \mathcal{N}^{\varepsilon\xi^*}_{\mathcal{A}_{\omega_0}}$ and that

$$opt_\varepsilon(\mathbf{x},\mathbf{y}) = \sum_{z \in (\mathbf{XY})_{\omega_0}} u(z) \cdot (\hat{\pi}^{\omega_0}_X(\{z\}) - \hat{\pi}^{\omega_0}_Y(\{z\})) \tag{1}$$

(to see this, note that the right side of the equation is a simple reformulation of the objective function with $\underline{v}^*$ plugged-in).

We have to show that

$$opt_\varepsilon(\mathbf{x},\mathbf{y}) = d^{\varepsilon}_{\mathbf{X},\mathbf{Y}}(\omega_0).$$

Assume, for contradiction, the above equality does not hold. We distinguish two cases:

Case 1: $opt_\varepsilon(\mathbf{x},\mathbf{y}) > d^{\varepsilon}_{\mathbf{X},\mathbf{Y}}(\omega_0)$. This would imply that there exists an $u' \in \mathcal{N}^{\varepsilon\xi^*}_{\mathcal{A}_{\omega_0}}$ that – if it was set in the right-hand side of the above Equation (1) (in the supplementary material) instead of u – would produce a value strictly smaller than $opt_\varepsilon(\mathbf{x},\mathbf{y})$. This contradicts the optimality of $\underline{v}^*$, since every $u' \in \mathcal{N}^{\varepsilon\xi^*}_{\mathcal{A}_{\omega_0}}$ produces an admissible solution to the linear program with objective value given by the right-hand side of the above Equation (1).

Case 2: $opt_\varepsilon(\mathbf{x},\mathbf{y}) < d^{\varepsilon}_{\mathbf{X},\mathbf{Y}}(\omega_0)$. This would be an immediate contradiction to the above Equation (1) (in the supplementary material), since $d^{\varepsilon}_{\mathbf{X},\mathbf{Y}}(\omega_0)$ is by definition the infimum over all the expressions on the equation's right-hand side.

This completes the proof of i). To see ii), note that i) implies $opt_0(\mathbf{x},\mathbf{y}) = d^{0}_{\mathbf{X},\mathbf{Y}}(\omega_0)$. Thus, we have $opt_0(\mathbf{x},\mathbf{y}) \geq 0$ if and only if $d^{0}_{\mathbf{X},\mathbf{Y}}(\omega_0) \geq 0$, which – by definition – is true if and only if there is in-sample GSD of X over Y. □

A.3 PROOFS OF PROPOSITION 5 AND 6: COMPUTATIONS FOR ROBUSTIFIED TESTING

We now give proofs of Proposition 5 resp. 6 from Section 6 concerning the computation of the robustified test statistic resp. its simplification under the special case of a γ-contamination model (with $\gamma \in [0,1]$).

Proposition 5 *For samples* $\mathbf{x}$ *and* $\mathbf{y}$ *of the form (6) and (7),* $\varepsilon \in [0,1]$*, and* $(\pi_1, \pi_2) \in \mathcal{E}(\mathcal{M}_X^{\omega_0}) \times \mathcal{E}(\mathcal{M}_Y^{\omega_0})$*, we consider the following linear program:*

$$\sum_{\ell=1}^{s} v_\ell \cdot (\pi_1(\{z\}) - \pi_2(\{z\})) \longrightarrow \min_{(v_1,\ldots,v_s)}$$

with constraints $(v_1, \ldots, v_s) \in C_{\varepsilon\xi^*}(\mathbf{x}, \mathbf{y})$*, where* ξ^* *denotes the optimal value of (8). Denote by* $opt_\varepsilon(\mathbf{x}, \mathbf{y}, \pi_1, \pi_2)$ *its optimal value and by* $\underline{opt}_\varepsilon(\mathbf{x}, \mathbf{y})$ *the minimal optimum over all combinations of* $(\pi_1, \pi_2) \in \mathcal{E}(\mathcal{M}_X^{\omega_0}) \times \mathcal{E}(\mathcal{M}_Y^{\omega_0})$*. It then holds:*

i) $\underline{opt}_\varepsilon(\mathbf{x}, \mathbf{y}) = \underline{d}^\varepsilon_{\mathbf{X},\mathbf{Y}}(\omega_0)$.

ii) There is in-sample GSD of X *over* Y *for any* π *with* $\hat{\pi}_X^{\omega_0} \in \mathcal{M}_X^{\omega_0}$ *and* $\hat{\pi}_Y^{\omega_0} \in \mathcal{M}_Y^{\omega_0}$ *if* $\underline{opt}_0(\mathbf{x}, \mathbf{y}) \geq 0$.

Proof. i) Since nothing in the proof of Proposition 4 hinges on the concrete structure of the involved empirical image measures, Proposition 4 is still valid if we replace $\hat{\pi}_X^{\omega_0}$ and $\hat{\pi}_Y^{\omega_0}$ by arbitrary $\pi_1 \in \mathcal{M}_X^{\omega_0}$ and $\pi_2 \in \mathcal{M}_Y^{\omega_0}$, respectively. This specifically implies

$$opt_\varepsilon(\mathbf{x}, \mathbf{y}, \pi_1, \pi_2) = \inf_{u \in \mathcal{N}_{\mathcal{A}\omega_0}^{\delta_\varepsilon(\omega_0)}} \sum_{z \in (\mathbf{XY})_{\omega_0}} u(z) \cdot (\pi_1(\{z\}) - \pi_2(\{z\})). \tag{2}$$

In order to show i), we now need to verify that

$$\inf_{(\pi_1,\pi_2) \in \mathcal{E}(\mathcal{M}_X^{\omega_0}) \times \mathcal{E}(\mathcal{M}_Y^{\omega_0})} opt_\varepsilon(\mathbf{x}, \mathbf{y}, \pi_1, \pi_2) = \underline{d}^\varepsilon_{\mathbf{X},\mathbf{Y}}(\omega_0).$$

Due to the above Equation (2) (in the supplementary material) and the fact that iterated infima can be equivalently replaced by one global infimum, we know that

$$\inf_{(\pi_1,\pi_2) \in \mathcal{M}_X^{\omega_0} \times \mathcal{M}_Y^{\omega_0}} opt_\varepsilon(\mathbf{x}, \mathbf{y}, \pi_1, \pi_2) = \underline{d}^\varepsilon_{\mathbf{X},\mathbf{Y}}(\omega_0). \tag{3}$$

We then can compute:

$$\begin{aligned}
\underline{d}^\varepsilon_{\mathbf{X},\mathbf{Y}}(\omega_0) &\overset{(3)}{=} \inf_{(\pi_1,\pi_2) \in \mathcal{M}_X^{\omega_0} \times \mathcal{M}_Y^{\omega_0}} opt_\varepsilon(\mathbf{x}, \mathbf{y}, \pi_1, \pi_2) \\
&= \inf_{(\pi_1,\pi_2) \in \mathcal{M}_X^{\omega_0} \times \mathcal{M}_Y^{\omega_0}} \inf_{u \in \mathcal{N}_{\mathcal{A}\omega}^{\delta_\varepsilon(\omega)}} \sum_{z \in (\mathbf{XY})_{\omega_0}} u(z) \cdot (\pi_1(\{z\}) - \pi_2(\{z\})) \\
&= \inf_{u \in \mathcal{N}_{\mathcal{A}\omega}^{\delta_\varepsilon(\omega)}} \inf_{(\pi_1,\pi_2) \in \mathcal{M}_X^{\omega_0} \times \mathcal{M}_Y^{\omega_0}} \sum_{z \in (\mathbf{XY})_{\omega_0}} u(z) \cdot (\pi_1(\{z\}) - \pi_2(\{z\})) \\
&\overset{(\star)}{=} \inf_{u \in \mathcal{N}_{\mathcal{A}\omega}^{\delta_\varepsilon(\omega)}} \left(\inf_{\pi_1 \in \mathcal{M}_X^{\omega_0}} \sum_{z \in (\mathbf{XY})_{\omega_0}} u(z) \cdot \pi_1(\{z\}) - \sup_{\pi_2 \in \mathcal{M}_Y^{\omega_0}} \sum_{z \in (\mathbf{XY})_{\omega_0}} u(z) \cdot \pi_2(\{z\}) \right) \\
&\overset{(\star\star)}{=} \inf_{u \in \mathcal{N}_{\mathcal{A}\omega}^{\delta_\varepsilon(\omega)}} \left(\inf_{\pi_1 \in \mathcal{E}(\mathcal{M}_X^{\omega_0})} \sum_{z \in (\mathbf{XY})_{\omega_0}} u(z) \cdot \pi_1(\{z\}) - \sup_{\pi_2 \in \mathcal{E}(\mathcal{M}_Y^{\omega_0})} \sum_{z \in (\mathbf{XY})_{\omega_0}} u(z) \cdot \pi_2(\{z\}) \right) \\
&= \inf_{(\pi_1,\pi_2) \in \mathcal{E}(\mathcal{M}_X^{\omega_0}) \times \mathcal{E}(\mathcal{M}_Y^{\omega_0})} \inf_{u \in \mathcal{N}_{\mathcal{A}\omega}^{\delta_\varepsilon(\omega)}} \sum_{z \in (\mathbf{XY})_{\omega_0}} u(z) \cdot (\pi_1(\{z\}) - \pi_2(\{z\})) \\
&= \inf_{(\pi_1,\pi_2) \in \mathcal{E}(\mathcal{M}_X^{\omega_0}) \times \mathcal{E}(\mathcal{M}_Y^{\omega_0})} opt_\varepsilon(\mathbf{x}, \mathbf{y}, \pi_1, \pi_2)
\end{aligned}$$

Here, ($\star$) follows since – for u fixed – the infimum of the differences of the two sums is attained if the first sum is smallest possible and the second sum is largest possible (note that all sums involved are finite). Further, ($\star\star$) follows since – again for u fixed – the sums are linear functions on the compact sets $\mathcal{M}_X^{\omega_0}$ resp. $\mathcal{M}_Y^{\omega_0}$ and, therefore, attain their optima on $\mathcal{E}(\mathcal{M}_X^{\omega_0})$ resp. $\mathcal{E}(\mathcal{M}_Y^{\omega_0})$. The fith and sixth equalities are just reversing the computation done in the first three equalities.

To see ii), note that i) implies $\underline{opt}_0(\mathbf{x},\mathbf{y}) = \underline{d}^0_{\mathbf{X},\mathbf{Y}}(\omega_0)$. Thus, $\underline{opt}_0(\mathbf{x},\mathbf{y}) \geq 0$ if and only if $\underline{d}^0_{\mathbf{X},\mathbf{Y}}(\omega_0) \geq 0$. But – by definition – the latter is true if and only if

$$\inf_{u \in \mathcal{N}^0_{\mathcal{A}_{\omega_0}}} \sum_{z \in (\mathbf{XY})_{\omega_0}} u(z) \cdot (\pi_1(\{z\}) - \pi_2(\{z\})) \geq 0$$

for all $(\pi_1, \pi_2) \in \mathcal{M}_X^{\omega_0} \times \mathcal{M}_Y^{\omega_0}$. This obviously implies in-sample GSD of X over Y for any π with $\hat{\pi}_X^{\omega_0} \in \mathcal{M}_X^{\omega_0}$ and $\hat{\pi}_Y^{\omega_0} \in \mathcal{M}_Y^{\omega_0}$, since $\mathcal{N}^0_{\mathcal{A}_{\omega_0}} = \mathcal{N}_{\mathcal{A}_{\omega_0}}$. □

Proposition 6 *Consider again the situation of Proposition 5 with the additional assumption that $\mathcal{M}_X^{\omega_0}$ and $\mathcal{M}_Y^{\omega_0}$ are of the form (11) with extreme points as in (12). It then holds:*

$$\underline{opt}_\varepsilon(\mathbf{x},\mathbf{y}) = opt_\varepsilon(\mathbf{x},\mathbf{y},\pi_*,\pi^*)$$

where

$$\pi_* = \gamma\delta_{a_*} + (1-\gamma)\hat{\pi}_X^{\omega_0}$$

and

$$\pi^* = \gamma\delta_{a^*} + (1-\gamma)\hat{\pi}_Y^{\omega_0}.$$

Proof. By again utilizing Equation (2) (of the supplementary material), the claim modifies to showing that

$$\underline{opt}_\varepsilon(\mathbf{x},\mathbf{y}) = \inf_{u \in \mathcal{N}^{\delta_\varepsilon(\omega_0)}_{\mathcal{A}_{\omega_0}}} \sum_{z \in (\mathbf{XY})_{\omega_0}} u(z) \cdot (\pi_*(\{z\}) - \pi^*(\{z\})).$$

Since, by Proposition 2, we know that $\underline{d}^\varepsilon_{\mathbf{X},\mathbf{Y}}(\omega_0) = \underline{opt}_\varepsilon(\mathbf{x},\mathbf{y})$ and $\underline{d}^\varepsilon_{\mathbf{X},\mathbf{Y}}(\omega_0)$ is by definition the infimum over all the expressions on the right-hand side, the direction $\leq$ is immediate. So, it remains to show the direction $\geq$. To do so, choose $(\pi_1,\pi_2) \in \mathcal{M}_X^{\omega_0} \times \mathcal{M}_Y^{\omega_0}$ arbitrarily. Since both $\mathcal{M}_X^{\omega_0}$ and $\mathcal{M}_Y^{\omega_0}$ are of the form (11), we then know that there exist probability measures ν_1 and ν_2 such that

$$\pi_1 = \gamma \cdot \nu_1 + (1-\gamma) \cdot \hat{\pi}_X^{\omega_0}$$

and

$$\pi_2 = \gamma \cdot \nu_2 + (1-\gamma) \cdot \hat{\pi}_Y^{\omega_0}.$$

Here, we utilized the fact that credal sets of the form (11) can be equivalently characterized as

$$\mathcal{M}_Z^\omega = \left\{\pi : \pi \geq (1-\gamma)\cdot\hat{\pi}_Z^\omega\right\} = \left\{\gamma\cdot\nu + (1-\gamma)\cdot\hat{\pi}_Z^{\omega_0} : \nu \text{ probability measure}\right\}.$$

For $u \in \mathcal{N}^{\delta_\varepsilon(\omega_0)}_{\mathcal{A}_{\omega_0}}$ fixed (but arbitrary), we then can compute:

$$\begin{aligned}
\sum_{z \in (\mathbf{XY})_{\omega_0}} u(z)\cdot\pi_1(\{z\}) &= \gamma \cdot \sum_{z \in (\mathbf{XY})_{\omega_0}} u(z)\cdot\nu_1(\{z\}) + (1-\gamma)\sum_{z \in (\mathbf{XY})_{\omega_0}} u(z)\cdot\hat{\pi}_X^{\omega_0}(\{z\}) \\
&\geq \gamma\cdot u(a_*) + (1-\gamma)\sum_{z \in (\mathbf{XY})_{\omega_0}} u(z)\cdot\hat{\pi}_X^{\omega_0}(\{z\}) \\
&= \gamma \cdot \sum_{z \in (\mathbf{XY})_{\omega_0}} u(z)\cdot\delta_{a_*}(\{z\}) + (1-\gamma)\sum_{z \in (\mathbf{XY})_{\omega_0}} u(z)\cdot\hat{\pi}_X^{\omega_0}(\{z\}) \\
&= \sum_{z \in (\mathbf{XY})_{\omega_0}} u(z)\cdot\pi_*(\{z\})
\end{aligned}$$

Analogous reasoning yields:

$$\sum_{z\in(\mathbf{XY})_{\omega_0}} u(z)\cdot\pi_2(\{z\}) \le \sum_{z\in(\mathbf{XY})_{\omega_0}} u(z)\cdot\pi^*(\{z\})$$

Putting the two together, we arrive at:

$$\sum_{z\in(\mathbf{XY})_{\omega_0}} u(z)\cdot(\pi_1(\{z\})-\pi_2(\{z\})) \ge \sum_{z\in(\mathbf{XY})_{\omega_0}} u(z)\cdot(\pi_*(\{z\})-\pi^*(\{z\}))$$

As π_1, π_2, and u were chosen arbitrarily, the inequality remains valid for the infimum, i.e.

$$\inf_{(\pi_1,\pi_2,u)\in\mathcal{M}_X^{\omega_0}\times\mathcal{M}_Y^{\omega_0}\times\mathcal{N}_{\mathcal{A}_{\omega_0}}^{\delta_\varepsilon(\omega_0)}} \sum_{z\in(\mathbf{XY})_{\omega_0}} u(z)\cdot(\pi_1(\{z\})-\pi_2(\{z\})) \ge \inf_{u\in\mathcal{N}_{\mathcal{A}_{\omega_0}}^{\delta_\varepsilon(\omega_0)}} \sum_{z\in(\mathbf{XY})_{\omega_0}} u(z)\cdot(\pi_*(\{z\})-\pi^*(\{z\}))$$

Observing that the left side of this inequality by definition equals $\underline{d}^{\varepsilon}_{\mathbf{X},\mathbf{Y}}(\omega_0)$ and, therefore, by Proposition 4, also $\underline{opt}_{\varepsilon}(\mathbf{x},\mathbf{y})$ completes the direction $\geq$ and thus the proof. □

A.4 PROOFS OF PROPOSITIONS 7 AND 8: MULTI-DIMENSIONAL SPACES

Finally, we give proofs of Propositions 7 and 8 from Section 7 concerning several different characterizing properties of the GSD-order for the special case of preferences systems arising from multi-dimensional spaces with differently scaled dimensions. For this, recall that in Section 4 for a preference system $\mathcal{A}$ and a probability measure π we defined

$$\mathcal{F}_{(\mathcal{A},\pi)} := \Big\{ X\in A^\Omega : u\circ X\in\mathcal{L}^1(\Omega,\mathcal{S}_1,\pi)\ \forall u\in\mathcal{U}_{\mathcal{A}}\Big\}.$$

This definition is needed for stating the next proposition.

Proposition 7 *Let π be a probability measure on $(\Omega,\mathcal{S}_1)$, and $X=(\Delta_1,\dots,\Delta_r), Y=(\Lambda_1,\dots,\Lambda_r)\in\mathcal{F}_{(\mathsf{pref}(\mathbb{R}^r),\pi)}$, where the first $0\le z\le r$ dimensions of $\mathsf{pref}(\mathbb{R}^r)$ are of cardinal scale. Then, the following holds:*

i) $\mathsf{pref}(\mathbb{R}^r)$ is consistent.

ii) If $z=0$, then $R_{(\mathsf{pref}(\mathbb{R}^r),\pi)}$ coincides with (first-order) stochastic dominance w.r.t. π and R_1^* (short: FSD(R_1^*,π)).

iii) If $(X,Y)\in R_{(\mathsf{pref}(\mathbb{R}^r),\pi)}$ and $\Delta_j,\Lambda_j\in\mathcal{L}^1(\Omega,\mathcal{S}_1,\pi)$ for all $j=1,\dots,r$, then

I. $\mathbb{E}_\pi(\Delta_j)\ge\mathbb{E}_\pi(\Lambda_j)$ for all $j=1,\dots,r$, and

II. $(\Delta_j,\Lambda_j)\in$FSD$(\geq,\pi)$ for all $j=z+1,\dots,r$.

Additionally, in the special case where all components of X are jointly independent and all components of Y are jointly independent, properties I. and II. imply $(X,Y)\in R_{(\mathsf{pref}(\mathbb{R}^r),\pi)}$ (i.e. also the converse implication holds).

Proof. i) Let $\alpha_1,\dots,\alpha_r\in\mathbb{R}^+$ and $\phi_{z+1},\dots,\phi_r:\mathbb{R}\to\mathbb{R}$ strictly isotone functions. Define $u:\mathbb{R}^r\to\mathbb{R}$ by setting

$$u(x) := \sum_{s=1}^{z}\alpha_s\cdot x_s + \sum_{s=z+1}^{r}\alpha_s\cdot\phi_s(x_s).$$

Then one easily verifies that u defines a representation of $\mathsf{pref}(\mathbb{R}^r)$, proving its consistency.

ii) Assume $z=0$, i.e. all considered dimensions are purely ordinal. We claim that for $\mathcal{A}_0 := [\mathbb{R}^r, R_1^*, \emptyset]$ it holds $\mathcal{U}_{\mathsf{pref}(\mathbb{R}^r)} = \mathcal{U}_{\mathcal{A}_0}$. The direction $\subseteq$ is trivial, so assume $u\in\mathcal{U}_{\mathcal{A}_0}$ arbitrary. It suffices to show that u represents arbitrary pairs of pairs in R_2^*. As R_2^* is antisymmetric for $z=0$, this reduces to show that u strictly represents arbitrary pairs of pairs in $P_{R_2^*}$. So, let $((v,w),(x,y))\in P_{R_2^*}$. This means that for all $j\in\{1,\dots,r\}$ we have $v_j\ge x_j\ge y_j\ge w_j$ and that there is $j_0\in\{1,\dots,r\}$ such that either $v_{j_0}>x_{j_0}$ or $y_{j_0}>w_{j_0}$. Together, this implies $u(v)>u(x)\ge u(y)\ge u(w)$ or $u(v)\ge u(x)\ge u(y)>u(w)$, either way implying $u(v)-u(w)>u(x)-u(y)$. Thus $u\in\mathcal{U}_{\mathsf{pref}(\mathbb{R}^r)}$. As $R_{(\mathcal{A}_0,\pi)}$ coincides with (first-order) stochastic dominance by definition and we have $\mathcal{U}_{\mathsf{pref}(\mathbb{R}^r)}=\mathcal{U}_{\mathcal{A}_0}$ also $R_{(\mathsf{pref}(\mathbb{R}^r),\pi)}$ coincides with (first-order) stochastic dominance.

iii) Let $(X,Y) \in R_{(\mathsf{pref}(\mathbb{R}^r),\pi)}$. We start by showing I, so choose $j \in \{1,\dots,r\}$ arbitrary. By part i) of the proof, for every $n \in \mathbb{N}$, the function $u_n : \mathbb{R}^r \to \mathbb{R}$ defined by

$$u_n(x) := x_j + \frac{1}{n} \cdot \sum_{s \neq j} x_s$$

is a representation of $\mathsf{pref}(\mathbb{R}^r)$, that is $u_n \in \mathcal{U}_{\mathsf{pref}(\mathbb{R}^r)}$. Thus, by our assumption $(X,Y) \in R_{(\mathsf{pref}(\mathbb{R}^r),\pi)}$, we know that we have $\mathbb{E}_\pi(u_n \circ X) \geq \mathbb{E}_\pi(u_n \circ Y)$. This implies (by the linearity of the expectation operator)

$$\mathbb{E}_\pi(\Delta_j) + \frac{1}{n} \cdot \sum_{s \neq j} \mathbb{E}_\pi(\Delta_s) \geq \mathbb{E}_\pi(\Lambda_j) + \frac{1}{n} \cdot \sum_{s \neq j} \mathbb{E}_\pi(\Lambda_s).$$

Letting $n \to \infty$ on both sides gives $\mathbb{E}_\pi(\Delta_j) \geq \mathbb{E}_\pi(\Lambda_j)$.

