Monika Jelizarow

Global Tests of Association for Multivariate Ordinal Data

Knowledge-based Statistical Analysis Strategies for Studies using the International Classification of Functioning, Disability and Health (ICF)

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

Global Tests of Association for Multivariate Ordinal Data

Knowledge-based Statistical Analysis Strategies for Studies using the International Classification of Functioning, Disability and Health (ICF)

Dissertation
zur Erlangung des akademischen Grades Doctor rerum naturalium
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians Universität München

vorgelegt von
Monika Jelizarow
aus Allenstein (Olsztyn)
am 28.11.2014

Erster Berichterstatter:	Prof. Dr. rer. nat. Ulrich Mansmann Ludwig-Maximilians Universität München
Zweiter Berichterstatter:	Prof. Dr. rer. nat. Jörg Rahnenführer Technische Universität Dortmund
Datum der Disputation:	17.04.2015

Omnia, Lucili, aliena sunt, tempus tantum nostrum est[.]

Lucius Annaeus Seneca
(aus: Epistulae morales ad Lucilium, Epistula I)

Zusammenfassung

Globale Tests werden immer dann relevant, wenn es von Interesse ist statistische Inferenz über Variablensets als Ganzes zu betreiben. Die vorliegende Arbeit unternimmt den Versuch solche Tests für den Fall potenziell hochdimensionaler multivariater ordinaler Daten zu entwickeln. Motiviert wurde sie hauptsächlich durch Forschungsfragen, die sich aus Daten ergeben, welche mit Hilfe der 'International Classification of Functioning, Disability and Health' erhoben wurden.

Im Wesentlichen umfasst die Arbeit zwei Teile. Im ersten Teil werden zunächst zwei Tests diskutiert, von denen sich jeder einem speziellen Problem im klassischen Fall zweier Gruppen widmet. Da beide Permutationstests sind, setzt ihre Validität voraus, dass die gemeinsame Verteilung der Variablen im zu testenden Set unter der Nullhypothese in beiden Gruppen identisch ist. Umfassende Simulationsstudien auf Basis der diskutierten Tests deuten jedoch darauf hin, dass Verletzungen dieser Bedingung aus rein praktischer Sicht nicht automatisch zu invaliden Tests führen müssen. Vielmehr scheint das Scheitern von Zwei-Stichproben-Permutationstests von zahlreichen Parametern abzuhängen, darunter dem Verhältnis zwischen den Gruppengrößen, der Anzahl der Variablen im interessierenden Set und nicht zuletzt der verwendeten Teststatistik. Im zweiten Teil werden zwei weitere Tests entwickelt; beide können verwendet werden, um im Kontext generalisierter linearer Modelle auf Assoziation zwischen einem Set aus ordinal skalierten Kovariablen und einer Zielvariable zu testen, falls erwünscht nach Adjustierung für bestimmte weitere Kovariaben. Der erste Test basiert auf expliziten Annahmen hinsichtlich der Abstände zwischen den Kategorien der Variablen, und es wird gezeigt, dass dieser Test den traditionellen Cochran-Armitage-Test auf höhere Dimensionen, kovariablenadjustierte Szenarien und Zielvariablen im Spektrum generalisierter linearer Modelle verallgemeinert. Der zweite Test wiederum parametrisiert diese Abstände und schenkt ihnen damit Flexibilität. Basierend auf den Powereigenschaften der Tests werden praktische Empfehlungen hinsichtlich ihrer Verwendung besprochen, und Verbindungen mit den im ersten Teil der Arbeit diskutierten Permutationstests werden aufgezeigt. Illustriert werden die entwickelten Methoden anhand der Analyse von Daten aus zwei Studien, welche die 'International Classification of Functioning, Disability and Health' verwenden. The Resultate versprechen ein breites Potenzial der vorgeschlagenen Tests in diesem Datenkontext ebenso wie darüber hinaus.

Summary

Global tests are in demand whenever it is of interest to draw inferential conclusions about sets of variables as a whole. The present thesis attempts to develop such tests for the case of multivariate ordinal data in possibly high-dimensional set-ups, and has primarily been motivated by research questions that arise from data collected by means of the 'International Classification of Functioning, Disability and Health'.

The thesis essentially comprises two parts. In the first part two tests are discussed, each of which addresses one specific problem in the classical two-group scenario. Since both are permutation tests, their validity relies on the condition that, under the null hypothesis, the joint distribution of the variables in the set to be tested is the same in both groups. Extensive simulation studies on the basis of the tests proposed suggest, however, that violations of this condition, from the purely practical viewpoint, do not automatically lead to invalid tests. Rather, two-sample permutation tests' failure appears to depend on numerous parameters, such as the proportion between group sizes, the number of variables in the set of interest and, importantly, the test statistic used. In the second part two further tests are developed which both can be used to test for association, if desired after adjustment for certain covariates, between a set of ordinally scaled covariates and an outcome variable within the range of generalized linear models. The first test rests upon explicit assumptions on the distances between the covariates' categories, and is shown to be a proper generalization of the traditional Cochran-Armitage test to higher dimensions, covariate-adjusted scenarios and generalized linear model-specific outcomes. The second test in turn parametrizes these distances and thus keeps them flexible. Based on the tests' power properties, practical recommendations are provided on when to favour one or the other, and connections with the permutation tests from the first part of the thesis are pointed out. For illustration of the methods developed, data from two studies based on the 'International Classification of Functioning, Disability and Health' are analyzed. The results promise vast potential of the proposed tests in this data context and beyond.

List of abbreviations

BDI	Beck depression inventory
BMI	Body mass index
CA	Cochran-Armitage
cdf	Cumulative distribution function
df	Degrees of freedom
FWER	Familywise error rate
GLM	Generalized linear model
ICD	International Classification of Diseases
ICF	International Classification of Functioning, Disability and Health
i.i.d.	Independent and identically distributed
IJD	Identical joint distributions
LR	Likelihood ratio
MI	Marginal inhomogeneity
MO	Marginal order
MS	Multiple sclerosis
NJD	Non-identical joint distribution
NPC	Non-parametric combination
pdf	Probability density function
PLS	Partial least squares
PP MS	Primary progressive multiple sclerosis
SMH	Simultaneous marginal homogeneity
SNP	Single-nucleotide polymorphism
SP MS	Secondary progressive multiple sclerosis
WHO	World Health Organization
WHO-FIC	World Health Organization Family of International Classifications

Contents

0. Motivation, scope and overview

Global hypothesis tests are in demand whenever an application involves a vast number of variables which can be meaningfully structured into sets by prior knowledge and researchers wish to draw inferential conclusions about the sets as a whole rather than about the individual variables. The main argument put forward in favour of such set-based analyses is that they may be worthwhile in view of both interpretability of results and power. Interpretability of results may improve because sets are usually defined based on substantial expert knowledge, and power may increase because multiplicity issues do not occur unless several sets are to be tested simultaneously, yet even then the multiplicity penalty will be far less severe than in the case of variable-wise tests.

In the past decade, global tests have become an important topic in statistical research. This has predominantly been driven by the need for statistical tools that allow to test predefined sets of microarray-based gene expression levels for association with some clinical parameter (Draghici et al., 2003; Goeman et al., 2004, 2005; Mansmann and Meister, 2005; Kong et al., 2006; Goeman and Bühlmann, 2007; Hummel et al., 2008). From the statistical viewpoint, gene expression levels are metrically scaled (or, to be more precise, ratio scaled) variables. Consequently, the vast number of tests proposed in this context (see for example Ackermann and Strimmer (2009), Fridley et al. (2010) or Maciejewski (2013) for a review) may likewise be applied to sets of metric variables stemming from any other context. The potential benefit of global tests, however, reaches beyond research problems on the metric scale.

The present thesis is concerned with global tests of association for sets of multivariate ordinal variables in potentially high-dimensional scenarios. Primarily, it has been motivated by research problems that arise from data collected by means of the International Classification of Functioning, Disability and Health (World Health Organization, 2001a,b), or briefly ICF, which over the last decade has established itself world-wide as a basis for the collection of data on human functioning and disability. A frequent objective of ICF-based studies is to assess the presence of an association between individuals' profile of functional limitations and disabilities and some other factor, for example some experimental condition or phenotypic feature. Global tests are relevant in this context because the ordinally scaled variables that underlie such functioning and disability profiles have the special feature that they can be grouped into sets by

superordinate aspects, if desired even at different levels of detail. In summary, the primary objective of this thesis is to enable ICF-based data to be analyzed in a way that makes use of the prior knowledge on their structure and, thereby, to contribute to a sound statistical analysis of such data.

In total, the thesis comprises five chapters. The main contents and objectives of the individual chapters are as follows. **Chapter 1** provides the application context of this thesis. It introduces the ICF, illustrates the special characteristics of ICF-based data, and broadly defines the problem addressed. **Chapter 2** provides the statistical and methodological context of this thesis. The main conclusion of this chapter is that, compared to standard univariate analyses of ICF-based data, both interpretability of results and power could be enhanced if global tests of association for potentially high-dimensional multivariate ordinal data were available. The next two chapters are then devoted to the development and discussion of such tests. They are self-contained and can in principle be read independently of each other. **Chapter 3** is devoted to research questions that can be framed as two-group comparisons. Two different questions are addressed, and hence two different tests are proposed. Both are permutation tests and as such rely on the rather restrictive assumption that, under the null hypothesis, the ordinal variables' joint distribution is identical between the two groups to be compared. Particular attention is therefore paid to the simulation-based examination of the tests' robustness properties in situations in which this assumption is not met, where robustness is meant with respect to type I error rate control. The tests' application is finally illustrated with data from an ICF-based stroke study. One limitation of the tests from Chapter 3 is that they do not allow for adjustment for the effect of other variables (e.g. potential confounders). This limitation is overcome in **Chapter 4** which is devoted to research questions that can be formulated within the context of generalized linear models (GLMs), with the ordinal variables playing the role of the covariates and the 'other' factor of interest playing the role of the outcome variable. In particular, two tests are developed, both within the framework of the 'global test' methodology of Goeman et al. (2004, 2006, 2011). The tests are based on different assumptions and hence are useful in different practical situations, as is confirmed by means of simulation studies. The tests' application is illustrated with data from an ICF-based multiple sclerosis (MS) study. Overall, the present thesis thus suggests four statistical hypothesis tests, although an intimate connection between two of them will be shown. As a side remark, Chapters 3 and 4 contain so-called excursuses, either in the form of a section or a subsection. The information offered to the reader by each excursus is relevant to the topic covered by the respective chapter and deserves separate mention, yet it plays a rather subordinate role in the overall context upon first reading. Sections and subsections that are marked as excursus can therefore be skipped without any problems. Finally, **Chapter 5** closes the thesis with a short summary and examination of its con-

tributions to the available literature, and with the key conclusions drawn therefrom. In this context, it furthermore addresses the limitations of the work presented and briefly sketches possible directions for future research.

While this thesis focuses on ICF-based applications, all global tests developed herein can likewise be used to analyze any other type of possibly high-dimensional multivariate ordinal data that can be structured into sets by external knowledge. Examples include realizations of items in psychodiagnostic tests (e.g. structured into sets by the subdimension they describe), side or adverse effects in drug safety or toxicity studies (e.g. structured into sets by the body function they affect) and single-nucleotide polymorphisms (SNPs) in next-generation sequencing studies (e.g. structured into sets by genes). As with ICF-based data, here it may likewise be preferable to shift the unit of analysis from individual variables to whole sets of variables. Because of this broad range of potential applications, the methodical issues of this thesis shall be presented mostly without particular reference to ICF-based data. Readers who, from the purely methodical viewpoint, wish the most efficient possible approach to the global tests developed may in principle skip Chapter 1 and Sections 2.1 – 2.3 in Chapter 2.

Partly, the contents of this thesis have already been published in a peer-reviewed statistical journal or as a technical report. Information as to the extent to which these manuscripts contribute to each of the five chapters summarized above are provided at the appropriate places in the text. Specifically, the manuscripts, and the respective authors' contributions to their contents, are:

- Jelizarow et al. (2014a): *M. Jelizarow, A. Cieza and U. Mansmann, 2014. Global permutation tests for multivariate ordinal data: alternatives, test statistics and the null dilemma. Journal of the Royal Statistical Society, Series C (Applied Statistics), doi: 10.1111/rssc.12070.*

  All central ideas presented herein were formulated and worked out independently by Monika Jelizarow. Furthermore, Monika Jelizarow implemented the methods discussed in the language `R` (R Development Core Team, 2014) (available for use from `http://wileyonlinelibrary.com/journal/rss-datasets`), conducted the simulation studies, performed the data analysis and wrote the manuscript. Ulrich Mansmann pointed out the potential of global tests for the statistical analysis of ICF-based data and thereby initiated the project. He supervised the respective research activities and contributed to the presentation of the manuscript. Alarcos Cieza provided the ICF-based data, supervised their analysis and contributed to the data example part of the manuscript.

- Jelizarow et al. (2014b): *M. Jelizarow, U. Mansmann and J. J. Goeman, 2014. A Cochran-Armitage-type and a score-free global test for multivariate ordinal data. Under revision. Preliminary version: Technical Report 168, Department of Statis-*

tics, LMU Munich.

All central ideas presented herein were worked out independently by Monika Jelizarow. She furthermore formulated the ideas regarding the Cochran-Armitage-type test, conducted the simulation studies, performed the data analysis and wrote the manuscript. Jelle Goeman suggested to extend the 'global test' to ordinally scaled covariates and thereby initiated the project. He, with small contributions made by Monika Jelizarow, implemented the methods discussed (provided for use in the `R` package `globaltest` (Goeman and Oosting, 2012) which can be obtained from `http://www.bioconductor.org`) and, together with Ulrich Mansmann, supervised the project and contributed to the presentation of the manuscript. Alarcos Cieza, who is mentioned in the acknowledgements of the latter, provided the ICF-based data.

As a final remark, the dissertation project outlined above was predominantly funded by a doctoral scholarship of the Studienstiftung des deutschen Volkes (German National Academic Foundation). The scholarship included a study abroad scholarship for a three-month research visit with Jelle Goeman at the Leiden University Medical Center in the Netherlands. I would like to thank the Studienstiftung des deutschen Volkes for the unique opportunity to have been one of its scholars; it was considerably more than the financial support from which I could benefit. I would furthermore like to sincerely thank my main supervisor Ulrich Mansmann, Alarcos Cieza and Jelle Goeman for discussions, valuable feedback and for sense of humour, and Ulrich Mansmann for his great support especially towards the end of the thesis work.

1. International Classification of Functioning, Disability and Health (ICF)

This chapter has the objective to clarify within which particular context the present thesis falls in application-related respects. Section 1.1 provides a short introduction to the ICF and ICF-based data. Subsequently, Section 1.2 discusses the research questions that commonly arise from ICF-based data, and further points out the importance of statistical hypothesis tests for the analysis of the latter. Parts of Section 1.1 are based on Jelizarow et al. (2014a).

1.1. The ICF in brief

Background

As noted previously, the work presented in this thesis has primarily been motivated by research problems related to data that have been collected by means of the ICF (World Health Organization, 2001b). The latter was officially endorsed by all 191 member states of the World Health Organization (WHO) in the 54th World Health Assembly on 22 May 2001 (World Health Organization, 2001a). As one of the classifications from the WHO Family of International Classifications (WHO-FIC) (Madden et al., 2007), the ICF provides a unified and comprehensive framework for the description of functioning and disability both across health conditions and for specific health conditions such as depression, MS, obesity and stroke. In particular, going beyond a purely medical approach, it allows to take into account biological, individual, social and environmental aspects of functioning and disability. Because the ICF shifts the focus from medical diagnoses to the lived health experience of individuals (Stucki et al., 2008), it can be understood as a complement to the International Classification of Diseases (ICD) (World Health Organization, 1992) which is used world-wide to monitor the incidence and prevalence of diseases. The WHO in fact encourages the combined utilization of ICD and ICF, wherever applicable (Ustün et al., 2003). As the combination of ICD and

ICF accounts for the fact that individuals with the same disease or health problem can experience very different functional limitations and disabilities (World Health Organization and The World Bank, 2011), it promises to provide a comprehensive picture of the health status of both individuals and populations. Further background information on the ICF can be found in World Health Organization (2001b) and Ustün et al. (2003).

Describing functioning and disability: ICF items and ICF core sets

The ICF-based description of different aspects of functioning and disability is realized by means of health-related items called ICF categories (e.g. 'memory functions', 'orientation functions' and 'sleep functions'), henceforth referred to as ICF items. From the statistical viewpoint, ICF items are ordinally scaled variables with either five or nine categories. Overall, the ICF comprises more than 1400 such ICF items. Aside from the fact that numerous ICF items may not be relevant in certain situations, data collection based on this entire volume is not feasible in practice owing to time and cost constraints. This is where so-called ICF core sets come into play (Stucki and Grimby, 2004; Ustün et al., 2004; Cieza et al., 2006; Rauch et al., 2008), initiated by the WHO in collaboration with the ICF Research Branch with the aim to operationalize the ICF for clinical practice and research. In brief, ICF core sets are health condition-specific selections from the overall pool of ICF items. They thus facilitate the implementation of the ICF in clinical practice and research on the one hand and link the ICF with the ICD on the other hand. ICF core sets are defined by health experts (e.g. physicians and physiotherapists) at international ICF consensus conferences, based on qualitative and quantitative evidence from preliminary studies (Cieza et al., 2004; Selb et al., 2014). Among the ICF core sets that have been developed up to now (see `http://www.icf-research-branch.org` for an overview), the total number of ICF items included varies from about 80 to 140. The statistical analysis of the resultant 80- to 140-dimensional profiles of functional limitations and disabilities may pose some challenges. This is because, in many ICF studies, the number of individuals involved is small, sometimes considerably smaller than the number of ICF items. The data situation may thus be high-dimensional. In the two ICF studies presented later on in this thesis, for example, the number of individuals is 104 and 93, wheras the number of ICF items amounts to 130 and 129, respectively. High-dimensional data situations like these call for non-standard statistical analysis strategies, since standard strategies often yield deficient results in such situations or even become inapplicable. We come back to this issue in Chapter 2.

It has already been mentioned that, from the statistical viewpoint, ICF items are ordinally scaled variables. This means that their possible realizations, represented by

either five or nine distinct categories, are naturally ordered but usually of unknown distance. Which ordinal scale is used for which ICF item depends on what the respective item specifically describes. In particular, each ICF item can be attributed to one of the four so-called ICF components

- b: body functions,
- s: body structures,
- d: activities and participation or
- e: environmental factors.

The WHO defines 'body functions' as the physiological and psychological functions of body systems, 'body structures' as the anatomical parts of the body, 'activity' as the execution of tasks or actions by an individual, 'participation' as an individual's involvement in life situations and 'environmental factors' as the physical, social and attitudinal environment in which individuals live and conduct their lives (World Health Organization, 2001b). ICF items that describe 'body functions' (e.g. ICF item 'memory functions'), 'body structures' (e.g. ICF item 'structure of cardiovascular system') and 'activities and participation' (e.g. ICF item 'doing housework') are now measured on an ordinal scale with five possible categories which, for reasons of practicability, are labelled with numbers 0 to 4. This scale is:

- 0: no problem
- 1: mild problem
- 2: moderate problem
- 3: severe problem
- 4: complete problem

For ICF items that describe 'environmental factors' (e.g. ICF item 'immediate family') it is differentiated between barriers and facilitators. In particular, the respective ICF items are measured on an ordinal scale with nine possible categories which are labelled with numbers -4 to 4. This scale is:

- -4: complete barrier
- -3: severe barrier
- -2: moderate barrier
- -1: mild barrier
- 0: neither barrier nor facilitator

1: mild facilitator

2: moderate facilitator

3: severe facilitator

4: complete facilitator

When it comes to the statistical analysis of ICF-based data, both the five-level and the nine-level ordinal scale are typically coarsened *ex post*, both for reasons of convenience and because evaluations (Cieza et al., 2009; Algurén et al., 2011; Bostan et al., 2012; Prodinger et al., 2012; Røe et al., 2013) have shown the need to collapse some categories. We elaborate on this issue further below.

Moreover, it should be mentioned that for each ICF item, irrespective of which ICF component it comes from, there are two additional answer options: 8 (not specified) and 9 (not applicable). 8 is used when the available information does not suffice to quantify the severity of the problem. 9 is used when the respective ICF item is not applicable to an individual; for example, the ICF item 'family relationships' is not applicable to an individual without family. Both answer options can be very useful from the clinician's point of view, but they may pose certain problems from the statistician's point of view, since they cannot be embedded in the ordinal scales from above (Cieza et al., 2006). Hence, practical strategies to handle such observations are needed. One strategy is to treat observed 8s and 9s as missing values and then replace them by imputed values on the relevant ordinal scale. Another strategy is to replace only observed 8s by imputed values and recode observed 9s into 0s (no problem/neither barrier nor facilitator). While the first strategy has the drawback that it does not respect the obvious difference between 8s and 9s, the second strategy may seem somewhat ad-hoc. Because the results obtained may vary between different strategies, they should always be interpreted with caution, especially when the data to be analyzed exhibit a large number of 8s and 9s. This potential bias due to the strategy that is employed to handle observed 8s and 9s is in fact one concern with ICF-based applications. In the two ICF-based applications that will be presented in this thesis, this issue is of limited relevance, since in both instances 8s have not been observed at all and 9s have been observed only rarely (see Sections 3.7 and 4.5 for more detailed information). To eliminate these 9s, we recoded them into 0s (no problem/neither barrier nor facilitator), as is often done in practice.

Scale coarsening in practice

In most ICF studies, both the five-level ordinal scale of ICF items of the ICF components 'body functions' (b), 'body structures' (s) and 'activities and participation' (d) and

the nine-level ordinal scale of ICF items of the ICF component 'environmental factors' (e) are used in their original form when it comes to data collection. When it comes to the statistical analysis of ICF-based data, however, the scales are typically coarsened. Suggestions to collapse some categories have been made by several researchers, based on results from evaluations via the Rasch model (Cieza et al., 2009; Algurén et al., 2011; Bostan et al., 2012; Prodinger et al., 2012; Røe et al., 2013). As an appreciable side effect, the number of ICF items for which one or more categories have remained unobserved in the sample can potentially be reduced, and data analysis becomes less challenging.