We use a very similar argument to see II: Choose $j \in \{z+1,\dots,r\}$ arbitrarily and let $\phi : \mathbb{R} \to \mathbb{R}$ be strictly isotone. By part i) of the proof, for every $n \in \mathbb{N}$, the function $u'_n : \mathbb{R}^r \to \mathbb{R}$ defined by

$$u'_n(x) := \phi(x_j) + \frac{1}{n} \cdot \sum_{s \neq j} x_s$$

is a representation of $\mathsf{pref}(\mathbb{R}^r)$, that is $u_n \in \mathcal{U}_{\mathsf{pref}(\mathbb{R}^r)}$. Thus, by our assumption $(X,Y) \in R_{(\mathsf{pref}(\mathbb{R}^r),\pi)}$, we know that we have $\mathbb{E}_\pi(u_n \circ X) \geq \mathbb{E}_\pi(u_n \circ Y)$. This implies (by the linearity of the expectation operator)

$$\mathbb{E}_\pi(\phi \circ \Delta_j) + \frac{1}{n} \cdot \sum_{s \neq j} \mathbb{E}_\pi(\Delta_s) \geq \mathbb{E}_\pi(\phi \circ \Lambda_j) + \frac{1}{n} \cdot \sum_{s \neq j} \mathbb{E}_\pi(\Lambda_s).$$

Letting $n \to \infty$ gives $\mathbb{E}_\pi(\phi \circ \Delta_j) \geq \mathbb{E}_\pi(\phi \circ \Lambda_j)$. As ϕ was chosen arbitrarily, this implies $(\Delta_j, \Lambda_j) \in$FSD$(\geq, \pi)$.

To see the addition to part iii), let $X = (\Delta_1, \dots \Delta_r)$ and $Y = (\Lambda_1, \dots, \Lambda_r)$ have both jointly independent components, respectively, and let I. and II. of iii) be true. Let furthermore $u \in \mathcal{U}_{\mathsf{pref}(\mathbb{R}^r)}$ be an arbitrary utility function that represents the preference system $\mathsf{pref}(\mathbb{R}^r)$. We now show that $\mathbb{E}_\pi(u \circ X) \geq \mathbb{E}_\pi(u \circ Y)$ holds: Because of independence we can compute the expectations of $u \circ X$ and $u \circ Y$ by using Fubini's theorem. To prove the inequality, we first integrate over the ordinal part and use isotonicity of u in every integration. Then we integrate over the cardinal parts and iteratively use the fact that the corresponding functions are representing the corresponding cardinal subsystem built by the components we did not integrate over before. Formally, we arrive at:

$$\begin{aligned}
\mathbb{E}_\pi(u \circ X) &= \int_\Omega u \circ X d\pi \\
&\overset{(ind.)}{=} \int_{\Delta_1(\Omega)} \cdots \int_{\Delta_r(\Omega)} u(\delta_1, \dots, \delta_z, \delta_{z+1}, \dots \delta_r) d\pi_{\Delta_r} \dots d\pi_{\Delta_{z+1}} d\pi_{\Delta_z} \dots d\pi_{\Delta_1} \\
&\overset{(\star)}{\geq} \int_{\Delta_1(\Omega)} \cdots \int_{\Lambda_r(\Omega)} u(\delta_1, \dots, \delta_z, \lambda_{z+1}, \dots \lambda_r) d\pi_{\Lambda_r} \dots d\pi_{\Lambda_{z+1}} d\pi_{\Delta_z} \dots d\pi_{\Delta_1} \\
&\overset{(\star\star)}{\geq} \int_{\Lambda_1(\Omega)} \cdots \int_{\Lambda_r(\Omega)} u(\lambda_1, \dots, \lambda_z, \lambda_{z+1}, \dots \lambda_r) d\pi_{\Lambda_r} \dots d\pi_{\Lambda_{z+1}} d\pi_{\Lambda_z} \dots d\pi_{\Lambda_1} \\
&\overset{(ind.)}{=} \mathbb{E}_\pi(u \circ Y)
\end{aligned}$$

Here, $(\star)$ is valid because, for fixed cardinal components, u is isotone in every ordinal component and we have first order stochastic dominance, which means that the iterated integrals gets smaller if one switches from π_{Δ_k} to π_{Λ_k}.

Similarly, $(\star\star)$ is valid because e.g., for the mapping

$$\psi : \mathbb{R}^{z-1} \to \mathbb{R} \quad , \quad (\delta_1, \dots, \delta_{z-1}) \mapsto \int_{\Delta_z(\Omega)} u(\delta_1, \dots, \delta_r) d\pi_{\Delta_z}$$

is a positive (affine) linear transformation w.r.t. the corresponding subsystem. □

Corollary 1 *If $\mathcal{C} = [C, R_1^c, R_2^c]$ is a bounded subsystem of $\mathsf{pref}(\mathbb{R}^r)$ and $X, Y \in \mathcal{F}_{(\mathcal{C},\pi)}$, then $\mathcal{C}$ is 0-consistent and ii) and iii) from Prop. 7 hold, if we replace $R_{(\mathsf{pref}(\mathbb{R}^r),\pi)}$ by $R_{(\mathcal{C},\pi)}$, FSD(R_1^*, π) by FSD(R_1^c, π), and $(X, Y) \in R_{(\mathsf{pref}(\mathbb{R}^r),\pi)}$ by $\forall u \in \mathcal{N}_\mathcal{C} : \mathbb{E}_\pi(u \circ X) \geq \mathbb{E}_\pi(u \circ Y)$.*

Proof. As, according to Proposition 7 i), we know that $\mathsf{pref}(\mathbb{R}^r)$ is consistent, the same holds true for all of its subsystems. Hence, $\mathcal{C}$ is consistent. Since $\mathcal{C}$ is assumed to be bounded, it then is 0-consistent by Proposition 1. The rest of the Corollary follows, since – by Proposition 2 – for bounded preference systems it suffices to check for dominance only over all normalized representations. □

Proposition 8 *Let $z = 1$ and denote by $\mathcal{U}_{sep}$ the set of all $u : \mathbb{R}^r \to \mathbb{R}$ such that, for $(x_2, \ldots, x_r) \in \mathbb{R}^{r-1}$ fixed, the function $u(\cdot, x_2, \ldots, x_r)$ is strictly increasing and (affine) linear and such that, for $x_1 \in \mathbb{R}$ fixed, the function $u(x_1, \cdot, \ldots, \cdot)$ is strictly isotone w.r.t. the the componentwise partial order on $\mathbb{R}^{r-1}$. Then $\mathcal{U}_{sep} = \mathcal{U}_{\mathsf{pref}(\mathbb{R}^r)}$.*

Proof. First, let $u \in \mathcal{U}_{\mathsf{pref}(\mathbb{R}^r)}$. One easily verifies that, for $x_- := (x_2, \ldots, x_r) \in \mathbb{R}^{r-1}$ fixed, the preference system $Z := [\mathbb{R}, R_1^{x_-}, R_2^{x_-}]$, where $R_1^{x_-} := \geq$ and $R_2^{x_-}$ is defined by

$$\left\{((t,u),(v,w)) : \left(\left(\begin{pmatrix} t \\ x_- \end{pmatrix}, \begin{pmatrix} u \\ x_- \end{pmatrix}\right), \left(\begin{pmatrix} v \\ x_- \end{pmatrix}, \begin{pmatrix} w \\ x_- \end{pmatrix}\right)\right) \in R_2^*\right\}$$

is a complete positive-difference structure in the sense of Krantz et al. [1971, Definition 1, p. 147]. According to Krantz et al. [1971, Theorem 1, p. 147] this implies that any two representations of Z are positive (affine) linear transformations of each other. But it is immediate that both $u(\cdot, x_2, \ldots, x_r)$ and $id_\mathbb{R}(\cdot)$ are representations of Z. Thus, $u(\cdot, x_2, \ldots, x_r) = \alpha \cdot id_\mathbb{R}(\cdot) + \beta$ for some $\alpha \in \mathbb{R}^+$ and $\beta \in \mathbb{R}$, proving the first claim of this direction. The second claim – i.e., the strict isotony of the function $u(x_1, \cdot, \ldots, \cdot)$ w.r.t. the the componentwise partial order on $\mathbb{R}^{r-1}$ for fixed $x_1 \in \mathbb{R}$ – is also immediate. Thus, $u \in \mathcal{U}_{sep}$.

For the other direction, assume that $u \in \mathcal{U}_{sep}$. It follows directly from the assumptions that u is strictly isotone w.r.t. R_1^*. To see that u also strictly represents R_2^*, choose $((x,y),(x',y')) \in R_2^*$ arbitrary. We have two cases:

Case 1: $((x,y),(x',y')) \in I_{R_2^*}$. This implies that $x_1 - y_1 = x'_1 - y'_1$ and therefore also $x_1 - x'_1 = y_1 - y'_1$. Moreover, one easily verifies that the restriction of R_2^* to the ordinal dimensions is antisymmetric . Since we have that x_- componentwise dominates x'_- and vice versa and that y_- componentwise dominates y'_- and vice versa, this antisymmetry then implies that $x_- = x'_-$ and $y_- = y'_-$. Therefore, there are common $\alpha_1, \alpha_2 \in \mathbb{R}^+$ and $\beta_1, \beta_2 \in \mathbb{R}$ such that

$$u(x) = \alpha_1 \cdot x_1 + \beta_1 \quad , \quad u(x') = \alpha_1 \cdot x'_1 + \beta_1$$
$$u(y) = \alpha_2 \cdot y_1 + \beta_2 \quad , \quad u(y') = \alpha_2 \cdot y'_1 + \beta_2$$

Moreover, observe that $\alpha_1 = \alpha_2$, since otherwise there wolud be $x^* \in \mathbb{R}$ with $u(x^*, x_-) < u(x^*, y_-)$, which is not possible, since u is strictly isotone w.r.t. R_1^*. Define

$$D := (u(x) - u(y)) - (u(x') - u(y')).$$

Simple computations then yield

$$D = \alpha_1 \cdot (x_1 - x'_1) - \alpha_2 \cdot (y_1 - y'_1) = (x_1 - x'_1) \cdot (\alpha_1 - \alpha_2)$$

which, as $\alpha_1 = \alpha_2$, implies $D = 0$.

Case 2: $((x,y),(x',y')) \in P_{R_2^*}$. This implies $x_- \geq x'_- \geq y'_- \geq y_-$, where $\geq$ is to be understood componentwise. Using the same argument as seen before, this implies that there exists a $\alpha \in \mathbb{R}^+$ and $\beta_1, \beta_2, \beta_3, \beta_4 \in \mathbb{R}$ such that

$$u(x) = \alpha \cdot x_1 + \beta_1 \quad , \quad u(x') = \alpha \cdot x'_1 + \beta_3$$
$$u(y) = \alpha \cdot y_1 + \beta_2 \quad , \quad u(y') = \alpha \cdot y'_1 + \beta_4$$

Thus, computing D defined as above yields:

$$D = \alpha \cdot ((x_1 - y_1) - (x'_1 - y'_1)) + \beta_1 - \beta_2 - \beta_3 + \beta_4$$

Sub-Case 2.1: $x_1 - y_1 > x'_1 - y'_1$. Observe that, as u is isotone w.r.t. R_1^*, we have that $u(y'_1, y'_-) \geq u(y'_1, y)$. However, this implies $\beta_4 \geq \beta_2$. Analogous reasoning yields $\beta_1 \geq \beta_3$. Using the assumptions of the sub-case, this implies $D > 0$.

Sub-Case 2.2: $x_1 - y_1 = x'_1 - y'_1$. Using the case assumption, this implies that either $x_- > x'_-$ or $y'_- > y_-$, where the $>$ is to be understood as the strict part of the componentwise $\geq$. As u is strictly isotone w.r.t. R_1^*, this implies that either $u(y'_1, y'_-) > u(y'_1, y)$ or $u(x'_1, x_-) > u(x'_1, x'_-)$, which itself implies either $\beta_4 > \beta_2$ or $\beta_1 > \beta_3$. As we know $\beta_4 \geq \beta_2$ and $\beta_1 \geq \beta_3$, this, together with the sub-case assumption, implies $D > 0$. □

B DETAILS ON IMPLEMENTATION AND REPRODUCIBILITY

In Section 8.1 we stated that the implementation of the constraint matrix has worst-case complexity $\mathcal{O}(s^4)$. This worst case occurs when everything in R_1^* and R_2^* is comparable and then

$$s \cdot (s-1) + (s \cdot (s-1)) \cdot ((s \cdot (s-1)) - 1) = s^4 - 2s^3 + s^2$$

many pairwise comparisons have to be considered. Note that we omit the reflexive part of the pre-orders R_1^* and R_2^*.

We are interested in the non-regularized test statistic as well as the regularized test statistic with $\varepsilon \in \{0.25, 0.5, 0.75, 1\}$, see Section 8. For all these cases, we compute the test statistics based on the sample, as well as 1000 times on a permuted version of that sample. Note that the linear programs for computing the test statistics based on the permuted data are identical to that for the non-permuted data except for the objective function, see Section 5.2. In Section C (in the supplementary material), we prove that the robustified test statistics are a shift of the non-robustified test statistic. Thus, the robustified test statistics are immediately given.

The simulation is based on a random sample of the data set. Two of the data sets and the corresponding R-code can be found here:

`https://github.com/hannahblo/Robust_GSD_Tests`

The data set used for the poverty analysis (ALLBUS) is freely accessible, but registration in the corresponding online portal is needed.[1]

For the computation of the linear programs, we used the R interface of Gurobi optimizer, which is documented in Gurobi Optimization, LLC [2020]. This is a commercial solver that offers free academic licenses[2]. In particular, the computation of linear programs is faster than using the free and open source solvers known to us, see Meindl and Templ [2012]. We also used the R-packages *purrr*, *dplyr*, *slam*, *readr*, *tidyr, forcast, ggplot2, reshape2, tidyverse, ggridges, latex2exp, RColorBrewer*, *rcartocolor* and *foreign* for our implementation, see Mailund [2022], Yarberry and Yarberry [2021], Wickham et al. [2022], Hornik et al. [2022], Wickham et al. [2023], Hyndman et al. [2023], Wickham and Chang [2014], Wickham [2022], Wickham and RStudio [2022], Wilke [2022], Meschiari [2022], Neuwirth [2022], Nowosad [2022], R Core Team et al. [2022].

The computation was done for

- ALLBUS data set, see GESIS [2018], on a commodity desktop laptop with a 8-core Intel(R) Core(TM) i7-8665U CPU @ 1.90GHz processor and 16 GB RAM in R version 4.2.2.
- dermatology data set, see Demiroz et al. [1998] accessed via Dua and Graff [2017], on a commodity desktop computer with a 32-core Intel(R) Xeon(R) CPU E5-2650 v2 @ 2.60GHz processor and 64 GB RAM in R version 4.2.1
- German credit data set, see Dua and Graff [2017], on a commodity desktop laptop with a 8-core Intel(R) Core(TM) i7-10510U CPU @ 1.80GHz processor and 16 GB RAM in R version 4.2.2.

C CALCULATIONS FOR ROBUSTIFIED TEST STATISTICS

In Section 8 we show a graph visualizing the fraction of resamples in favor of **non**-rejection of H_0 (i.e., the p-values) as a function of the size of the contamination γ of the underlying linear-vacuous model (see Figure 3). We will briefly show here how the exact function is calculated. For general (polyhedral) credal sets, a resample I is in favor of rejection of H_0 under the robustified resampling scheme, if $\underline{d}^{\varepsilon}_{\mathbf{X},\mathbf{Y}}(\omega_0) > \overline{d}^{\varepsilon}_I$. Hence, the fraction of resamples in favor of rejection of H_0 is given by

$$\frac{1}{N} \cdot \sum_{I \in \mathcal{I}_N} \mathbb{1}_{\left\{\underline{d}^{\varepsilon}_{\mathbf{X},\mathbf{Y}}(\omega_0) > \overline{d}^{\varepsilon}_I\right\}}$$

where N denotes the number of resamples and $\mathcal{I}_N$ is the corresponding set of resamples. In the special case that the credal sets involved are γ-contamination models, we can use Proposition 6 (and a slight variation of it with π_* and π^* in reversed

[1] Further information on the survey and the data set itself can be found here: `https://search.gesis.org/research_data/ZA5240` (accessed: Febr 16, 2023)

[2] Further details can be found here: `https://www.gurobi.com/academia/academic-program-and-licenses/` (accessed: Febr 16, 2023)

roles) to obtain

$$\underline{d}^{\varepsilon}_{\mathbf{X},\mathbf{Y}}(\omega_0) = (1-\gamma)\cdot d^{\varepsilon}_{\mathbf{X},\mathbf{Y}}(\omega_0) - \gamma$$

and

$$\overline{d}^{\varepsilon}_{I} = (1-\gamma)\cdot d^{\varepsilon}_{I} + \gamma$$

and, therefore, the condition in the indicator above is satisfied if and only if

$$d^{\varepsilon}_{\mathbf{X},\mathbf{Y}}(\omega_0) - d^{\varepsilon}_{I} > \frac{2\gamma}{(1-\gamma)}.$$

Finally, if we interpret ε as a function parameter, then we can write the fraction of resamples in favor of **non**-rejection of H_0 (i.e., the observed p-values) as a function of the size γ of the contamination of the underlying linear-vacuous model:

$$f_{\varepsilon}(\gamma) := 1 - \frac{1}{N}\cdot \sum_{I\in\mathcal{I}_N} \mathbb{1}_{\left\{d^{\varepsilon}_{\mathbf{X},\mathbf{Y}}(\omega_0) - d^{\varepsilon}_{I} > \frac{2\gamma}{(1-\gamma)}\right\}}.$$

D FURTHER DETAILS ON THE APPLICATIONS

D.1 DATA SETS

We applied our analysis to three different data sets:

- For the poverty analysis, see Section 8, we used the ALLBUS data set. The data set is described by GESIS [2018] and Breyer and Danner [2015]. As mentioned already in the previous section, the data set is freely accessible, but only after registration in the corresponding online portal: `https://search.gesis.org/research_data/ZA5240` (accessed: 08.02.2023). Please download the file ZA5240_v2-2-0.sav (5.31MB) there.
  The analysis was done on a sample consisting of 100 female and 100 male observations.
- We analyzed the dermatology data set, see Demiroz et al. [1998] accessed via Dua and Graff [2017].
  The analysis was performed on a sample of 46 individuals with family history of eryhemato-squamous disease and 100 individuals without.
- We analyzed the German credit data set, see Dua and Graff [2017].
  The analysis was performed on a sample of 100 credit risks classified as good and 100 credit risks classified as poor individuals.

D.2 APPLICATION ON CREDIT DATA

We focus on three variables (features) in the German credit data set Dua and Graff [2017]: credit amount (numeric), credit history (ordinal, 5 levels ranging from "delay in paying off in the past" to "all credits paid back duly") and employment status (ordinal, 5 levels ranging from "unemployed" to "present employment longer than 7 years"). We use a subsample with $n = m = 100$ high-risk applicants and low-risk applicants each. We are interested in the hypothesis that high-risk applicants are dominated by low-risk applicants w.r.t. GSD. The test results (see Figures 1 and 2 in the supplementary material) can be interpreted analogously to Section 8: For $\varepsilon \in \{0.75, 1\}$ we reject for the common significance level of $\alpha \approx 0.05$. This time, we do not reject in case of $\varepsilon = 0.5$.

Similar to the example of poverty analysis in Section 8, rejecting H_0 does not necessarily mean that high-risk applicants are dominated by low-risk applicants. They could also be incomparable, see also Section 5. However, our tests with reversed variables give no evidence of incomparability: The observed p-values for all these reversed tests are all 1.

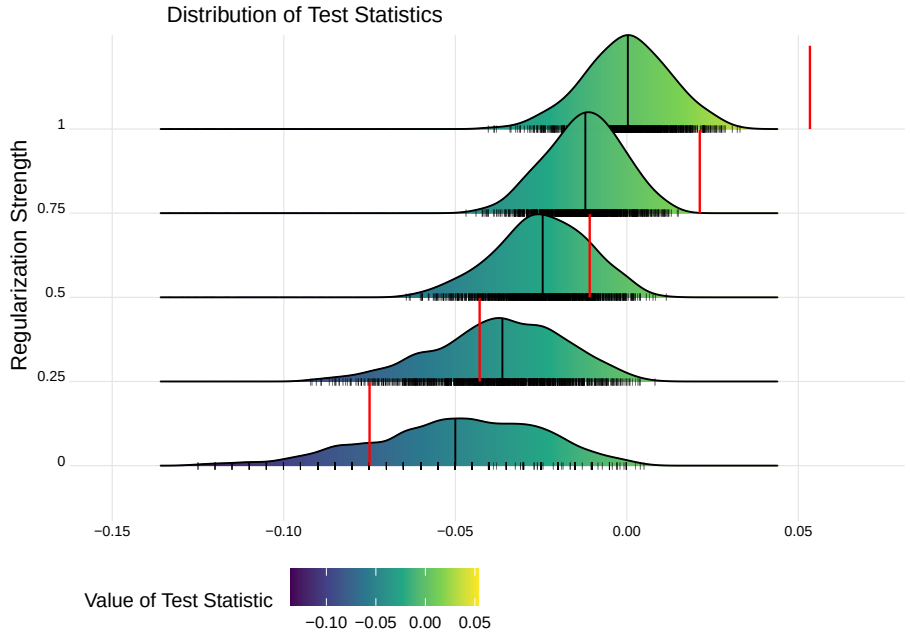

Figure 1: Distributions of d_I^ε with $\varepsilon \in \{0, 0.25, 0.5, 0.75, 1\}$ obtained from $N = 1000$ resamples of Credit data. Black stripes show exact positions of d_I^ε values. Vertical black line marks median. Red line shows value of the respective observed test statistics $d^\varepsilon_{\mathbf{X},\mathbf{Y}}(\omega)$.

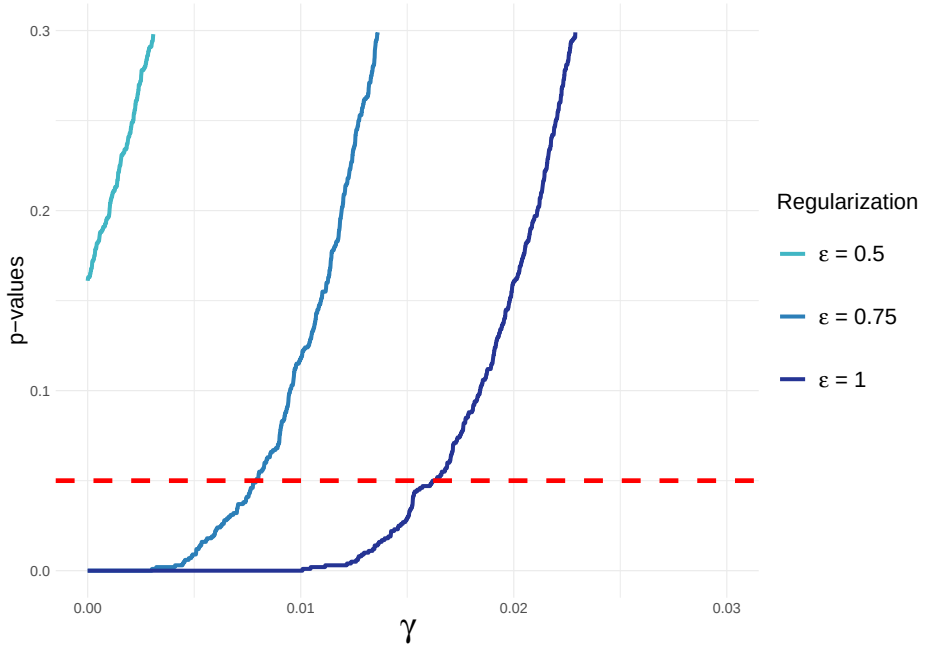

Figure 2: P-values as function of the contamination γ (see Supp. C) for tests with different regularization strength ε performed nd on credit data set. Dotted red line marks significance level $\alpha = 0.05$.

D.3 APPLICATION ON DERMATOLOGICAL DATA

We focus on three variables (features) in the dermatology data set Demiroz et al. [1998], Dua and Graff [2017]: age of skin (numeric), the intensity of itching (ordinal, 4 levels ranging from "no itching" to "strong itching") and erythema (redness of skin) (ordinal, 4 levels again ranging from no to highest intensity). We use a subsample with $n = 46$ patients with a family history of eryhemato-squamous disease and $m = 100$ without. We are interested in the hypothesis that patients without a family history of the disease are dominated by patients without a family history with respect to GSD. The test results (see Figures 3 and 4 in the supplementary material) can be interpreted analogously to Section 8: For $\varepsilon \in \{0.75, 1\}$ we again reject for the common significance level of $\alpha \approx 0.05$. However, the p-values are much higher than in the other two applications, see also Figure 4 (in the supplementary material).

Similar to the example of poverty analysis in Section 8, rejecting H_0 does not necessarily mean that patients with a family history of eryhemato-squamous disease are dominated by patients without. They could also be incomparable; see also Section 5. However, our tests with reversed variables give no evidence of incomparability: The observed p-values for all these reversed tests are all 1.

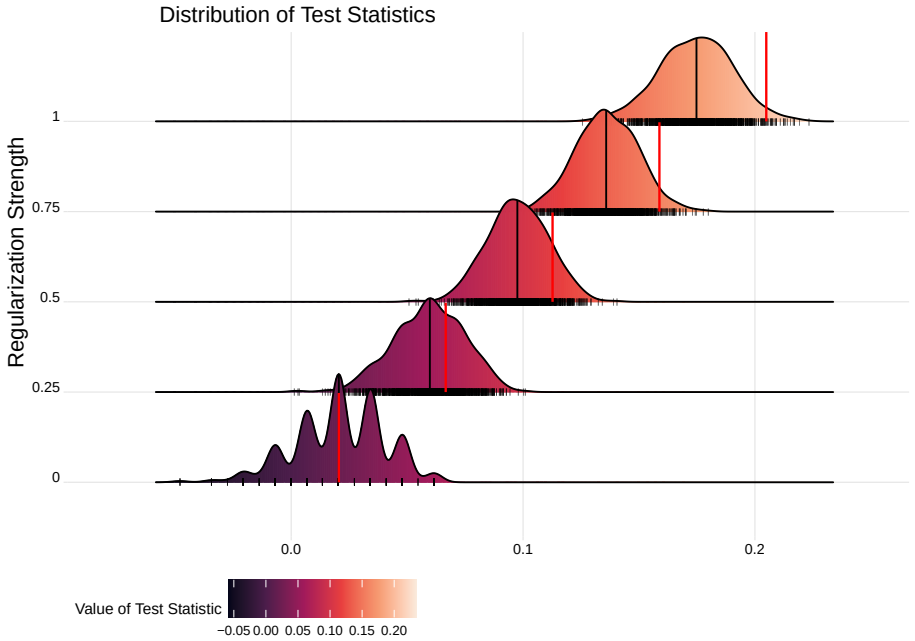

Figure 3: Distributions of d_I^ε with $\varepsilon \in \{0, 0.25, 0.5, 0.75, 1\}$ obtained from $N = 1000$ resamples of dermatology data. Black stripes show exact positions of d_I^ε values. Vertical black line marks median. Red line shows value of the respective observed test statistics $d^\varepsilon_{\mathbf{X},\mathbf{Y}}(\omega)$.

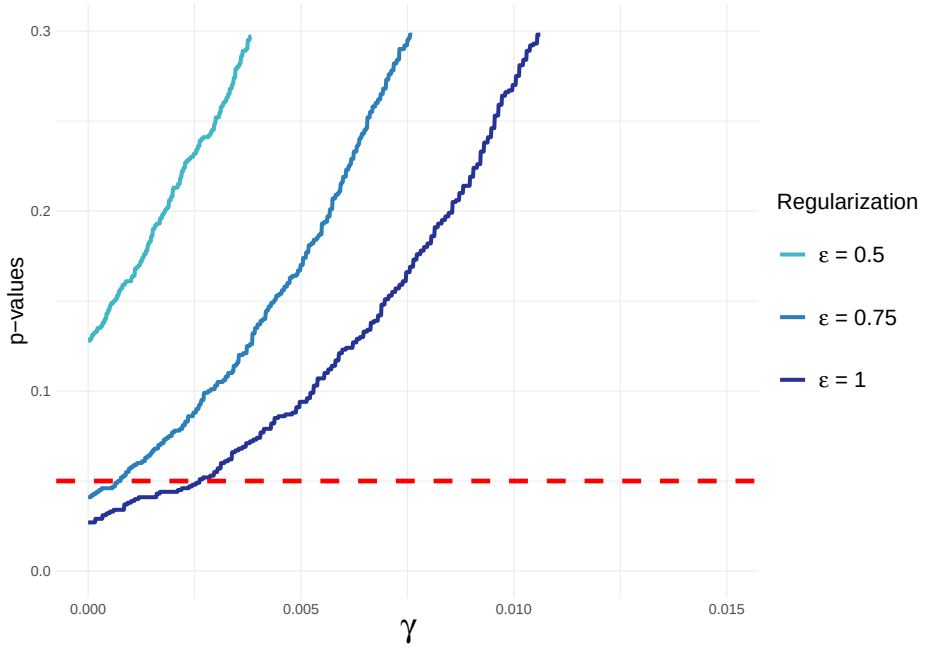

Figure 4: P-values as function of the contamination γ (see Supp. C) for tests with different regularization strength ε performed on Dermatology data set. The dotted red line marks significance level $\alpha = 0.05$.

References

Gurobi Optimization, LLC. Gurobi optimizer reference manual, 2020. URL `https://www.gurobi.com/wp-content/plugins/hd_documentations/documentation/9.0/refman.pdf`. [Online; Accessed 08.02.2023].

B. Breyer and D. Danner. Skala zur Erfassung des Lebenssinns (ALLBUS). In *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS) (GESIS – Leibniz-Institut für Sozialwissenschaften)*, volume 10, 2015.

G. Demiroz, H. Govenir, and N. Ilter. Learning differential diagnosis of Eryhemato-Squamous diseases using voting feature intervals. *Artif Intell Med*, 13:147–165, 1998.

D. Dua and C. Graff. UCI machine learning repository, 2017. `http://archive.ics.uci.edu/ml`.

GESIS. Allgemeine Bevölkerungsumfrage der Sozialwissenschaften ALLBUS 2014. GESIS Datenarchiv, Köln. ZA5240 Datenfile Version 2.2.0, https://doi.org/10.4232/1.13141, 2018.

K. Hornik, D. Meyer, and C. Buchta. Package 'slam', October 2022. URL `https://cran.r-project.org/web/packages/slam/slam.pdf`. [Online; Accessed 08.02.2023].

R. Hyndman, G. Athanasopoulos, C. Bergmeir, G. Caceres, L. Chhay, K. Kuroptev, M. O'Hara-Wild, F. Petropoulos S. Razbash, E. Wang, F. Yasmeen, F. Garza, D. Girolimetto, R. Ihaka, R Core Team, D. Reid, D. Shaub, Y. Tang, X. Wang, and Z. Zhou. Package 'forcast', January 2023. URL `https://cran.r-project.org/web/packages/forecast/forecast.pdf`. [Online: Accessed 09.02.2023].

D. Krantz, R. Luce, P. Suppes, and A. Tversky. *Foundations of Measurement. Volume I: Additive and Polynomial Representations*. Academic Press, 1971.

T. Mailund. Functional programming: purrr. In *R 4 Data Science Quick Reference: A Pocket Guide to APIs, Libraries, and Packages*, pages 89–110. Springer, 2022.

B. Meindl and M. Templ. Analysis of commercial and free and open source solvers for linear optimization problems. *Eurostat and Statistics Netherlands within the project ESSnet on common tools and harmonised methodology for SDC in the ESS*, 20, 2012.

S. Meschiari. Package 'latex2exp', November 2022. URL https://cran.r-project.org/web/packages/latex2exp/latex2exp.pdf. [Online: Accessed 09.02.2023].

E. Neuwirth. Package 'rcolorbrewer', October 2022. URL https://cran.r-project.org/web/packages/RColorBrewer/RColorBrewer.pdf. [Online: Accessed 09.02.2023].

J. Nowosad. Package 'rcartocolor', October 2022. URL https://cran.r-project.org/web/packages/rcartocolor/rcartocolor.pdf. [Online: Accessed 09.02.2023].

R Core Team, R. Bivand, V. Carey, S. DebRoy, S. Eglen, R. Guha, S. Herbrandt, N. Lewin-Koh, M. Myatt, M. Nelson, B. Pfaff, B. Quistorff, F. Warmerdam, S. Weigand, and Inc. Free Software Foundation. Package 'foreign', December 2022. URL https://cran.r-project.org/web/packages/foreign/foreign.pdf. [Online: Accessed 09.02.2023].

H. Wickham. Package 'reshape', October 2022. URL https://cran.r-project.org/web/packages/reshape/reshape.pdf. [Online: Accessed 09.02.2023].

H. Wickham and W. Chang. Package 'ggplot2', December 2014. URL https://cran.microsoft.com/snapshot/2015-01-06/web/packages/ggplot2/ggplot2.pdf. [Online: Accessed 09.02.2023].

H. Wickham and RStudio. Package 'tidyverse', October 2022. URL https://cran.r-project.org/web/packages/tidyverse/tidyverse.pdf. [Online: Accessed 09.02.2023].

H. Wickham, J. Hester, R. Francois, J. Bryan, and S. Bearrows. Package 'readr', October 2022. URL https://cran.r-project.org/web/packages/readr/readr.pdf. [Online; Accessed 08.02.2023].

H. Wickham, D. Vaughan, M. Girlich, K. Ushey, and PBC Posit. Package 'tidyr', January 2023. URL https://www.vps.fmvz.usp.br/CRAN/web/packages/tidyr/tidyr.pdf. [Online: Accessed 09.02.2023].

C. O. Wilke. Package 'ggridges', October 2022. URL https://cran.r-project.org/web/packages/ggridges/ggridges.pdf. [Online: Accessed 09.02.2023].

W. Yarberry and W. Yarberry. Dplyr. *CRAN Recipes: DPLYR, Stringr, Lubridate, and RegEx in R*, pages 1–58, 2021.

Contribution 7

Christoph Jansen, Georg Schollmeyer, Julian Rodemann, Hannah Blocher, and Thomas Augustin (2024). "Statistical Multicriteria Benchmarking via the GSD-Front". In: *38th Conference on Neural Information Processing System (NeurIPS 2024)*. Ed. by Amir Globerson, Lester Mackey, Angela Fan, Ulrich Paquet, Jakub Tomczak, Cheng Zhang, and Lam Nguyen. Vancouver: OpenReview.net

Statistical Multicriteria Benchmarking via the GSD-Front

Christoph Jansen[1,*]
c.jansen@lancaster.ac.uk

Georg Schollmeyer[2,*]
georg.schollmeyer@stat.uni-muenchen.de

Julian Rodemann[2,*]
julian@stat.uni-muenchen.de

Hannah Blocher[2,*]
hannah.blocher@stat.uni-muenchen.de

Thomas Augustin[2]
thomas.augustin@stat.uni-muenchen.de

[1]School of Computing & Communications
Lancaster University Leipzig
Leipzig, Germany

[2]Department of Statistics
Ludwig-Maximilians-Universität München
Munich, Germany

Abstract

Given the vast number of classifiers that have been (and continue to be) proposed, reliable methods for comparing them are becoming increasingly important. The desire for reliability is broken down into three main aspects: (1) Comparisons should allow for different quality metrics simultaneously. (2) Comparisons should take into account the statistical uncertainty induced by the choice of benchmark suite. (3) The robustness of the comparisons under small deviations in the underlying assumptions should be verifiable. To address (1), we propose to compare classifiers using a generalized stochastic dominance ordering (GSD) and present the GSD-front as an information-efficient alternative to the classical Pareto-front. For (2), we propose a consistent statistical estimator for the GSD-front and construct a statistical test for whether a (potentially new) classifier lies in the GSD-front of a set of state-of-the-art classifiers. For (3), we relax our proposed test using techniques from robust statistics and imprecise probabilities. We illustrate our concepts on the benchmark suite PMLB and on the platform OpenML.

1 Introduction

The comparison of classifiers in machine learning is usually carried out using *quality metrics* $\phi : \mathcal{C} \times \mathcal{D} \to [0, 1]$, i.e., bounded functions assigning a real number to every pair (C, D) of classifier and data set from a suitable domain $\mathcal{C} \times \mathcal{D}$, where, by construction, higher numbers indicate better quality. However, in many applications, the choice of a unique quality metric used for the comparison is not self-evident. Instead, competing quality metrics are available, each of which can be well-motivated but may lead to a different ranking of the analyzed classifiers. One attempt to safeguard against this effect is to use *multidimensional quality metrics*: instead of a single metric, one chooses a set of metrics $\Phi := (\phi_1, \ldots, \phi_n) : \mathcal{C} \times \mathcal{D} \to [0, 1]^n$ that – taken together – provide a balanced

[*]marks equal contribution.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

foundation for assessing the quality of classifiers. Generally, we distinguish two (related but) different motivations for choosing multidimensional quality metrics:

Performance is a latent construct: The application at hand suggests a very clear evaluation concept, which, however, is too complex to be expressed in terms of a single metric. In this case, the *latent* construct to evaluate is operationalized with a set of quality metrics (that serve as an approximation). For example, the latent construct of *robust accuracy* can be operationalized by taking together the following three quality metrics: *Accuracy* of a classifier (i.e., the proportion of correctly predicted labels), and *robustness* of this proportion under weak perturbations of the data in either the features or the target variable. This will be exemplified in Section 5.2 using the PMLB benchmark suite.

Quality is a multidimensional concept: Even if the application at hand suggests evaluation criteria that can be perfectly expressed using quality metrics, it can still be desirable to compare the classifiers under consideration in terms of various contentual dimensions. For example, one can be interested in how well a classifier performs in the trade-off between *accuracy* and *computation time* in the training and the test phase: Clearly distinguishable contentual dimensions are included and the analysis aims at investigating how the different classifiers under consideration trade-off between these dimensions. This will be exemplified in Section 5.1 using one of OpenML's benchmark suites.

Regardless of the motivation for considering multidimensional quality metrics, their interpretative advantage naturally comes at a price: Without further assumptions, classifiers will often be incomparable, as the quality metrics in the different dimensions contradict each other in their ranking.[2] Already on one data set, a multidimensional quality metric only induces a (natural yet potentially incomplete) *preorder*: a classifier is rated at least as good as a competitor if (and only if) it receives at least the same metric value in each dimension. The problem of incomparability becomes even more severe for multiple data sets (as considered here). In this case, one of the following analysis paths is often chosen: (I) An expected *weighted sum* (for example, weighted by importance) of the individual quality metrics is considered and the problem is then analyzed on this new pooled quantity.[3] (II) The problem is analyzed based on the *Pareto-front* $\text{par}(\Phi)$, i.e., the set of all classifiers that are not component-wise (strictly) dominated by any competitor, whose definition followed by an illustrative example are included for reference.

Definition 1. *Let $\tilde{\mathcal{D}} \subseteq \mathcal{D}$ be some set of data sets. The $\tilde{\mathcal{D}}$-**Pareto front** $par(\Phi, \tilde{\mathcal{D}})$ of Φ is given by*

$$\left\{C \in \mathcal{C} | \nexists C' \in \mathcal{C}\ \forall D \in \tilde{\mathcal{D}} :\ \Phi(C', D) \succcurlyeq \Phi(C, D)\right\},$$

where $\succcurlyeq$ is the strict part of the component-wise $\geq$-relation on $\mathbb{R}^n$. Set $par(\Phi) := par(\Phi, \mathcal{D})$.

Example 1. *Consider the following schematic example of three classifiers $\mathcal{C} = \{C_1, C_2, C_3\}$ evaluated for a fictitious population of four data sets $\mathcal{D} = \{D_1, D_2, D_3, D_4\}$. Every entry gives the two-dimensional evaluation $\Phi(C, D)$ of a classifier on a data set w.r.t. predictive accuracy and the computation time for training in three ordinal categories `fast`, `medium` and `slow`.*

classifier \ *data set*	D_1	D_2	D_3	D_4
C_1	$(0.7, \texttt{slow})$	$(0.8, \texttt{medium})$	$(0.9, \texttt{fast})$	$(0.95, \texttt{slow})$
C_2	$(0.75, \texttt{slow})$	$(0.85, \texttt{fast})$	$(0.91, \texttt{fast})$	$(0.96, \texttt{slow})$
C_3	$(0.99, \texttt{slow})$	$(0.91, \texttt{fast})$	$(0.85, \texttt{fast})$	$(0.75, \texttt{slow})$

Here, it holds that $\Phi(C_2, D_i) \succcurlyeq \Phi(C_1, D_i)$ for all $i = 1, 2, 3, 4$, i.e., C_1 is component-wise (strictly) dominated by C_2. Classifiers C_2 and C_3 are not component-wise (strictly) dominated. Thus, the Pareto-front is given by $par(\Phi) = \{C_2, C_3\}$.

Both approaches are extreme in a certain sense: (I) reduces the multidimensional information structure of the problem to one single real-valued score. Any selection of classifier based on this score will heavily depend on the choice of the weights in the sum score and, therefore, becomes dubious once

[2] This effect is usually more pronounced under the second motivation: Whereas in the first case the different metrics attempt to formalize the same latent construct, here different quality dimensions are actually to be covered. E.g., an improvement in accuracy may often be accompanied by a deterioration in computation time.

[3] There, one interprets the data sets as realizations of a random variable $T : \Omega \to \mathcal{D}$ on some probability space $(\Omega, \mathcal{S}, \pi)$, chooses weights $w_1, \ldots, w_n \in \mathbb{R}_+$ and assigns each $C \in \mathcal{C}$ the value $\sum_{i=1}^{n} w_i \mathbb{E}_\pi(\phi_i(C, T))$.

this choice is not perfectly justified. This seems even more severe for problems where some of the involved quality metrics might only allow for an ordinal interpretation, e.g., feature sparseness as a proxy for interpretability [75], risk levels in the EU AI act [50] or other regulatory frameworks [73], robustness (see experiments in Section 5.2) or runtime levels (Section 5.1). Opposed to this, (II) seems to be very conservative: By considering classifiers that are in the Pareto-front par(Φ), one (potentially) completely ignores both information encoded in the cardinally interpretable dimensions and information about the distribution of the data sets. As a trade-off between these two extremes, which utilizes the complete available information but avoids the choice of weights, it has recently been proposed to compare classifiers under multidimensional quality metrics using *generalized stochastic dominance (GSD)* [45]. The rough idea of this approach is to first embed the range of the multivariate performance measure in a special type of relational structure, a so-called *preference system*, which then allows for also formalizing the entire information originating from the cardinal dimensions of the quality metric. A classifier is then judged at least as good as a competitor (similar to classic stochastic dominance), if its expected utility is at least as high with respect to every utility function representing (both the ordinal and the cardinal parts of) the preference system (also see Definition 5). Although GSD also induces only a preorder, the set of not strictly dominated classifiers will generally be considerably smaller than under the classical Pareto analysis. Furthermore, it avoids potentially difficult to justify assumptions about the weighting of the different quality metrics. Therefore, working with the GSD-front, as introduced below, will prove to be a very promising analysis option; it combines the advantages of the conservative Pareto analysis with those of the liberal comparison of weighted sums.

1.1 Our contribution

GSD-Front: We introduce the concept of the GSD-front (see, after some preparations, Definition 6) and characterize it in Theorem 2 as more discriminative than the Pareto-front. In this sense, the GSD-front is an information-efficient way to handle the multiplicity/implicit multidimensionality of quality criteria, powerfully exploiting their ordinal and quantitative components.

Proper handling of statistical uncertainty; estimating and testing: Since typically the available data sets are just a sample of the corresponding universe, empirical counterparts of the major concepts are needed to do justice to the underlying statistical uncertainty. In particular, we give a sound inference framework: Firstly, we propose a set-valued estimator for the GSD-front and provide sufficient conditions for its consistency (see Theorem 1 and Remark 3). Secondly, we develop static and dynamic statistical permutation-tests if a classifier is in the GSD-front and prove their level-$\alpha-$validity and their consistency (see Theorem 3).

Robustification: Additionally, we recognize the fact that the underlying assumption of identically and independently distributed *(i.i.d.)* sampling is questionable in many benchmarking studies. Thus, in Section 4.2 we quantify how robust the test decisions are under such deviations.

Experiments with benchmark suites and implementation: We illustrate the concepts and corroborate their relevance with experiments run over two benchmark suites (PMLB and OpenML, see Section 5), based on an implementation that is freely available and easily adaptable to comparable problems.[4]. We consider experiments with *mixed-scaled* (ordinal and cardinal) multidimensional quality metrics, also incorporating (potentially) ordinal criteria.

1.2 Related work

Benchmarks are the foundations of applied machine learning research [27, 90, 78, 65, 91]. Specifically, benchmarking classifiers over multiple data sets is a much-studied problem in machine learning, as it enables practitioners to make informed choices about which methods to consider for a given data set. Furthermore, also proposals for novel classifiers must often first demonstrate their potential for improvement in benchmark studies. Examples include [58, 40, 31, 57, 12]. In recent years, in recognition of the fact that the benchmark suite under consideration is only a sample of data sets, especially focusing on *statistically significant* differences between classifiers has received great interest (see, e.g., [24, 35, 34, 19, 45] or, e.g., [9, 22, 8] for Bayesian variants). An R implementation of some of these tests is described in [15], whereas use-cases in the context of time series and

[4]Implementations of all methods and scripts to reproduce the experiments:`https://github.com/hannahblo/Statistical-Multicriteria-Benchmarking-via-the-GSD-Front`.

neural networks for regression are discussed in [44, 36]. The diversity and the associated problem of selecting quality metrics (e.g., [51]) is currently attracting a great deal of interest (e.g., [89]). Consequently, finding ways for comparing classifiers in terms of *multidimensional quality metrics* is intensively studied, ranging from multidimensional interpretability measures (e.g., [59]) over classical Pareto-analyses (e.g., [31]) to embeddings in the theory of data depth (e.g., [13, 71]). While utilizing variants of *stochastic dominance* in statistics is quite common (e.g., [56, 61, 6, 76, 67]), the same seems not to hold for machine learning. Exceptions include [23] in an optimization context, [47, 48], who investigate special types of stochastic orders, and [45], utilizing already GSD-relations for classifier comparisons without the GSD-front. Finally, relying on imprecise probabilities (e.g., [85, 86, 3]) to robustify statistical hypotheses follows the tradition of [66, 42, 41, 2], see also, e.g., [5, 25, 4, 60, 48]. For application to Bayesian networks, see, e.g, [55, 14, 54], and [81, 69, 18, 80, 1, 70, 52, 30, 16, 17], among others, for robustified machine learning in this spirit.

2 Decision-theoretic preliminaries

The relevant basic concepts in order theory are collected in Appendix A.1. Based on these we can make the following definition, originating from the decision-theoretic context discussed in [46].