All ICF-based data considered in this thesis have been preprocessed as follows. As has been recommended by Bostan et al. (2012) for the five-level ordinal scale originally used in the ICF components b, s and d, we coarsened both the five-level and the nine-level ordinal scale originally used in the ICF component e to three levels: the scale 0 1 2 3 4 was coarsened to 0 1 1 2 2, whereas the scale -4 -3 -2 -1 0 1 2 3 4 was coarsened to -1 0 1. Given that the numbers with which an ordinal variable's categories are labelled are arbitrary, we subsequently relabelled the latter such that the lowest category is labelled with 1 (rather than with 0 or -1). This corresponds to how ordinal variables' categories shall be labelled in the remainder of this thesis. For ICF items of the ICF component e, we furthermore reversed the roles of the lowest and the highest category, in order that the highest category be most negatively connotated, as is the case for ICF items of the other ICF components. The coarsening and relabelling strategies are depicted in Figures 1.1 and 1.2.

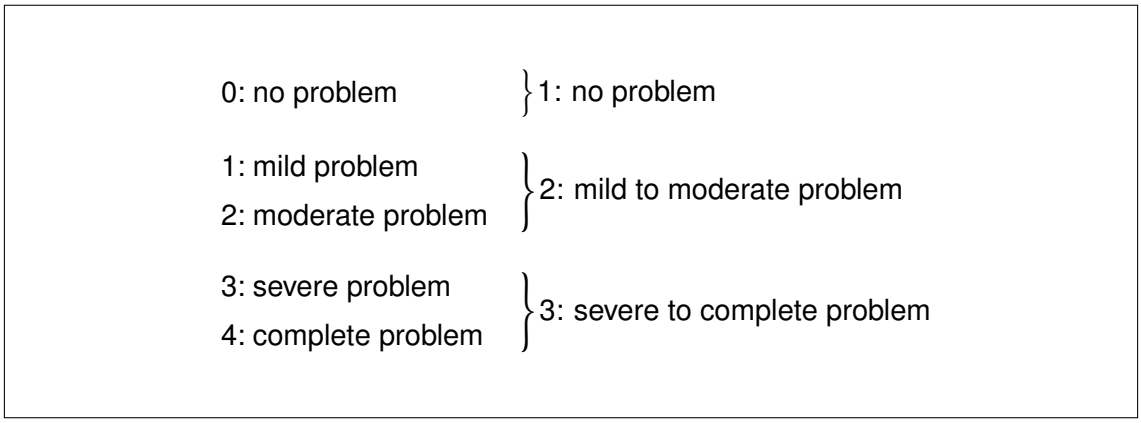

Figure 1.1.: Coarsening and relabelling strategy for the five-level ordinal scale of ICF items of the ICF components b, s and d.

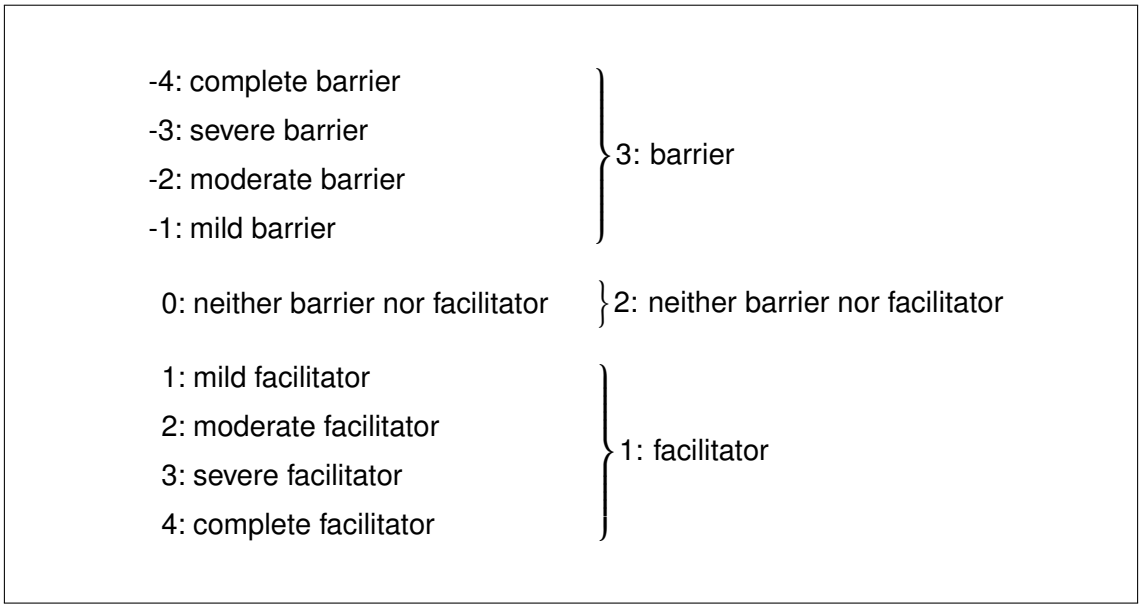

Figure 1.2.: Coarsening and relabelling strategy for the nine-level ordinal scale of ICF items of the ICF component e.

Tree structure

As has been said above, each ICF item can be attributed to one of the ICF components b, s, d or e. Within each ICF component, there are first-, second-, third- and fourth-level ICF items, with the level depth indicating how precise the measured information is. First-level ICF items are called ICF chapters and are designated by the letters b, s, d or e, followed by a one-digit number, the chapter number (e.g. b2 for the ICF chapter 'sensory functions and pain'). Second-level ICF items are designated by the letters b, s, d or e, followed by the one-digit chapter number and a two-digit number (e.g. b210 for the ICF item 'seeing functions'). Third- and fourth-level ICF items receive one further digit each (e.g. b2102 for the third-level ICF item 'quality of vision' and b21022 for the fourth-level ICF item 'contrast sensitivity'). The way in which ICF items are designated thus well reflects how precise the information that they measure is. This thesis focuses on ICF-based applications where all (or nearly all) ICF items considered are two-level ICF items, which is the standard case in practice. Henceforth, the term ICF item shall therefore refer solely to two-level ICF items.

ICF items' designation represents prior knowledge. Let us consider the ICF item 'seeing functions', for example. Its designation b210 tells us that it belongs to the ICF component 'body functions' (b) and, to be more specific, to those ICF items within b that describe 'sensory functions and pain' (b2). Hence, any pool or set of ICF items considered in an ICF study may be organized hierarchically by available expert knowl-

edge. In the first step, the overall set of ICF items can be structured or divided by ICF components. ICF items included in the resultant disjoint sets b, s, d and e can, in the second step, be divided further by their ICF chapter number. We shall refer to the more specific sets that arise from the second step as ICF chapters. As a result of the division of ICF items in the way just described, a classical tree structure is obtained.

Figure 1.3 illustrates the natural four-level tree structure in which ICF items can be arranged, exemplarily for an arbitrary selection of 20 ICF items. Each tree level reflects a particular level of detail at which functioning and disability can be looked at. The level of detail increases the more similar ICF items from the same set are or, equivalently, the more dissimilar ICF items from different sets are with respect to the aspect that they describe. It thus increases from the top to the bottom of the tree: the first tree level which is made up of the complete pool or set of ICF items considered has the lowest level of detail, whereas the fourth tree level which is made up of individual ICF items has the highest level of detail. Given that, on each tree level, ICF items from the same set describe more similar aspects than ICF items from different sets (e.g. the ICF items 'memory functions' (b144) and 'attention functions' (b140) measure more similar aspects than the ICF items 'memory functions' (b144) and 'washing oneself' (d510)), it is only reasonable to assume that ICF items' realizations are not independent but rather come from a multivariate distribution with a complex dependence structure.

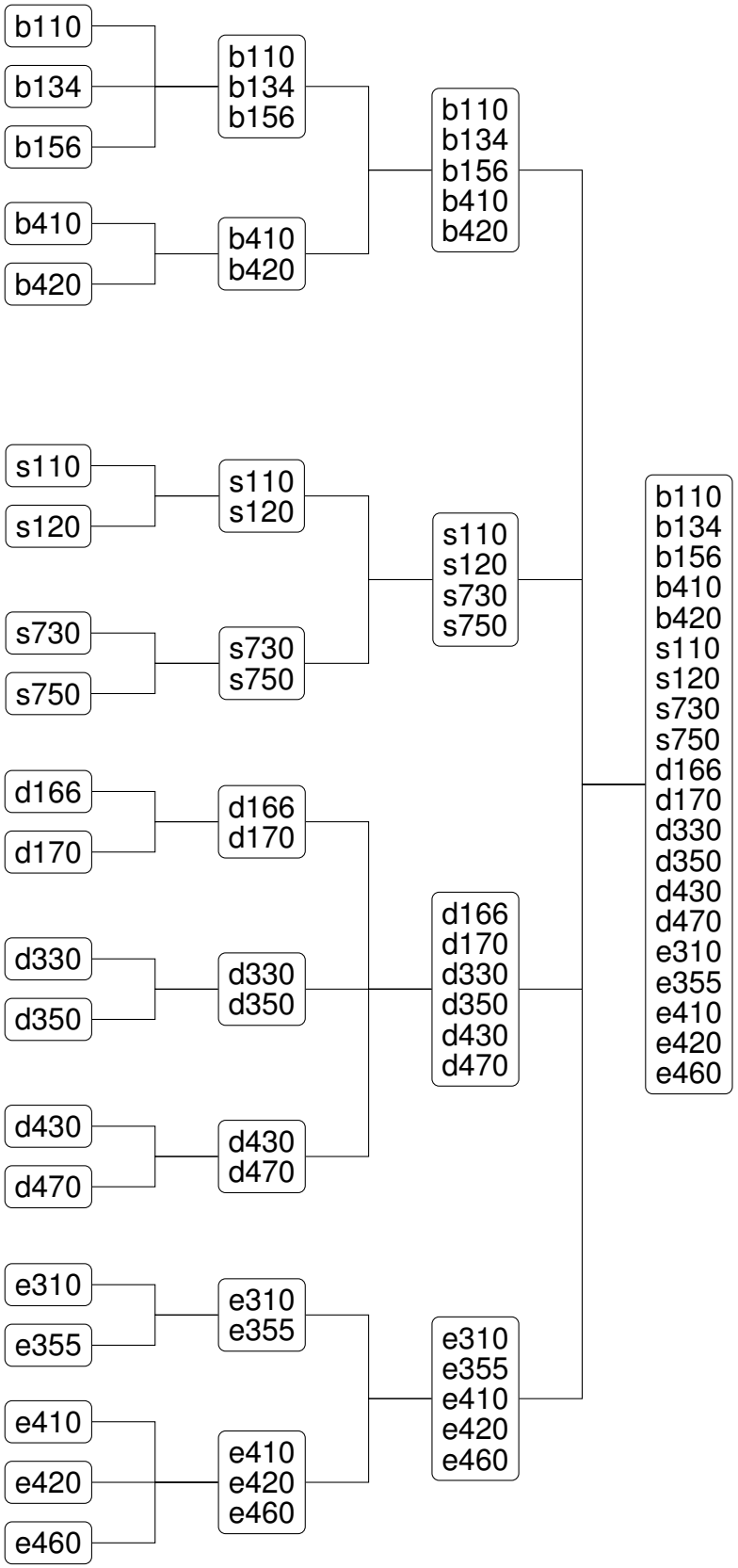

Figure 1.3.: Tree structure of ICF items, exemplarily for an arbitrary selection of 20 ICF items. **1st tree level**: root set or complete set of ICF items considered (lowest level of detail); **2nd tree level**: level of ICF components (b, s, d and e); **3rd tree level**: level of ICF chapters (here b1, b4, s1, s7, d1, d3, d4, e3 and e4); **4th tree level**: level of individual ICF items (highest level of detail). A description of the ICF items and ICF chapters involved can be found in World Health Organization (2001b) or in Tables 3.2 and 4.4.

1.2. Typical objectives of ICF-based studies

As a relatively new type of possibly high-dimensional multivariate ordinal data, ICF-based data have so far received only little attention in statistical and methodical research. Outside the latter, for instance in rehabilitation sciences, the interest and public investment in ICF-related subjects are large. This is well reflected by the number of ICF-related PubMed records per year which has increased almost steadily since 2001 when the ICF was officially endorsed (U. S. National Library of Medicine, 2014). As all 191 member states of the WHO have agreed to use the ICF in their clinical practice, research, surveillance and reporting and many have already started, it is expected by the WHO that the number of ICF-based studies and thus the amount of ICF-based data collected will rapidly increase over the years to come. Hence, the need for statistical tools to answer the research questions that arise from such data will continue to rise.

The vast majority of ICF-related research questions either concerns relationships between realizations of ICF items among themselves or relationships between realizations of ICF items and some other data. 'Some other data' may be different experimental conditions (e.g. inpatient rehabilitation treatment versus outpatient rehabilitation treatment) or phenotypic features in the widest sense (e.g. disease subtypes, body mass index (BMI) or some subjective quality-of-life score). To answer research questions of the first kind, graphical models have proven to be a useful statistical tool (Kalisch et al., 2010; Fellinghauer et al., 2010; Fellinghauer, 2011; Fellinghauer et al., 2013). The aim of ICF studies with such research questions is usually to provide a deeper understanding of human functioning and disability in itself (Kalisch et al., 2010). ICF studies with research questions of the second kind, in contrast, are conducted with the aim to better understand the interplay between human functioning and disability and other factors. It is research questions of the second kind that are addressed in this thesis. A common ICF-related problem that falls within this class is the statistical validation of ICF core sets; here one is often interested in whether ICF core set-based data are related to or associated with some general health or quality-of-life score. Gertheiss (2011), Gertheiss et al. (2011) and Oberhauser et al. (2013) have demonstrated that, in this specific situation, predictive modeling techniques can be very useful. In general, however, it is particularly statistical hypothesis tests that lend themselves well to address research questions of the second kind (Holper et al., 2010; Herrmann et al., 2011; Tschiesner et al., 2011). Chapter 2 discusses how the fact that ICF items can be structured by prior knowledge can be exploited in this context and, thereby, the development of global tests of association for potentially high-dimensional multivariate ordinal data is motivated.

2. Statistical hypothesis testing of ICF-based data

This chapter aims to clarify within which particular context the present thesis falls in statistical and methodological respects. Section 2.1 provides some basics which are needed throughout the chapter. Readers familiar with the multiplicity problem in simultaneous statistical inference may skip this section. Section 2.2 outlines classical ways to perform statistical hypothesis tests of ICF-based data. Section 2.3 discusses alternative ways and, in this context, motivates the development of global tests of association for ICF-based or, formulated in statistical terms, for potentially high-dimensional multivariate ordinal data. Section 2.4 briefly reviews the literature that is relevant to this subject. Parts of Sections 2.3 and 2.4 are based on Jelizarow et al. (2014a).

2.1. Simultaneous testing of multiple hypotheses

Let us start with some basics. Consider Table 2.1 which displays the four possible events that can happen when a statistical hypothesis test is performed.

Table 2.1.: Two-by-two table showing the four possible events that can happen when a statistical hypothesis test is performed.

	Null hypothesis is maintained	Null hypothesis is rejected
Null hypothesis is true	Correct test decision ('True negative')	Type I error ('False positive')
Null hypothesis is false	Type II error ('False negative')	Correct test decision ('True positive')

A type I error thus occurs when a true null hypothesis is rejected, and a type II error occurs when a false null hypothesis is not rejected. Here we focus on type I errors, since false positive findings are usually considered more problematic in scientific research than false negative findings. For a statistical hypothesis test that controls the probability of making a type I error at a significance level $\alpha = 0.05$, we can now say that the test will correctly maintain the null hypothesis with a probability of $1 - 0.05 = 0.95$.

Suppose now that not only one but two hypotheses are being tested, and that each test is performed at level $\alpha = 0.05$. Then, under the assumption that the two tests are independent of each other, the probability that at least one type I error is committed is $1 - 0.95^2 = 0.0975$. For three independent tests the probability that at least one type I error is committed is larger than 0.1426, and for ten independent tests it is even larger than 0.4012. The probability

$$\Pr(\textit{commit one or more type I errors among all hypotheses tested})$$

is the so-called **familywise error rate (FWER)**. Under the independence assumption, the FWER equals $1 - (1-\alpha)^m$, with m the number of hypotheses tested, and it is easy to see that it approaches 1 as m increases:

$$1 - (1-\alpha)^m \overset{m\to\infty}{\longrightarrow} 1,\ \alpha \in (0,1].$$

In practice, the m tests performed are typically not independent. In the context of ICF-based data, for instance, this becomes immediately clear from the fact that many ICF items describe related aspects (e.g. the ICF items 'memory functions' (b144) and 'attention functions' (b140)), so the respective test statistics will be correlated. In such situations, the FWER will be smaller than $1 - (1-\alpha)^m$, but it may still substantially exceed α. When multiple hypotheses are to be tested simultaneously and the FWER is to be controlled at, for example, level $\alpha = 0.05$, it will thus not be possible to test each individual hypothesis at level $\alpha = 0.05$. Consequently, to ensure FWER control at some prespecified level α, we need to decrease the hypothesis-specific significance levels appropriately or, alternatively, increase the hypothesis-specific P-values obtained at level α. The number of so-called multiplicity adjustment procedures that can be used for this purpose is vast; in Sections 2.2 and 2.3 we discuss those procedures that are particularly suitable for ICF-based problems.

2.2. Procedures ignoring prior knowledge

Classical multiple testing

As has been said in Section 1.2, this thesis is concerned with a frequent objective of ICF studies: to assess whether there is an association between individuals' functioning and disability pattern or profile and some other factor of interest (e.g. some experimental condition or phenotypic feature). Typically, the prior knowledge on ICF items' structure is not exploited for this purpose. The classical approach is in fact to conduct a well-established univariate test for each ICF item (Holper et al., 2010; Herrmann et al., 2011; Tschiesner et al., 2011). When the research question can be framed

as a two-sample problem, for example, the most widely used univariate test for ordinally scaled variables is the two-sided **Cochran-Armitage (CA) test** for trend (Cochran, 1954; Armitage, 1955) which, at least in medical statistics, is often better known in the one-sided formulation of Freidlin et al. (2002). The simplest procedure to then adjust the univariate P-values for multiplicity such that the FWER is controlled at the prespecified level α is the Bonferroni procedure. Let p be the number of variables considered, which corresponds to the number of univariate hypothesis tests performed. With $P\text{-value}_k^{\text{raw}}$ the raw P-value obtained for the kth variable, $k = 1, \ldots, p$, the Bonferroni-adjusted P-value, $P\text{-value}_k^{\text{Badj}}$, is given by

$$P\text{-value}_k^{\text{Badj}} = \min\left(p \cdot P\text{-value}_k^{\text{raw}}, 1\right).$$

The Bonferroni-adjusted P-value is thus the raw P-value multiplied by the number of tests performed (or 1 if this product exceeds 1). Due to its simplicity, the Bonferroni procedure is widely used in practice. A major concern with it is, however, that it is conservative, which means that the FWER is smaller than α. While the conservativeness is minor when the individual test statistics are independent, it can be rather serious when the individual test statistics are positively correlated (Goeman and Solari, 2014). Less conservative yet more complex multiplicity adjustment procedures that provide FWER control have been proposed by Holm (1979), Hochberg (1988) and Hommel (1988). It should be noted, however, that while the procedures of Bonferroni and Holm are valid under any dependence structure of the univariate test statistics, the procedures of Hochberg and Hommel are valid only if the univariate test statistics are positively correlated. In the ICF context, the assumption that the ICF item-specific test statistics are positively correlated may not always be justified. For this reason, here it seems reasonable to use Holm's procedure, in order that FWER control can be ensured. Holm's procedure is a sequential variant of Bonferroni's procedure. In the first step, it penalizes the raw P-values in the same way as does Bonferroni's procedure: it multiplies them by the number of hypotheses tested. In the second step, the multiplicity penalty equals the number of hypotheses that could not be rejected in the first step and, in the third step, it equals the number of hypotheses that could not be rejected in the second step, and so on. The process is terminated when a step fails to result in further rejections or, trivially, when all hypotheses have been rejected.

Although it is simple to use, the procedure just described has potentially low power in the data situation that we consider, both because the multiplicity penalty becomes rather severe when the number of hypotheses tested is large and because it does not take into account the unknown dependencies between the ICF item-specific test statistics. An alternative procedure which respects such dependencies is the permutation-based max-T procedure of Westfall and Young (1993). Given that permutation tests are discussed later on in Chapter 3, here we omit details on it for reasons of clar-

ity. For the moment, it is sufficient to keep in mind that, under certain conditions, permutation tests both preserve the dependence structure in the data and yield exact rather than only asymptotic α-level tests, irrespective of the specific distribution of the data. When compared with the Holm-based procedure described above, the max-T procedure often has more power; for certain dependence structures, it is even asymptotically optimal (Meinshausen et al., 2012). The max-T procedure uses the closure test principle of Marcus et al. (1976), which we now sketch. With H_k the kth hypothesis of interest, $k = 1,\ldots,p$, let $H_M = \bigcap_{k\in M} H_k$ denote the intersection hypothesis for $M \subseteq \{1,\ldots,p\}$. The closure test principle says: each individual hypothesis H_k can be rejected at FWER level α if this hypothesis, and every intersection hypothesis that contains it, have been rejected by an appropriate α-level test. Overall, there are thus $2^p - 1$ hypotheses to be tested, which becomes computationally infeasible for large p. If, however, the test statistic used to test each hypothesis H_M is $\max_{k\in M} T_k$, where T_k is the kth non-negative univariate test statistic (e.g. two-sided CA test statistic), the number of hypotheses to be tested reduces to p. This short-cut is the max-T procedure. For further information and algorithmic details we refer to Westfall and Young (1993), Westfall et al. (2001), Westfall and Troendle (2008) or the tutorial by Goeman and Solari (2014).