Definition 2. *Let A be a non-empty set, $R_1 \subseteq A \times A$ a preorder on A, and $R_2 \subseteq R_1 \times R_1$ a preorder on R_1. The triplet $\mathcal{A} = [A, R_1, R_2]$ is then called a **preference system** on A. The preference system $\mathcal{A}' = [A', R_1', R_2']$ is called **subsystem** of $\mathcal{A}$ if $A' \subseteq A$, $R_1' \subseteq R_1$, and $R_2' \subseteq R_2$.*

In our context, R_1 formalizes the ordinal information, i.e., the information about the ranking of the objects in A, whereas R_2 describes the cardinal information, i.e., the information about the intensity of certain rankings. To ensure that R_1 and R_2 are compatible, we use a consistency criterion relying on the idea of simultaneous representability of both relations. For this, for a preorder R, we denote by I_R its indifference and by P_R its strict part (see A.1).

Definition 3. *The preference system $\mathcal{A} = [A, R_1, R_2]$ is **consistent** if there exists a **representation** $u : A \to \mathbb{R}$ such that for all $a, b, c, d \in A$ we have:*

i) $(a, b) \in R_1 \Rightarrow u(a) \geq u(b)$ *with equality iff* $(a, b) \in I_{R_1}$

ii) $((a, b), (c, d)) \in R_2 \Rightarrow u(a) - u(b) \geq u(c) - u(d)$ *with equality iff* $((a, b), (c, d)) \in I_{R_2}$

The set of all representations of $\mathcal{A}$ is denoted by $\mathcal{U}_\mathcal{A}$.

Finally, we need to recall the concept of *generalized stochastic dominance (GSD)* (see, e.g., [48]), which is crucial for the concepts presented in this paper: For a probability space $(\Omega, \mathcal{S}, \pi)$ and a consistent preference system $\mathcal{A}$, we define by $\mathcal{F}_{(\mathcal{A},\pi)}$ the set of all $X \in A^\Omega$ such that $u \circ X \in \mathcal{L}^1(\Omega, \mathcal{S}, \pi)$ for all $u \in \mathcal{U}_\mathcal{A}$. We then can define the GSD-preorder on $\mathcal{F}_{(\mathcal{A},\pi)}$ as follows.

Definition 4. *Let $\mathcal{A} = [A, R_1, R_2]$ be consistent. For $X, Y \in \mathcal{F}_{(\mathcal{A},\pi)}$, say X $(\mathcal{A}, \pi)$-**dominates** Y if $\mathbb{E}_\pi(u \circ X) \geq \mathbb{E}_\pi(u \circ Y)$ for all $u \in \mathcal{U}_\mathcal{A}$. Denote the induced **GSD-preorder** on $\mathcal{F}_{(\mathcal{A},\pi)}$ by $R_{(\mathcal{A},\pi)}$.*

3 GSD for classifier comparison

We return to the initial problem: Assume we are given a finite set $\mathcal{C}$ of classifiers, an arbitrary set $\mathcal{D}$ of data sets and n quality metrics $\phi_1, \ldots, \phi_n : \mathcal{C} \times \mathcal{D} \to [0, 1]$, combined to the multidimensional quality metric $\Phi := (\phi_1, \ldots, \phi_n) : \mathcal{C} \times \mathcal{D} \to [0, 1]^n$. As we also want to allow ordinal quality metrics, we assume that, for $0 \leq z \leq n$, the metrics $\phi_1, \ldots, \phi_z$ are of cardinal scale (differences may be interpreted), while the remaining ones are purely ordinal (differences are meaningless apart from sign). We embed the range $\Phi(\mathcal{C} \times \mathcal{D})$ of Φ in the following preference system:

$$\mathcal{P} = [[0, 1]^n, R_1^*, R_2^*] \text{ , where} \tag{1}$$

$$R_1^* = \Big\{(x, y) : x_j \geq y_j \; \forall j \leq n\Big\}, \text{ and } R_2^* = \left\{((x, y), (x', y')) : \begin{array}{l} x_j - y_j \geq x'_j - y'_j \;\; \forall j \leq z \\ x_j \geq x'_j \geq y'_j \geq y_j \;\forall j > z \end{array} \right\}.$$

R_1^* is the usual component-wise $\geq$-relation. For R_2^*, one pair of consequences is preferred to another if, in the ordinal dimensions, the exchange associated with the first pair is not a deterioration to the

exchange associated with the second pair and, in addition, there is component-wise dominance of the differences of the cardinal dimensions. In order to transfer the GSD-relation from Definition 4 to the case of comparing classifiers under multidimensional performance metrics, we interpret the data sets in $\mathcal{D}$ as realizations of a random variable $T : \Omega \to \mathcal{D}$ on some probability space $(\Omega, \mathcal{S}, \pi)$. We then associate each classifier $C \in \mathcal{C}$ with the random variable $\Phi_C := \Phi(C, T(\cdot))$ on Ω and compare classifiers by comparing the associated random variables by means of GSD.

Definition 5. *Denote by $\mathcal{P}_\Phi$ the preference system obtained by restricting $\mathcal{P}$ to $\Phi(\mathcal{C} \times \mathcal{D})$. Further, let $\mathcal{C}$ be such that $\{\Phi_C : C \in \mathcal{C}\} \subseteq \mathcal{F}_{(\mathcal{P}_\Phi, \pi)}$. For $C, C' \in \mathcal{C}$, say that C **dominates** C', abbreviated with $C \succsim C'$, whenever $(\Phi_C, \Phi_{C'}) \in R_{(\mathcal{P}_\Phi, \pi)}$.*

In the application situation, instead of the true GSD-order $\succsim$ among classifiers, we will often have to get along with its *empirical analogue*, i.e., the GSD-relation where a sample of data sets is treated like the underlying population and the true probability measure is replaced by the corresponding empirical ones. More precisely, we assume that we have sampled *i.i.d.* copies $T_1, \ldots, T_s$ of T and then define the set $Z_s := \big\{\Phi(C, T_i) : i \leq s \wedge C \in \mathcal{C}\big\}$, of (random) observations under the different classifiers. We then use $\mathcal{W}$ to denote the (random) subsystem of $\mathcal{P}$ that arises when $\mathcal{P}$ is restricted to the (random) set Z_s. For $C, C' \in \mathcal{C}$ we define the random variable

$$d_s(C, C') := \inf_{u \in \mathcal{U}_\mathcal{W}} \sum\nolimits_{z \in Z_s} u(z)(\hat{\pi}_C(\{z\}) - \hat{\pi}_{C'}(\{z\})),$$

where, for $M \subseteq [0,1]^n$, we set $\hat{\pi}_C(M) := \frac{1}{s}\big|\big\{i : i \leq s \wedge \Phi(C, T_i) \in M\big\}\big|$. For a concrete sample associated to $\omega_0 \in \Omega$, we then say that C *empirically GSD-dominates* C', if $d_s(C, C')(\omega_0) \geq 0$. Intuitively, d_s can thus be used to check whether the classifier C empirically dominates the classifier C' with respect to GSD in the samples at hand (i.e., in the benchmark suite under investigation).

Based on these concepts, we can now define the sets of (empirically) GSD-undominated classifiers.

Definition 6. *Let $\mathcal{C}$ be such that $\{\Phi_C : C \in \mathcal{C}\} \subseteq \mathcal{F}_{(\mathcal{P}_\Phi, \pi)}$. Let denote $T_1, \ldots, T_s$ i.i.d. copies of T.*

*i) The **GSD-front** is the set*

$$\mathit{gsd}(\mathcal{C}) := \big\{C \in \mathcal{C} : \nexists C' \in \mathcal{C} \text{ s.t. } C' \succ C\big\},$$

where $\succ$ denotes the strict part of $\succsim$.

*ii) Let $\varepsilon \in [0, 1]$. The ε-**empirical GSD-front** is the (random) subset of $\mathcal{C}$ defined by*

$$\mathit{egsd}_s^\varepsilon(\mathcal{C}) = \Big\{C : \nexists C' \in \mathcal{C} \text{ s.t. } \begin{array}{l} d_s(C', C) \geq -\varepsilon \\ d_s(C, C') < 0 \end{array}\Big\}.$$

Remark 1. *$\mathit{egsd}_s^0(\mathcal{C})$ is always non-empty. In contrast, $\mathit{egsd}_s^\varepsilon(\mathcal{C})$ may very well be empty if $\varepsilon > 0$. Note that choosing values of $\varepsilon > 0$ is intended to make $\mathit{egsd}_s^\varepsilon(\mathcal{C})$ less prone to sampling noise.*

Remark 2. *Some words on the semantics of the GSD-front: From a decision-theoretic point of view, classifier C strictly GSD-dominates classifier C' iff C has at least as high expected utility as C regarding any compatible utility representation of all the metrics considered, and stricly higher for at least one such utility. The GSD-front then simply collects all classifiers from $\mathcal{C}$ which are not strictly GSD-dominated by any competitor, i.e., which potentially can be optimal in expectation.*

Example 2. *Consider again the situation of Example 1 and recall that $\mathit{par}(\Phi) = \{C_2, C_3\}$ leaves C_2 and C_3 incomparable. However, if considering only the distribution of the (multivariate) performance of the classifiers (while assuming a uniform distribution over $\mathcal{D}$), C_3 is clearly dominating C_2 w.r.t. GSD: Matching dataset D_i with dataset D_{5-i} creates a (strict) pointwise dominance of C_3 over C_2 (where the strict dominance is due to D_1 and D_4). Thus, $\mathit{gsd}(\mathcal{C}) = \{C_3\} \subsetneq \mathit{par}(\Phi) = \{C_2, C_3\}$.*

The following two theorems show that the ε-empirical GSD-front fulfills two very natural requirements: First, under some regularity conditions, it is a consistent statistical estimator for the true GSD-front (Theorem 1). This is important because in practical benchmarking we almost never have access to the GSD-front of the whole population, i.e., the benchmarking results on all possible datasets from a specific problem class $\mathcal{D}$. Second, it is ensured that neither the ε-empirical nor the true GSD-front can ever become larger than the respective Pareto-front, irrespective of the choice of ε (Theorem 2). This is important as it guarantees our analysis does never conflict with, but is potentially more information-efficient than a Pareto-type analysis. Proofs are given in B.1 and B.2.

Theorem 1. *Denote by $\mathcal{I}_\Phi$ the set of all sets $\{a : u(a) \geq c\}$, where $c \in [0,1]$ and $u \in \mathcal{U}_{\mathcal{P}_\Phi}$. Assume that $\succsim$ is antisymmetric. If the VC-dimension[5] of $\mathcal{I}_\Phi$ is finite and if $\varepsilon : \mathbb{N} \to [0,1]$ converges to 0 with rate at most $\Theta(1/\sqrt[4]{s})$, then $(\mathrm{egsd}_s^{\varepsilon(s)}(\mathcal{C}))_{s\in\mathbb{N}}$ is a consistent statistical estimator, i.e.,*

$$\pi\Big(\Big\{\omega \in \Omega : \lim_{s\to\infty} \mathrm{egsd}_s^{\varepsilon(s)}(\mathcal{C}) = \mathrm{gsd}(\mathcal{C})\Big\}\Big) = 1,$$

where set convergence is defined via the trivial metric.

Remark 3. *The assumption of a finite VC dimension is only necessary to ensure that the ε-empirical GSD front does not become too large. In particular, the following does hold **without** this assumption:*

$$\pi\Big(\Big\{\omega \in \Omega : \lim_{s\to\infty} \mathit{egsd}_s^{\varepsilon(s)}(\mathcal{C}) \supseteq \mathit{gsd}(\mathcal{C})\Big\}\Big) = 1.$$

Thus, the ε-empirical GSD-front almost surely converges to a superset of the true GSD-front.

Theorem 2. *Assume $\mathcal{C}$ with $\{\Phi_C : C \in \mathcal{C}\} \subseteq \mathcal{F}_{(\mathcal{P},\pi)}$. Let further denote $T_1, \ldots, T_s$ i.i.d. copies of T and let $\varepsilon_1 \leq \varepsilon_2 \in [0,1]$. It then holds that i)* $\mathrm{gsd}(\mathcal{C}) \subseteq \mathrm{par}(\Phi)$. *Moreover, it holds that ii)* $\mathrm{egsd}_s^{\varepsilon_2}(\mathcal{C}) \subseteq \mathrm{egsd}_s^{\varepsilon_1}(\mathcal{C}) \subseteq \mathrm{par}(\Phi, \{T_1, \ldots, T_s\})$.

4 Statistical testing

We saw the ε-empirical GSD-front can be a consistent statistical estimator and that both the empirical and the true GSD-front are compatible with the Pareto-front. We now address statistical testing.

4.1 A test for the GSD-front

From now on, we make the (technical) assumption that the order $\succsim$ among the classifiers from $\mathcal{C}$ is additionally *antisymmetric*, transforming it from a preorder into a partial order.[6] Equipped with this assumption, we want to address the question how to *statistically test* if a given classifier $C \in \mathcal{C}$ is an element of the true GSD-front $\mathrm{gsd}(\mathcal{C})$. To achieve this, we formulate the question of actual interest as the alternative hypothesis of the test, i.e., we obtain the hypothesis pair:

$$H_0 : C \notin \mathrm{gsd}(\mathcal{C}) \quad \textbf{vs.} \quad H_1 : C \in \mathrm{gsd}(\mathcal{C})$$

A possible motivation for developing tests on the hypothesis pair $(H_0, \neg H_0)$ is the following: One would like to compare the quality of a newly developed classifier C for a problem class $\mathcal{D}$ with the classifiers in $\mathcal{C} \setminus \{C\}$ that are considered state-of-the-art for this problem class, see application in Section 5.2. If a suitable statistical test would allow the above null hypothesis to be rejected, then one could draw the conclusion (subject to statistical uncertainty) that the new classifier C on the problem class $\mathcal{D}$ could potentially improve the state-of-the-art. As first step, note that (under asymmetry) the null hypothesis H_0 can be equivalently rewritten as $H_0 : \exists C' \in \mathcal{C} \setminus \{C\} : C' \succsim C$. This reformulation makes obvious that H_0 is false if and only if *for every* $C' \in \mathcal{C} \setminus \{C\}$ the auxiliary hypothesis $H_0^{C'} : C' \succsim C$ is false. Statistical tests for hypothesis pairs of the form $(H_0^{C'}, \neg H_0^{C'})$ were proposed (in the context of statistical inequality analysis) in [48]: The authors there showed how exact statistical tests under *i.i.d.* sampling can be constructed by using a (non-parametric) permutation test based on a regularized version $d_s^\delta(C', C)$ of $d_s(C', C)$ as a test statistic. The strength of regularization of the test statistic is there controlled by a parameter $\delta \in [0,1]$, whose increase reduces the number of representation functions over which the infimum in the test statistic is formed, while equally attenuating all quality metrics.[7] Due to space limitations, we omit to recall an exact description of the testing scheme in the main text and instead refer to Appendix A.2.

The idea is then to replace the global test for $(H_0, \neg H_0)$ with $c := |\mathcal{C}| - 1$ tests of hypotheses $(H_0^{C'}, \neg H_0^{C'})$ and to reject the null hypothesis at significance level α if all tests reject their individual null hypotheses $H_0^{C'}$ at the same significance level α. Call this the **static GSD-test**. Clearly, this test tends to be conservative, as it ignores potential correlations of the test statistics for different pairs of classifiers. Moreover, a slightly modified test in the context of the GSD-front is directly

[5] The VC-dimension of a set system $\mathcal{S}$ is the largest cardinality of a set A with $2^A = \{A \cap S : S \in \mathcal{S}\}$.

[6] This is not very restrictive, it only assumes to consider classifiers that are not already equivalent w.r.t. GSD.

[7] In both applications in Section 5 the tests are based on the unregularized statistics $d_s^0(C', C)$, as the regularization performed in [48] aims at reaching a goal which is not primarily relevant for our paper (see A.2.5).

derivable: If one is rather interested in identifying the maximal subset $\mathcal{S}_{\max}$ of $\mathcal{C}$ for which C significantly lies in the GSD-front, i.e., in testing $\tilde{H}_0^{\mathcal{S}} : C \notin \text{gsd}(\mathcal{S})$ **vs.** $\tilde{H}_1^{\mathcal{S}} : C \in \text{gsd}(\mathcal{S})$ for all $\mathcal{S} \subseteq \mathcal{C}$ with $C \in \mathcal{S}$ *simultaneously*, the following alternative test is a statistically valid level-α test: First, perform all individual tests for $(H_0^{C'}, \neg H_0^{C'})$ with level $\frac{\alpha}{c}$. Then identify $\mathcal{S}_{\max}$ as the set of all classifiers from $\mathcal{C}$ for which the individual hypotheses are rejected. The (random) alternative hypothesis $\tilde{H}_1^{\mathcal{S}_{\max}} : C \in \text{gsd}(\mathcal{S}_{\max})$ is then statistically valid in the sense of being false only with a probability bounded by α. Call this the **dynamic GSD-test**. We have the following theorem, demonstrating that the proposed tests are indeed reasonable statistical tests (see B.3 for the proof).

Theorem 3. *Let the assumptions of Theorem 1 hold. Then, both the static and dynamic GSD-test are valid level-α tests. Additionally, both tests are consistent in the sense that under the corresponding alternative hypothesis, i.e., $H_1 : C \in gsd(\mathcal{C})$ resp. $\tilde{H}_1 : \exists \mathcal{S} \subseteq \mathcal{C} : C \in \mathcal{S}, |\mathcal{S}| \geq 2, C \in gsd(\mathcal{S})$, the probability of rejecting the corresponding null hypothesis converges to 1 as $s \to \infty$.*

4.2 Checking robustness under non-i.i.d.-scenarios

We argue that meaningful benchmark studies should abstain from treating the sample of data sets in the suite as a *complete survey*. That is, benchmark analyses should aim at statements about a well-defined population and regard the benchmark suite as a non-degenerate sample thereof. A major practical problem in this context is that often little is known about the inclusion criteria for data sets or test problems in the respective benchmark suite (see, e.g., the discussions in [83, 53, 37]). For instance, the popular platform OpenML [82] allows users to upload benchmark results for machine learning models with varying hyperparameters, harming representativity, see Section 5.1 and Appendix C.1. The absence of methods to randomly sample from the set of all problems or data sets is identified as an unsolved issue in [57, Section 2]. This calls the common *i.i.d.* sampling assumption into question, which our (and most other) tests are based upon, and raises the issue as to what extent statistically significant results depend on this assumption. We now address precisely this question.

In [48] it was shown how the binary tests on the hypothesis pairs $(H_0^{C'}, \neg H_0^{C'})$ discussed in Section 4.1 can be checked for robustness against deviations from the underlying *i.i.d.*-assumption. The idea here is to deliberately perturb the empirical distributions of the performances for the different classifiers and to analyze the permutation test used under the most extreme yet compatible worst-case. The perturbation of the empirical distribution is carried out here using a γ-contamination model (see, e.g., [85, p. 147]), which is widely used in robust statistics. We now want to adapt a similar robustness check for the global hypothesis pair $(H_0, \neg H_0)$ discussed here. For this, suppose we have a sample $T_1, \ldots, T_s$ of data sets (i.e., the benchmark suite). We further assume that $k \leq s$ of these variables (where it is not known which ones) are not sampled *i.i.d.*, but come from an arbitrary distribution about which nothing else is known. We then know, for every fixed $C \in \mathcal{C}$, that its associated true empirical measure $\hat{\pi}_C^{true}$ based on the true (uncontaminated) sample would have to be contained in

$$\mathcal{M}_C = \left\{(1 - \tfrac{k}{s})\hat{\pi}_C^{cont} + \tfrac{k}{s}\mu : \mu \text{ probability measure}\right\}, \tag{2}$$

where $\hat{\pi}_C^{cont}$ denotes the empirical measure based on the contaminated sample $T_1, \ldots, T_s$. Note that $\mathcal{M}_C$ is by definition a γ-contamination model with central distribution $\hat{\pi}_C^{cont}$ and contamination degree $\gamma := \frac{k}{s}$. In this setting, [48] show that to ensure that their permutation tests used for hypothesis pairs $(H_0^{C'}, \neg H_0^{C'})$ only advise rejection of the null hypothesis if this is justifiable for any empirical distribution compatible with the contaminated sample, i.e., for every combination of measures $(\pi_1, \pi_2) \in \mathcal{M}_C \times \mathcal{M}_{C'}$, one has to compare the most pessimistic value of the test statistic for the concrete sample at hand with the most optimistic value of the test in each of the resamples. Moreover, they show that the (approximate) *observed p-values* for a concrete contaminated sample $T_1(\omega_0), \ldots, T_s(\omega_0) \in \mathcal{D}$ associated with $\omega_0 \in \Omega$ of this robustified test can be expressed by a function in the number of contaminations k, given by

$$f_{(C',C)}(k) := 1 - \tfrac{1}{N} \cdot \sum\nolimits_{I \in \mathcal{I}_N} \mathbb{1}_{\left\{d_I^\delta - d_s^\delta(C',C)(\omega_0) > \frac{2k}{(s-k)}\right\}},$$

where N denotes the number of resamples, $\mathcal{I}_N$ is the corresponding set of resamples, and d_I^δ is the test statistic evaluated for the resample associated to I. Due to space limitations, we omit an exact description of the robustness check for the test on the hypothesis pairs $(H_0^{C'}, \neg H_0^{C'})$ as well as a derivation of the function $f_{(C',C)}$ in the main text and instead refer to Appendix A.3.

Similar as shown in Section 4.1, it is straightforward to calculate an (approximate) observed p-value for the static GSD-test for $(H_0, \neg H_0)$: We calculate the maximal observed p-value among all $C' \in \mathcal{C}\setminus\{C\}$, i.e. set $F_C(k) := \max\{f_{(C',C)}(k) : C' \in \mathcal{C}\setminus\{C\}\}$. The **robustified static GSD-test** for the degree of contamination k can be carried out as follows: Calculate $F_C(k)$ and reject H_0 if $F_C(k) \leq \alpha$. This indeed gives us a valid level-α-test for the desired global hypothesis $H_0 : C \notin \text{gsd}(\mathcal{C})$ under the additional freedom that up to k of the variables in the sample might be contaminated. Note, however, that also this test tends to be conservative as both performing the individual tests at level α as well as the adapted resampling scheme of the permutation test are worst-case analyses. Finally, also the **robustified dynamic GSD-test** can be obtained straightforwardly: Under up to k contaminations, the (random) alternative hypothesis $\tilde{H}_1^{\mathcal{S}_{\max}} : C \in \text{gsd}(\mathcal{S}_{\max})$ is statistically valid with level α if all individual robustified tests reject $H_0^{C'}$ at level $\frac{\alpha}{c}$, i.e., if $F_C(k) \leq \frac{\alpha}{c}$.

We end the section with a short comment on computation: The test statistics for the permutation test and the robustified variant can be calculated using linear programming. We are guided here by the linear programs proposed in [48, Propositions 4 and 5]. There are two computational bottlenecks in the actual evaluation: (1) the creation and storage of the constraint matrices of the linear programs and (2) the repeated need to solve large linear programs. An efficient, well-commented implementation that can be quickly transferred to similar applications is made available on GitHub (see Footnote 4).