Classical global testing

A second approach that does not make use of the prior knowledge on ICF items' structure and which, compared to the classical approach from above, promises a gain in power is to treat the overall set of ICF items considered in a study as one entity and perform only one test. Here the null hypothesis is that none of the ICF items in the overall set is associated with the other factor of interest, and the alternative hypothesis is that at least one of the ICF items in the overall set shows such an association. Let us suppose for the moment that a test suitable for the particular problem at hand is available. Then this approach eliminates the need for adjustment for multiplicity, since only one hypothesis is tested, yet it has the drawback that the inferential conclusion that may be drawn from a significant test result is rather unspecific. For illustration, let us consider the two-sample case. When, in this situation, the null hypothesis of no association is rejected, this tells us that the profile of functional limitations and disabilities is different in one group as opposed to the other, but no information is provided on which parts of the profile the significant difference can be attributed to. Given that ICF-based applications often involve more than 100 ICF items, this will not be satisfactory, and therefore such an approach is in general regarded as irrelevant.

Both approaches from above are somewhat extreme. While the first one tests at the highest possible level of detail where power is lowest because the multiplicity penalty

is most severe and '[...] the effect of highly correlated variables can be very difficult to separate [...]' (Meinshausen, 2008), the second one tests at the lowest possible level of detail where power is highest but test results are little informative. In the context of ICF-based data, however, it is possible to achieve a compromise between these extremes, as we shall now discuss.

2.3. Procedures exploiting prior knowledge

Towards a compromise between classical global and classical multiple testing by exploiting prior knowledge

In the American Heritage Dictionary (American Heritage Dictionary, 2014), a compromise is defined as 'something that combines qualities or elements of different things'. When the different things are classical global and classical multiple testing as described in Section 2.2, then a compromise between the two should be both powerful, which is one quality or element of the former, and informative, which is one quality or element of the latter. In ICF-based applications, such a compromise can in fact be achieved if the prior knowledge on the structure of ICF items is exploited inferentially. In this thesis, we shall differentiate between a *user-driven* and a *method-driven* compromise.

Procedures leading to a user-driven compromise

In some instances, researchers may consider it worthwhile and meaningful to perform their statistical analysis at the level of ICF components or ICF chapters. This particularly means that the individual ICF components or ICF chapters are tested separately and that, subsequently, the respective set-specific P-values are adjusted for multiplicity. For illustration, let us consider the ICF stroke study that will be presented in detail in Chapter 3. The study overall involves 130 ICF items which can be divided into four ICF components, which is the standard case, and 24 ICF chapters, respectively. Hence, the Bonferroni penalty for ICF component-specific tests equals 4, whereas for ICF chapter-specific tests it equals 24. This is considerably less severe than 130, which would be the Bonferroni penalty for ICF item-specific tests.

Obviously, the compromise that is achieved when ICF component-specific or ICF chapter-specific tests are performed is user-driven, since the user needs to decide at which level of detail the research problem shall be looked at. In situations where this decision is arbitrary rather than well-founded, however, it seems desirable to, on the one hand, exploit the prior knowledge on the structure of ICF items inferentially and, on the other hand, to dispense with any — to some extent subjective — input from the user. Procedures that enable such a user-independent compromise are discussed next.

Procedures leading to a method-driven compromise: Meinshausen's top-down procedure and improvements

When the final result is expected to be a compromise, it is often prudent to start from an extreme position.

John Maynard Keynes

The procedures that lead to a user-driven compromise focus either on the tree level of ICF components or on the tree level of ICF chapters, and hence exploit the available information on the structure of ICF items merely to a partial extent. Alternatively, it is possible to use the entire information inferentially, as recent advances in simultaneous inference have shown.

For tree-structured hypotheses such as depicted in Figure 1.3, Meinshausen (2008) introduced a simple top-down multiplicity adjustment procedure, henceforth called **Meinshausen's procedure**, which offers FWER control simultaneously over all tree levels. The procedure starts with testing the root set, that is, the overall or complete set of variables at the prespecified level α. If the null hypothesis is rejected, it continues by testing the child sets at the subsequent tree level and descends only into child sets of rejected null hypotheses. This means that child sets of sets whose null hypotheses could not be rejected are *not* tested. For any set

$$M \subseteq \{1, \ldots, p\}$$

that is tested in the top-down approach, the adjusted P-value, $P\text{-value}_M^{\text{adj}}$, is

$$P\text{-value}_M^{\text{adj}} = \min\left(\frac{p}{|M|} \cdot P\text{-value}_M^{\text{raw}}, 1\right), \tag{2.1}$$

where $P\text{-value}_M^{\text{raw}}$ is the raw P-value for set M, $|M|$ denotes the cardinality of set M, and p denotes the cardinality of the root set. It is easy to see that the P-value of the root set is unadjusted, whereas univariate P-values receive the Bonferroni adjustment which has been explained in the previous section. For an illustrative example of Meinshausen's adjustment procedure see Figure 2.1. Each tree level can thus be tested at level α, even though the FWER is controlled simultaneously over *all* tree levels at level α. Recently, Goeman and Solari (2010) and Goeman and Finos (2012) developed more elaborate sequential multiplicity adjustment procedures for tree structures which are uniformly more powerful than that of Meinshausen. For clarity and simplicity, however, their procedures are not considered in this thesis.

Provided that an effect has been ascertained in the root set, Meinshausen's procedure thus tries to attribute this effect to more specific sets or even individual variables.

Figure 2.2 provides further clarification. In particular, for the arbitrary selection of 20 ICF items from Figure 1.3, it shows an example of one possible final test result with Meinshausen's procedure: three significant ICF items (b110, b156 and d470), one significant ICF chapter (d1) and one significant ICF component (e). This well exemplifies that Meinshausen's procedure opens the door to a method-driven compromise between classical global and classical multiple testing, since here it is not determined a priori at which level of detail it will be possible to draw inferential conclusions.

As stated above, in Meinshausen's procedure the multiplicity penalty for any tested set M is $p/|M|$. Sets that comprise many variables will thus be easier to reject than sets that comprise few variables. In some applications, such an implicit prioritization of large sets may be inconvenient. In most ICF-based applications, however, this will even be desirable because it reflects the expert opinion based on which the overall sets of ICF items are composed. The ICF stroke study considered later on in Chapter 3, for example, is based on the ICF core set for stroke which comprises 130 ICF items (Geyh et al., 2004). Of this total, 5 ICF items belong to the ICF component 'body structures' (s) and 33 to the ICF component 'environmental factors' (e). In Meinshausen's procedure, the ICF components s and e will thus receive the multiplicity penalties 130/5 and 130/33; this is plausible because social and attitudinal aspects are considered more relevant for stroke patients than anatomical aspects (Geyh et al., 2004). (Otherwise, more than just five ICF items describing anatomical aspects would have been included by the health experts in the core set.) This is different for patients suffering from ankylosing spondylitis, for example. In the respective ICF core set, the ICF components s and e therefore receive the multiplicity penalties 80/19 and 80/14 (Boonen et al., 2010).

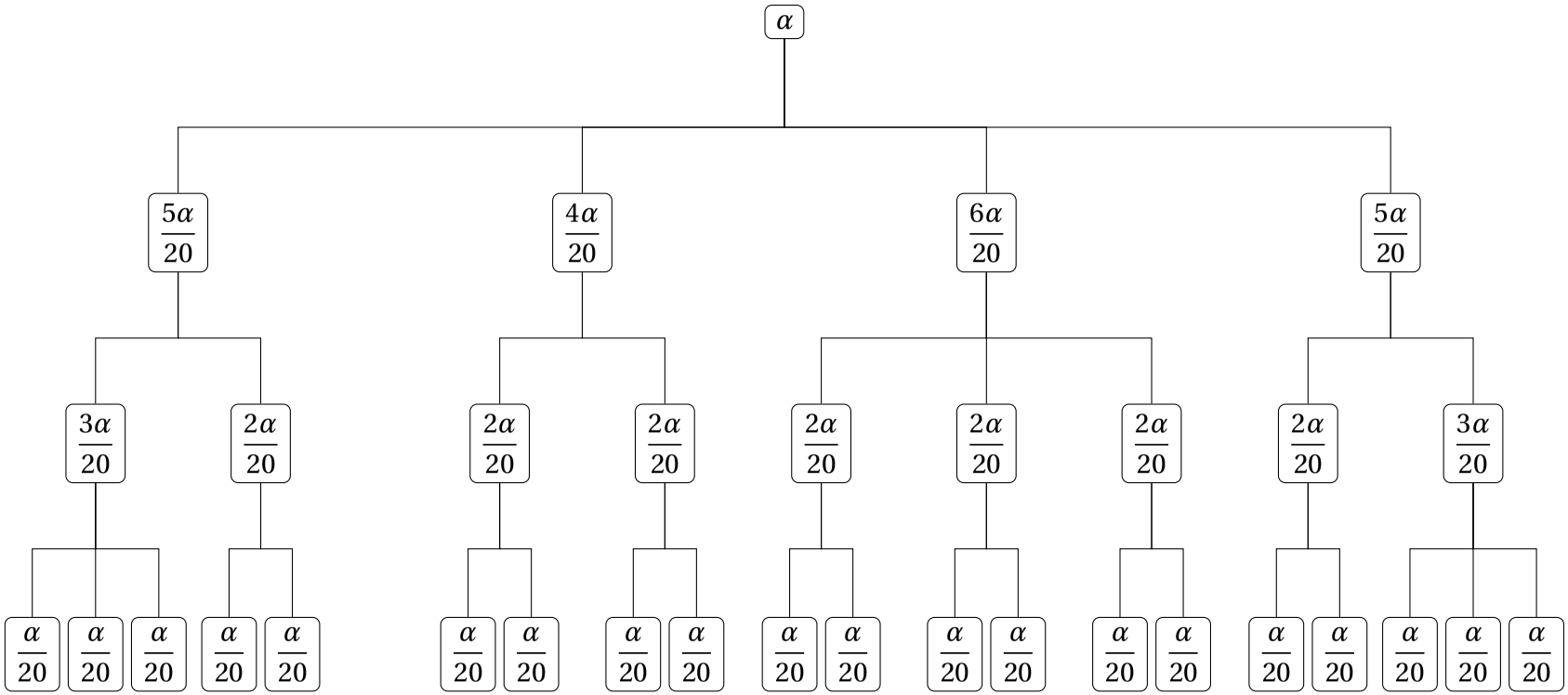

Figure 2.1.: Tree structure that corresponds to that from Figure 1.3. Instead of information on which ICF items are included in the respective sets, here the significance levels are given at which the sets are tested when using Meinshausen's top-down procedure.

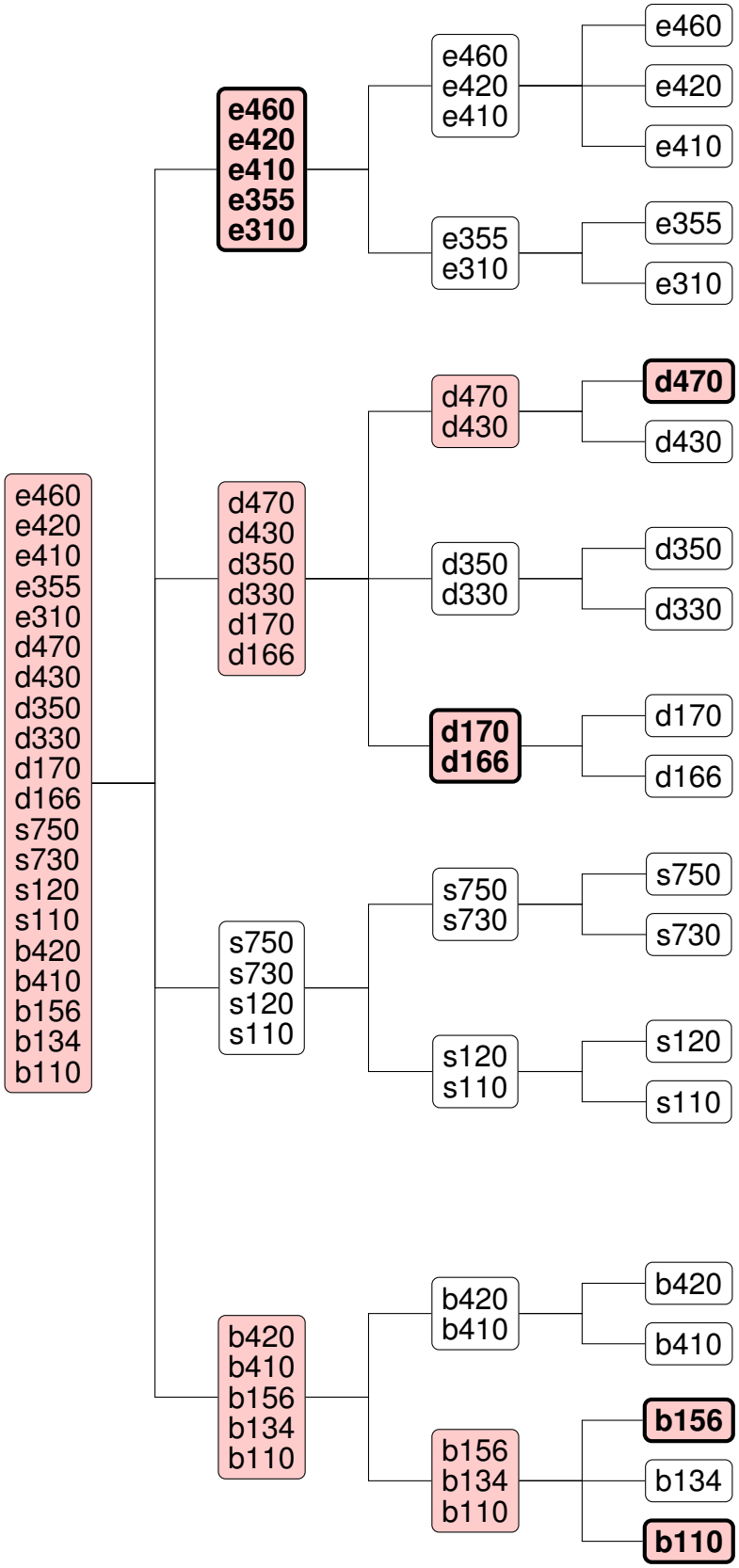

Figure 2.2.: Example of one possible test result when applying Meinshausen's top-down procedure to the tree structure from Figure 1.3. Coloured sets indicate those sets for which the null hypothesis has been rejected. The collection of sets written in **bold** constitutes the final result of the procedure.

2.4. Need for global tests for multivariate ordinal data

We have seen in Section 2.3 that, as soon as the prior knowledge on the structure of ICF items is to be exploited inferentially, this will rest on the availability of an appropriate test that provides set-specific P-values. One requirement of this test is that it remains feasible in high-dimensional data situations. This requirement becomes particularly relevant when it is Meinshausen's procedure that is used to exploit the external information inferentially, since, in its first step, Meinshausen's procedure tests the root set, and the number of ICF items included in the root set often exceeds the number of subjects in the sample. The construction of such global tests is intricate in itself and becomes particularly challenging when the data are multivariate ordinal. For illustration of one of the major issues, let us consider the data from the ICF stroke study that will be presented later on in Chapter 3. The overall 104 individuals that have participated in this study can be divided into two groups of sizes 46 and 58, and these two groups are to be compared on the basis of the respective individuals' ICF profiles. Provided that all 130 ICF items included in the root set can take three distinct values, the two 130-way contingency tables that cross-classify the 46 and 58 multivariate observations have $3^{130} \approx 1.06 \times 10^{62}$ cells; they are thus very sparse, which does not allow us to consider the full multivariate structure of the data. This shows that test statistics based on the maximum likelihood will be impossible to compute because here the maximum likelihood relies on the two 3^{130} joint distributions. Hence, test statistics are needed that involve fewer parameters.

One way to reduce the number of parameters involved is to dichotomize the multivariate ordinal data. The situation from above, however, will then not substantially improve, aside from the fact that dichotomization usually results in a loss of information. Another way to reduce the number of parameters involved is to treat the ordinally scaled data as metrically scaled and assume a multivariate normal distribution for them. However, even if we do so, test statistics that take into account the covariances between the variables in the set of interest, such as Hotelling's T^2 (Hotelling, 1931) which is the two-sample t-statistic's multivariate analogue, will still not be computable when the data are high-dimensional. This is because such test statistics will require the $p \times p$ sample covariance matrix to be inverted. For the construction of global tests for possibly high-dimensional data, and in particular when the data have been measured on an ordinal scale, it thus seems reasonable to use test statistics that dispense with the covariances between variables. Test statistics that fall within this class are **sum statistics** (Chung and Fraser, 1958; Pesarin, 2001; Ackermann and Strimmer, 2009; Pesarin and Salmaso, 2010) and **max-T-statistics** (Westfall and Young, 1993).

A sum statistic is the sum of variable-specific test statistics over a set, i.e. $\sum_{k=1}^{p} T_k$, where T_k is the kth variable-specific test statistic. The construction of global test statis-

tics in such a way is in the spirit of Pesarin's permutation-based non-parametric combination (NPC) methodology (Pesarin, 2001): the NPC methodology combines univariate P-values from traditional univariate tests through some well-chosen combination function (e.g. Fisher's product method (Fisher, 1932)) into one test statistic for the entire set. A prominent counter-concept to sum statistics are max-T-statistics which, in a different context, have already been discussed in Section 2.2. Provided that large values of the T_ks support the alternative hypothesis, a max-T-statistic is the maximum over the variable-specific test statistics over a set, i.e. $\max_k T_k$. As has been said in Section 2.2, max-T-statistics enable a short-cut of the closure test principle of Marcus et al. (1976) and hence are useful when multiple tests are to be conducted at the level of individual variables. For the assessment of *set effects*, however, sum statistics are more suitable, for two reasons. Firstly, they can be interpreted conveniently as the accumulated effect of variables over a whole set. Thus, they reflect the whole set's effect more adequately than do max-T-statistics which focus solely on the strongest individual effect. Secondly, sum statistics usually lead to more powerful tests in the presence of many weak or moderate individual effects.

Despite the fact that sum statistics lend themselves well for the construction of global hypothesis tests in diverse data situations, remarkably, explorations of their usefulness have only focused on the case of metrically scaled data so far. In fact, the literature concerned with global tests for ordinally scaled data is sparse, irrespective of whether the data situation considered is low- or high-dimensional. To the best of our knowledge, the only authors who have explicitly addressed ordinal data situations are Klingenberg et al. (2009). In particular, for the two-sample case, they proposed a one-sided permutation test for stochastic order between the marginal distributions of the ordinal variables in the set of interest. The research questions that arise from ICF-based and other multivariate ordinal data, however, are so diverse that further tests need to be developed. In this thesis we shall use the results of, inter alia, Klingenberg et al. (2009) to construct global tests of association for potentially high-dimensional multivariate ordinal data, and we will see that each of the respective test statistics falls into the sum statistics framework. We start with tests of global hypotheses in the two-group scenario in Chapter 3, and subsequently move on to tests of global hypotheses in the GLM in Chapter 4.

3. Testing global hypotheses in the two-group scenario

This chapter is concerned with two-sample global tests for sets of ordinally scaled variables in possibly high-dimensional set-ups; it is thus devoted to research questions that can be framed as two-group comparisons. Such comparisons constitute an important problem in statistical practice. In the ICF context, for instance, two-group comparisons of ICF profiles or patterns have been the major objective of numerous studies conducted world-wide (Holper et al., 2010; Herrmann et al., 2011; Tschiesner et al., 2011). Section 3.1 provides an overview of the particular contents of this chapter. The chapter, apart from Sections 3.2.4, 3.4.1, 3.4.3 and 3.6, is mainly based on Jelizarow et al. (2014a).

3.1. Guideline through the chapter

The structure of this chapter is as follows. Section 3.2 defines and discusses the two closely related problems that are addressed. In both instances, the null hypothesis is that the ordinal variables' marginal distributions are identical between the two groups to be compared. Joint distributions are left unspecified. The alternative hypotheses are that, for at least one of the ordinal variables in the set to be tested, there is between-group inhomogeneity and, as a special case thereof, direction-independent stochastic order of the respective marginal distributions. In Section 3.3 simple test statistics that are sensitive towards the alternative hypotheses from Section 3.2 are proposed. In this context we shall see that, under the working assumption of independence between variables, the test statistic of Klingenberg et al. (2009) reduces to the sum of univariate one-sided CA test statistics, which provides important insight into the power properties of the respective test. For inference, we focus on the popular permutation procedure. The latter is known to be valid only if the ordinal variables' *joint* distributions are identical under the null hypothesis, which is not necessarily so under the null hypothesis that we consider. This issue is addressed in Section 3.4, and the so-called **null dilemma** that arises when no superior inference method is available is discussed. By means of simulations, Section 3.5 subsequently examines the permutation procedure's

robustness properties under theoretically unfavourable conditions. Section 3.6 briefly presents a bootstrap-based procedure which, however, turns out not to be an appropriate alternative to the permutation procedure. Section 3.7 illustrates the proposed tests' application and practical benefits with data from an ICF stroke study. Finally, Section 3.8 closes the chapter with a short summary and discussion of its contents.

3.2. Global hypotheses

3.2.1. Notation and preliminaries

We address the scenario in which two independent groups of sizes n_1 and n_2, $n_1 + n_2 = n$, are to be compared on the basis of p-dimensional ordinal data vectors, and we assume that the p ordinal variables that underlie the data have the same number $c \geq 2$ of categories. (The case of possibly unequal numbers of categories will be discussed briefly in Section 3.8.) For convenience of notation, let the ordered categories of unknown distance be labelled with numbers 1 to c. Suppose that the n_g multivariate observations in group g, $g = 1,2$, form an independent and identically distributed (i.i.d.) sample of a $p \times 1$ random vector

$$\boldsymbol{X}_g = \left(X_{g1}, \ldots, X_{gp}\right)^\top$$

which has a multivariate multinomial distribution $\boldsymbol{\Pi}_g$ with unknown dependence structure. Let $\pi_g(v_1, \ldots, v_p)$ denote the joint probability $\Pr(X_{g1} = v_1, \ldots, X_{gp} = v_p)$ for an entire profile or pattern in group g, where $v_k \in \{1, \ldots, c\}$ is the category that has been observed for the kth ordinal variable X_{gk}, $k = 1, \ldots, p$. Unless further specified when the two groups are considered different, it seems natural to test the null hypothesis

$$H_0 : \boldsymbol{X}_1 \stackrel{d}{=} \boldsymbol{X}_2$$

against the alternative hypothesis

$$H_1 : \boldsymbol{X}_1 \stackrel{d}{\neq} \boldsymbol{X}_2,$$

where '$\stackrel{d}{=}$' means equality in distribution. H_0 (i.e. $\pi_1(v_1, \ldots, v_p) = \pi_2(v_1, \ldots, v_p)$ for all c^p possible sequences $(v_1, \ldots, v_p) \in \{1, \ldots, c\}^p$) is referred to as **identical joint distribution (IJD)**, and H_1 (i.e. $\pi_1(v_1, \ldots, v_p) \neq \pi_2(v_1, \ldots, v_p)$ for at least one $(v_1, \ldots, v_p) \in \{1, \ldots, c\}^p$) as **non-identical joint distribution (NJD)**. However, because confirmation of NJD carries little information as to *why* it has been confirmed, the problem 'IJD against NJD' is seldom of interest in practice.

3.2.2. Marginal inhomogeneity

Intuitively, rather than to test their joint distribution, it seems preferable to test the one-way multinomial distributions $\Pi_{gk} = \{\pi_{gk}(v)\}_{v=1}^{c}$ of the random variables X_{gk}, with $\pi_{gk}(v)$ denoting the marginal probability $\Pr(X_{gk} = v)$, $v \in \{1,\ldots,c\}$. The associated hypotheses are

$$H_0^m : \bigcap_{k=1}^{p} H_{0k} = \bigcap_{k=1}^{p} \{X_{1k} \overset{d}{=} X_{2k}\} \tag{3.1}$$

and

$$H_1^m : \bigcup_{k=1}^{p} H_{1k} = \bigcup_{k=1}^{p} \{X_{1k} \overset{d}{\neq} X_{2k}\}, \tag{3.2}$$

where the intersection null hypothesis H_0^m in (3.1) (i.e. $\{\pi_{1k}(v)\}_{v=1}^{c} = \{\pi_{2k}(v)\}_{v=1}^{c}$ simultaneously for all k) is referred to as **simultaneous marginal homogeneity (SMH)**, and the alternative hypothesis H_1^m in (3.2) (i.e. $\{\pi_{1k}(v)\}_{v=1}^{c} \neq \{\pi_{2k}(v)\}_{v=1}^{c}$ for at least one k) as **marginal inhomogeneity (MI)**. For $c = 2$, this problem was tackled by Agresti and Klingenberg (2005). Evidently,

$$\text{IJD} \Rightarrow \text{SMH}.$$

IJD is thus more restrictive than SMH. For $p = 1$, IJD and SMH are equivalent because, unlike the normal distribution for which mean μ and variance σ^2 need to be specified separately, the multinomial distribution is fully determined by its mean. We come back to the distinction between both null hypotheses and its importance in permutation-based inference in Section 3.4.

3.2.3. Marginal order

As in the ICF stroke study presented later on in Section 3.7, it will be the primary aim in many other studies to detect MI. In some instances, however, the information provided under MI may be too unspecific and the research question may focus on special cases of MI. The most important special case of MI is marginal stochastic order. The random variables X_{1k} and X_{2k} are stochastically ordered if either

(a) $\Pr(X_{1k} \leq v) \geq \Pr(X_{2k} \leq v)$, which is written $X_{1k} \preceq X_{2k}$, or

(b) $\Pr(X_{1k} \leq v) \leq \Pr(X_{2k} \leq v)$, which is written $X_{1k} \succeq X_{2k}$, for all $v \in \{1,\ldots,c\}$.

Without loss of generality, if the inequality in (a) is strict for at least one v, X_{1k} and X_{2k} are said to be stochastically strictly ordered, which is written $X_{1k} \prec X_{2k}$. Let the narrower alternative hypothesis be

$$\tilde{H}_1^m : \bigcup_{k=1}^{p} \tilde{H}_{1k} = \bigcup_{k=1}^{p} \{\{X_{1k} \prec X_{2k}\} \cup \{X_{1k} \succ X_{2k}\}\}, \tag{3.3}$$

where $\{X_{1k} \prec X_{2k}\}$ and $\{X_{1k} \succ X_{2k}\}$ are mutually exclusive for all k. Under $\tilde{H}_1^m$ in (3.3), we thus have either $\Pr(X_{1k} \leq v) > \Pr(X_{2k} \leq v)$ or $\Pr(X_{1k} \leq v) < \Pr(X_{2k} \leq v)$ for at least one k and v, and we shall refer to this two-sided alternative as **marginal order (MO)**, noting that

$$\text{MO} \Rightarrow \text{MI},$$

i.e. (3.3) $\Rightarrow$ (3.2). The one-sided counterpart of MO (i.e. $\bigcup_{k=1}^{p}\{X_{1k} \prec X_{2k}\}$) was tackled by Klingenberg et al. (2009), motivated by the statistical analysis of ordinally scaled adverse effects data from toxicity studies. Here it is plausible to assume that there is equal or greater chance of observing severe effects (i.e. high categories) in the treatment group (group 2) than in the placebo group (group 1). For ICF studies, a similar assumption will be rarely plausible. In the above-mentioned ICF stroke study, for instance, the two groups to be compared are Asian and European stroke patients, and it would not come as a surprise if some body functions were more severely impaired among Asian patients than among European patients, while the opposite holds for other functions. Because we are in fact usually equally interested in '$X_{1k} \prec X_{2k}$' and '$X_{1k} \succ X_{2k}$' contributions to set effects, it is sensible to consider the direction-independent stochastic order alternative MO. Compared to MI, it is the more appropriate choice if we wish to explicitly take into account the natural ordering of the variables' c categories. If this is not essential in the application at hand, it seems reasonable to choose MI which is broader in the sense that it includes but is not restricted to stochastically ordered one-way multinomial distributions. Given that the problems 'SMH against MI' and 'SMH against MO' are closely related and similarly widespread in and beyond ICF-based applications, both are discussed in the present thesis, and the former is exemplified in Section 3.7.

3.2.4. *Excursus:* the case of ordered joint distributions

In the previous section it has been mentioned that the one-sided counterpart of MO, namely $\bigcup_{k=1}^{p}\{X_{1k} \prec X_{2k}\}$, was tackled by Klingenberg et al. (2009) within the context of toxicity studies in which the presence and severity of adverse effects is typically captured through ordinal variables, and that the authors assumed severe effects (i.e. high categories) to be observed equally or more likely in the treatment group (group 2) than in the placebo group (group 1). This assumption, provided that it can be made *jointly* for all adverse effects, is the joint stochastic order assumption. Unlike marginal stochastic order which refers to the distribution of random variables (see Section 3.2.3), joint stochastic order refers to the distribution of entire random vectors. In more explicit terms, the random vectors $\boldsymbol{X}_1 = (X_{11}, ..., X_{1p})^T$ and $\boldsymbol{X}_2 = (X_{21}, ..., X_{2p})^T$ are stochastically ordered if either

(a) $\Pr(X_{11} \leq v_1, \ldots, X_{1p} \leq v_p) \geq \Pr(X_{21} \leq v_1, \ldots, X_{2p} \leq v_p)$, which is written $\boldsymbol{X}_1 \preceq \boldsymbol{X}_2$, or

(b) $\Pr(X_{11} \leq v_1, \ldots, X_{1p} \leq v_p) \leq \Pr(X_{21} \leq v_1, \ldots, X_{2p} \leq v_p)$, which is written $\boldsymbol{X}_1 \succeq \boldsymbol{X}_2$, for all $(v_1, \ldots, v_p) \in \{1, \ldots, c\}^p$.

Without loss of generality, if the inequality in (a) is strict for at least one $(v_1, \ldots, v_p)$, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are said to be stochastically strictly ordered, which is written $\boldsymbol{X}_1 \prec \boldsymbol{X}_2$ (Shaked and Shanthikumar, 2006). Under the assumption $\boldsymbol{X}_1 \preceq \boldsymbol{X}_2$ or $\boldsymbol{X}_1 \succeq \boldsymbol{X}_2$, Klingenberg et al. (2009) have now shown that

$$H_0 : \boldsymbol{X}_1 \stackrel{d}{=} \boldsymbol{X}_2 \;\Leftrightarrow\; H_0^m : \bigcap_{k=1}^{p} \{X_{1k} \stackrel{d}{=} X_{2k}\},$$

i.e. IJD $\Leftrightarrow$ SMH. Stated differently, under the assumption that the group-specific joint distributions of the p ordinal variables in the set to be tested are stochastically ordered, there is no difference anymore between the null hypotheses IJD and SMH. This connection becomes important in permutation-based inference which, as will be elaborated on in Section 3.4, is valid only if the null hypothesis is IJD.

3.3. Global test statistics

3.3.1. Testing for marginal inhomogeneity

Generalization of Agresti and Klingenberg's test statistic to more than two categories

To test for MI in the case $c = 2$, Agresti and Klingenberg (2005) proposed a test statistic that is a quadratic form in the vector of differences in sample means. We shall see below that their test statistic can easily be generalized to the case $c \geq 2$, even though in most practical situations it will not be computable without additional assumptions on the covariance structure between variables.

Let n_{gkv} be the number of subjects with observed category v of the kth ordinal variable in group g, with respective sample proportion

$$\hat{\pi}_{gk}(v) = \frac{n_{gkv}}{n_g}.$$

As $\hat{\pi}_{gk}(c) = 1 - \sum_{v=1}^{c-1} \hat{\pi}_{gk}(v)$, the truncated $(c-1)p \times 1$ vector of marginal sample proportions or, equivalently, the vector of sample means for group g is

$$\hat{\boldsymbol{\pi}}_g = \left(\hat{\pi}_{g1}(1), \ldots, \hat{\pi}_{g1}(c-1), \ldots\ldots, \hat{\pi}_{gp}(1), \ldots, \hat{\pi}_{gp}(c-1)\right)^\top.$$

Let now

$$\boldsymbol{d} = \hat{\boldsymbol{\pi}}_2 - \hat{\boldsymbol{\pi}}_1$$

denote the vector of differences in marginal sample proportions with entries $d_k(v) = \hat{\pi}_{2k}(v) - \hat{\pi}_{1k}(v)$. From basic multinomial theory it is known that $\mathrm{E}(\boldsymbol{d}) = \boldsymbol{\pi}_2 - \boldsymbol{\pi}_1$, and that the $(c-1)p \times (c-1)p$ covariance matrix

$$\mathrm{Cov}(\boldsymbol{d}) = \boldsymbol{\Sigma}$$

has the entries

$$\mathrm{Var}(d_k(v)) = \sum_{g=1}^{2} \frac{\pi_{gk}(v)\left(1-\pi_{gk}(v)\right)}{n_g}, \tag{3.4}$$

$$\mathrm{Cov}(d_k(v), d_k(\tilde{v}))_{v\neq\tilde{v}} = -\sum_{g=1}^{2} \frac{\pi_{gk}(v)\pi_{gk}(\tilde{v})}{n_g}, \tag{3.5}$$

$$\mathrm{Cov}(d_k(v), d_{\tilde{k}}(\tilde{v}))_{k\neq\tilde{k}} = \sum_{g=1}^{2} \frac{\pi_{gk\tilde{k}}(v,\tilde{v}) - \pi_{gk}(v)\pi_{g\tilde{k}}(\tilde{v})}{n_g}. \tag{3.6}$$

A test statistic sensitive towards MI can now be constructed as the simple quadratic form $\boldsymbol{d}^\top \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{d}$, with $\hat{\boldsymbol{\Sigma}}$ being the sample version of $\boldsymbol{\Sigma}$. As becomes apparent from (3.4) and (3.5), the variances and covariances within variables can easily be estimated from the sample proportions $\hat{\pi}_{gk}(v)$. Under the null hypothesis SMH, we can pool the data to obtain the more efficient pooled estimator $\hat{\boldsymbol{\pi}}_0$ with entries

$$\hat{\pi}_{0k}(v) = \frac{n_{1kv} + n_{2kv}}{n_1 + n_2}.$$

The covariances between variables from (3.6), however, depend on the two-way multinomial distributions $\Pi_{gk\tilde{k}} = \left\{\pi_{gk\tilde{k}}(v,\tilde{v})\right\}_{v,\tilde{v}=1}^{c}$, where $\pi_{gk\tilde{k}}(v,\tilde{v}) = \Pr(X_{gk} = v, X_{g\tilde{k}} = \tilde{v})$, $k \neq \tilde{k}$. Their estimation proves to be problematic. Firstly, when we pool the data for this purpose, we additionally assume that the two groups have the same $\binom{p}{2}$ two-way multinomial distributions under the null hypothesis, which is more restrictive than SMH. This assumption was made by Agresti and Klingenberg (2005), rendering their test statistic an analogue of Hotelling's T^2 for multivariate binary data. Secondly, even pooled data often lead to sparse two-way contingency tables unless n is very large and/or $c = 2$. In most ICF-based applications, an approach along the lines of Agresti and Klingenberg (2005) is therefore bound to fail. In the previously mentioned ICF stroke study, for instance, even with the coarser three-level ordinal scale for all ICF items, 4818 of the $\binom{130}{2} = 8385$ (3×3) tables have one or more empty cells, rendering numerous $\pi_{0k\tilde{k}}(v,\tilde{v})$s inestimable. As a result, we may obtain an estimate of $\boldsymbol{\Sigma}$ that is not positive definite. To prevent this, one needs to simplify the covariance structure between variables considerably. Here we assume working independence, which results

in an estimated covariance matrix $\hat{\boldsymbol{\Sigma}}_0$ that is block-diagonal. The kth null-estimated $(c-1)\times(c-1)$ block and its inverse are given by

$$\hat{\boldsymbol{\Sigma}}_{0k} = \frac{n_1+n_2}{n_1 n_2}\left(\operatorname{diag}(\hat{\boldsymbol{\pi}}_{0k}) - \hat{\boldsymbol{\pi}}_{0k}\hat{\boldsymbol{\pi}}_{0k}^\top\right)$$

and

$$\hat{\boldsymbol{\Sigma}}_{0k}^{-1} = \frac{n_1 n_2}{n_1+n_2}\left[\operatorname{diag}(\hat{\boldsymbol{\pi}}_{0k})^{-1} + \left(1-\sum_{v}^{c-1}\hat{\pi}_{0k}(v)\right)^{-1}\mathbf{1}\mathbf{1}^\top\right],$$

respectively, where $\mathbf{1}$ is a $(c-1)\times 1$ vector of ones. Then, the quadratic form can be written as $\sum_{k=1}^{p}\boldsymbol{d}_k^\top\hat{\boldsymbol{\Sigma}}_{0k}^{-1}\boldsymbol{d}_k$, which is the sum of variable-specific test statistics. It can readily be verified that the p summands are equivalent to univariate χ^2 test statistics (of independence if the marginal totals of the respective two-by-c contingency tables are considered random or of homogeneity if certain marginal totals of the respective two-by-c contingency tables are considered fixed), each with an asymptotic χ^2 null distribution with degrees of freedom df $= c-1$. We shall therefore refer to the overall test statistic as χ^2 **sum statistic**, and we write

$$\sum_{k=1}^{p}\boldsymbol{d}_k^\top\hat{\boldsymbol{\Sigma}}_{0k}^{-1}\boldsymbol{d}_k =: Q_{\chi^2}. \tag{3.7}$$

Under independence between variables, this χ^2 sum statistic has an asymptotic χ^2 null distribution with df $= p(c-1)$. However, independence rarely holds and is particularly questionable in the ICF context where ICF items from the same set describe more similar aspects than ICF items from different sets (e.g. the ICF items 'memory functions' (b144) and 'attention functions' (b140) measure more similar aspects than the ICF items 'memory functions' (b144) and 'washing oneself' (d510)). As mentioned earlier, we will therefore turn our attention to null distributions derived via the permutation procedure which accounts for the dependence between variables by resampling entire multivariate observations. For further information on this issue see Section 3.4.

As a side remark, it should be noted that, from a broader perspective, the sum statistic Q_{χ^2} may be seen as **Hotelling-type** test statistic for multivariate ordinal data, under the assumption that the variables in the set to be tested are independent in both groups. Alternatively, one could likewise construct Hotelling-type test statistics under less stringent assumptions; for example, one could assume that the $\binom{p}{2}$ two-way multinomial distributions are uniform, paired with equality in the two groups to be compared. In high-dimensional data situations, however, test statistics of this kind will not necessarily be computable, and will therefore not be discussed in this thesis.

Excursus: relationship between the test statistic Q_{χ^2} and the partial least squares method

In the special case $c = 2$, there is a noteworthy connection between the test statistic Q_{χ^2} and the partial least squares (PLS) method[1] (Martens and Naes, 1989; Stone and Brooks, 1990; Brown, 1993; Frank and Friedman, 1993; Garthwaite, 1994; Martens, 2001; Boulesteix, 2005). This excursus briefly looks at this connection, despite the fact that, in the methodical context of this thesis, the case $c = 2$ is merely of little relevance.

The PLS method can be used for prediction of some outcome variable on the basis of continuous covariates, and it has the valuable feature that it remains feasible in high-dimensional data situations. The method summarizes the possibly many covariates of interest into a small number of so-called PLS components. In particular, PLS components are uncorrelated linear transformations of the covariates. The first PLS component is the most important one: among all PLS components, it has maximum covariance with the outcome of interest. Consider now the two-group problem addressed in this chapter, and let the variables of interest be binary. Suppose that the PLS method is applied to this problem, where the group membership is treated as binary outcome and the binary variables are treated as continuous covariates. Then, under the condition that both the outcome and the covariates are scaled to have unit variance, it can be shown that the covariance between the first PLS component and the binary outcome is, up to some factor, equivalent to the χ^2 sum statistic Q_{χ^2}. Under the assumption of working independence between variables, this implies equivalence with the test statistic of Agresti and Klingenberg (2005).

3.3.2. Testing for marginal order

Generalization of Klingenberg's test statistic to two-sided alternatives

To construct a test statistic that is sensitive towards MO, we can exploit the results from Section 3.3.1. Let

$$\hat{\boldsymbol{\pi}}'_g = (\hat{\pi}_{g1}(1),\ldots,\hat{\pi}_{g1}(c),\ldots\ldots,\hat{\pi}_{gp}(1),\ldots,\hat{\pi}_{gp}(c))^\top$$

denote the non-truncated $cp \times 1$ vector of marginal sample proportions for group g, and be

$$\boldsymbol{d}' = \hat{\boldsymbol{\pi}}'_2 - \hat{\boldsymbol{\pi}}'_1.$$

[1] This connection was both found and proved by Anne-Laure Boulesteix from the LMU Munich after we had discussed the test statistic Q_{χ^2}. I would like to sincerely thank her for the permission to mention the results of her work in this thesis.

In order to take into account that the variables' categories are naturally ordered, we can multiply $\boldsymbol{d}'$ with a $p \times cp$ matrix

$$\boldsymbol{U} = \operatorname{diag}\left(\boldsymbol{u}_1^\top, \ldots, \boldsymbol{u}_p^\top\right),$$

where $\boldsymbol{u}_k^\top = (u_k(1), \ldots, u_k(c))$ is the kth vector of scores that need to be assigned a priori to the respective variable's c categories. Typically, the scores are chosen such that they increase or decrease monotonically. In the ICF context, for example, one can assign 1 to 'no problem', 2 to 'mild to moderate problem' and 4 to 'severe to complete problem' if one believes that the distance between 'mild to moderate problem' and 'severe to complete problem' is twice the distance between 'no problem' and 'mild to moderate problem'. (A more profound discussion of the choice of scores will be provided later on in Section 4.3.2.) What we obtain is

$$\boldsymbol{s} = \boldsymbol{U}\boldsymbol{d}',$$

which is the $p \times 1$ vector of mean score differences with covariance matrix

$$\operatorname{Cov}(\boldsymbol{s}) = \boldsymbol{\Delta} = \boldsymbol{U}\operatorname{Cov}(\boldsymbol{d}')\boldsymbol{U}^\top.$$

It is sensible to estimate $\operatorname{Cov}(\boldsymbol{d}')$ under SMH based on the pooled $\hat{\boldsymbol{\pi}}_0'$ and, eventually for the same reasons outlined in Section 3.3.1, the assumption of working independence between variables. Then, the estimated $p \times p$ covariance matrix $\hat{\boldsymbol{\Delta}}_0$ is block-diagonal, and the kth null-estimated block is given by the scalar

$$\hat{\delta}_{0k} = \boldsymbol{u}_k^\top \hat{\operatorname{Cov}}(\boldsymbol{d}_k')\boldsymbol{u}_k.$$

To test for the one-sided counterpart of MO (i.e. $\bigcup_{k=1}^p \{X_{1k} \prec X_{2k}\}$), Klingenberg et al. (2009) employed the test statistic $p^{-1}\mathbf{1}^\top \hat{\boldsymbol{\Delta}}_0^{-\frac{1}{2}} \boldsymbol{s} = p^{-1}\sum_{k=1}^p \hat{\delta}_{0k}^{-\frac{1}{2}} s_k$, which is equivalent to the sum of variable-specific standardized mean score differences (up to the factor p^{-1}). Hence, to test for MO where stochastic order but not its direction is relevant, we propose to use the direction-independent test statistic $\sum_{k=1}^p \hat{\delta}_{0k}^{-1} s_k^2$, which is the sum of squared variable-specific standardized mean score differences. As with the χ^2 sum statistic Q_{χ^2} from (3.7), here the p summands likewise turn out to be well-known in the literature: a closer look at them reveals that the test statistic proposed is equivalent to the sum of traditional univariate CA trend test statistics (Cochran, 1954; Armitage, 1955), for any choice of scores. The proof is a simple calculation and is provided further below. Hence, we shall refer to the test statistic proposed as **CA sum statistic**, and we write

$$\sum_{k=1}^p \hat{\delta}_{0k}^{-1} s_k^2 =: Q_{\mathrm{CA}}. \tag{3.8}$$

The CA sum statistic is thus a generalization of the traditional CA test statistic to higher dimensions. As the univariate CA test statistic has an asymptotic χ^2 null distribution with df = 1, under independence between variables, the CA sum statistic has a χ^2 null distribution with df = p.

Finally, it should be pointed out that the connection between the traditional CA test statistic and the CA sum statistic Q_{CA} from (3.8) implies one further connection. As has already been said in Section 2.2, the traditional test of Cochran (1954) and Armitage (1955), which is two-sided, is often better known in the one-sided formulation of Freidlin et al. (2002). Given that the test statistic of the two-sided CA test is the square of the direction-dependent test statistic of the one-sided CA test, we can thus say that, up to the factor p^{-1}, Klingenberg's test statistic from above is equivalent to the sum of univariate one-sided CA test statistics.