5 Benchmarking experiments

We demonstrate our concepts on two well-established benchmark suites: OpenML [82, 11] and PMLB [64]. While for PMLB we compare classifiers w.r.t. the latent quality metric robust accuracy (see the first motivation in Section 1), for OpenML we use a multidimensional metric that includes accuracy and computation time as unidimensional metrics (see the second motivation in Section 1). The analysis of PMLB is kept short in the main text and detailed in Appendix C. Since the metrics in both applications are composed of one continuous and two (finitely) discrete metrics, we have (see B.4):

Corollary 1. *In both applications, the ε-empirical GSD-front is a consistent estimator for the true GSD-front (provided ε is chosen as in Theorem 1).*

5.1 Experiments on OpenML

We select 80 binary classification datasets (according to criteria detailed in Appendix C.1) from OpenML [82] to compare the performance of *Support Vector Machine* (SVM) with *Random Forest* (RF), *Decision Tree* (CART), *Logistic Regression* (LR), *Generalized Linear Model with Elastic net* (GLMNet), *Extreme Gradient Boosting* (xGBoost), and *k-Nearest Neighbors* (kNN).[8] Our multidimensional quality metric is composed of *predictive accuracy*, *computation time on the test data*, and *computation time on the training data*. Since the computation time depends strongly on the used computing environment (e.g. number of cores or free memory), we discretize the time-related metrics and treat them as ordinal. Accuracy is not affected by this and is therefore treated as cardinal. For details, see Appendix C.1. To gain a purely descriptive impression, we computed the empirical GSD relation. For this, we calculated $d_{80}(C, C')$ for $C \neq C' \in \mathcal{C} :=$ {SVM, RF, CART, LR, GLMNet, xGBoost, kNN} (see Hasse graph in Figure 2 in Appendix C.1). We see that CART (strictly) empirically GSD-dominates xGBoost, SVM, LR, and GLMNet. All other classifiers are pairwise incomparable. Three classifiers are not strictly empirically GSD-dominated by any other, namely RF, CART, and kNN. Thus, the 0-empirical GSD-front is formed by these. While at first glance this result might seem rather unexpected, a closer look on the performance evaluations provided by OpenML indeed confirms the dominance structure found, see Appendix C.1 for details.

To move to reliable inferential statements that take into account the statistical uncertainty, we exemplarily test (at level $\alpha = 0.05$) if SVM significantly lies in the GSD-front of some subset of $\mathcal{C}$. As described in Section 4.1, we therefore perform six pairwise permutation tests for the hypothesis pairs $(H_0^{C'}, \neg H_0^{C'})$ (where $C :=$ SVM and $C' \in \mathcal{C}\setminus\{\text{SVM}\}$) at level α in case of the **static GSD-test** or at level $\frac{\alpha}{6}$ in case of the **dynamic GSD-test**.[9] That is, we test six auxiliary null hypotheses each stating that SVM is GSD-dominated by kNN, xGBoost, RF, CART, LR, and GLMNet, respectively.

[8] For benchmarking deep learning classifiers or optimizers, we refer to future work discussed in Section 6.

[9] As explained in Footnote 7, we base the tests in Sections 5.1 and 5.2 on the unregularized $d_s^0(C', C)$.

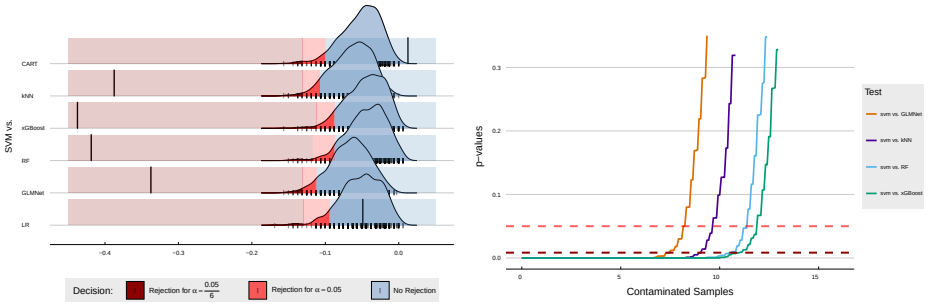

Figure 1: Left: Densities of resampled test statistics for pairwise permutation tests of SVM vs. six other classifiers on 80 datasets from OpenML. Big (small) vertical lines depict observed (resampled) test statistics. Rejection regions for the static (dynamic) GSD-test are highlighted red (dark red). Right: Effect of Contamination: p-values for pairwise tests of SVM versus GLMNet, kNN, RF and xGBoost.Red lines mark significance levels of $\alpha = 0.05$ (dark red: $\alpha = \frac{0.05}{6}$). Significance of SVM being in the GSD-front remains stable under contamination of up to 7 of 80 datasets.

The distributions of the test statistics are visualized on the left of Figure 1 (densities) and Figure 3 (CDFs) in C.1. They show that the pairwise tests of SVM versus kNN, xGBoost, RF, and GLMNet reject at level $\frac{\alpha}{6}$ and, thus, that SVM significantly (at level α) lies in the GSD-front of the subset of $\mathcal{C}$ composed of SVM and these four classifiers. In other words, we conclude that SVM is significantly ($\alpha = 0.05$) not outperformed by kNN, xGBoost, RF, and GLMNet regarding all compatible utility representation of accuracy, training and test runtime. Finally, as discussed in Section 4.2, we turn to the third aspect of reliability (besides multiple criteria and statistical uncertainty): We analyze how robust this test decision is under contamination of the benchmark suite, i.e., deviations from *i.i.d.*. The results are visualized on the right of Figure 1. It can be seen that the tests at level $\frac{0.5}{6}$ of SVM against GLMNet, kNN, RF and xGBoost cease to be significant from a contamination of (approximately) 7, 8, 11, and 11 of 80 data sets, respectively. That is, the results on up to 7, 8, 11, and 11 datasets could be arbitrarily redistributed, while maintaining significance of rejection. Since the significance of the dynamic GSD-test's decision depends on all pairwise tests being significant at level $\frac{0.5}{6}$, we can conclude that SVM would still have been significantly in the GSD-front of {SVM, kNN, xGBoost, RF, GLMNet}, even if 7 out of 80 data sets had been contaminated. Summing up, our proposed testing scheme not only allowed for meaningful statistical benchmarking of SVM versus competitors regarding accuracy, test time, and train time; it also enabled us to quantify as to what degree our conclusions remained stable under contamination of the benchmark suite.

Method comparison: The results highlight the advantages of the GSD-front over existing approaches. Applying *first-order stochastic dominance* (a special case of GSD where R_2^* is the trivial preorder) on the same set-up, yields that no classifier is significantly larger than (or incomparable to) any other classifier, based on a 5% significance level. This illustrates that the GSD-approach accounts for accuracy being a *cardinal* measure. In contrast, the *Pareto-front* here contains all considered classifiers. Thus, the Pareto front is much less informative than the GSD-front, which is also reflected in Theorem 2. Unlike the Pareto-front, the GSD-front is based on the distribution of the multidimensional quality metric and not only on the pairwise comparisons, and can use this knowledge to define the front. Thus, the GSD front is a balance between the conservative Pareto analysis and the liberal weighted sum comparison. Finally, we want to compare our method with an approach based on extending the test for single quality metrics proposed in [24] to the multiple metric setting. We therefore perform all possible single-metric tests as in [24] and define the *marginal front* as those classifiers that are not statistically significantly worse than another classifier on *all* metrics. However, this procedure can not be used to define a hypothesis test. Therefore, only a comparison with the empirical GSD-front is meaningful. For OpenML, this marginal front consists of all classifiers and is less exploratory than the empirical GSD-front. More details on the results of these other approaches and how these compare to the GSD front can be found in Appendix C.1.

5.2 Experiments on PMLB

We select 62 datasets from the Penn Machine Learning Benchmark (PMLB) suite [64] according to criteria explained in Appendix C.2. The following analysis shall exemplify how our proposed statistical tests can aid researchers in benchmarking newly developed classifiers against state-of-the-art ones. To this end, we compare a recently proposed classifier based on compressed rule ensembles of trees (CRE) [62] w.r.t. robust accuracy against five well-established classifiers, namely CART, RF, SVM with radial kernel, kNN and GLMNet. We operationalize the latent quality criterion of robust accuracy through i) classical accuracy (metric), ii) robustness of accuracy w.r.t. noisy features (ordinal), and iii) robustness of accuracy w.r.t. noisy classes (ordinal). Computation of i) is straightforward; in order to retrieve ii) and iii), we follow [92, 93] by randomly perturbing a share (here: 20 %) of both classes and features and computing the accuracy subsequently, as detailed in Appendix C.2. Since there exist competing definitions of robustness [43, 10, 72] and due to the share's arbitrary size, we treat ii) and iii) as ordinal and discretize the perturbated accuracy in the same way as for the runtimes in the openML experiments. Detailed results and visualization thereof can be found in Appendix C.2. In a nutshell, we find no evidence to reject the null of both the static and the dynamic GSD-test at significance level $\alpha = 0.05$. In particular, we do not reject any of the pairwise auxiliary tests for hypothesis pairs $(H_0^{C'}, \neg H_0^{C'})$ with $C :=$ CRE and $C' \in \mathcal{C} \setminus \{\text{CRE}\}$) for neither α nor $\frac{\alpha}{5}$. Our analysis hence concludes that we cannot rule out at significance level $\alpha = 0.05$ that the newly proposed classifier CRE is dominated by the five state-of-the-art classifiers w.r.t. all compatible utility representation of the latent criterion robust accuracy.

5.3 Additional recommendations for the end-user

We end the section with a few brief general notes for end-users of our benchmark methodology. This should make it easy to decide whether a GSD-based analysis is appropriate in a given use-case.

1. GSD-based studies do not primarily aim to identify the best algorithm for a given benchmark suite. Often, the GSD front contains more than one element. They are rather intended for checking whether a newly proposed classifier for a certain problem class can potentially improve on the state-of-the-art classifiers, or whether it disqualifies itself from the outset.
2. GSD-based studies allow statements with inferential guarantees by providing appropriate statistical tests: Assuming an *i.i.d.* benchmark suite, a judgment about an algorithm represents a statement about an underlying population and not just this specific suite.
3. GSD-based studies enable the robustness of the results to be quantified under the deviation from the *i.i.d.* assumption: It can be checked which share of the benchmark suite may be contaminated without affecting the obtained inferential statements.
4. GSD-based studies allow algorithms to be compared w.r.t. multiple metrics simultaneously. They enable the full exploitation of the information contained in differently scaled metrics.

6 Concluding remarks

Summary: We introduced the GSD-front for multicriteria comparisons of classifiers, gave conditions for its consistent estimability and proposed a statistical test for checking if a classifier belongs to it. We illustrated our concepts using two well-established benchmark suites. The results came with threefold reliability: They included several quality metrics, representation of statistical uncertainty, and a quantification of robustness under deviations from the assumptions.

Limitations and future research: Two specific limitations open promising avenues: 1.) *Comparing other types of algorithms:* We restricted ourselves to comparing classifiers. However, any situation in which objects are to be compared on the basis of different (potentially differently scaled) metrics over a random selection of instances can be analyzed using these ideas. For instance, applications of our framework to the multicriteria deep learning benchmark suite `DAWNBench` [21] or the bi-criteria optimization benchmark suite `DeepOBS` [74] appear straighforward. 2.) *Extension to regression-type analysis:* Analyses based on the GSD-front do not account for meta properties of the data sets. A straightforward extension to the case of additional covariates for the data sets is to stratify by these for the GSD-comparison. This would allow for a situation-specific GSD-analysis, presumably yielding more informative results.

Acknowledgements

We thank the anonymous reviewers and the area chair for providing valuable feedback. HB sincerely thanks the Evangelisches Studienwerk Villigst e.V. for the funding and support of her doctoral studies. Support by the Federal Statistical Office of Germany within the co-operation project "Machine Learning in Official Statistics" (JR and TA), by the Bavarian Academy of Sciences (BAS) through the Bavarian Institute for Digital Transformation (bidt, JR) and by the LMU Mentoring Program (JR and HB) is gratefully acknowledged.

References

[1] J. Abellan, C. Mantas, J. Castellano, and S. Moral-Garcia. "Increasing diversity in random forest learning algorithm via imprecise probabilities". In: *Expert Systems with Applications* 97 (2018), pp. 228–243.

[2] T. Augustin. "Neyman-Pearson testing under interval probability by globally least favorable pairs: Reviewing Huber-Strassen theory and extending it to general interval probability". In: *Journal of Statistical Planning and Inference* 105 (2002), pp. 149–173.

[3] T. Augustin, F. Coolen, G. de Cooman, and M. Troffaes, eds. *Introduction to Imprecise Probabilities*. Wiley, 2014.

[4] T. Augustin and G. Schollmeyer. "Comment: On focusing, soft and strong revision of Choquet capacities and their role in statistics". In: *Statistical Science* 36.2 (2021), pp. 205–209.

[5] T. Augustin, G. Walter, and F. Coolen. "Statistical Inference". In: *Introduction to Imprecise Probabilities*. Ed. by T. Augustin, F. Coolen, G. de Cooman, and M. Troffaes. Wiley, 2014, pp. 135–189.

[6] G. Barrett and S. Donald. "Consistent tests for stochastic dominance". In: *Econometrica* 71.1 (2003), pp. 71–104.

[7] D. Bates and M. Maechler. *Package 'Matrix'*. [Accessed: 13.05.2024]. 2010. URL: `http://cran.%20r-project.%20org/package=%20Matrix`.

[8] A. Benavoli, G. Corani, J. Demšar, and M. Zaffalon. "Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis". In: *Journal of Machine Learning Research* 18.77 (2017), pp. 1–36.

[9] A. Benavoli, G. Corani, and F. Mangili. "Should we really use post-hoc tests based on mean-ranks?" In: *Journal of Machine Learning Research* 17.1 (2016), pp. 152–161.

[10] D. Bertsimas, J. Dunn, C. Pawlowski, and Y. Zhuo. "Robust classification". In: *INFORMS Journal on Optimization* 1.1 (2019), pp. 2–34.

[11] B. Bischl, G. Casalicchio, M. Feurer, P. Gijsbers, F. Hutter, M. Lang, R. Mantovani, J. van Rijn, and J. Vanschoren. "OpenML: A benchmarking layer on top of OpenML to quickly create, download, and share systematic benchmarks". In: *NeurIPS – Track on Datasets and Benchmarks* (2021).

[12] B. Bischl, P. Kerschke, L. Kotthoff, M. Lindauer, Y. Malitsky, A. Fréchette, H. Hoos, F. Hutter, K. Leyton-Brown, K. Tierney, and J. Vanschoren. "ASlib: A benchmark library for algorithm selection". In: *Artificial Intelligence* 237 (2016), pp. 41–58.

[13] H. Blocher, G. Schollmeyer, C. Jansen, and M. Nalenz. "Depth functions for partial orders with a descriptive analysis of machine learning algorithms". In: *Proceedings of the Thirteenth International Symposium on Imprecise Probability: Theories and Applications*. Vol. 215. Proceedings of Machine Learning Research. PMLR, 2023, pp. 59–71.

[14] R. Cabanas, A. Antonucci, D. Huber, and M. Zaffalon. "CREDICI: A Java library for causal inference by credal networks". In: *International Conference on Probabilistic Graphical Models*. Ed. by M. Jaeger and T. Nielsen. Vol. 138. PMLR. 2020, pp. 597–600.

[15] B. Calvo and G. Santafé. "scmamp: Statistical comparison of multiple algorithms in multiple problems". In: *The R Journal* 8.1 (2016), pp. 248–256.

[16] M. Caprio, Y. Sale, E. Hüllermeier, and I. Lee. "A Novel Bayes' Theorem for Upper Probabilities". In: *Epistemic Uncertainty in Artificial Intelligence – First International Workshop, Epi UAI 2023, Pittsburgh, PA, USA, August 4, 2023, Revised Selected Papers*. Ed. by F. Cuzzolin and M. Sultana. Vol. 14523. Lecture Notes in Computer Science. Springer, 2024, pp. 1–12.

[17] M. Caprio, M. Sultana, E. Elia, and F. Cuzzolin. *Credal Learning Theory*. 2024. arXiv: 2402.00957. URL: https://arxiv.org/abs/2402.00957.

[18] Y. Carranza and S. Destercke. "Imprecise Gaussian discriminant classification". In: *Pattern Recognition* 112 (2021), p. 107739.

[19] L. Chang. "Partial order relations for classification comparisons". In: *Canadian Journal of Statistics* 48.2 (2020), pp. 152–166.

[20] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, M. Li, J. Xie, M. Lin, Y. Geng, Y. Li, and J. Yuan. *Package 'xgboost'*. [Accessed: 13.05.2024]. 2023. URL: https://cran.r-project.org/web/packages/xgboost/xgboost.pdf.

[21] C. Coleman, D. Narayanan, D. Kang, T. Zhao, J. Zhang, L. Nardi, P. Bailis, K. Olukotun, C. Ré, and M. Zaharia. "Dawnbench: An end-to-end deep learning benchmark and competition". In: *Training* 100.101 (2017), p. 102.

[22] G. Corani, A. Benavoli, J. Demšar, F. Mangili, and M. Zaffalon. "Statistical comparison of classifiers through Bayesian hierarchical modelling". In: *Machine Learning* 106.11 (2017), pp. 1817–1837.

[23] H. Dai, Y. Xue, N. He, Y. Wang, N. Li, D. Schuurmans, and B. Dai. "Learning to optimize for stochastic dominance constraints". In: *International Conference on Artificial Intelligence and Statistics*. Ed. by F. Ruiz, J. Dy, and J. van de Meent. Vol. 206. PMLR. 2023, pp. 8991–9009.

[24] J. Demšar. "Statistical comparisons of classifiers over multiple data sets". In: *Journal of Machine Learning Research* 7 (2006), pp. 1–30.

[25] S. Destercke, I. Montes, and E. Miranda. "Processing distortion models: A comparative study". In: *International Journal of Approximate Reasoning* 145 (2022), pp. 91–120.

[26] E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, and A. Weingessel. *Package 'e1071'*. [Accessed: 13.05.2024]. 2010. URL: https://cran.r-project.org/web/packages/e1071/e1071.pdf.

[27] D. Donoho. "Data Science at the Singularity". In: *2023 IMS International Conference on Statistics and Data Science (ICSDS)*. Ed. by R. Liu and A. Qu. 2023, p. 3.

[28] R. Dudley. "Central limit theorems for empirical measures". In: *The Annals of Probability* 6.6 (1978), pp. 899–929.

[29] R. Durrett. *Probability: Theory And Examples*. Vol. 49. Cambridge University Press, 2019.

[30] S. Dutta, M. Caprio, V. Lin, M. Cleaveland, K.J. Jang, I. Ruchkin, O. Sokolsky, and I. Lee. *Distributionally Robust Statistical Verification with Imprecise Neural Networks*. 2023. arXiv: 2308.14815 [cs.AI]. URL: https://arxiv.org/abs/2308.14815.

[31] M. Eugster, T. Hothorn, and F. Leisch. "Domain-based benchmark experiments: Exploratory and inferential analysis". In: *Austrian Journal of Statistics* 41.1 (2012), pp. 5–26.

[32] J. Friedman, T. Hastie, R. Tibshirani, B. Narasimhan, K. Tay, N. Simon, and J. Qian. *Package 'glmnet'*. [Accessed: 13.05.2024]. 2021. URL: https://cran.r-project.org/web/packages/glmnet/glmnet.pdf.

[33] M. Friedman. "The use of ranks to avoid the assumption of normality implicit in the analysis of variance". In: *Journal of the American Statistical Association* 32.200 (1937), pp. 675–701.

[34] S. García, A. Fernández, J. Luengo, and F. Herrera. "Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power". In: *Information Sciences* 180.10 (2010), pp. 2044–2064.

[35] S. García and F. Herrera. "An extension on "Statistical comparisons of classifiers over multiple cata sets" for all pairwise comparisons". In: *Journal of Machine Learning Research* 9 (2008), pp. 2677–2694.

[36] M. Graczyk, T. Lasota, Z. Telec, and B. Trawiński. "Nonparametric statistical analysis of machine learning algorithms for regression problems". In: *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*. Ed. by R. Setchi, I. Jordanov, R. Howlett, and L. Jain. Springer. 2010, pp. 111–120.

[37] N. Hansen, A. Auger, D. Brockhoff, and T. Tušar. "Anytime performance assessment in blackbox optimization benchmarking". In: *IEEE Transactions on Evolutionary Computation* 26.6 (2022), pp. 1293–1305.

[38] K. Hechenbichler and K. Schliep. *Weighted k-Nearest-Neighbor Techniques and Ordinal Classification*. Technical Report, LMU. 2004. URL: http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-1769-9.

[39] K. Hornik, C. Buchta, T. Hothorn, A. Karatzoglou, D. Meyer, and A. Zeileis. *Package 'rweka'*. [Accessed: 13.05.2023]. 2007. URL: https://cran.r-project.org/web/packages/RWeka/index.html.

[40] T. Hothorn, F. Leisch, A. Zeileis, and K. Hornik. "The design and analysis of benchmark experiments". In: *Journal of Computational and Graphical Statistics* 14.3 (2005), pp. 675–699.

[41] P. Huber. *Robust Statistics*. New York: Wiley, 1981.

[42] P. Huber. "The use of Choquet capacities in statistics". In: *Proceedings of the 39th Session of the International Statistical Institute* 45 (1973), pp. 181–191.

[43] S. Ishii and D. Ljunggren. "A comparative analysis of robustness to noise in machine learning classifiers". PhD thesis. KTH Royal Institute of Technology, 2021.

[44] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller. "Deep learning for time series classification: a review". In: *Data Mining and Knowledge Discovery* 33.4 (2019), pp. 917–963.

[45] C. Jansen, M. Nalenz, G. Schollmeyer, and T. Augustin. "Statistical comparisons of classifiers by generalized stochastic dominance". In: *Journal of Machine Learning Research* 24 (2023), pp. 1–37.

[46] C. Jansen, G. Schollmeyer, and T. Augustin. "Concepts for decision making under severe uncertainty with partial ordinal and partial cardinal preferences". In: *International Journal of Approximate Reasoning* 98 (2018), pp. 112–131.

[47] C. Jansen, G. Schollmeyer, and T. Augustin. "Multi-target decision making under conditions of severe uncertainty". In: *Modeling Decisions for Artificial Intelligence*. Ed. by V. Torra and Y. Narukawa. Springer, 2023, pp. 45–57.

[48] C. Jansen, G. Schollmeyer, H. Blocher, J. Rodemann, and T. Augustin. "Robust statistical comparison of random variables with locally varying scale of measurement". In: *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*. Ed. by R. Evans and I. Shpitser. Vol. 216. Proceedings of Machine Learning Research. PMLR, 2023, pp. 941–952.

[49] M. Kuhn. *Package 'caret'*. [Accessed: 13.05.2023]. 2015. URL: https://cran.r-project.org/web/packages/caret/index.html.

[50] J. Laux, S. Wachter, and B. Mittelstadt. "Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk". In: *Regulation & Governance* 18.1 (2024), pp. 3–32.

[51] N. Lavesson and P. Davidsson. "Evaluating learning algorithms and classifiers". In: *International Journal of Intelligent Information and Database Systems* 1 (2007), pp. 37–52.

[52] J. Lienen and E. Hüllermeier. "Credal Self-Supervised Learning". In: *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*. Ed. by MA. Ranzato, A. Beygelzimer, Y.N. Dauphin, P. Liang, and J.W. Vaughan. 2021, pp. 14370–14382.

[53] D. Mattos, L. Ruud, J. Bosch, and H. Holmström Olsson. *On the assessment of benchmark suites for algorithm comparison*. 2021. arXiv: 2104.07381 [cs.NE].