Relationship between Klingenberg's test statistic and the one-sided Cochran-Armitage test statistic: proof and practical consequences

We now prove the above-mentioned relationship between Klingenberg's test statistic and the one-sided CA test statistic. Under the assumption of independence between variables, the test statistic of Klingenberg et al. (2009) equals $p^{-1}\sum_{k=1}^{p}\hat{\delta}_{0k}^{-\frac{1}{2}}s_k$. It is therefore sufficient to examine the univariate case. We refer to the notation that was introduced in Section 3.3. With $n_{.kv} = n_{1kv} + n_{2kv}$, for the kth component we obtain

$$
\begin{aligned}
\hat{\delta}_{0k}^{-\frac{1}{2}} s_k &= \left(\boldsymbol{u}_k^\top \hat{\text{Cov}}(\boldsymbol{d}_k')\boldsymbol{u}_k\right)^{-\frac{1}{2}} \boldsymbol{u}_k^\top \boldsymbol{d}_k' \\
&= \left\{\frac{n}{n_1 n_2}\left[\boldsymbol{u}_k^\top\left(\text{diag}(\hat{\boldsymbol{\pi}}_{0k}') - \hat{\boldsymbol{\pi}}_{0k}'\hat{\boldsymbol{\pi}}_{0k}'^\top\right)\boldsymbol{u}_k\right]\right\}^{-\frac{1}{2}} \boldsymbol{u}_k^\top \boldsymbol{d}_k' \\
&= \left\{\frac{n}{n_1 n_2}\left[\frac{1}{n}\sum_{v=1}^{c} u_k(v)^2 n_{.kv} - \left(\frac{1}{n}\sum_{v=1}^{c} u_k(v) n_{.kv}\right)^2\right]\right\}^{-\frac{1}{2}} \times \\
&\quad \sum_{v=1}^{c} u_k(v)\left(\frac{n_{2kv}}{n_2} - \frac{n_{1kv}}{n_1}\right) \\
&= \left\{\frac{1}{n n_1 n_2}\left[n\sum_{v=1}^{c} u_k(v)^2 n_{.kv} - \left(\sum_{v=1}^{c} u_k(v) n_{.kv}\right)^2\right]\right\}^{-\frac{1}{2}} \times \\
&\quad \left(\frac{n_1^2 n_2^2}{n^2}\right)^{-\frac{1}{2}} \sum_{v=1}^{c} u_k(v)\left(\frac{n_1 n_{2kv}}{n} - \frac{n_2 n_{1kv}}{n}\right) \\
&= \left\{\frac{n_1 n_2}{n^3}\left[n\sum_{v=1}^{c} u_k(v)^2 n_{.kv} - \left(\sum_{v=1}^{c} u_k(v) n_{.kv}\right)^2\right]\right\}^{-\frac{1}{2}} \times \\
&\quad \sum_{v=1}^{c} u_k(v)\left(\frac{n_1 n_{2kv}}{n} - \frac{n_2 n_{1kv}}{n}\right).
\end{aligned}
$$

Since the right-hand side corresponds to $\frac{T_k}{\sqrt{\mathrm{Var}(T_k)}}$ with

$$T_k = \sum_{v=1}^{c} u_k(v)\Big(\frac{n_1 n_{2kv}}{n} - \frac{n_2 n_{1kv}}{n}\Big),$$

which is the most common formulation of the one-sided CA test statistic (Freidlin et al., 2002; Neuhäuser, 2010), equivalence of Klingenberg's test statistic and the sum of one-sided variable-specific CA test statistics follows directly (up to the factor p^{-1}). It is now easy to see that the CA sum statistic Q_{CA} must be equivalent to the sum of traditional two-sided variable-specific CA test statistics.

The connection between Q_{CA} and the traditional CA test statistic deserves special attention because it provides important information on which inferences may or may not be drawn from a test result. The crux is that the CA test statistic is intended to test for some suspected trend in the binomial proportions across the c ordered categories. Which particular trend the test statistic will be sensitive towards is determined by scores which are in one-to-one correspondence with the scores $u_k(v)$ from above. For the CA sum statistic Q_{CA}, we suppose that the scores are uniform over all k (i.e. $u_k(v) = u(v)$) and that they increase or decrease monotonically. Note that uniform scores are not compulsory, but they are a convenient choice in most applications. It is now easily verfied that MO is fulfilled if there is some monotonic trend in the binomial proportions for at least one variable in the set to be tested. The reverse, however, is not true. To further clarify this point, let us consider one simple example. Suppose that, for the kth ICF item in some ICF study, we have observed the two-by-three frequency table

	1	2	3	
Group 1	6	5	5	16
Group 2	4	6	5	15
	10	11	10	31

,

where the numbers 1, 2 and 3 label the ordered categories 'no problem', 'mild to moderate problem' and 'severe to complete problem'. From this table we can calculate $\hat{\Pr}(X_{1k} = 1) = 6/16 = 0.3750$, $\hat{\Pr}(X_{1k} = 2) = 5/16 = 0.3125$, $\hat{\Pr}(X_{1k} = 3) = 5/16 = 0.3125$, $\hat{\Pr}(X_{2k} = 1) = 4/15 = 0.266\bar{6}$, $\hat{\Pr}(X_{2k} = 2) = 6/15 = 0.4000$ and $\hat{\Pr}(X_{2k} = 3) = 5/15 = 0.333\bar{3}$. It is now easy to see that

$$\begin{aligned}
\hat{\Pr}(X_{1k} \leq 1) &> \hat{\Pr}(X_{2k} \leq 1),\\
\hat{\Pr}(X_{1k} \leq 2) &> \hat{\Pr}(X_{2k} \leq 2),\\
\hat{\Pr}(X_{1k} \leq 3) &= \hat{\Pr}(X_{2k} \leq 3).
\end{aligned}$$

The data observed thus suggest that it holds $X_{1k} \prec X_{2k}$, and hence that MO is true. A monotonic trend in the kth ICF item's binomial proportions now requires that either

$$\frac{n_{2k1}}{n_{\cdot k1}} \leq \frac{n_{2k2}}{n_{\cdot k2}} \leq \cdots \leq \frac{n_{2kc}}{n_{\cdot kc}}$$

with $\frac{n_{2k1}}{n_{\cdot k1}} < \frac{n_{2kc}}{n_{\cdot kc}}$ or

$$\frac{n_{2k1}}{n_{\cdot k1}} \geq \frac{n_{2k2}}{n_{\cdot k2}} \geq \cdots \geq \frac{n_{2kc}}{n_{\cdot kc}}$$

with $\frac{n_{2k1}}{n_{\cdot k1}} > \frac{n_{2kc}}{n_{\cdot kc}}$, where

$$n_{\cdot kv} = n_{1kv} + n_{2kv}.$$

With the data from the above table, however, we obtain $(6/10 = 0.6000, 5/11 = 0.4545, 5/10 = 0.5000)$ for group 1 and $(4/10 = 0.4000, 6/11 = 0.5455, 5/10 = 0.5000)$ for group 2, which suggests a non-monotonic rather than a monotonic trend in the binomial proportions across the ICF item's three categories. This shows that the presence of MO does not automatically come along with the presence of a monotonic trend in the binomial proportions. We can thus say that a monotonic trend in the binomial proportions constitutes an alternative hypothesis that is narrower than MO. Consequently, because statistical hypothesis tests that rest upon the CA sum statistic Q_{CA} are essentially designed to detect such monotonic trends, they may have low power to detect MO if there is no such trend. This should be kept in mind whenever MO, perhaps unexpectedly, could not be confirmed.

Compared to the χ^2 sum statistic Q_{χ^2}, the CA sum statistic Q_{CA} will result in more power when the suspected trend or its inverse is correct for all k for which H_{1k} in (3.2) is fulfilled, but it is likely to result in considerably less power otherwise. In the special case $c = 2$, the two sum statistics Q_{χ^2} and Q_{CA} are equivalent for any choice of scores and will therefore result in equally powerful tests. In the case $c > 2$, Q_{χ^2} and Q_{CA} are equivalent only if we use the data-driven scores of Zheng et al. (2009) which, however, do not necessarily increase or decrease monotonically.

3.3.3. Multivariate versus marginal perspective

The sum statistic Q_{χ^2} has been constructed to test for MI, and the sum statistic Q_{CA} has been constructed to test for the narrower alternative MO. Both sum statistics have been presented as special cases of multivariate quadratic forms, under the assumption that the variables in the set of interest are independent. This multivariate perspective is beneficial, particularly because it immediately clarifies why the independence assumption will be difficult to circumvent in real-life applications where n is typically

small to moderate, p is moderate to large and $c > 2$. Nevertheless, the fact that both Q_{χ^2} and Q_{CA} have turned out to be composed of well-known traditional univariate test statistics suggests to look at them from the less sophisticated yet popular marginal perspective which, already from the outset, frees the researcher from the need to model the multivariate dependence structure of the variables in the set to be tested. This is in the spirit of the permutation-based NPC methodology of Pesarin (2001) which has previously been sketched in Section 2.4 and which includes tests based on test statistics of the form $\sum_{k=1}^{p} T_k$ as a special case, where T_k is the kth variable-specific test statistic.

Irrespective of which of the above perspectives we wish to adopt, to be able to perform statistical hypothesis tests based on the sum statistics proposed, we still need their distributions under the null hypothesis. For inferences to be valid, the latter should take the multivariate dependence structure in the data into account, even if the sum statistics do not so. Permutation-based null distributions, which are the subject of Section 3.4, can accomplish this, albeit only at the price of an assumption that may not be justified in practice.

3.4. Permutation-based global inference about marginal distributions

3.4.1. Permutation null distribution of a test statistic

For completeness, let us start with how a test statistic's permutation null distribution is obtained. Let

$$G_i \in \{1, 2\}$$

be the group label of the ith subject, $i = 1, \ldots, n$. The permutation null distribution of any test statistic, say Q, is then derived as follows.

1. Permute the G_is of the p-dimensional observations R times. This provides R permutation resamples.
2. For each permutation resample, calculate the test statistic Q. This provides the resampled test statistics $Q^{(1)}, \ldots, Q^{(R)}$.
3. The empirical distribution of the resampled test statistics $Q^{(1)}, \ldots, Q^{(R)}$ is the permutation null distribution of the test statistic Q.

Provided that all

$$\frac{n!}{n_1!n_2!} =: R_{\text{pmax}}$$

possible permutation resamples are considered, the permutation null distribution of Q will be exact. Given that R_{pmax} may be exceedingly large, however, it may be rather

difficult to compute this exact distribution. For this reason, it is common practice to sample $R < R_{\text{pmax}}$ times with replacement from the finite population of possible permutation resamples and approximate the permutation null distribution of Q on the basis of these R resamples. For more profound information on permutation null distributions we refer to Pesarin (2001), Good (2005) and Pesarin and Salmaso (2010).

3.4.2. The null dilemma

In high-dimensional multivariate scenarios, permutation null distributions of test statistics have become popular since, apart from being easy to calculate, they automatically preserve the dependence structure in the data and can yield exact α-level tests. The price to pay for these appealing properties to be provided is that the multivariate observations must be **exchangeable** within and between groups under the null hypothesis, which means that the observations' joint distribution must be invariant to group label permutation. In our context, this condition is fulfilled under IJD, but not necessarily under SMH. Permutation tests for MI or MO will thus not be valid unless the null hypothesis is IJD, where validity refers to whether the type I error rate tends to the prespecified level α. In practice, however, IJD is unrealistic or at least questionable. Perhaps the only scenario where it appears realistic is that encountered in randomized studies. ICF studies, however, are often non-randomized. In the ICF stroke study mentioned previously, for example, the dependence structure between the ICF items in the ICF chapter 'attitudes' (e4) is expected to be different for Asian and European patients, rendering IJD untenable. Whether we test SMH against MI or against MO, this inevitably leads to what we call here the null dilemma: we can either use the permutation null distribution despite its deficiency under SMH, but then the test result must be interpreted carefully because it may be conservative or anticonservative, or we can attempt to derive an alternative bootstrap null distribution, but bootstrap tests are only asymptotic α-level tests (Efron and Tibshirani, 1993) and usually come with their own problems, especially when $n < p$ (Troendle et al., 2004). Note that further options may exist in specific situations, yet the two mentioned are most common in statistical practice. Because the permutation procedure is preferred whenever it appears applicable, it is desirable to understand its robustness properties under SMH. Several authors have established conditions under which permutation tests remain valid even under non-exchangeability, at least in an asymptotic sense (Romano, 1990; Good, 2002; Pollard and van der Laan, 2004; Huang et al., 2006; Xu and Hsu, 2007; Westfall and Troendle, 2008; Kaizar et al., 2011). For test statistics that rely on differences in sample mean vectors, Huang et al. (2006) compared the permutation distribution and true distribution in terms of cumulants. Unless the cumulants are equal in the two multivariate distributions to be compared (trivial case), it turned out that the even-order cumulants

of the test statistic's permutation and true distribution will be asymptotically equal if $n_1 = n_2$, while the odd-order cumulants will be different irrespective of how n_1 and n_2 relate to each other. In the multivariate normal case where merely the first two cumulants (i.e. mean vector and covariance matrix) are non-zero, the permutation and true distribution thus coincide asymptotically if $n_1 = n_2$, rendering the permutation procedure asymptotically valid. In the multivariate ordinal case, however, there may be infinitely many non-zero cumulants. Hence, even if $n_1 = n_2$, here the permutation procedure will be invalid.

Although the validity constraints of permutation tests have been well studied on the theoretical side, it is unclear yet what effect they have on the practical side. In the simulation experiments of Klingenberg et al. (2009), the permutation procedure appeared to remain applicable under SMH, even for $n_1 \neq n_2$. Kaizar et al. (2011), in contrast, found scenarios under SMH in which the max-T permutation test based on Fisher test statistics fails. More systematic simulation experiments on this issue will be presented in Section 3.5.

3.4.3. *Excursus:* recap of cumulants

In the previous section the validity constraints of permutation tests have been explained by means of cumulants. This rather comprehensive excursus is meant to help readers to develop a more general intuition for the latter.

General definitions

Suppose that we have a random variable, say X, with probability density function (pdf) f, which for our purposes is unbound and extends from $-\infty$ to ∞. Suppose further that our task is to characterize this pdf in the most efficient way.

The first natural guess would be to expand any test function in terms of power of X, and characterize f from there. This means: we would like to compute $< X^m >$, where

$$< X^m >:= \int_{-\infty}^{\infty} x^m f_X(x)\,\mathrm{d}x.$$

These quantities are the moments of f. Rather than to calculate all $< X^m >$s for arbitrary m, however, we would like to have a compact mathematical form which allows to easily obtain any moment we want. This is accomplished by the so-called moment-generating function

$$M_X(t) :=< e^{tX} >,\ t \in \mathbb{R}.$$

It is now easy to see that any moment of f can be obtained by derivatives of M at $t = 0$. In explicit terms, under the condition that the operations of integration with respect to

x and differentiation with respect to t can be interchanged, it holds

$$< X^m >= \left(\frac{\partial}{\partial t}\right)^m M_X(t)\big|_{t=0}.$$

Hence, via the moment-generating function perspective, only one single integral has to be worked out to obtain any moment. However, most pdfs cannot be sufficiently characterized by a finite number of moments, which is intuitive from the viewpoint of regularity: one would need quite a large polynomial expression to compose a pdf that is regular for all x and gives a finite value for the integral over x. An improvement is provided by the cumulant-generating function, which is simply

$$C_X(t) := \log M_X(t),$$

and, accordingly, the mth cumulant is defined as

$$C_m := \left(\frac{\partial}{\partial t}\right)^m C_X(t)\big|_{t=0}.$$

For illustration, let us work out two explicit examples. For the first cumulant we obtain

$$\begin{aligned} C_1 &= \frac{\partial}{\partial t} C_X(t)\big|_{t=0} \\ &= \frac{\partial}{\partial t} \log M_X(t)\big|_{t=0} \\ &= \frac{1}{M_X(t)} < X > \big|_{t=0} \\ &=< X > . \end{aligned}$$

The first cumulant is thus the expectation for f. For the second cumulant, with the quotient rule for derivatives and the previous result for the first cumulant, we obtain

$$\begin{aligned} C_2 &= \left(\frac{\partial}{\partial t}\right)^2 C_X(t)\big|_{t=0} \\ &= \left(\frac{\partial}{\partial t}\right)^2 \log M_X(t)\big|_{t=0} \\ &= \frac{< e^{tX} >< X^2 e^{tX} > - < X e^{tX} >< X e^{tX} >}{< e^{tX} >^2}\Big|_{t=0} \\ &=< X^2 > - < X >^2 . \end{aligned}$$

This tells us that the second cumulant is the variance for f.

The normal distribution as a concrete example

Let us now particularize the discussion from above to the pdf of the normal or, equivalently, Gaussian distribution, which is the limit distribution for many other probability

distributions. Its pdf is

$$f^*(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

where μ is the expectation and σ^2 is the variance for f^*, as we will prove now. To begin with, we define

$$F_X(t) :=< e^{itX} >,$$

which is exactly the moment-generating function from before for $it \to t$. The reason why we choose this representation is that it allows for an immediate interpretation as a Fourier transform: if $f^*(x)$ describes the distribution of a random variable in real space, $F_X(t)$ describes the distribution of a random variable in momentum space, where the momentum variable is now given by t. It turns out that, by completing the square, we can immediately compute $F_X(t)$ analytically:

$$\begin{aligned} F_X(t) =& < e^{itX} > \\ &= \int_{-\infty}^{\infty} \mathrm{d}x e^{itx} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \\ &= \int_{-\infty}^{\infty} \mathrm{d}x \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left[x-(\mu+\sigma^2 it)\right]^2}{2\sigma^2}} e^{-\frac{\left[\mu^2-(\mu+\sigma^2 it)^2\right]}{2\sigma^2}}, \end{aligned}$$

where we have completed the square to separate one part in the integral that does depend on x and another one that does not. The latter we can move in front of the integral, while for the former we notice that, compared with $f^*(x)$ which we started with, only the mean has changed but not the variance. Consequently, the integral has not changed, such that the first term from above gives 1, and the Fourier function becomes $F_X(t) = e^{\mu i t - \sigma^2 t^2/2}$. From there we transform back $t \to -it$ to obtain the moment-generating function of the normal distribution, which is

$$M^*(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}.$$

The respective cumulant-generating function then is

$$C^*(t) = \mu t + \frac{\sigma^2 t^2}{2},$$

and we see immediately that it is only quadratic in t. This means that

$$\begin{aligned} C_1^* &= \mu, \\ C_2^* &= \sigma^2, \\ C_{m\geq 3}^* &= 0. \end{aligned}$$

As opposed to any other probability distribution, the normal distribution thus has the remarkable property that it has only two non-zero cumulants. The same can be shown for the multivariate normal distribution, as has already been noted in Section 3.4.2. For any other probability distribution this means that computation of C_1 and C_2 basically leads to the optimal Gauss fit of this distribution.

3.4.4. Significance assessment under discreteness

When the permutation procedure is used for significance assessment, provided that large values of the test statistic support the alternative hypothesis, the permutation-based P-value is usually calculated as the proportion of resampled test statistics more extreme than the observed one plus the proportion of resampled test statistics equal to the observed one. When the data are discrete, repeated values of the observed test statistic may occur and, as a consequence, P-values may be overly conservative. One way to adjust for discreteness is the mid-P-value approach (Lancester, 1961): mid-P-values are calculated as the proportion of resampled test statistics more extreme than the observed one plus half (instead of all) of the proportion of resampled test statistics equal to the observed one. Although, theoretically, this approach does not guarantee type I error rate control, various numerical evaluations have shown that null mid-P-values tend to be conservative yet come closer to the desired level than ordinary P-values, and that they tend to be more uniformly distributed than ordinary P-values (Hirji, 1991; Agresti, 2001; Klingenberg et al., 2009). Unless stated otherwise, the P-values provided in this chapter are mid-P-values.

3.5. Robustness properties of the permutation procedure under non-exchangeability: a simulation study

3.5.1. Simulation set-up

We conducted an extensive simulation study with the aim to better understand, for small to moderate sample sizes, the behaviour of permutation tests under SMH, that is, in case of possible violations of exchangeability. In particular, we considered tests based on our sum statistics Q_{χ^2} and Q_{CA} (with equally spaced scores $u(v) = v$) as well as their max-T-counterparts (i.e. the maximum univariate χ^2 and traditional two-sided CA test statistic). Systematic power comparisons under MI without MO and/or MO were outside the scope of this study. Multivariate ordinal data were generated using the 'mean mapping method' from the `R` package `orddata` (Kaiser and Leisch, 2010; Kaiser, 2011; Kaiser et al., 2011), which is based on cutting multivariate normal distributions at quantiles defined by the ordinal variables' marginal distributions. (One needs to

specify p vectors of c marginal probabilities adding up to 1 and a positive semi-definite $p \times p$ correlation matrix.) As a result of this technique, it was not possible to examine the effect of non-exchangeability in cumulants of order higher than two.

We considered the set sizes

$$p = \{20, 100\}$$

with $c = 4$ and the overall sample sizes $n = \{20, 40, 60, 80\}$ which were split into

$$(n_1, n_2) = \{(10,10), (20,20), (30,30), (40,40)\} \text{ (balanced groups)},$$
$$(n_1, n_2) = \{(8,12), (16,24), (26,34), (32,48)\} \text{ (unbalanced groups) and}$$
$$(n_1, n_2) = \{(5,15), (12,28), (18,42), (24,56)\} \text{ (very unbalanced groups)}.$$

In order to reflect SMH, we set the marginal probabilities to $(0.25, 0.25, 0.25, 0.25)$ for all variables in both groups. We generated (non-)exchangeability between groups by means of 16 pairs of uniform correlation matrices:

$$(\rho_1, \rho_2) = \left\{(0, 0.25, 0.5, 0.75)^2\right\},$$

with ρ_i denoting the correlation parameter in group g. Thus, the number of different combinations of set sizes, group sizes and correlation parameters was

$$2 \times 3 \times 4 \times 16 = 384.$$

(For completeness, it should be mentioned that, because for equal group sizes there is no difference between, for example, $(\rho_1, \rho_2) = (0, 0.25)$ and $(\rho_1, \rho_2) = (0.25, 0)$, such scenarios were not generated individually.) For each such parameter constellation, the type I error rate was estimated from 1000 random data sets as the average rejection rate of true null hypotheses, and the desired significance level was $\alpha = 0.05$. The simulation margin of error thus amounted to $\pm 2\{0.05\,(1 - 0.05)\,/1000\}^{1/2} \approx \pm 0.0138$. The test statistics' permutation null distributions were approximated on the basis of 5000 permutation resamples. It is important to note that, because the margins of the p one-way tables are invariant to group label permutation, the respective type I error rates are to be interpreted *conditional* on the observed table margins.

3.5.2. Simulation results

All simulation results are reported in detail in Tables A.1–A.4 in the appendix. For the 384 parameter constellations, the heat maps in Figure 3.1 illustrate the deviations of the actual type I error rate from the nominal type I error rate ($\alpha = 0.05$) with the permutation null distribution of the sum statistic Q_{χ^2}. Values < 0 indicate conservative behaviour (shown in violet) and values > 0 anticonservative behaviour (shown in red).

To spot possible biases (i.e. systematic fluctuations around the ideal value 0 (shown in white)) more easily, values outside the simulation margin of error of approximately ±1.38% are additionally highlighted. For $p = 20$ (Figure 3.1A), the actual type I error rate is close to the nominal one in the scenarios with balanced group sizes, regardless of whether under exchangeability (i.e. when $\rho_1 = \rho_2$) or non-exchangeability (i.e. when $\rho_1 \neq \rho_2$). For unbalanced and very unbalanced group sizes, this applies only under exchangeability. Under non-exchangeability, it seems crucial to distinguish which group the higher correlation is combined with: higher correlation in the larger group (i.e. $\rho_1 < \rho_2$) entails conservative behaviour (the actual type I error rate ranges from 0.025 to 0.056 for unbalanced and from 0.011 to 0.046 for very unbalanced group sizes), whereas higher correlation in the smaller group (i.e. $\rho_1 > \rho_2$) entails overly anticonservative behaviour (the actual type I error rate ranges from 0.051 to 0.081 for unbalanced and from 0.048 to 0.122 for very unbalanced group sizes). Perhaps unexpectedly, the permutation procedure's robustness properties seem not to vary systematically with the overall sample size, as has already been observed by Kaizar et al. (2011). For $p = 100$ (Figure 3.1B), we come to basically the same conclusions, but the deviations from the nominal type I error rate are partly considerably more pronounced than for $p = 20$, which is readily visible from Figure 3.1B. For very unbalanced group sizes, for example, the actual type I error rate ranges from 0.005 to 0.046 when $\rho_1 < \rho_2$ and from 0.066 to 0.200 when $\rho_1 > \rho_2$. With the permutation null distribution of the sum statistic Q_{CA}, we arrive at very similar results throughout, which becomes evident when we compare the heat maps in Figure 3.2 with those in Figure 3.1. When our sum statistics are employed, it thus seems that the permutation procedure cannot be recommended under SMH unless it holds $n_1 = n_2$. One should note, however, that many scenarios in which the permutation procedure seriously fails are unlikely to be encountered in practice (e.g. those with $\rho_1 = 0$ and $\rho_2 = 0.75$ or vice versa), whereas its failure in more realistic scenarios (e.g. those with $\rho_1 = 0.25$ and $\rho_2 = 0.5$ or vice versa) seems to be less dramatic, in particular for moderately unbalanced group sizes. Therefore, if potentially some more type I errors than desired do not pose enormous problems in the application at hand and the group sizes are not exceedingly unbalanced, we believe that the permutation procedure may still be used.

Similarly, the heat maps in Figure 3.3 now illustrate the results obtained with the permutation null distribution of the max-T based on χ^2 test statistics. Remarkably, here the permutation null distribution seems to remain 'practically valid' even under non-exchangeability and unbalancedness, with nearly all deviations from the nominal type I error rate lying within the simulation margin of error. In contrast to that, Figure 3.4 suggests that the permutation null distribution of the max-T based on CA test statistics is less robust. For very unbalanced group sizes, it is particularly prone to anticonservative behaviour when the higher correlation is combined with the larger

group: for $p = 20$ (Figure 3.4A), the actual type I error rate ranges from 0.049 to 0.075 across the respective scenarios, while its range for $p = 100$ (Figure 3.4B) is from 0.046 to 0.085. Compared to the extent to which permutation tests based on our sum statistics may fail, this appears almost negligible. Nevertheless, max-T-tests are not per se the better choice, especially when many weak rather than few strong individual effects are expected in the set of interest.

In follow-up simulations, we repeated the complete study with $c = 2$ to see whether the robustness properties that were identified above depend on the number of categories per variable. The respective results are reported in detail in Tables A.5 and A.6 in the appendix. To reflect SMH, here we set the marginal probabilities to (0.5,0.5) for all variables in both groups. In the context of this thesis, the case $c = 2$ is of relatively little interest, but it is computationally convenient because here our sum statistics and their max-T-counterparts, respectively, are equivalent. Hence, it is sufficient to examine one sum and one max-T-statistic. For the sum statistic (Figure 3.5), we find that the results are similar to those in the case $c = 4$ (Figures 3.1 and 3.2). For the max-T-statistic (Figure 3.6), the results are similar to those for the CA-based max-T-statistic in the case $c = 4$ (Figure 3.4). The max-T permutation test based on χ^2 test statistics thus has robustness properties for $c = 2$ that are different from those for $c = 4$. We expect permutation tests that rest upon traditional CA test statistics to have similar robustness properties for any choice of c because, unlike the χ^2 test statistic which has df $= c - 1$, the traditional test statistic has df = 1 independent of c.

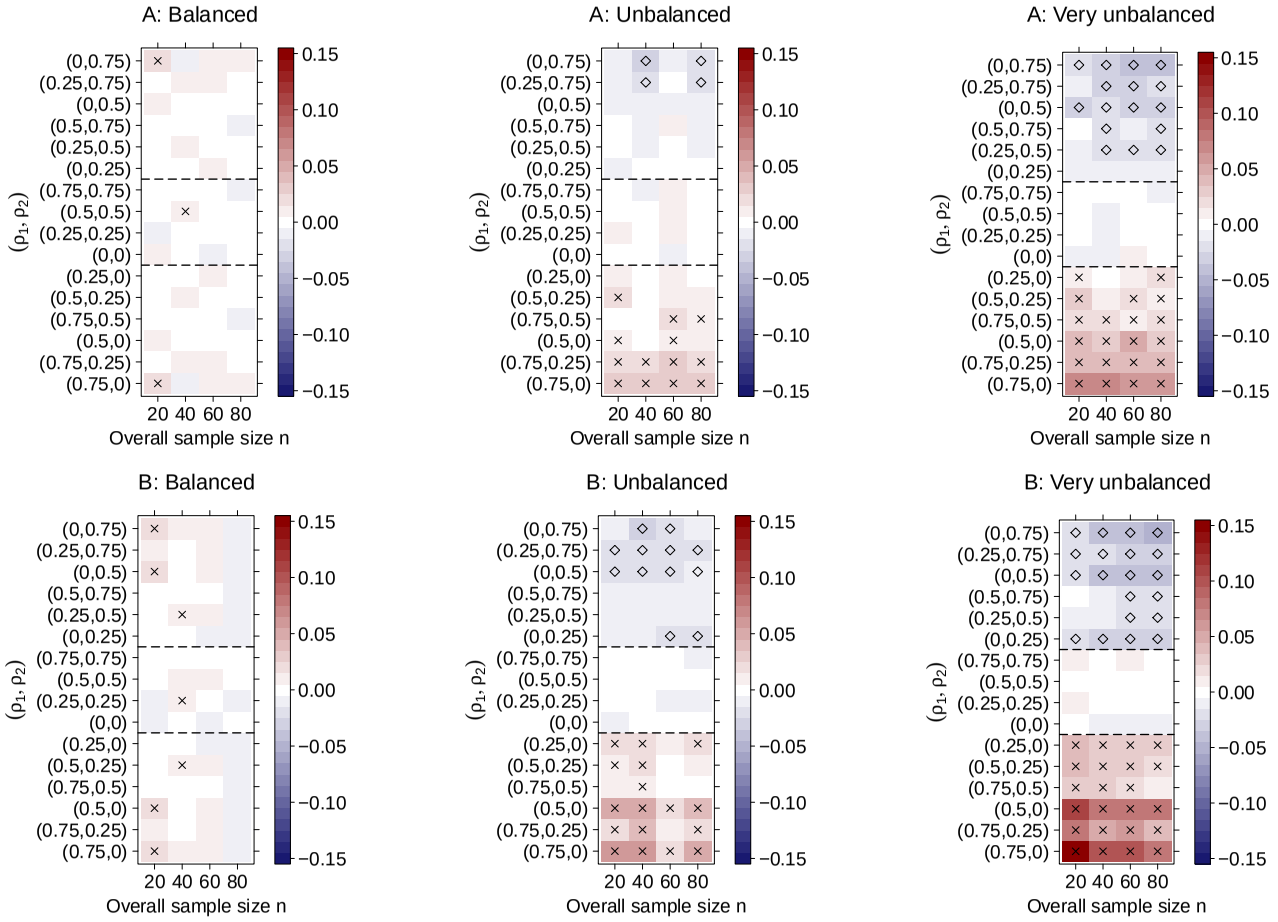

Figure 3.1.: Actual minus nominal type I error rate with the permutation null distribution of the χ^2 sum statistic Q_{χ^2} for A: $p = 20$ and B: $p = 100$ in the case $\mathbf{c} = \mathbf{4}$. Each heat map cell corresponds to one of the 384 simulation scenarios. Simulation margin of error for $\alpha = 0.05$: ± 0.0138. Values outside the margin of error are marked: diamonds indicate systematic conservativeness and crosses systematic anticonservativeness. The colour scale has been chosen such that a direct visual comparison of Figures 3.1–3.6 is enabled.

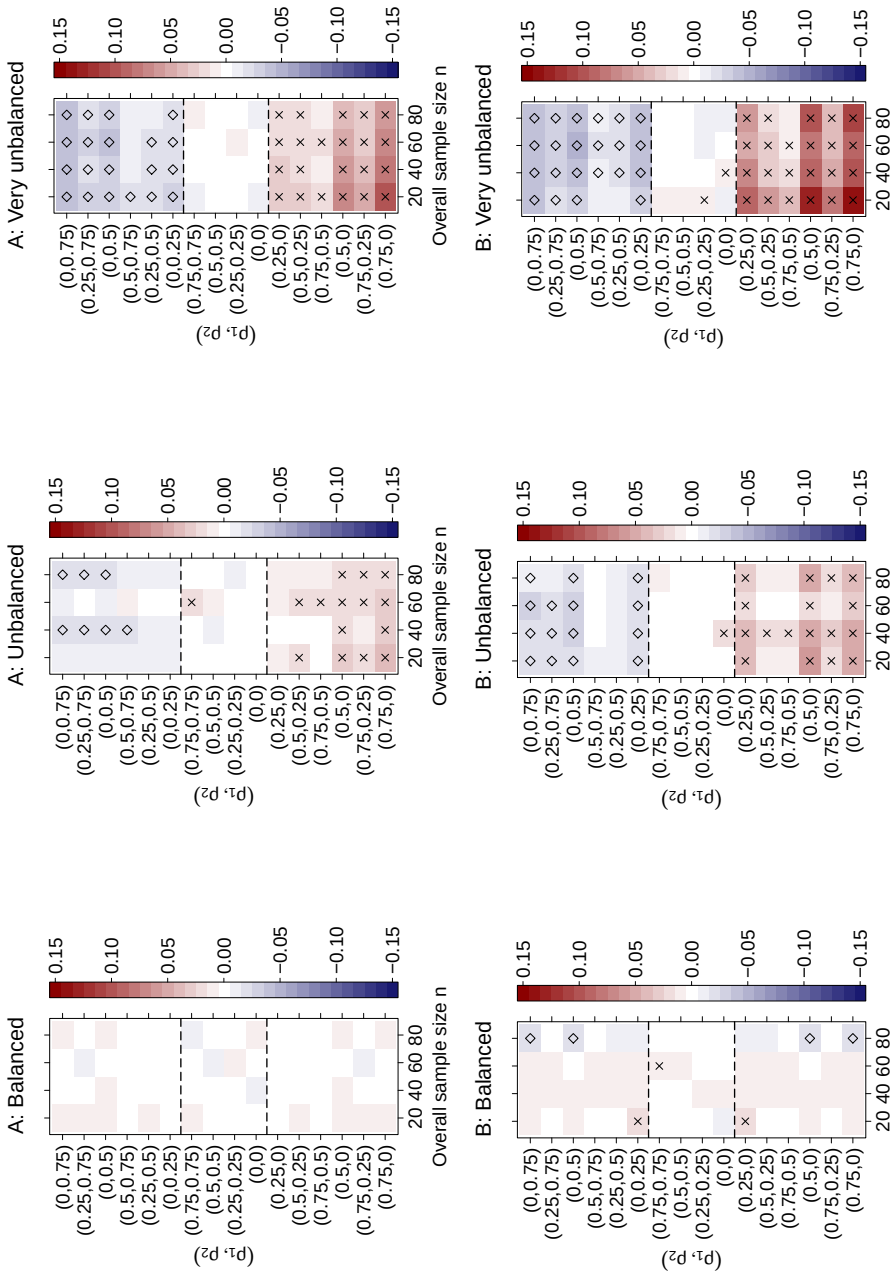

Figure 3.2.: Actual minus nominal type I error rate with the permutation null distribution of the CA sum statistic Q_{CA} for A: $p = 20$ and B: $p = 100$ in the case $\mathbf{c} = \mathbf{4}$. For further explanations see the caption for Figure 3.1.

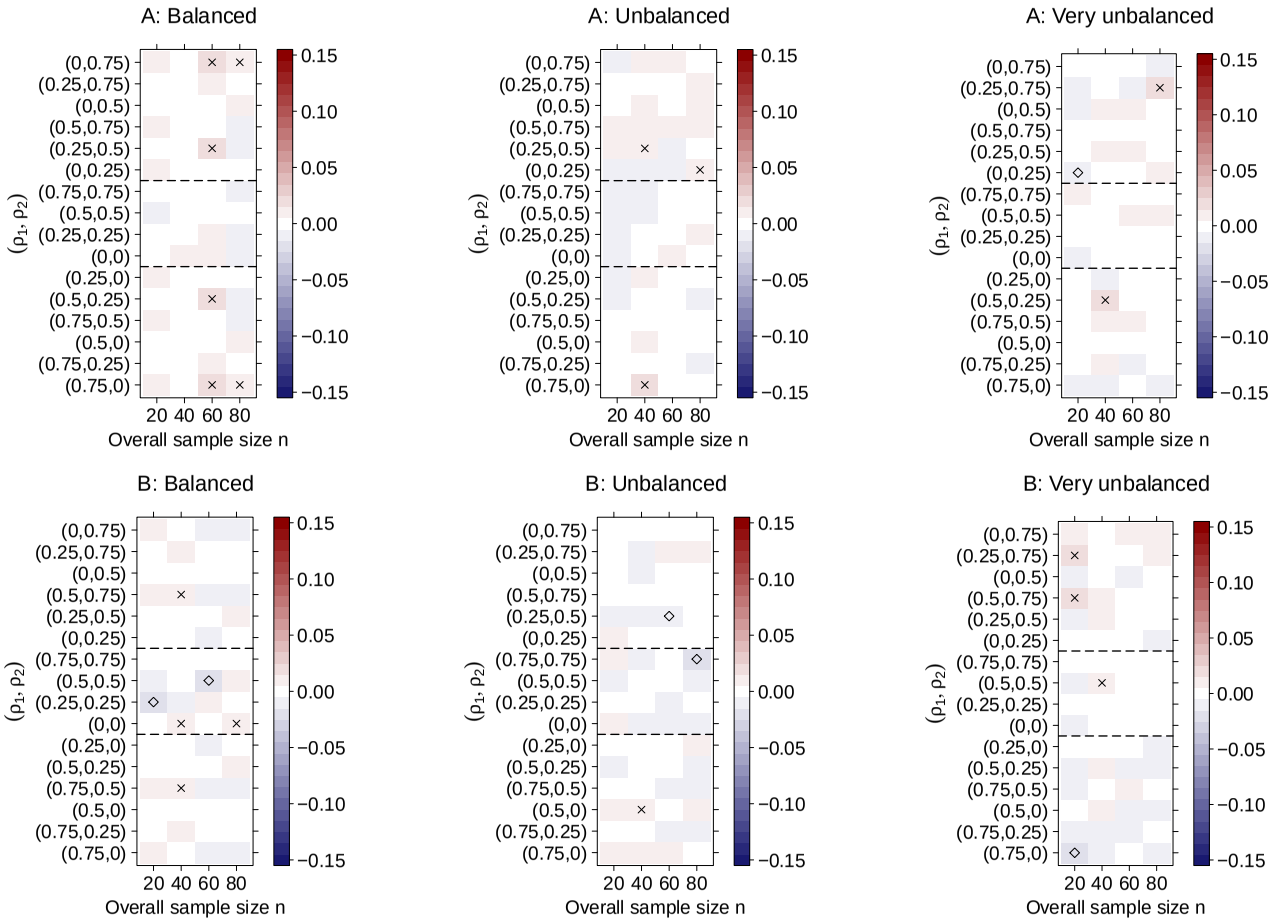

Figure 3.3.: Actual minus nominal type I error rate with the permutation null distribution of max-T (χ^2) for A: $p = 20$ and B: $p = 100$ in the case $\mathbf{c} = \mathbf{4}$. For further explanations see the caption for Figure 3.1.

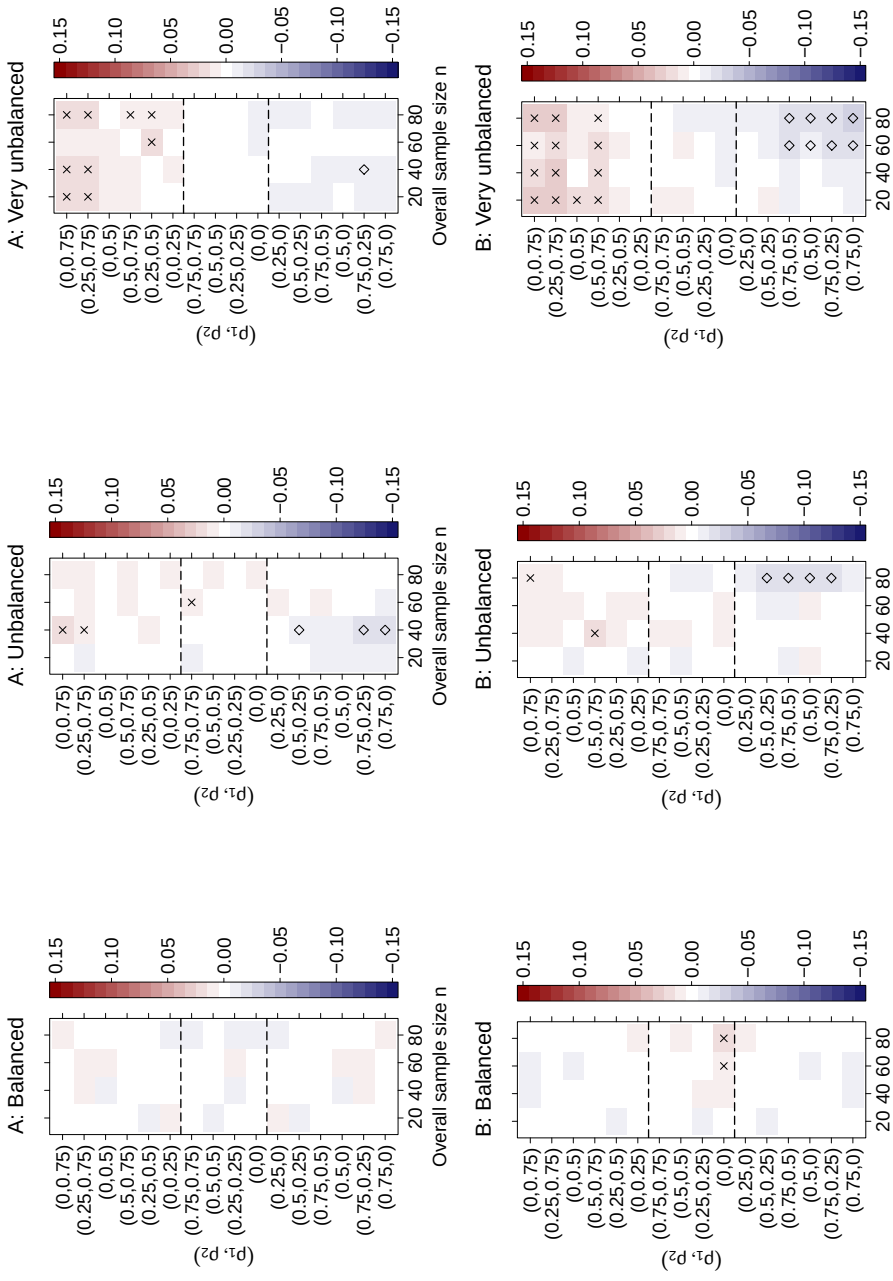

Figure 3.4.: Actual minus nominal type I error rate with the permutation null distribution of max-T (CA) for A: $p = 20$ and B: $p = 100$ in the case $\mathbf{c} = \mathbf{4}$. For further explanations see the caption for Figure 3.1.