[54] D. Maua and F. Cozman. "Thirty years of credal networks: Specification, algorithms and complexity". In: *International Journal of Approximate Reasoning* 126 (2020), pp. 133–157.

[55] D. Maua and C. de Campos. "Editorial to: Special issue on robustness in probabilistic graphical models". In: *International Journal of Approximate Reasoning* 137 (2021), p. 113.

[56] D. McFadden. "Testing for stochastic dominance". In: *Studies in the Economics of Uncertainty*. Ed. by T. Fomby and T. Seo. Springer, 1989, pp. 113–134.

[57] O. Mersmann, M. Preuss, H. Trautmann, B. Bischl, and C. Weihs. "Analyzing the BBOB results by means of benchmarking concepts". In: *Evolutionary Computation* 23 (2015), pp. 161–185.

[58] D. Meyer, F. Leisch, and K. Hornik. "The support vector machine under test". In: *Neurocomputing* 55.1 (2003), pp. 169–186.

[59] C. Molnar, G. Casalicchio, and B. Bischl. "Quantifying model complexity via functional decomposition for better post-hoc interpretability". In: *Machine Learning and Knowledge Discovery in Databases*. Ed. by P. Cellier and K. Driessens. Springer International Publishing, 2020, pp. 193–204.

[60] I. Montes, E. Miranda, and S. Destercke. "Unifying neighbourhood and distortion models: Part II – new models and synthesis". In: *International Journal of General Systems* 49 (2020), pp. 636–674.

[61] K. Mosler. "Testing whether two distributions are stochastically ordered or not". In: *Grundlagen der Statistik und ihre Anwendungen: Festschrift für Kurt Weichselberger*. Ed. by H. Rinne, B. Rüger, and H. Strecker. Physica-Verlag, 1995, pp. 149–155.

[62] M. Nalenz and T. Augustin. "Compressed rule ensemble learning". In: *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. Ed. by G. Camps-Valls, F. Ruiz, and I. Valera. Vol. 151. Proceedings of Machine Learning Research. PMLR, 2022, pp. 9998–10014.

[63] P. Nemenyi. "Distribution-free Multiple Comparisons". PhD thesis. Princeton University, 1963.

[64] R. Olson, W. La Cava, P. Orzechowski, R. Urbanowicz, and J. Moore. "PMLB: a large benchmark suite for machine learning evaluation and comparison". In: *BioData Mining* 10 (2017), p. 36.

[65] S. Ott, A. Barbosa-Silva, K. Blagec, J. Brauner, and M. Samwald. "Mapping global dynamics of benchmark creation and saturation in artificial intelligence". In: *Nature Communications* 13.1 (2022), p. 6793.

[66] Huber P. and V. Strassen. "Minimax tests and the Neyman-Pearson lemma for capacities". In: *The Annals of Statistics* 1 (1973), pp. 251–263.

[67] T. Range and L. Østerdal. "First-order dominance: stronger characterization and a bivariate checking algorithm". In: *Mathematical Programming* 173 (2019), pp. 193–219.

[68] B. Ripley and W. Venables. *Package 'nnet'*. [Accessed: 13.05.2024]. 2016. URL: `https://staff.fmi.uvt.ro/~daniela.zaharie/dm2019/RO/lab/lab3/biblio/nnet.pdf`.

[69] J. Rodemann and T. Augustin. "Accounting for Gaussian process imprecision in Bayesian optimization". In: *International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making (IUKM)*. Springer. 2022, pp. 92–104.

[70] J. Rodemann, C. Jansen, G. Schollmeyer, and T. Augustin. "In all likelihoods: Robust selection of pseudo-labeled data". In: *Proceedings of the Thirteenth International Symposium on Imprecise Probability: Theories and Applications*. Ed. by E. Miranda, I. Montes, E. Quaeghebeur, and B. Vantaggi. Vol. 215. Proceedings of Machine Learning Research. PMLR, 2023, pp. 412–425.

[71] Julian Rodemann and Hannah Blocher. "Partial Rankings of Optimizers". In: *International Conference on Learning Representations (ICLR), Tiny Papers Track*. 2024.

[72] J. Sáez, J. Luengo, and F. Herrera. "Evaluating the classifier behavior with noisy data considering performance and robustness: The equalized loss of accuracy measure". In: *Neurocomputing* 176 (2016), pp. 26–35.

[73] L. Schmitt. "Mapping global AI governance: a nascent regime in a fragmented landscape". In: *AI and Ethics* 2.2 (2022), pp. 303–314.

[74] F. Schneider, L. Balles, and P. Hennig. "DeepOBS: A deep learning optimizer benchmark suite". In: *International Conference on Learning Representations*. 2018.

[75] L. Schneider, B. Bischl, and J. Thomas. "Multi-ojective optimization of performance and interpretability of tabular supervised machine learning Mmodels". In: *Proceedings of the Genetic and Evolutionary Computation Conference*. 2023, pp. 538–547.

[76] G. Schollmeyer, C. Jansen, and T. Augustin. *Detecting stochastic dominance for poset-valued random variables as an example of linear programming on closure systems*. [Accessed: 13.05.2024]. 2017. URL: `https://epub.ub.uni-muenchen.de/40416/13/TR_209.pdf`.

[77] M. Shaked and G. Shanthikumar. *Stochastic orders*. Springer, 2007.

[78] A. Shirali, R. Abebe, and M. Hardt. "A theory of dynamic benchmarks". In: *The Eleventh International Conference on Learning Representations*. 2023.

[79] T. Therneau, B. Atkinson, and B. Ripley. *Package 'rpart'*. [Accessed: 15.02.2023]. 2015. URL: `http://cran.ma.ic.ac.uk/web/packages/rpart/rpart.pdf`.

[80] L. Utkin. "An imprecise deep forest for classification". In: *Expert Systems with Applications* 141 (2020), p. 112978.

[81] L. Utkin and A. Konstantinov. "Attention-based random forest and contamination model". In: *Neural Networks* 154 (2022), pp. 346–359.

[82] J. Van Rijn, B. Bischl, L. Torgo, B. Gao, V. Umaashankar, S. Fischer, P. Winter, B. Wiswedel, M. Berthold, and J. Vanschoren. "OpenML: A collaborative science platform". In: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13*. Springer. 2013, pp. 645–649.

[83] H. Vandierendonck and K. De Bosschere. "Experiments with subsetting benchmark suites". In: *IEEE International Workshop on Workload Characterization, 2004. WWC-7. 2004*. 2004, pp. 55–62.

[84] V. Vapnik. *The Nature Of Statistical Learning Theory*. Springer, 1999.

[85] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. London: Chapman and Hall, 1991.

[86] K. Weichselberger. *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I: Intervallwahrscheinlichkeit als umfassendes Konzept [Elementary Foundations of a More General Calculus of Probability I: Interval Probability as a Comprehensive Concept]*. Physica, Heidelberg, 2001.

[87] H. Wickham, R. François, L. Henry, and K. Müller. *Package 'dplyr'*. [Accessed: 13.05.2024]. 2019. URL: `https://cran.r-project.org/web/packages/dplyr/index.html`.

[88] M. Wright and A. Ziegler. "ranger: A fast implementation of random gorests for high dimensional data in C++ and R". In: *Journal of Statistical Software* 77.1 (2017), pp. 1–17.

[89] B. Yu and K. Kumbier. "Veridical data science". In: *Proceedings of the National Academy of Science* 117.8 (2020), pp. 3920–3929.

[90] G. Zhang and M. Hardt. *Inherent Trade-Offs between Diversity and Stability in Multi-Task Benchmark*. 2024. arXiv: `2405.01719 [cs.LG]`.

[91] J. Zhang, M. Harman, L. Ma, and Y. Liu. "Machine learning testing: Survey, landscapes and horizons". In: *IEEE Transactions on Software Engineering* 48.1 (2020), pp. 1–36.

[92] X. Zhu and X. Wu. "Class noise vs. attribute noise: A quantitative study". In: *Artificial Intelligence Review* 22 (2004), pp. 177–210.

[93] X. Zhu, X. Wu, and Y. Yang. "Error detection and impact-sensitive instance ranking in noisy datasets". In: *Proceedings of the National Conference on Artificial Intelligence*. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999. 2004, pp. 378–384.

A Mathematical background

A.1 Basic definitions from order theory

A binary relation R on a set M is a subset of the Cartesian product of M with itself, i.e., $R \subseteq M \times M$. R is called *reflexive*, if $(a, a) \in R$, *transitive*, if $(a, b), (b, c) \in R \Rightarrow (a, c) \in R$, *antisymmetric*, if $(a, b), (b, a) \in R \Rightarrow a = b$, and *complete*, if $(a, b) \in R$ or $(b, a) \in R$ (or both) for arbitrary elements $a, b, c \in M$. A *preference relation* is a binary relation that is complete and transitive; a *preorder* is a binary relation that is reflexive and transitive; a *linear order* is a preference relation that is antisymmetric; a *partial order* is a preorder that is antisymmetric. If R is a preorder, we denote by $P_R \subseteq M \times M$ its *strict part* and by $I_R \subseteq M \times M$ its *indifference part*, defined by $(a, b) \in P_R \Leftrightarrow (a, b) \in R \wedge (b, a) \notin R$, and $(a, b) \in I_R \Leftrightarrow (a, b) \in R \wedge (b, a) \in R$.

A.2 Detailed description of the permutation test from Section 4.1

In this section we describe in detail the statistical test for the hypothesis pair $(H_0^{C'}, \neg H_0^{C'})$ discussed in Section 4.1 and first introduced in [48]. Moreover, we give further details on our proposed extension of this test to the global hypothesis pair $(H_0, \neg H_0)$ in both the static and the dynamic variant.

A.2.1 Preliminaries

Before we can describe the test from Section 4.1 in detail, we first need to recall two more definitions.

Definition 7. *Let $\mathcal{A} = [A, R_1, R_2]$ be a preference system. We call $\mathcal{A}$ **bounded**, if there exist $a_*, a^* \in A$ such that $(a^*, a) \in R_1$, and $(a, a_*) \in R_1$ for all $a \in A$, and $(a^*, a_*) \in P_{R_1}$.*

Definition 8. *Let $\mathcal{A} = [A, R_1, R_2]$ be a consistent and bounded preference system with a_*, a^* as before. Define*

$$\mathcal{N}_{\mathcal{A}} := \big\{u \in \mathcal{U}_{\mathcal{A}} : u(a_*) = 0 \ \wedge \ u(a^*) = 1\big\}.$$

For $\delta \in [0, 1)$, denote by $\mathcal{N}_{\mathcal{A}}^{\delta}$ the set of all $u \in \mathcal{N}_{\mathcal{A}}$ with

$$u(a) - u(b) \geq \delta \quad \wedge \quad u(c) - u(d) - u(e) + u(f) \geq \delta$$

for all $(a, b) \in P_{R_1}$ and for all $((c, d), (e, f)) \in P_{R_2}$.

We now start by describing an adapted version of the permutation test for the hypothesis pairs $(H_0^{C'}, \neg H_0^{C'})$ proposed in [48]. For a concrete realization of the *i.i.d.*-sample of data sets $D_1 := T_1(\omega_0), \ldots, D_s := T_s(\omega_0) \in \mathcal{D}$ with $s \in \mathbb{N}$ associated with $\omega_0 \in \Omega$, we define the set

$$(C, C')_{\omega_0} = \{\Phi(C, D_i) : i \leq s\} \cup \{\Phi(C', D_i) : i \leq s\} \cup \{\mathbf{0}, \mathbf{1}\},$$

where $\mathbf{0}$ is the vector containing n zeros and $\mathbf{1}$ is the vector containing n ones. Denote by $\mathcal{P}_{\omega_0}$ the restriction of $\mathcal{P}$ to $(C, C')_{\omega_0}$. It is then easy to verify that $\mathcal{P}_{\omega_0}$ is a consistent and bounded preference system with $a_* := \mathbf{0}$ and $a^* := \mathbf{1}$. For testing the hypothesis pair $(H_0^{C'}, \neg H_0^{C'})$ defined and discussed in Section 4.1 of the main text, we then use the following regularized test statistic for the specific sample induced by ω_0:

$$d_s^{\delta}(C', C)(\omega_0) := \inf_{u \in \mathcal{N}_{\mathcal{P}_{\omega_0}}^{\mu_\delta}} \sum_{z \in (C, C')_{\omega_0}} u(z) \cdot (\hat{\pi}_{C'}^{\omega_0}(\{z\}) - \hat{\pi}_{C}^{\omega_0}(\{z\}))$$

with $\delta \in [0, 1]$ and $\mu_\delta := \delta \cdot \sup\{\xi : \mathcal{N}_{\mathcal{A}_\omega}^{\xi} \neq \emptyset\}$, and $\hat{\pi}_{C}^{\omega_0}$ resp. $\hat{\pi}_{C'}^{\omega_0}$ are the emiprical probability measures of the performances of C resp. C' for the specific sample induced by ω_0.

A.2.2 Testing scheme for $(H_0^{C'}, \neg H_0^{C'})$

We denote our samples as follows:

$$\begin{aligned} \mathbf{x} &:= (x_1, \ldots, x_s) := (\Phi(C, D_1), \ldots, \Phi(C, D_s)) \\ \mathbf{y} &:= (y_1, \ldots, y_s) := (\Phi(C', D_1), \ldots, \Phi(C', D_s)) \end{aligned}$$

The concrete testing scheme for the permutation test for hypothesis pair $(H_0^{C'}, \neg H_0^{C'})$ then looks as follows:

Step 1: Take the pooled data sample: $\mathbf{w} := (w_1, \ldots, w_{2s}) := (x_1, \ldots, x_s, y_1, \ldots, y_s)$

Step 2: Take all $r := \binom{2s}{s}$ index sets $I \subseteq \{1, \ldots, 2s\}$ of size s. Evaluate $d_s^\delta(C', C)$ for $(w_i)_{i \in I}$ and $(w_i)_{i \in \{1,\ldots,n+m\} \setminus I}$ instead of $\mathbf{x}$ and $\mathbf{y}$. Denote the evaluations by d_I^δ.

Step 3: Sort all d_I^δ in increasing order to get $d_{(1)}^\delta, \ldots, d_{(r)}^\delta$.

Step 4: Reject $H_0^{C'}$ if $d_s^\delta(C', C)(\omega_0)$ is strictly smaller than $d_{(\ell)}^\delta$, with $\ell := \lfloor \alpha \cdot r \rfloor$ and α the significance level.

Note that, for large $\binom{2s}{s}$, we can approximate the above resampling scheme by computing d_I^δ only for a large number N of randomly drawn I. Moreover, note that only the *i.i.d.* assumption is needed for the above test to be valid.

A.2.3 Static GSD-test

As argued in the Section 4.1 of the main part of the paper, if we want to obtain a valid statistical test at the significance level $\alpha \in [0, 1]$ for hypothesis pair $(H_0, \neg H_0)$, we can simply perform all pairwise tests of hypothesis pairs $(H_0^{C'}, \neg H_0^{C'})$ at this same significance level α. We can then reject the hypothesis H_0 at level α if we can reject each hypothesis $H_0^{C'}$ at level α or, in other words, if

$$\min\left\{ \tfrac{1}{N} \cdot \sum\nolimits_{I \in \mathcal{I}_N} \mathbb{1}_{\left\{ d_s^\delta(C', C)(\omega_0) < d_I^\delta \right\}} : C' \in \mathcal{C} \setminus \{C\} \right\} \geq 1 - \alpha.$$

We call this the static GSD-test.

To see that this procedure indeed gives a valid level-α test for the global hypothesis pair $(H_0, \neg H_0)$, observe that – assuming H_0 to be true – the probability of H_0 being rejected equals the probability of *all* hypothesis $H_0^{C'}$ being rejected simultaneously. The latter probability – still assuming H_0 to be true – is obviously bounded from above by the probability that *one specific* hypothesis $H_0^{C^*}$ is rejected, which itself is bounded from above by the significance level α by construction.

A.2.4 Dynamic GSD-test

As discussed in the main text and reprinted here again for convenience of the reader, a slightly modified test in the context of the GSD-front is directly derivable: If one is rather interested in identifying the maximal subset $\mathcal{S}_{\max}$ of $\mathcal{C}$ for which C significantly lies in the GSD-front, i.e., in testing the hypothesis pairs $(\tilde{H}_0^{\mathcal{S}}, \neg \tilde{H}_0^{\mathcal{S}})$ for all $\mathcal{S} \subseteq \mathcal{C}$ *simultaneously*, the following alternative test would be a statistically valid level-α test: First, perform all individual tests for $(H_0^{C'}, \neg H_0^{C'})$ with level $\frac{\alpha}{c}$. Then identify $\mathcal{S}_{\max}$ as the set of all classifiers from $\mathcal{C}$ for which the individual hypotheses are rejected. The (random) alternative hypothesis $\tilde{H}_1^{\mathcal{S}_{\max}} : C \in \text{gsd}(\mathcal{S}_{\max})$ is then statistically valid in the sense of being false only with a probability bounded by α. We call this the dynamic GSD-test.

To see that this procedure indeed gives a valid level-α test for the (random) hypothesis pair $(\tilde{H}_0^{\mathcal{S}_{\max}}, \neg \tilde{H}_0^{\mathcal{S}_{\max}})$, observe that – under the null hypothesis – the probability of C lying in the GSD-front of some random subset $\mathcal{S}$ of $\mathcal{C}$ is bounded from above by the sum of probabilities of C lying in the GSD-front of $\{C, S\}$, where summation is over all $S \in \mathcal{S}$. As each of these probabilities is bounded from above by $\frac{\alpha}{c}$ by construction, the corresponding sum is bounded from above by $|\mathcal{S}| \cdot \frac{\alpha}{c}$. Finally, as $|\mathcal{S}| \leq c$, this gives the desired upper bound of α.

A.2.5 Computation and regularization

Note that the test statistic $d_s^\delta(C', C)(\omega_0)$ can be computed by solving a linear optimization problem (see [48, Proposition 4]) and, hence, the test just described is computationally tractable.

Moreover, note that in both applications in Section 5 the tests are based on the unregularized statistics $d_s^0(C', C)$, as the regularization performed in [48] aims at reaching a goal which is not primarily relevant for the present paper: The authors there are primarily interested in significantly detecting GSD of one variable over the other. Consequently, their regularization aims at making the test more sensitive for exactly this purpose. In contrast, in our study we are primarily interested in significantly detecting *incomparabilities* between variables, making the regularization by far less natural.

A.3 Detailed description of the robustness check from Section 4.2

In this section we describe in detail the robustification of the statistical test for the hypothesis pair $(H_0^{C'}, \neg H_0^{C'})$ discussed in Section 4.2 and first introduced in [48]. Moreover, we give further details on our proposed extension of this robustified statistical test to the global hypothesis pair $(H_0, \neg H_0)$ in both the static and the dynamic variant.

A.3.1 Preliminiaries

If we assume, as done in Section 4.2, that up to $k \leq s$ of the observations in our sample $T_1, \ldots, T_s$ might be contaminated and, accordingly, follow any arbitrary distribution, then we have to base the permutation test for hypothesis pair $(H_0^{C'}, \neg H_0^{C'})$ on a worst-case analysis of between the measures contained in $\mathcal{M}_C$ and $\mathcal{M}_{C'}$ defined instead of the true empirical measures of the two samples induced by the classifiers C and C'. If again $D_1 := T_1(\omega_0), \ldots, D_s := T_s(\omega_0) \in \mathcal{D}$ is a concrete (now potentially contaminated) sample associated with $\omega_0 \in \Omega$, and we again define $\mathbf{x}$ and $\mathbf{y}$ as in Section A.2, then the observed contamination models of C and C' look as follows:

$$\mathcal{M}_C(\omega_0) = \Big\{(1 - \tfrac{k}{s})\hat{\pi}_C^{cont,\omega_0} + \tfrac{k}{s}\mu : \mu \text{ probability measure}\Big\},$$
$$\mathcal{M}_{C'}(\omega_0) = \Big\{(1 - \tfrac{k}{s})\hat{\pi}_{C'}^{cont,\omega_0} + \tfrac{k}{s}\mu : \mu \text{ probability measure}\Big\}.$$

A.3.2 Testing scheme for robustified test on $(H_0^{C'}, \neg H_0^{C'})$

If we set

$$\overline{d_s^\delta}(C', C)(\omega_0) := \sup_{\pi_1 \in \mathcal{M}_{C'}(\omega_0), \pi_2 \in \mathcal{M}_C(\omega_0)} \left(\inf_{u \in \mathcal{N}_{\mathcal{P}_{\omega_0}}^{\mu_\delta}} \sum_{z \in (C,C')_{\omega_0}} u(z) \cdot (\pi_1(\{z\}) - \pi_2(\{z\})) \right),$$

$$\underline{d_s^\delta}(C', C)(\omega_0) := \inf_{\pi_1 \in \mathcal{M}_{C'}(\omega_0), \pi_2 \in \mathcal{M}_C(\omega_0)} \left(\inf_{u \in \mathcal{N}_{\mathcal{P}_{\omega_0}}^{\mu_\delta}} \sum_{z \in (C,C')_{\omega_0}} u(z) \cdot (\pi_1(\{z\}) - \pi_2(\{z\})) \right),$$

then the concrete testing scheme for the permutation test for hypothesis pair $(H_0^{C'}, \neg H_0^{C'})$ *under at most k contaminated sample members* looks as follows:

Step 1: Take the pooled data sample: $\mathbf{w} := (w_1, \ldots, w_{2s}) := (x_1, \ldots, x_s, y_1, \ldots, y_s)$

Step 2: Take all $r := \binom{2s}{s}$ index sets $I \subseteq \{1, \ldots, 2s\}$ of size s. Evaluate $\underline{d_s^\delta}(C', C)$ for $(w_i)_{i \in I}$ and $(w_i)_{i \in \{1,\ldots,n+m\} \setminus I}$ instead of $\mathbf{x}$ and $\mathbf{y}$. Denote the evaluations by $\underline{d_I^\delta}$.

Step 3: Sort all $\underline{d_I^\delta}$ in increasing order to get $\underline{d^\delta}_{(1)}, \ldots, \underline{d^\delta}_{(r)}$.

Step 4: Reject $H_0^{C'}$ if $\overline{d_s^\delta}(C', C)(\omega_0)$ is strictly smaller than $\underline{d^\delta}_{(\ell)}$, with $\ell := \lfloor \alpha \cdot r \rfloor$ and α the significance level.

The adapted testing scheme just described gives a valid (yet conservative) level-α-test for the hypothesis pair $(H_0^{C'}, \neg H_0^{C'})$ under at most k contaminated sample members.

Moreover, it directly follows from the discussions in Part C of the supplementary material to [48] that the (approximate) observed p-value of this test is given by

$$f_{(C',C)}(k) := 1 - \tfrac{1}{N} \cdot \sum_{I \in \mathcal{I}_N} \mathbb{1}_{\left\{d_I^\delta - d_s^\delta(C',C)(\omega_0) > \frac{2k}{(s-k)}\right\}},$$

where again N denotes the number of resamples, $\mathcal{I}_N$ is the corresponding set of resamples, and d_I^δ is the test statistic evaluated for the resample associated to I.

A.3.3 Robustified static GSD-test

As already argued in the main text, it is now easy to calculate an (approximate) observed p-value for our global hypothesis pair $(H_0, \neg H_0)$: We simply calculate the maximal observed p-value among all $C' \in \mathcal{C} \setminus \{C\}$, i.e. set

$$F_C(k) := \max\big\{f_{(C',C)}(k) : C' \in \mathcal{C} \setminus \{C\}\big\}.$$

The robustified test for the degree of contamination k can be carried out as follows: Calculate $F_C(k)$ and reject H_0 if $F_C(k) \leq \alpha$, i.e., if the maximal (approximate) p-value of the pairwise tests is still lower or equal than the significance level.

The argument that the testing procedure just described indeed produces a valid level-α test of the global hypothesis pair $(H_0, \neg H_0)$ under up to k contaminated data sets in the sample, can be carried out completely analogous as done in Appendix A.2.3.

A.3.4 Robustified dynamic GSD-test

Finally, as discussed in the main text and reprinted here again for convenience of the reader, also the robustified dynamic GSD-test can be obtained in a straightforward manner: Under up to k contaminated data sets in the sample, the (random) alternative hypothesis $\tilde{H}_1^{\mathcal{S}_{\max}} : C \in \text{gsd}(\mathcal{S}_{\max})$ from before is statistically valid with level α if all individual robustified tests reject $H_0^{C'}$ at level $\frac{\alpha}{c}$, i.e., if $F_C(k) \leq \frac{\alpha}{c}$.

The argument that the testing procedure just described indeed produces a valid level-α test for the (random) hypothesis pair $(\tilde{H}_0^{\mathcal{S}_{\max}}, \neg\tilde{H}_0^{\mathcal{S}_{\max}})$ under up to k contaminated data sets in the sample, can be carried out completely analogous as done in Appendix A.2.4.

A.3.5 Computation and regularization

Note that also the robustified test statistic $\overline{d_s^\delta}(C', C)(\omega_0)$ can be computed by solving a linear optimization problem (see [48, Proposition 6]) and, hence, the test just described is computationally tractable.

Again, note that the tests in Section 5 are based on the unregularized test statistics with $\delta = 0$. The reason for this is the same as discussed at the end of Appendix A.2.

B Proofs

B.1 Proof of Theorem 1

Proof. First, note that for $C, C' \in \mathcal{C}$, we have that $C \succsim C'$ if and only if

$$D(C, C') := \inf_{u \in \mathcal{U}_{\mathcal{P}_\Phi}} (\mathbb{E}_\pi(u \circ \Phi_C) - \mathbb{E}_\pi(u \circ \Phi_{C'})) \geq 0.$$

Thus, the GSD-front can equivalently be rewritten as

$$\text{gsd}(\mathcal{C}) = \left\{ C \in \mathcal{C} : \nexists C' \in \mathcal{C} \text{ s.t. } \begin{array}{l} D(C', C) \geq 0 \\ D(C, C') < 0 \end{array} \right\}.$$

Now, let $\varepsilon : \mathbb{N} \to [0, 1] : s \mapsto 1/\sqrt[4]{s}$. We show that:

$$C \in \text{gsd}(\mathcal{C}) \Rightarrow C \in \lim_{s \to \infty} \text{egsd}_s^{\varepsilon(s)}(\mathcal{C}) \quad \pi\text{-a.s. , and} \tag{3}$$

$$C \notin \text{gsd}(\mathcal{C}) \Rightarrow C \notin \lim_{s \to \infty} \text{egsd}_s^{\varepsilon(s)}(\mathcal{C}) \quad \pi\text{-a.s.} \tag{4}$$

Note that the proof immediately translates to the more general case of $\varepsilon(s) \in \Theta(1/\sqrt[4]{s})$ as stated in Theorem 1. Denote with $\hat{\mathbb{E}}$ the expectation w.r.t. the empirical measure associated with the i.i.d. sample[10] $(T_1, \ldots, T_s)$. For Implication (3), assume that $C \in \text{gsd}(\mathcal{C})$. Then for every other classifier C' there exists an utility function $u \in \mathcal{U}_{\mathcal{P}_\Phi}$ with $\mathbb{E}_\pi(u \circ \Phi_C) > \mathbb{E}_\pi(u \circ \Phi_{C'})$ (Otherwise we would have $D(C', C) \geq 0$ and $D(C, C') < 0$, where the second statement is due to antisymmetry). For these corresponding utility functions, because of the strong law of large numbers, we would get $d_s(C', C) \leq \hat{\mathbb{E}}(u \circ \Phi_{C'}) - \hat{\mathbb{E}}(u \circ \Phi_C) \xrightarrow{a.s.} c < 0$. Since $\mathcal{C}$ consists only of finitely many classifiers, $\text{egsd}_s^{\varepsilon(s)}(\mathcal{C})$ will almost surely not contain C asymptotically. Note that for Implication (3) to hold,

[10] Note that assuming only an exchangeable sample would also suffice. Note further that we have to assume the measurability of the involved infimum type statistics. For more details on this issue, see, e.g., [28].

it is only necessary that $\varepsilon(s)$ converges to zero as s goes to infinity. The order of convergency as $\Theta(1/\sqrt[4]{s})$ is only needed for Implication (4).

For Implication (4) assume that $C \notin \text{gsd}(\mathcal{C})$. Then there exists a classifier C' with $D(C', C) \geq 0$ and $D(C, C') < 0$. An analog argumentation like above shows that $d_s(C, C')$ converges almost surely to a value smaller than zero. It remains to analyze $D(C', C)$. For this, we have to show that $d_s(C', C) + \varepsilon(s) \xrightarrow{a.s.} c \geq 0$. We utilize uniform convergence: For arbitrary $\xi > 0$, [84, p. 192 Theorem 5.1] gives us

$$P\left(\sup_{u \in \mathcal{U}_{\mathcal{P}_\Phi}} \left|\mathbb{E}(u \circ \Phi_C) - \hat{\mathbb{E}}(u \circ \Phi_C)\right| > \xi\right) \leq 8\left(\frac{e \cdot 2s}{h}\right)^h \cdot \exp\left\{-\xi_*^2 s\right\},$$

where $\xi_* = \xi - 1/s$ and h is the VC dimension of $\mathcal{I}_\Phi$. The same holds for $\Phi_{C'}$. The triangle inequality then gives

$$P\left(\sup_{u \in \mathcal{U}_{\mathcal{P}_\Phi}} \left|\hat{\mathbb{E}}(u \circ \Phi_C) - \hat{\mathbb{E}}(u \circ \Phi_{C'})\right| > 2\xi\right) \leq 8\left(\frac{e \cdot 2s}{h}\right)^h \cdot \exp\left\{-\xi_*^2 s\right\}.$$

For $\varepsilon(s) = 1/\sqrt[4]{(1/s)}$ and s large enough, we have $\varepsilon_*(s) = \varepsilon(s) - 1/s \geq \varepsilon(s)/2$ and therefore

$$\begin{aligned} P\left(\sup_{u \in \mathcal{U}_{\mathcal{P}_\Phi}} \left|\hat{\mathbb{E}}(u \circ \Phi_C) - \hat{\mathbb{E}}(u \circ \Phi_{C'})\right| > 2\varepsilon(s)\right) &\leq 8\left(\frac{e \cdot 2s}{h}\right)^h \cdot \exp\left\{-\varepsilon_*(s)^2 s\right\} \\ &\leq 8\left(\frac{e \cdot 2s}{h}\right)^h \exp\left\{-\varepsilon(s)^2 s/4\right\}. \end{aligned}$$

This implies

$$P\left(\sup_{u \in \mathcal{U}_{\mathcal{P}_\Phi}} \left|\hat{\mathbb{E}}(u \circ \Phi_C) - \hat{\mathbb{E}}(u \circ \Phi_{C'})\right| > \varepsilon(s)/2\right) \leq 8\left(\frac{e \cdot 2s}{h}\right)^h \exp\left\{-\varepsilon(s)^2 s/64\right\} \quad (5)$$

$$= 8\left(\frac{e \cdot 2s}{h}\right)^h \exp\left\{-\sqrt{s}/64\right\}. \quad (6)$$

If the VC dimension h is finite, the term $8\left(\frac{e \cdot 2s}{h}\right)^h$ is polynomially growing in $\sqrt{s}$ (or s), whereas the term $\exp\left\{-\sqrt{s}/64\right\}$ is exponentially decreasing in $\sqrt{s}$ (or s). Therefore, the right hand side of Inequality (5) converges to zero, which shows that

$$\sup_{u \in \mathcal{U}_{\mathcal{P}_\Phi}} \left|\hat{\mathbb{E}}(u \circ \Phi_C) - \hat{\mathbb{E}}(u \circ \Phi_{C'})\right| - \varepsilon(s)$$

converges in probability to a value $c \leq 0$ or equivalently, that $d_s(C', C) + \varepsilon(s)$ converges to a value $c \geq 0$. Since the right hand side of Inequality (5) converges exponentially in s, the Borel-Cantelli theorem (cf., e.g., [29, p.67ff]) gives also strong convergency, which completes the proof. Note that it is not necessary to specify $\varepsilon(s)$ concretely as $1/\sqrt[4]{s}$. It would be sufficient to define $\varepsilon(s)$ as of the order of $\Theta(1/\sqrt[4]{s})$. □

B.2 Proof of Theorem 2

Proof. i) Assume that $C \notin \text{par}(\Phi)$. Then, by definition of $\text{par}(\Phi)$, there exists $C' \in \mathcal{C}$ such that for all $D \in \mathcal{D}$ it holds that $\Phi(C', D) \succcurlyeq \Phi(C, D)$. This implies that for all $D \in \mathcal{D}$ it holds that $(\Phi(C', D), \Phi(C, D)) \in P_{R_1^*}$. Now, choose $u \in \mathcal{U}_{\mathcal{P}_\Phi}$. Since u then, by definition, is strictly isotone with respect to $P_{R_1^*}$, this allows us to conclude that the function $u(\Phi(C', \cdot)) - u(\Phi(C, \cdot))$ is strictly positive, i.e., we have $u(\Phi(C', D)) - u(\Phi(C, D)) > 0$ for arbitrary $D \in \mathcal{D}$.

We compute:

$$\begin{aligned} \mathbb{E}_\pi(u \circ \Phi_{C'}) - \mathbb{E}_\pi(u \circ \Phi_C) &= \int_\Omega u(\Phi(C', T(\omega)))d\pi(\omega) - \int_\Omega u(\Phi(C, T(\omega)))d\pi(\omega) \\ &= \int_\Omega \underbrace{u(\Phi(C', T(\omega))) - u(\Phi(C, T(\omega)))}_{>0 \text{ for all } \omega \in \Omega, \text{ since } T(\omega) \in \mathcal{D}} d\pi(\omega) > 0 \end{aligned}$$

This gives $\mathbb{E}_\pi(u \circ \Phi_{C'}) > \mathbb{E}_\pi(u \circ \Phi_C)$. As u was chosen arbitrarily, this implies that $C' \succ C$. Hence, by definition of the GSD-front, we have $C \notin \text{gsd}(\mathcal{C})$.

ii) First, note that both postulates are statements involving random sets (i.e., sets dependent on the realizations of the variables $T_1, \dots, T_s$). Thus, we have to prove both statements for arbitrary realizations of these variables. So let $D_1 := T_1(\omega_0), \dots, D_s := T_s(\omega_0) \in \mathcal{D}$ be an arbitrary realisation. For this concrete realization of the sample, the first statement is immediate, since if there is no C' such that $d_s(C', C)(\omega_0) \geq -\varepsilon_2$ there is also no C' such that $d_s(C', C)(\omega_0) \geq -\varepsilon_1$ (as the latter is harder to satisfy due to $\varepsilon_1 \leq \varepsilon_2$).

Again for the chosen concrete realization of the sample, the second postulate is an immediate consequence of statement i) from above. As in both situations the realization of the variables was chosen arbitrarily, this implies the statement. □

B.3 Proof of Theorem 3

To see that the static test is a valid level-α test for the global hypothesis pair $(H_0, \neg H_0)$, observe that – assuming H_0 to be true – the probability of H_0 being rejected equals the probability of *all* hypothesis $H_0^{C'}$ being rejected simultaneously. The latter probability – still assuming H_0 to be true – is obviously bounded from above by the probability that *one specific* hypothesis $H_0^{C^*}$ is rejected, which itself is bounded from above by the significance level α by construction.

Furthermore, the reason for the consistency of the static test is the following: First, note that under the assumption of Theorem 1 (because of the finite VC dimension), we have that $d_s(C', C)$ converges to $D(C', C)$ in probability for every abrbitrary classifier $C' \neq C$. Therefore, for fixed C' and under the null hypothesis $H_0^{C'}$, we have $d_s(C', C)$ converges in probability to a value larger than or equal to zero. This implies that under this null hypothesis the implicit critical values of the permutation test become arbirarily close to a values larger than or equal to zero.

Now, let C be in the GSD-front. Then, due to antisymmetry of $\succsim$, for every other classifier C', there exists a utility for which the expectation of $u \circ \Phi_C$ is larger than the expectation of $u \circ \Phi_{C'}$. Because of the weak law of large numbers, this translates to the empirical expectations with an arbitrarily high probability if only the sample size is large enough. Therefore, all-together, as s converges to infinity, the test rejects the null hypothesis in this situation with arbitrary high probability. Finally, since we have only a finite number of hypothesis of the static test, this also translates to the static test itself. Therefore the static test is indeed a consistent level-α test.

To see that also the dynamic test is a valid level-α test for the (random) hypothesis pair $(\tilde{H}_0^{\mathcal{S}_{\max}}, \neg\tilde{H}_0^{\mathcal{S}_{\max}})$, observe that – under the null hypothesis – the probability of C lying in the GSD-front of some random subset $\mathcal{S}$ of $\mathcal{C}$ is bounded from above by the sum of probabilities of C lying in the GSD-front of $\{C, S\}$, where summation is over all $S \in \mathcal{S}$. As each of these probabilities is bounded from above by $\frac{\alpha}{c}$ by construction, the corresponding sum is bounded from above by $|\mathcal{S}| \cdot \frac{\alpha}{c}$. Finally, as $|\mathcal{S}| \leq c$, this gives the desired upper bound of α.

Finally, also the consistency of the dynamic test follows from the fact that it is constructed from a finite set of consistent tests for every possible set $\mathcal{S} \subseteq \mathcal{C}$. □

B.4 Proof of Corollary 1

Assume that $\Phi(\mathcal{C} \times \mathcal{D}) \subseteq M \times S_1 \times S_2$, where $S_1, S_2 \subset [0, 1]$ are finite, and $M \subseteq [0, 1]$ is arbitrary. This is possible since by definition of Φ we have $M \subseteq \phi_1(\mathcal{C} \times \mathcal{D})$ and $S_1 \subseteq \phi_2(\mathcal{C} \times \mathcal{D})$ and $S_2 \subseteq \phi_3(\mathcal{C} \times \mathcal{D})$, and the metrics ϕ_2 and ϕ_3 are assumed to be finitely discrete. We show that the width[11] of the restriction of R_1^* to $\Phi(\mathcal{C} \times \mathcal{D})$ is finite. It then follows directly from e.g. [76, Proposition 2] that the VC-dimension of

$$\mathcal{I}_\Phi := \left\{ \{a : u(a) \geq c\} : c \in [0, 1] \wedge u \in \mathcal{U}_{\mathcal{P}_\Phi} \right\}$$

is also finite. The claim then follows from Theorem 1.

[11] The *width* of a preordered set is the maximal cardinality of an antichain, i.e., the maximal number of pairwise incomparable elements.

To show the finiteness of the width, assume - wlog - that $|S_1| = g < \infty$ and $|S_2| = h < \infty$. Assume, for contradiction, that there exists an antichain[12] $Q \subseteq \Phi(\mathcal{C} \times \mathcal{D})$ within the restriction of R_1^* to $\Phi(\mathcal{C} \times \mathcal{D})$ of cardinality strictly greater than $g \cdot h$. Then there exist $x = (x_1, x_2, x_3), y = (y_1, y_2, y_3) \in Q$ such that $x_2 = y_2$ and $x_3 = y_3$ (as there are only $g \cdot h$ different combinations of the second and the third component). However, since the first component is completely ordered by $\geq$, this implies either (x, y) or (y, x) (or both) is contained in the restriction of R_1^* to $\Phi(\mathcal{C} \times \mathcal{D})$. This is a contradiction to x and y being elements of the same antichain Q, completing the argument. □

C Further results on the applications

This section provides further information on the benchmarking examples in Section 5.

C.1 Experiments with OpenML

This sections gives further insight to the example on the OpenML data analysed in Section 5.1. We start with giving more details on the data set with all the computation settings of the classifier algorithms. Afterwards, we provide more graphics and explanations of the analysis.

C.1.1 Data

Overall, we are comparing the performance of *Support Vector Machine* (SVM) to further 6 classifier algorithms on 80 data sets. The data sets as well as the performance evaluation is given by the OpenML library [82].[13] The analysis is restricted to binary classification problems. We selected those data sets of OpenML that evaluated the *predictive accuracy*, *train data time computation* and *test data time computation* (both measured in milliseconds) for all of the 7 algorithms. Since the computation times depend on the environment, i.e. the number of cores used or the free memory, we discretized the computation times and considered them as ordinal. Therefore, we divided each computation time into ten categories, where category one contains the 10% highest times, and so on. Moreover, we restricted our analysis on data sets with more than 450 and less than 10000 observations. This gives us in total 80 data sets.

The algorithms discussed are:

- *Support Vector Machine* (SVM) algorithm is implemented in the `e1071` library [26]
- *Random Forests* (RF) algorithm is implemented in the `ranger` library [88],
- *Decision Tree* (CART) algorithm is implemented via the `rpart` library [79],
- *Logistic regression* (LR) algorithm is implemented via the `nnet` library [68],
- *eXtreme Gradient Boosting* (xGBoost) algorithm is implemented in the `xgboost` library [20],
- *Elastic net* (GLMNet) algorithm is implemented through the `glmnet` library [32], and
- *k-nearest neighbors* (kNN) algorithm is implemebted via the `kknn` library [38].

C.1.2 Detailed results of the GSD-based analysis

We started our analysis in Section 5.1 by computing the empirical GSD-front. This gives the Hasse graph 2, where a top-down edge from C to C' states that $d_{80}(C, C') \geq 0$ holds.

In addition to the left of Figures 1 (densities of resampled test statistics) and the right of Figure 1 (effect of contamination on p-values) in the main paper, we include the cumulative distribution functions (CDFs) in Figure 3. Since we do not include the values of the observed test statistics here, the differences in distributions are visible to a greater extent. We observe the resampled test statistics' distributions for SVM vs. xGBoost and GLMNet to be left-shifted compared to SVM vs. CART, xGBoost, and LR. A visual analysis of the test decision, however, is not possible in the absence of the observed test statistics. This is why we include their values in the caption of Figure 3.

[12] An *antichain* of a preordered set (M, R) is a subset $A \subseteq M$ such that for all $m_1, m_2 \in A$ it holds $(m_1, m_2) \notin R$ and $(m_2, m_1) \notin R$.

[13] Last OpenML access: 24/10/2024

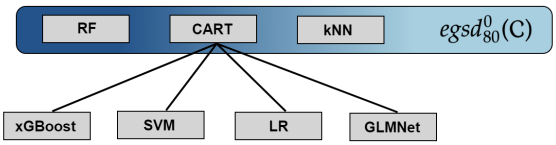

Figure 2: The blue shaded region symbolizes the 0-empirical GSD-front for the OpenML data sets.

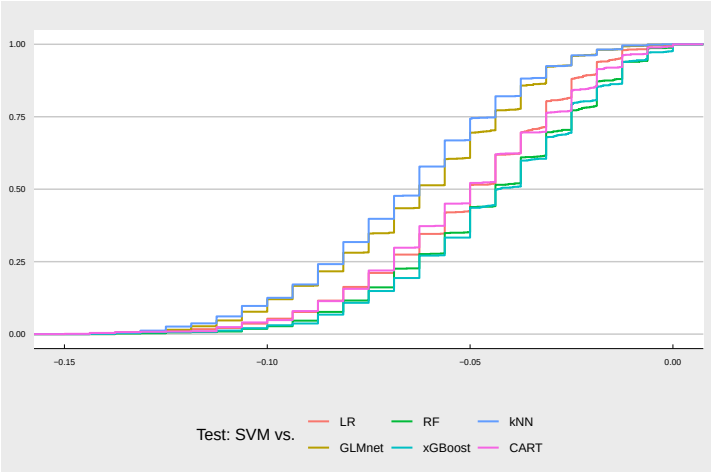

Figure 3: Cumulative Distribution Functions (CDFs) of resampled test statistics for hypothesis tests of SVM vs. LR, RF, kNN, GLMNet, xGBoost, and CART, respectively, on OpenML's benchmarking suite. As opposed to Figure 1 in the main paper, values of observed test statistics are not included. They are: 0.0125 (CART), -0.3875 (kNN), -0.4375 (xGBoost), -0.41875 (RF), -0.3375 (GLMNet), and -0.04897227 (LR). It becomes evident that the resampled test statistics' distributions for SVM vs. xGBoost and GLMNet are left-shifted compared to SVM vs. CART, xGBoost, and LR. This is also visible in Figure 1 in the main paper, albeit less clearly.

C.1.3 Detailed results of state-of-the-art analyses and comparison to GSD-front

This section provides the detailed computation of the state-of-the-art approaches and the comparison with the GSD approach. Here, we go step by step through all the methods touched in Section C.1.

First-order stochastic dominance Analogously to the GSD-front, one can define the front based on (multivariate) first-order stochastic dominance (see, e.g., [77]). Note that classical first-order stochastic dominance is a special case of our generalized stochastic dominance (GSD) in the case that all quality metrics are (treated as) of ordinal scale of measurement. Given the test logic followed by, for example, [6], for the OpenML data it turns out that no classifier is significantly stochastically larger than (or incomparable to) any other classifier (based on a significance level of 5 %). Compared to the results we obtain from our GSD-front analysis, this indiscriminative result is much less informative.

Pareto-front Both for the PMLB benchmark suite and the OpenML setup the Pareto-front contains all considered classifiers and is therefore not very informative. This shows the advantage of our approach because our approach is generally more informative (see Theorem 2). In particular, by using generalized stochastic dominance one refrains from solely relying on pointwise comparisons of classifiers over datasets. Instead one only looks at the distribution of the multidimensional quality metric. Beyond this, compared to both a Pareto analysis and a classical first order stochastic dominance analysis (see above), the GSD approach does justice to the fact that the dimension accuracy is cardinal and, at the same time, the fact that the other dimensions are of ordinal scale of measurement. Additionally, compared to an approach that only looks at the marginal distributions of every single quality metric separately, the GSD approach takes also the dependence structure

between the different quality metrics into account. This is of particular interest if one has different performance dimensions that are anticorrelated.

Weighted sum approach The GSD-based approach has advantages over an approach of weighted summation of the various quality metrics especially when it is not clear how specifically the weights are to be chosen. Specifically, each weighting leads to a total ordering among the classifiers under consideration. A clear best classifier can, therefore, be identified for each weighting. Thus, different weightings generally lead to different best classifiers. As a consequence, if one chooses a specific weighting, one should really be convinced that domain knowledge thoroughly justifies it, as even small changes in the weighting can completely change the resulting ranking. In contrast, the GSD-based approach can be used if no weighting of the involved metrics is available, but still more information (e.g., from the cardinal metrics) is available than required for a Pareto-type analysis. In summary, we emphasize that our method and the weighted summation should be used under different conditions and, therefore, complement rather than compete with each other.