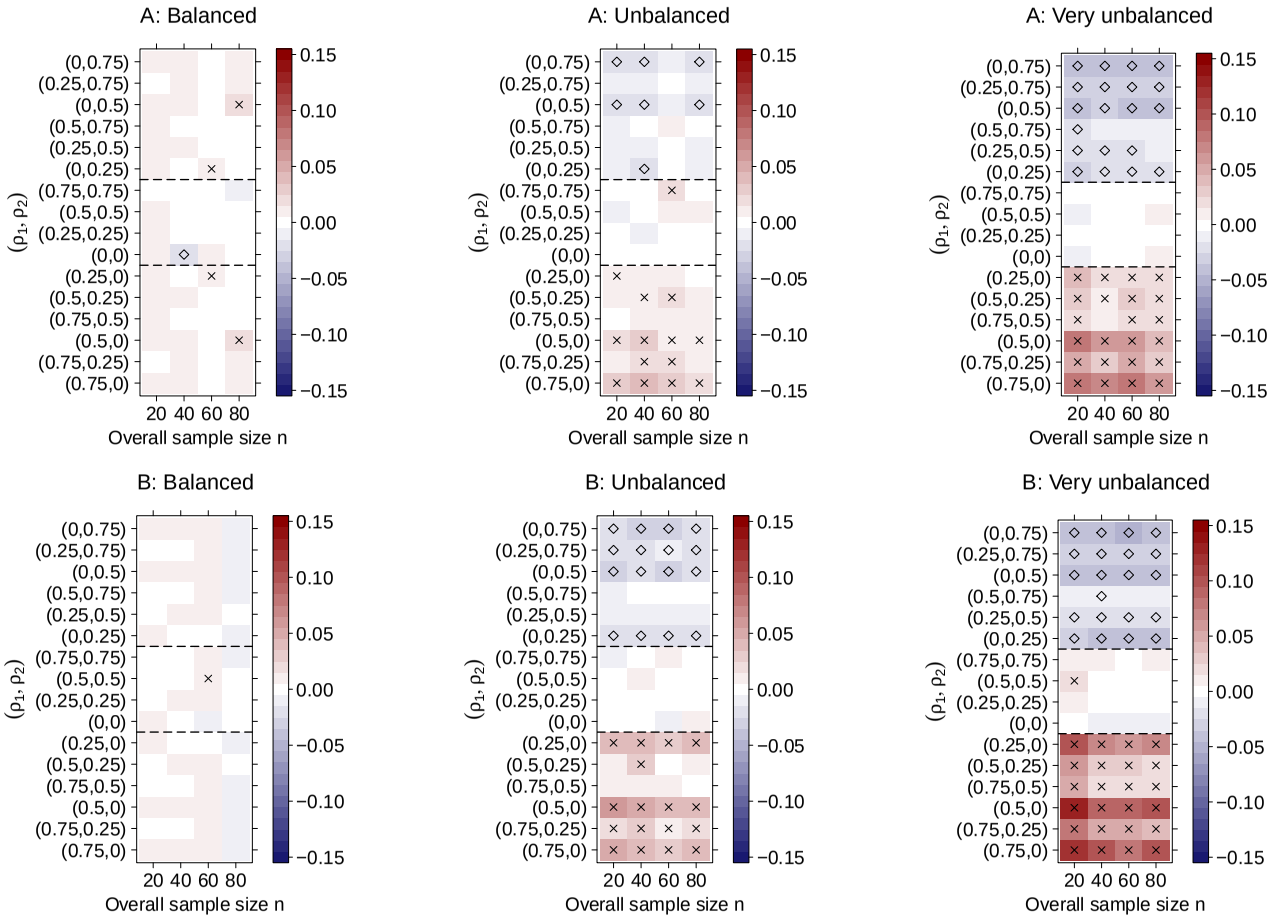

Figure 3.5.: Actual minus nominal type I error rate with the permutation null distribution of the χ^2 sum statistic Q_{χ^2} for A: $p = 20$ and B: $p = 100$ in the case $\mathbf{c} = \mathbf{2}$. For further explanations see the caption for Figure 3.1.

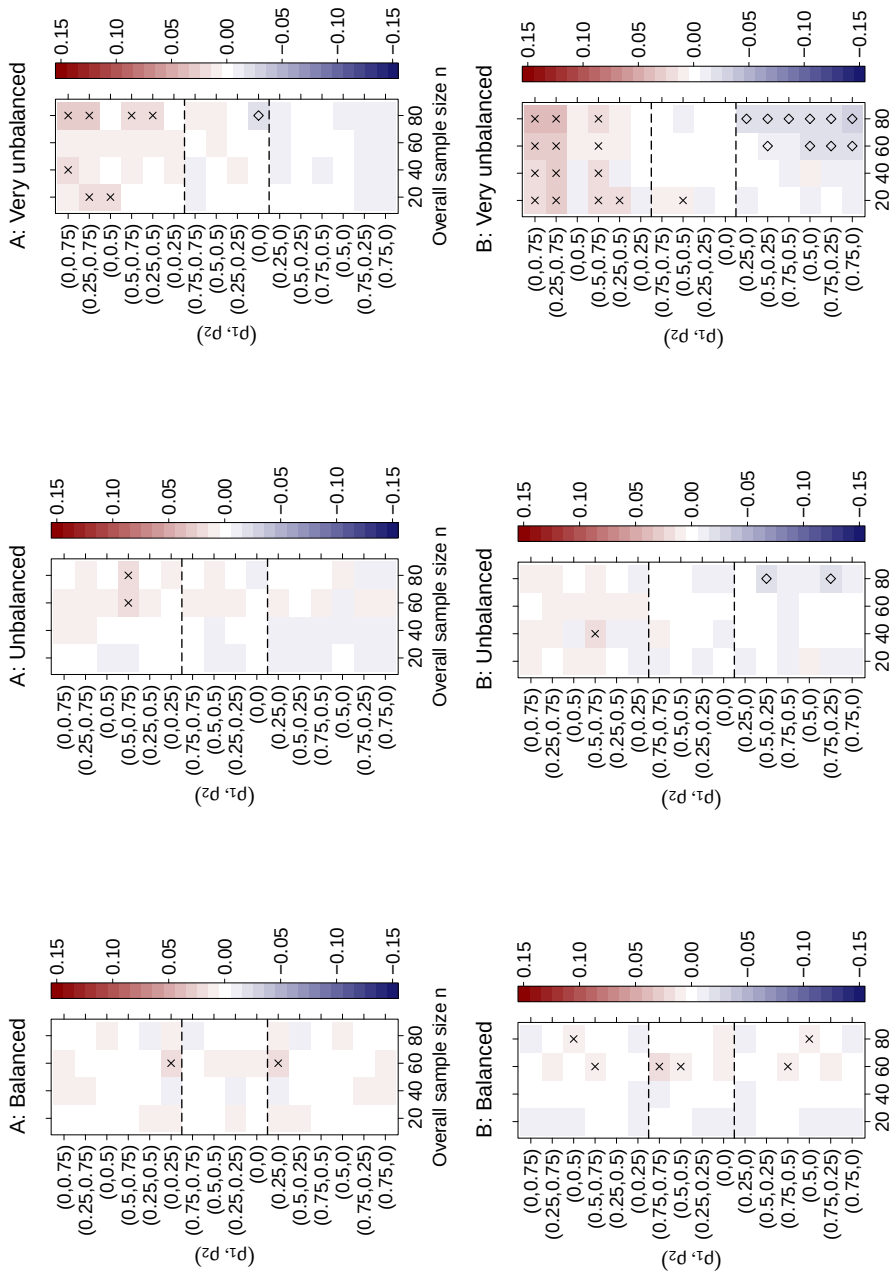

Figure 3.6.: Actual minus nominal type I error rate with the permutation null distribution of max-T (χ^2) for A: $p = 20$ and B: $p = 100$ in the case **c** = **2**. For further explanations see the caption for Figure 3.1.

3.6. *Excursus:* a bootstrap-based alternative to the permutation procedure

The bootstrap procedure

As has been discussed in Section 3.4 and illustrated in Section 3.5, when the null hypothesis is SMH, the permutation procedure may produce deficient results. This excursus discusses a bootstrap-based alternative to the permutation procedure. As opposed to permutation tests, bootstrap tests are only asymptotic α-level tests (Efron and Tibshirani, 1993), yet they are more flexible with respect to the situations in which they can be employed (Troendle et al., 2004).

The bootstrap null distribution that we particularly consider is based on ideas of Dudoit et al. (2004) and Dudoit and van der Laan (2008). For any sum statistic, say $\sum_{k=1}^{p} T_k$, where T_k denotes the kth variable-specific test statistic, this bootstrap null distribution is derived as follows.

1. Draw n units with replacement from the total sample, and repeat this R times. This provides R bootstrap resamples, where the total number of distinct bootstrap resamples is

$$\frac{(2n-1)!}{n!(n-1)!} =: R_{\text{bmax}}.$$

2. For each bootstrap resample, calculate the variable-specific test statistics T_k. This provides the resampled test statistics $T_k^{(1)}, \ldots, T_k^{(R)}$. Store these test statistics in a $p \times R$ matrix $\boldsymbol{R}$.

3. Compute the row means and row variances of $\boldsymbol{R}$ to provide the estimated means $\hat{m}(k)$ and variances $\hat{v}^2(k)$ of the test statistics T_k under the true distribution.

4. Null-shift and scale-transform the resampled test statistics $T_k^{(1)}, \ldots, T_k^{(R)}$ based on upper bounds for their means and variances, the so-called null means $m_0(k)$ and null variances $v_0^2(k)$. (For $m_0(k)$ and $v_0^2(k)$, which must be chosen with caution to ensure type I error rate control, Dudoit et al. (2004) suggest the means and variances of the asymptotic null distributions that one would use for univariate tests. For example, when each variable-specific test statistic equals the traditional CA test statistic, then $m_0(k) = 1$ and $v_0^2(k) = 2$.) Define the null-shifted and scale-transformed test statistics as

$$\min\left\{1, \frac{v_0(k)}{\hat{v}(k)}\right\}\left\{T_k^{(r)} - \hat{m}(k)\right\} + m_0(k) =: \tilde{T}_k^{(r)},$$

$r = 1, \ldots, R$. This provides the $p \times R$ matrix $\tilde{\mathbf{R}}$.

5. The empirical distribution of the column sums of $\tilde{\mathbf{R}}$ is the desired bootstrap null distribution of the sum statistic $\sum_{k=1}^{p} T_k$.

As a side remark, the bootstrap null distribution of the test statistic $\max_k T_k$ is obtained similarly: rather than to work with the sums of the column entries of $\tilde{\mathbf{R}}$ in step 5, one needs to work with their maxima. Furthermore, it should be noted that, in the two-group context of this chapter, it would actually be more appropriate to resample with replacement within each group than across the groups (see step 1 above). Simulation-based evaluations have indicated, however, that this bootstrap variant does not provide asymptotic FWER control at the level of individual variables. For this reason, it has not been considered further.

Robustness properties of the bootstrap procedure under non-exchangeability: a simulation study

To see whether, under SMH, tests based on the null-shifted and scale-transformed bootstrap null distribution from above perform better than permutation tests, we repeated the extensive simulation study described in Section 3.5.1 with the bootstrap null distribution. The respective simulation set-up thus corresponded to that described in Section 3.5.1; the only difference was that instead of the permutation null distribution the bootstrap null distribution was used.

All simulation results are reported in detail in Tables A.7–A.12 in the appendix. For brevity and clarity, here we merely focus on the key observations that we have made, and we suppose that the content of Section 3.5.2 is known to the reader. When the sum statistics Q_{χ^2} and Q_{CA} are used, we find from the heatmaps in Figures 3.7, 3.8 and 3.11 that, under non-exchangeability and unbalancedness, the bootstrap procedure's behaviour is in principle broadly similar to that of the permutation procedure. There is, however, one important difference: it seems to naturally tend towards fairly conservative behaviour. This becomes particularly obvious from the heatmaps in Figure 3.7: across the respective 384 simulation scenarios, the actual type I error rate ranges from 0.000 to 0.102; its minimum is thus even below the most conservative level that is reached with the permutation procedure. In contrast, when the max-T based on χ^2 or traditional CA test statistics is used, the bootstrap procedure appears to be prone to overly anticonservative behaviour. As is readily visible from the respective heat maps in Figures 3.9, 3.10 and 3.12, this anticonservative behaviour vanishes with increasing sample size, even though rather slowly. Taken together, the results from our simulation study show that the bootstrap procedure does not seem to be an appropriate alternative to the permutation procedure. This well illustrates the null dilemma discussed previously in Section 3.4.2.

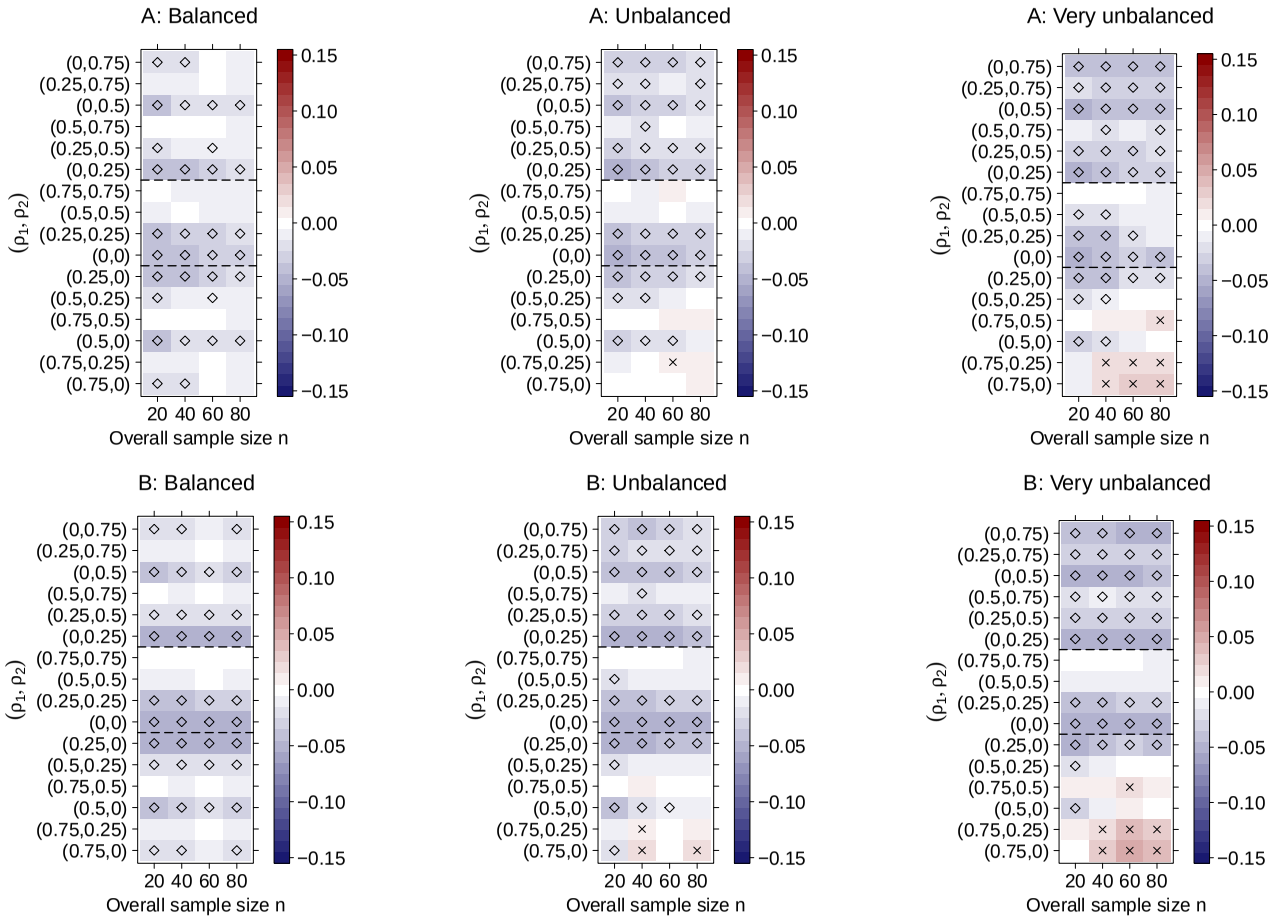

Figure 3.7.: Actual minus nominal type I error rate with the bootstrap null distribution of the χ^2 sum statistic Q_{χ^2} for A: $p = 20$ and B: $p = 100$ in the case $\mathbf{c} = \mathbf{4}$. Each heat map cell corresponds to one of the 384 simulation scenarios. Simulation margin of error for $\alpha = 0.05$: ± 0.0138. Values outside the margin of error are marked: diamonds indicate systematic conservativeness and crosses systematic anticonservativeness. The colour scale has been chosen such that a direct visual comparison of Figures 3.1–3.12 is enabled.

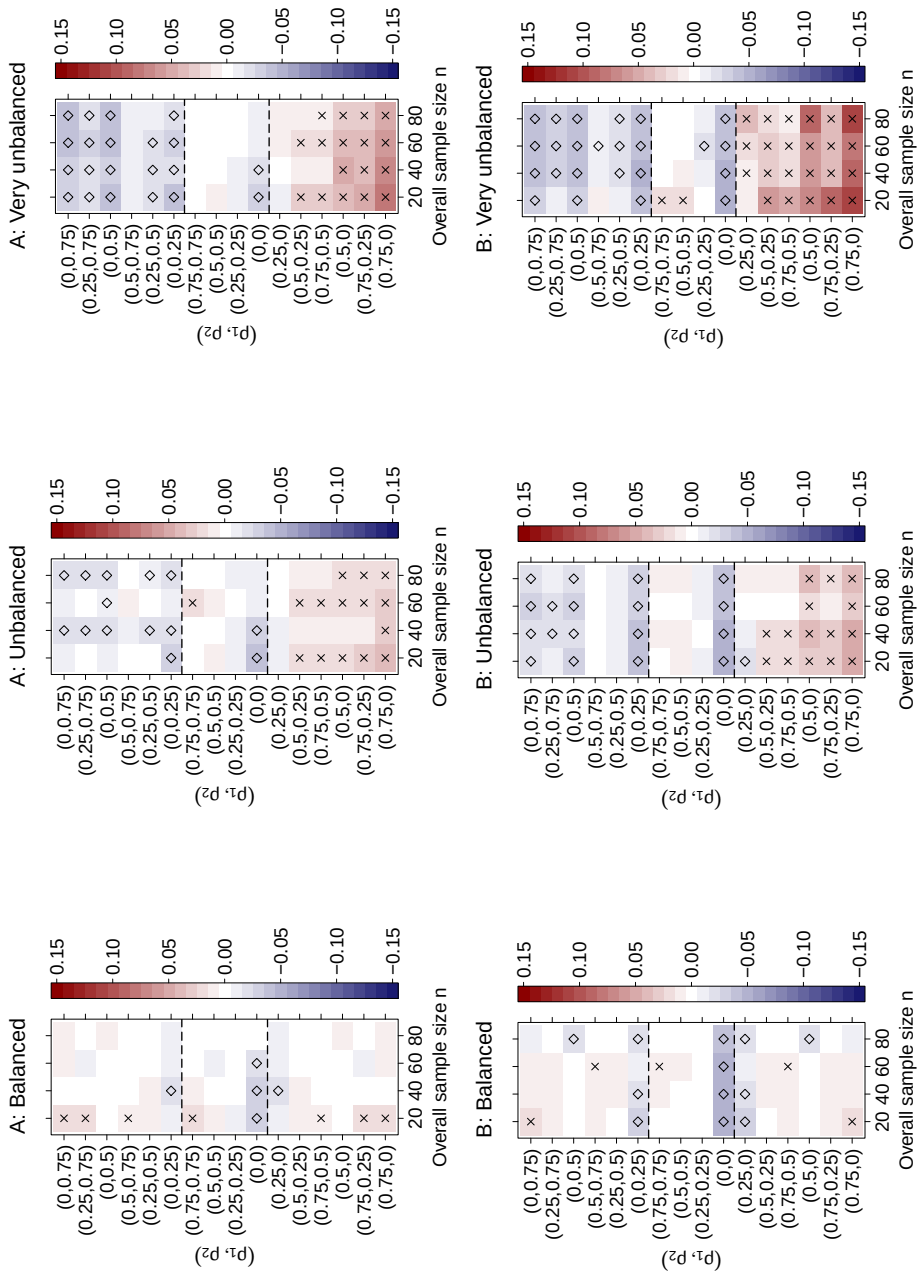

Figure 3.8.: Actual minus nominal type I error rate with the bootstrap null distribution of the CA sum statistic Q_{CA} for A: $p = 20$ and B: $p = 100$ in the case $\mathbf{c} = \mathbf{4}$. For further explanations see the caption for Figure 3.7.

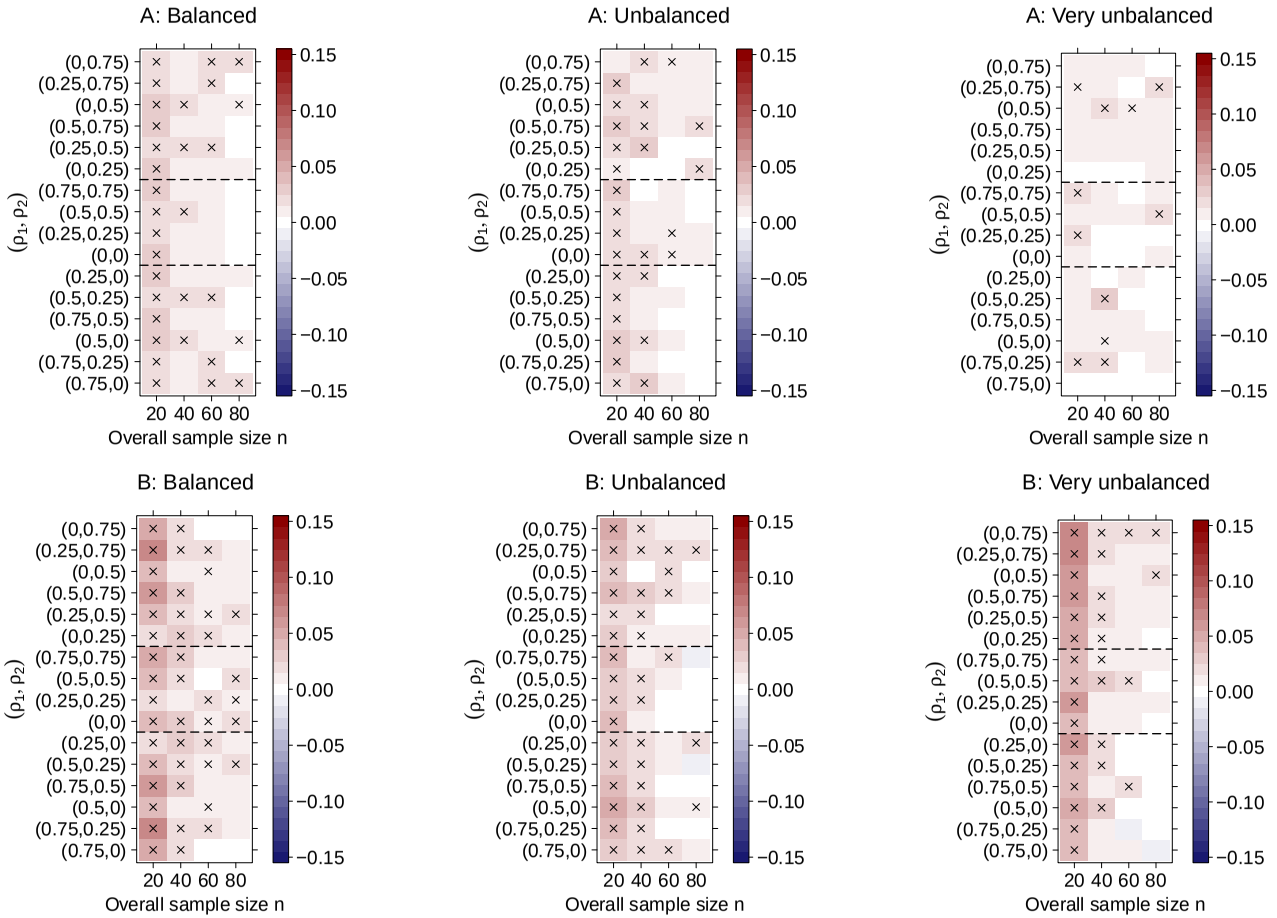

Figure 3.9.: Actual minus nominal type I error rate with the bootstrap null distribution of max-T (χ^2) for A: $p = 20$ and B: $p = 100$ in the case $\mathbf{c} = \mathbf{4}$. For further explanations see the caption for Figure 3.7.