Marginal-front A highly popular testing scheme for benchmark analysis is the one proposed by [24]. We compare our approach against using a marginal front that directly results from following this scheme. This marginal front is defined as a function of (a) statistical test(s), i.e. classifiers are in it depending on the test results. We emphasize that this front is not directly comparable to the GSD-front, since the GSD-front is a theoretical object (like the Pareto front) that can be used to formulate hypotheses that can then be tested by statistical tests like the ones proposed in this paper. Thus, we compare the marginal front to the *empirical* GSD front as reported in Fig. 1 of the paper for the application on OpenML.

We run multiple single-objective evaluations and include in the marginal-front the classifiers that are not statistically significantly worse than another classifier on all metrics. For the single-objective tests, we follow the well-established procedure of [24]. That is, we first run a global Friedman test (see [33]) for the null hypothesis that all classifiers have no differences with respect to the quality metric under investigation. In case we reject this null hypothesis, we can run post hoc Nemenyi pairwise tests (see [63]), comparing the performance of algorithms pairwise, with the null hypothesis being that there is no difference between their performances w.r.t. the multidimensional quality metric considered. We would like to emphasize that such an approach does not take into account the dependence structure among the quality metrics. In other words, it only considers the marginal distribution (hence the term *marginal-front*) of the classifiers w.r.t. the individual quality metrics separately, not their joint distribution. In the following, we conduct the suggested marginal analysis for OpenML w.r.t. the three-dimensional quality metric considered (accuracy, computation time on training data, computation time on test data):

Accuracy

- *Global Friedman Test:* Friedman rank sum test [33] rejects global null of no differences (p-value = 3.986e-14). This means we can conduct (two-sided) pairwise post hoc tests ($\alpha = 0.05$) with no difference as the null hypothesis.
- *Post Hoc Nemenyi Test:* Table 1 below shows the pairwise comparisons of algorithm performance with the Nemenyi test [63]. P-values below 0.05 are highlighted and indicate statistically significant differences in performance.

Table 1: Pairwise comparisons of algorithm performance with the Nemenyi test based on accuracy. Underlined values indicate differences significant at $\alpha = 0.05$ level.

	LR	RF	CART	SVM	xGBoost	GLMNet	kNN
RF	$\underline{3.9e-11}$	-	-	-	-	-	-
CART	0.19662	$\underline{6.9e-05}$	-	-	-	-	-
SVM	$\underline{0.00055}$	0.06513	0.55264	-	-	-	-
xGBoost	0.92896	$\underline{5.7e-08}$	0.85263	$\underline{0.03259}$	-	-	-
GLMNet	0.92341	$\underline{6.4e-08}$	0.86095	$\underline{0.03446}$	$\underline{1.00000}$	-	-
kNN	0.98454	$\underline{9.2e-09}$	0.68760	$\underline{0.01261}$	0.99995	0.99993	-

Computation time on training data

- *Global Friedman Test:* Friedman rank sum test [33] rejects the global null hypothesis of no differences (p-value < 2.2e-16). This means we can conduct pairwise post hoc tests ($\alpha = 0.05$) with the null hypothesis of no difference.
- *Post Hoc Nemenyi Test* [63], see Table 2.

Table 2: Pairwise comparisons of algorithm performance with the Nemenyi test based on computation time on the training data. Underlined values indicate differences significant at $\alpha = 0.05$ level.

	LR	RF	CART	SVM	xGBoost	GLMNet	kNN
RF	$9.1e-14$	-	-	-	-	-	-
CART	0.13788	$<2e-16$	-	-	-	-	-
SVM	$3.4e-05$	0.00037	$4.0e-12$	-	-	-	-
xGBoost	$5.9e-14$	0.97584	$<2e-16$	$5.1e-06$	-	-	-
GLMNet	0.03081	$5.1e-08$	$2.9e-07$	0.62723	$1.6e-10$	-	-
kNN	$1.3e-08$	$<2e-16$	0.00541	$5.8e-14$	$<2e-16$	$7.2e-14$	-

Computation time on test data

- *Global Friedman Test* Friedman rank sum test [33] rejects the global null hypothesis of no differences (p-value < 2.2e-16). This means we can conduct pairwise post hoc tests ($\alpha = 0.05$) with the null hypothesis of no difference.
- *Post Hoc Nemenyi Test* [63], see Table 3.

Table 3: Pairwise comparisons of algorithm performance with the Nemenyi test based on computing time on testing data. Underlined values indicate differences significant at $\alpha = 0.05$ level.

	LR	RF	CART	SVM	xGBoost	GLMNet	kNN
RF	$<2e-16$	-	-	-	-	-	-
CART	0.676	$<2e-16$	-	-	-	-	-
SVM	0.652	$6.5e-14$	0.019	-	-	-	-
xGBoost	$<2e-16$	0.996	$<2e-16$	$7.2e-14$	-	-	-
GLMNet	$3.2e-09$	$6.2e-06$	$1.2e-13$	$4.0e-05$	$1.9e-07$	-	-
kNN	$9.1e-14$	0.177	$6.8e-14$	$2.3e-12$	0.034	0.106	-

Table 4 provides the mean results of the classifier comparisons. (Recall that for train/test time: the lower, the better)

Table 4: Mean results of the classifier comparisons.

	LR	RF	CART	SVM	xGBoost	GLMNet	kNN
Accuracy	0.761	0.854	0.831	0.8113	0.820	0.763	0.789
Train Time	0.370	7.019	0.199	1.866	9.561	1.491	0.012
Test Time	0.062	0.458	0.055	0.106	0.407	0.184	0.291

As becomes evident from the mean values of the three quality metrics and the single-criterion test results presented above, there is no classifier that is significantly dominated by another classifer w.r.t. all three quality metrics. Hence, the marginal-front would contain all classifiers and would be rather indiscriminative compared to the empirical GSD-front that we present in the paper, see Figure 2 Appendix C.1, which contains random forest (RF), trees (CART), and k-nearest neighbor (kNN). This is in line with our explanation of OpenML results above. Since the quality metrics accuracy, train time, and test time are only weakly (if at all) correlated due to a trade-off between speed and accuracy, the marginal-front based on single-criterion comparisons does not facilitate practitioners' decision-making, while our empirical GSD-front provides valuable insights.

For the sake of completeness, we also report the results of these multiple single-objective evaluations on the PMLB benchmark suite in tables 5, 6, 7, and 8. The interpretation is completely analogous

Table 5: Post Hoc Nemenyi Test (Accuracy) on PMLB.

	cre	svmRadial	J48	ranger	knn	glmnet
svmRadial	0.74628	-	-	-	-	-
J48	0.78740	0.08257	-	-	-	-
ranger	0.00106	0.09912	2.2e-06	-	-	-
knn	0.00227	4.2e-06	0.13239	2.0e-13	-	-
glmnet	1.00000	0.67200	0.84844	0.00064	0.00360	-

Table 6: Post Hoc Nemenyi Test Summary (Accuracy with Noisy X) on PMLB.

	cre	svmRadial	J48	ranger	knn	glmnet
svmRadial	0.8130	-	-	-	-	-
J48	0.4971	0.0323	-	-	-	-
ranger	0.3063	0.9647	0.0019	-	-	-
knn	0.0072	3.8e-05	0.5290	5.1e-07	-	-
glmnet	0.7173	0.0826	0.9994	0.0067	0.3195	-

to the interpretation of the results on OpenML above. Note that the Friedman rank sum test rejects global null of no differences for all three criteria. This means we can conduct pairwise post hoc tests ($\alpha = 0.05$) with (two-sided) null of no difference.

C.1.4 Discussion of the unexpected results

Recall the discussion in Section 5.1 about the unexpected results. We want to emphasize that these have a high degree of originality and should be of particular interest to practitioners. This shows that experience and intuition with a method can also be misleading if only the evaluation framework is slightly modified: A multidimensional quality metric that seeks the optimal trade-off between different, potentially conflicting metrics will generally rank differently than a unidimensional one. In the following, we show that the dominance of CART over xGBoost, SVM, LR, and GLMNet is indeed consistent with the quality metrics provided by the OpenML repository.

First of all, here, we are interested in the trade-off between accuracy and computation time, (e.g., the better the accuracy, the higher/worse the computation time). We now look at the comparison between SVM and CART to demonstrate that the results are indeed in line with the data. We obtain:

- For 27 datasets, CART outperforms SVM on all dimensions (e.g., prediction accuracy, computation time on test data, and computation time on training data) at once.
- For 9 datasets, CART dominates SVM for at least one quality metric and for all other quality metrics the performance of CART is not worse.
- For 41 datasets, SVM's prediction accuracy is better than CART's. At the same time, CART outperforms SVM for at least one of the two computation times. The two classifiers are therefore incomparable for these datasets.
- For 3 datasets CART outperforms SVM based on accuracy, but at least one of the computation times of SVM is below the one of CART.

Overall, there exists no dataset where SVM dominates CART in all dimensions at once. Either the two classifiers are incomparable, or CART dominates SVM. Furthermore, CART dominates SVM

Table 7: Post Hoc Nemenyi Test Summary (Accuracy with Noisy Y) on PMLB.

	cre	svmRadial	J48	ranger	knn	glmnet
svmRadial	1.00000	-	-	-	-	-
J48	0.03722	0.04911	-	-	-	-
ranger	0.06405	0.04911	1.7e-07	-	-	-
knn	0.00096	0.00141	0.90728	2.3e-10	-	-
glmnet	0.67200	0.73193	0.68732	0.00031	0.12513	-

Table 8: Mean Results (Accuracy and Noisy Data) on PMLB

	cre	svmRadial	J48	ranger	knn	glmnet
Accuracy	0.7807	0.8494	0.8347	0.8629	0.7780	0.8106
Accuracy with noisy x	0.7307	0.7823	0.7679	0.7924	0.7339	0.7570
Accuracy with noisy y	0.7346	0.7776	0.7638	0.7984	0.7237	0.7640

for nearly half of the datasets (27 + 9 = 36 of 80). Thus, the dominance structure provided by our method is in line with the performance evaluation values provided by OpenML.

A second issue that may have influenced the unexpected performance structure obtained in the paper is the way performance is evaluated by OpenML. OpenML is based on the uploads of its users. Each user is free to decide which hyperparameter settings to use. Thus, as there might be a different goal on the hyperparameter setting in each dataset, the results are not representative for the best performance of each algorithm. This aspect should be included in any further discussion. Especially since some algorithms are more dependent on hyperparameter settings/tuning than others. For an example involving hyperparameter tuning that is fixed in advance, see Section 5.2.

C.2 Experiments with PMLB

This sections give further insight to the exemplary benchmarking analysis on the Penn Machine Learning Benchmarks (PMLB) in Section 5.2 in the main paper. We start by giving more details on the data sets with all the computation settings of the classifier algorithms. Afterwards, we provide more figures and explanations of the analysis.

C.2.1 Data

Penn Machine Learning Benchmarks (PMLB) is a collection of curated benchmark datasets for evaluating and comparing supervised machine learning algorithms [64]. We select all datasets from PMLB for binary classification tasks with 40 to 1000 observations[14] and less than 100 features. On these 62 datasets[15], a recently proposed classifier based on compressed rule ensemble learning [62] is compared w.r.t. robust accuracy against five well-established classifiers, namely classification tree (CART), random forest (RF), support vector machine with radial kernel (SVM), k-nearest neighbour (kNN), and generalized linear model with elastic net (GLMNet). In detail, we deploy

- *Support Vector Machine* (SVM) algorithm as implemented in the `e1071` library [26]
- *Random Forests* (RF) algorithm as implemented in the `ranger` library [88], requiring `e1071` library [26] and `dplyr` [87]
- *Decision Tree* (CART) algorithm (C4.5-like trees) as implemented in the `RWeka` library [39],
- *Elastic net* (GLMNet) algorithm is implemented through the `glmnet` library [32] requiring library `Matrix` [7], and
- *k-nearest neighbors* (kNN) algorithm as implemented in the `kknn` library [38].

Note that we used the respective methods in the `caret` library [49] for hyperparameter tuning and cross-validation to retrieve i) through iii), as detailed below.

We operationalize the latent quality criterion of robust accuracy through i) classical accuracy (metric), ii) robustness w.r.t. noisy features (ordinal), and iii) robustness w.r.t. noisy classes (ordinal). Computation of i) is straightforward; in order to retrieve ii) and iii), we follow [92, 93] by randomly perturbing a share (here: 20 %) of the data points. We randomly selected data points with a selection probability of 20% and replaced the values by a random draw from the marginal distribution of the corresponding variable. (This is a slight difference to [92, 93] who replaced the data points by a random draw from a uniform distribution of the corresponding support of the marginal distribution.)

We then tune the six classifiers' hyperparameters on a (multivariate) grid of size 10 following [49] for each of the 62 datasets and eventually compute i) to iii) through 10-fold cross validation.

[14] [49] requires at least 4 data points in the test set, which translates to a mininmal n of 40, since we deploy 10-fold cross validation.

[15] Last access of PMLB: 12/05/24.

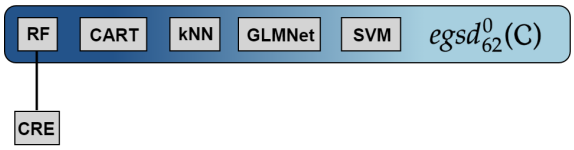

Figure 4: Hasse graph of the empirical GSD-relation for the PMLB data sets. The blue shaded region symbolizes the 0-empirical GSD-front, see Definition 6 ii).

C.2.2 Detailed results of the GSD-based analysis

To initially obtain a purely descriptive overview, we construct the Hasse graph illustrating the empirical GSD relation. In this process, we calculate the value $d_{62}(C, C')$ for $C \neq C' \in \mathcal{C} := \{\text{CRE, SVM, RF, CART, GLMNet, kNN}\}$ and connect C to C' with a top-down edge whenever $d_{62}(C, C') \geq 0$. The resulting graph is portrayed in Figure 4. It is evident from the graph that RF (strictly) empirically GSD-dominates the classifier CRE. All other classifiers are pairwise incomparable. Five classifiers, namely RF, CART, kNN, GLMNet, and SVM are not strictly empirically GSD-dominated by any other considered classifier and, thus, form the 0-empirical GSD-front.

This latter purely descriptive analysis already hints at the CRE not belonging to the GSD-Front. In order to transition to inferential statements, we aim to statistically test (at level $\alpha = 0.05$) whether CRE significantly lies in the GSD-front of some subset of $\mathcal{C}$. As detailed in Section 4.1, we conduct five pairwise tests for the hypothesis pairs $(H_0^{C'}, \neg H_0^{C'})$ (where $C :=$ CRE and $C' \in \mathcal{C} \setminus \{\text{CRE}\}$) at a level of $\frac{\alpha}{5}$, as explained in Section 4.[16] In other words, we test five auxiliary null hypotheses, each asserting that CRE is GSD-dominated by SVM, RF, CART, GLMNet, and kNN, respectively.

The results of these tests are visualized in Figure 5 (densities) and Figure 6 (cumulative distribution functions).[17] They indicate that the pairwise tests of CRE versus SVM, RF, CART, GLMNet, and kNN do not reject at a level of $\frac{\alpha}{5}$ nor at level α. Hence, we conclude that based on the observed benchmark results we cannot conclude at significance level $\alpha = 0.05$ that CRE lies in the GSD-front of any subset of $\mathcal{C}$. In other words, we have no evidence to rule out that CRE is in the GSD-front, i.e., we cannot confirm based on the data that CRE is not outperformed by SVM, RF, CART, GLMNet, and kNN with respect to all compatible utility representation of robust accuracy. As can be seen in Figure 5, testing CRE vs. CART results in the smallest p-value of all pairwise tests, which appears plausible, since CRE is a CART-based method. On the other hand, the observed test statistic of CRE vs. RF is far away from the critical value and the test cleary does not reject, even though RF is also a tree-based method.

Finally, as discussed in Section 4.2, we further analyze the robustness of this test decision under contamination of the benchmark suite, i.e., deviations from the *i.i.d.*-assumption. As opposed to our OpenML analysis in Section 5.1, see also Appendix C.1, contamination does not affect the test decisions here, since none of the tests rejects already for uncontaminated samples. Increasing contamination only drives p-values further. The results are visualized in Figure 7. It is observed that the tests are neither significant at a level of $\frac{0.05}{5}$ nor at 0.05 and this clearly does not change with growing size of contaminated benchmark data sets.

In summary, the PMLB experiments demonstrated how to apply our benchmarking framework to the problem of comparing a newly proposed classifier to a set of state-of-the-art ones. Furthermore, it illustrated our tests' applications to multiple criteria of mixed scales (ordinal and cardinal) that operationalize a latent performance measure, namely robust accuracy. It became evident that our framework allows to statistically assess whether the novel classifier CRE can compete with existing ones - that is, whether CRE lies in the GSD-front of some state-of-the-art classification algorithms. In

[16] As clarified in Footnote 7, the tests in Sections 5.1 and 5.2 are based on the unregularized test statistics $d_s^0(C', C)$.

[17] For generating these plots, we used quantile functions from both base r and the ggplot library. As the underlying quantile functions definitions differed slightly, we relied on the latter and corrected the quantiles from the other manually. Detailed documentation of all computations involved in generating the visualizations in the paper, we refer the interested reader to `https://github.com/hannahblo/Statistical-Multicriteria-Benchmarking-via-the-GSD-Front`.

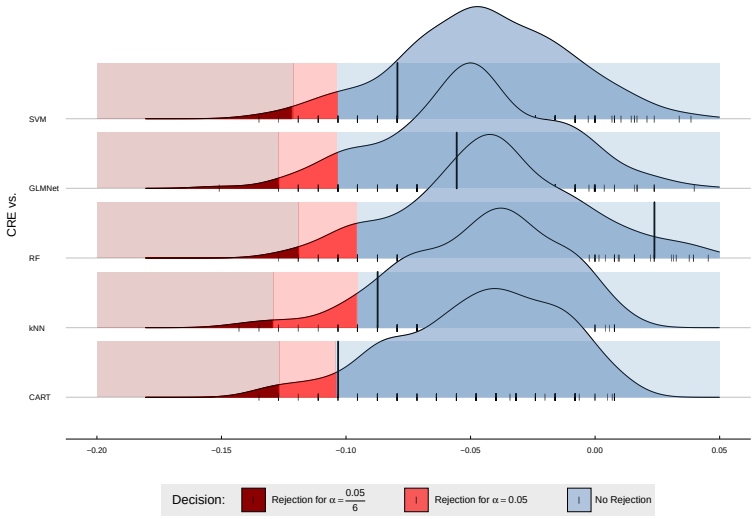

Figure 5: Densities of resampled test statistics for pairwise tests of CRE vs. six other classifiers on 62 datasets from PMLB. Big (small) vertical lines depict observed (resampled) test statistics. Rejection regions for the static (dynamic) GSD-test are highlighted red (dark red). As becomes evident, we cannot reject any of the pairwise tests for neither significance level.

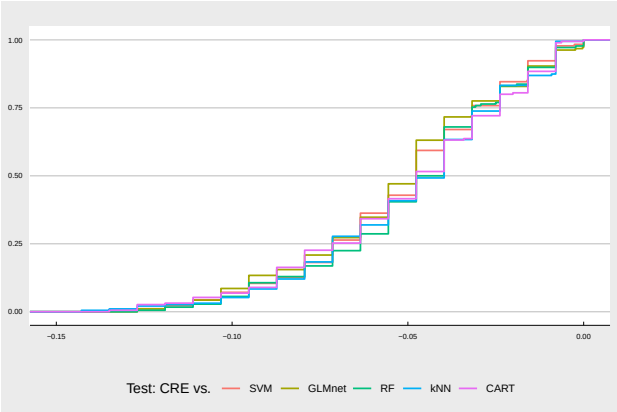

Figure 6: Cumulative Distribution Functions (CDFs) of resampled test statistics for hypothesis tests on PMLB benchmark suite of CRE vs. SVM, GLMNet, RF, kNN, and CART, respectively. As opposed to Figure 5 above, values of observed test statistics are not included. They are: -0.1031746 (CART), -0.08730159 (kNN), 0.02380952 (RF), -0.05555556 (GLMNet), -0.07936508 (SVM). It becomes evident that the resampled test statistics' distributions are more similar to each other than in the case of testing SVM vs. competitors in the OpenML benchmark suite.

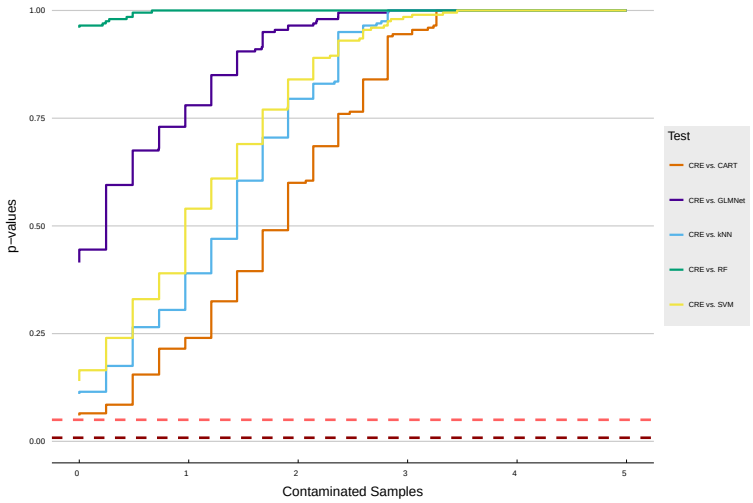

Figure 7: Effect of Contamination: p-values for pairwise tests of CRE versus the five competitors in PMLB benchmark suite application. Analogous to Figure 5, dotted red lines mark significance levels of $\alpha = 0.05$ (dark red: $\alpha = \frac{0.05}{6}$). Since none of the tests reject for $\alpha = 0.05$ under no contamination, this obviously does not change with contaminated samples.

this case, the test decisions of both static and dynamic GSD-tests was not to reject the null hypothesis of CRE being outperformed by RF, CART, SVM, GLMNet, and kNN w.r.t. to robust accuracy.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Abstract and introduction of the paper reflect the paper's contribution and scope, covering both theoretical and experimental results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of the work are discussed in Sections 4.2 and 6 of the main paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The main theoretical results of the paper are stated in Theorems 1, 2 and 3 as well as Corollary 1. For each of these theorems, full sets of assumptions are provided. Moreover, complete proofs for all these statements are provided in the paper's appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper fully discloses all information needed to reproduce the experimental results in Section 5 of the main paper and in Section C of the paper's appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All data and code needed to reproduce the paper's experiments are openly accessible via a GitHub repository. A link to this repository is provided in Footnote 3 of the main paper.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All necessary background information to understand the papers experimental results can be found in Section 5 of the main paper and Section C of the paper's appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All experimental results of the paper are statistically tested for significance. Moreover, the test decisions are checked with respect to their robustness towards the assumptions underlying the respective tests. Compare Section 4 for theoretical considerations on testing and Sections 5 and Appendix C for statistical test results of the applications.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Information on computer resources is provided in the GitHub repository referenced to in Footnote 3 of the main paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper is, in every aspect, conform with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: We do not foresee direct negative societal impact from the current work. Positive societal impact in form of making benchmarking results more robust to unjustified assumptions are discussed in Section 4.2 of the paper.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12.07.11, § 8, Abs. 2 Pkt. .5.)

Hiermit erkläre ich an Eidesstatt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

Blocher, Hannah

Name, Vorname

München, 28.11.2025

Ort, Datum

Hannah Blocher

Unterschrift Doktorand/in