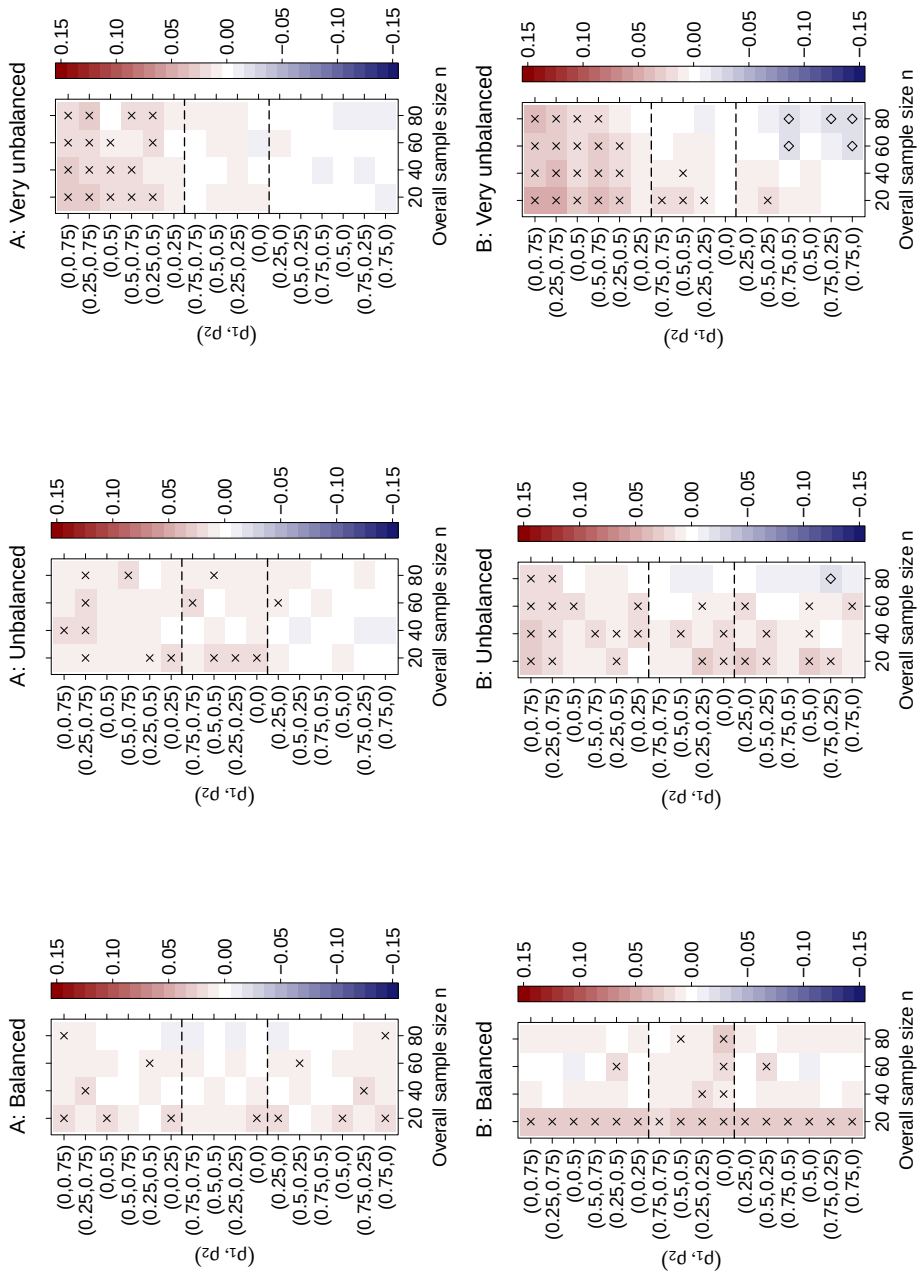

Figure 3.10.: Actual minus nominal type I error rate with the bootstrap null distribution of max-T (CA) for A: $p = 20$ and B: $p = 100$ in the case $\mathbf{c} = \mathbf{4}$. For further explanations see the caption for Figure 3.7.

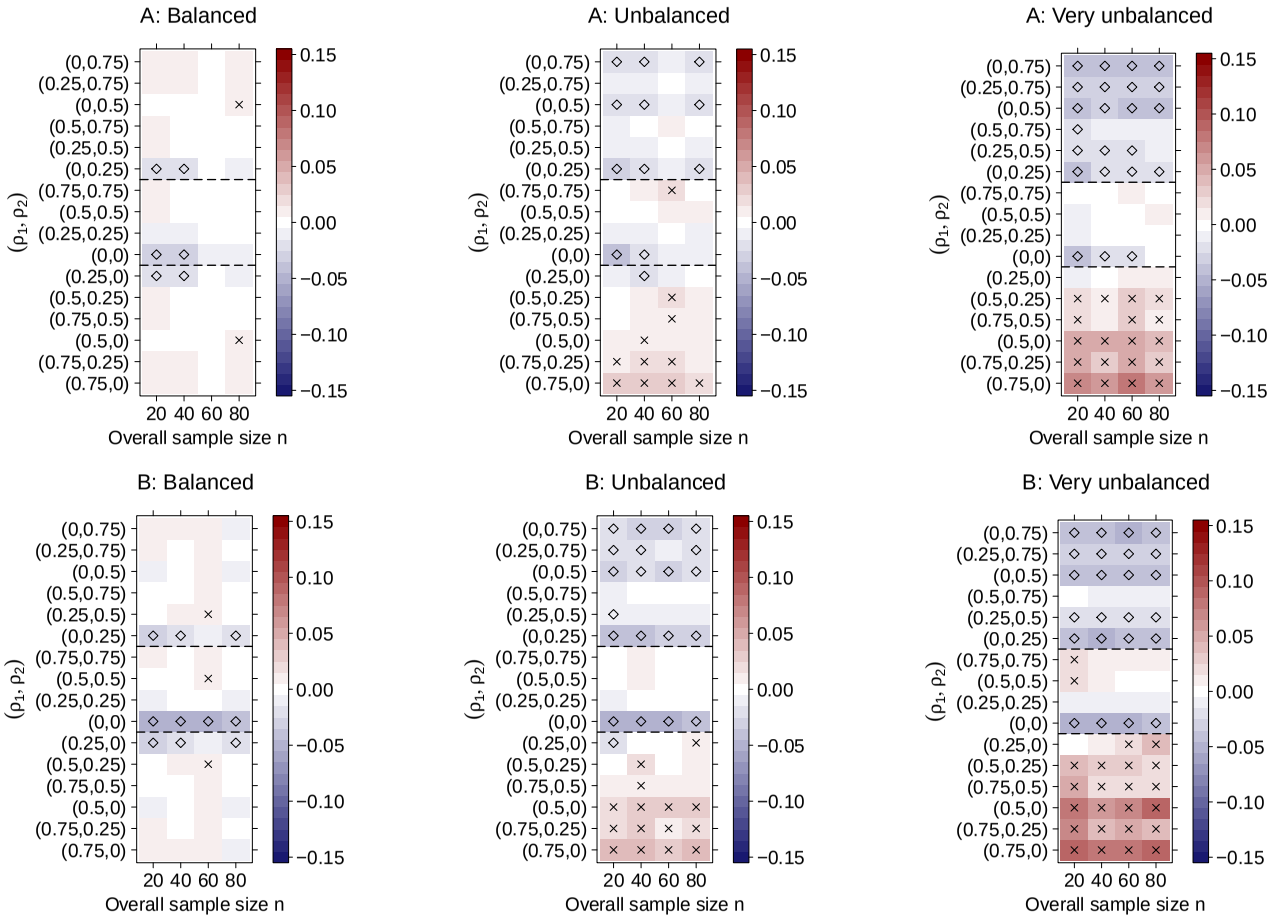

Figure 3.11.: Actual minus nominal type I error rate with the bootstrap null distribution of the χ^2 sum statistic Q_{χ^2} for A: $p = 20$ and B: $p = 100$ in the case $\mathbf{c} = \mathbf{2}$. For further explanations see the caption for Figure 3.7.

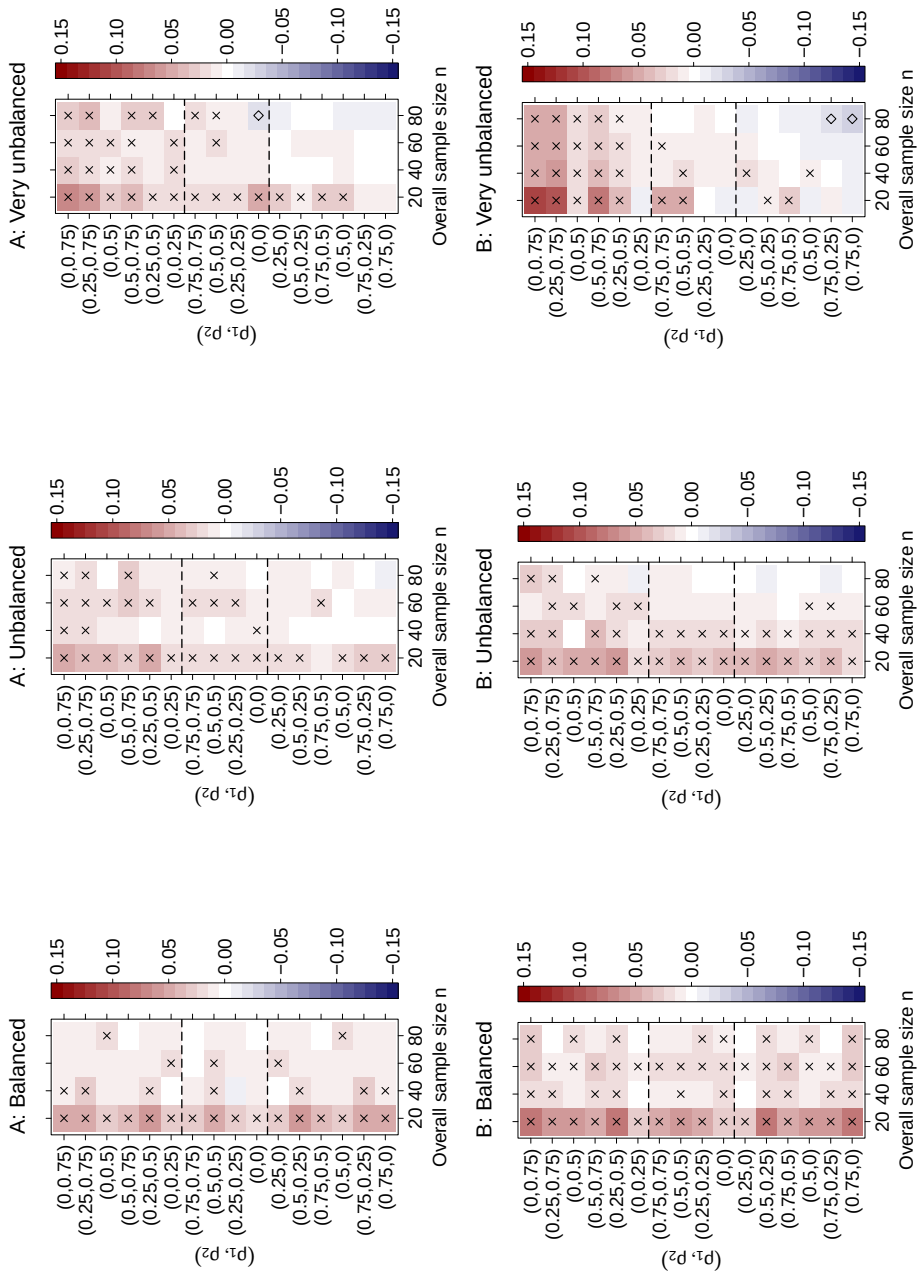

Figure 3.12.: Actual minus nominal type I error rate with the bootstrap null distribution of max-T (χ^2) for A: $p = 20$ and B: $p = 100$ in the case **c** = **2**. For further explanations see the caption for Figure 3.7.

3.7. Application 1: functioning and disability after first stroke

Data set and question of interest

To illustrate the application of the methods presented in this chapter, we analyzed data from a multi-centre cross-sectional study on functional limitations and disabilities after first stroke. The study was conducted in post-acute rehabilitation facilities from 2004 to 2007, and it was based on the ICF core set for stroke (Geyh et al., 2004) which comprises $p = 130$ ICF items (listed further below in this section in Table 3.2 which, in addition, provides information on the ICF items' particular tree structure). The recruitment of the individuals involved and the data collection were performed by physicians and other health professionals specifically trained for this purpose in ICF workshops. The respective data set includes $n = 104$ individuals of which $n_1 = 46$ underwent rehabilitation in high-income Asian countries (China, Malaysia, South Korea and Thailand) and $n_2 = 58$ in European countries (Austria, Germany, Italy, Norway, Sweden and Switzerland). At the time of data collection, all individuals were 50 years old or older and their BMI was 30 or less. Moreover, the distribution of the individuals' sex, age and BMI did not differ substantially between the two patient groups (rendering adjustment for these typical confounders unnecessary). The question of interest was now whether and, most notably, in which respects stroke patients from Asian versus European countries differ in their 130-dimensional ICF profile.

To answer the above question, the ICF-based data were first preprocessed as described in Section 1.1. In particular, this means that both the five-level ordinal scale of ICF items of the ICF components 'body functions' (b), 'body structures' (s) and 'activities and participation' (d) and the nine-level ordinal scale of ICF items of the ICF component 'environmental factors' (e) were coarsened to three levels (see Figures 1.1 and 1.2), and that the additional answer option 9 (not applicable), which has been observed merely 24 times, was recoded into the answer option 0 (no problem/neither barrier nor facilitator). As has already been noted in Section 1.1, the scale coarsening potentially reduces the number of ICF items for which one or more categories could not be observed in both groups, which is an appreciable side effect because such ICF items lead to degenerate $\hat{\mathbf{\Sigma}}_{0k}$s. With the three-level ordinal scale this occurs only for the ICF item 'blood pressure functions' (b420) where the third category (severe to complete problem) has never been observed. We set the associated univariate test statistic to zero. Less conservative strategies to handle the ICF item b420 are to exclude it from the analysis or to treat it as binary; both led to the same conclusions as our strategy.

Methods

For the comparison of Asian and European stroke patients with respect to their ICF profile, we contrasted five approaches (A1–A5) with each other as follows.

(a) A1: we combined Meinshausen's procedure, which has previously been explained in Section 2.3, with our permutation test based on the χ^2 sum statistic Q_{χ^2}. We approximated the permutation null distribution of Q_{χ^2} on the basis of 10000 resamples. A complete enumeration of all 104!/ (46!58!) $\approx 7.96 \times 10^{29}$ possible permutation resamples was too computationally intensive. In this context, one issue deserves particular mention: computationally, when permutation tests are used to test for set effects, Meinshausen's procedure seems to involve as many permutation rounds as there are sets in the tree. However, provided that the test statistic for any set can be calculated from the respective variable-specific test statistics, P-values for an entire tree structure can be computed efficiently on the basis of one permutation round for the root set (i.e. from the resultant $p \times R$ matrix that contains the p univariate test statistics for the R permutation resamples). This is beneficial, in particular when extensive tree structures are studied.

(b) A2: see approach A1, but with the max-T permutation test based on χ^2 test statistics to test any set considered in Meinshausen's procedure.

(c) A3: we carried out the traditional univariate χ^2 test for each ICF item and subsequently applied the Bonferroni-Holm procedure (Holm, 1979) to adjust for multiplicity. This approach is rather simplistic, yet it is widely used.

(d) A4: see approach A3, but with the permutation rather than with the analytically derived asymptotic null distribution of the univariate χ^2 test statistic, approximated on the basis of 10000 resamples.

(e) A5: we used the permutation-based 'discrete Bonferroni method' (Westfall and Wolfinger, 1997; Westfall and Troendle, 2008), again with 10000 resamples. This approach is similar to the popular max-T-based stepdown approach of Westfall and Young (1993), with the crucial difference being that it provides FWER control under SMH, at the price of potentially less power. Unlike the approaches A1–A4, this approach dispenses with mid-P-values.

Because, in this particular ICF-based application, it was of primary interest to detect MI rather than MO, analogous approaches based on the CA sum statistic Q_{CA} or its max-T-counterpart were not taken into consideration. The results obtained with approaches A1–A5 are summarized in Table 3.1 and discussed below. It should be emphasized that, unlike approaches A1 and A2, approaches A3–A5 do not exploit the tree structure of the data inferentially. However, we can exploit it *ex post* for interpretation by treating the smallest adjusted P-value in a set as set-specific test.

Results

Figure 3.13 shows, for the complete ICF core set, the permutation null distributions of the test statistics Q_{χ^2} and max-T (χ^2), together with their analytically derived asymptotic null distributions under the assumption of independence between the 130 ICF items. Strictly speaking, the permutation and analytically derived asymptotic null distributions are not fully comparable because the former are conditional on the observed table margins for each ICF item, whereas the latter are unconditional distributions. Under independence, however, the conditional and unconditional null distributions asymptotically behave similarly (or even the same under certain conditions (Romano, 1990)). For this reason, because $n = 104$ is sufficiently large, the comparison between the two null distributions in Figure 3.13 provides reliable information on how valid or invalid results based on the analytically derived asymptotic null distributions would be. As becomes evident from Figure 3.13, the analytically derived asymptotic null distributions are inappropriate in the present application. Regardless of which test statistic is chosen, we find that the ICF core set is significant (i.e. MI is confirmed between the overall ICF profile of stroke patients from Asian and European countries). Table 3.1 now tells us which sets (i.e. ICF components, chapters and items) this significant difference can be attributed to. For clarity, it contains only the ICF components, chapters and items that have been identified as significant by at least one of the five approaches A1–A5 that were described further above in this section. (An `R` script to produce the results in Table 3.1 as well as those omitted is available from `http://wileyonlinelibrary.com/journal/rss-datasets`.) We find that the results are fairly consistent across all approaches apart from A3. Mostly owing to ignored dependencies between the ICF items and thus between the associated test statistics, approach A3 yields the most conservative conclusion with four significant ICF items: 'structure of upper extremity' (s730), 'acquiring, keeping and terminating a job' (d845), 'products and technology for personal indoor and outdoor mobility and transportation' (e120) and 'architecture and construction services, systems and policies' (e515). These four ICF items are also found to be significant by approaches A1, A2, A4 and A5, together with the ICF items 'structure of lower extremity' (s750), 'housing services, systems and policies' (e525) and 'associations and organizational services, systems and policies' (e555). For the ICF item 'doing housework' (d640), MI is revealed merely by approaches A2, A4 and A5. As displayed in Table 3.1, approach A1, which is based on the sum statistic Q_{χ^2}, does not reject SMH for the ICF chapter 'domestic life' (d6); Meinshausen's procedure hence does not descend further into individual ICF items, one of which is d640. This potential type II error may be explained by the fact that approach A1 has power properties that are different from those of approaches A2–A5. Conversely, it is solely approach A1 which detects MI for the ICF chapters 'neuromusculoskeletal and movement-related functions' (b7) and 'support and relationships' (e3), whereas

none of the nine and seven ICF items contained is found to be marginally significant by either of the approaches. Apparently, the ICF items in b7 and e3 only *jointly* provide evidence against SMH. This result indicates that approach A1 outperforms the other approaches in the presence of many weak individual effects, as has been expected. Even though it is unlikely that approaches A1, A2 and A4 are theoretically valid in the present application, we assume that they are practically valid, for two reasons. Firstly, the group sizes are balanced to a rough approximation ($n_1 = 46, n_2 = 58$). Secondly, at the level of individual ICF items, approaches A1, A2 and A4 (which do not guarantee FWER control under SMH) lead to nearly the same conclusions as approach A5 (which guarantees FWER control under SMH).

The fact that many of the differences that we have found are in the environment suggests that the kind of support people receive after stroke differs between Asian and European countries, which in turn reflects that the two country groups differ in their health and social policies. This does not come as a surprise and supports the validity of our results which may now serve the WHO or other international organizations to uncover those inequalities in health service provision that directly affect stroke patients. Information of this kind may help policy makers to eliminate or reduce such inequalities and ultimately to improve the quality of post-stroke rehabilitation services. The difference that has been found in support and relationships is particularly noteworthy. Astin et al. (2008) reported that cardiac patients are more frequently cared at home by their family in Asian than in European countries where residential care is much more common. Both results put together form a good basis for more detailed studies on the role of family and non-family relationships in post-stroke rehabilitation. The differences that we have found in body functions and structures require additional explanation. The question is whether they are due to different evaluation approaches more than really due to differently affected body functions and structures. Further studies are needed to answer this question.

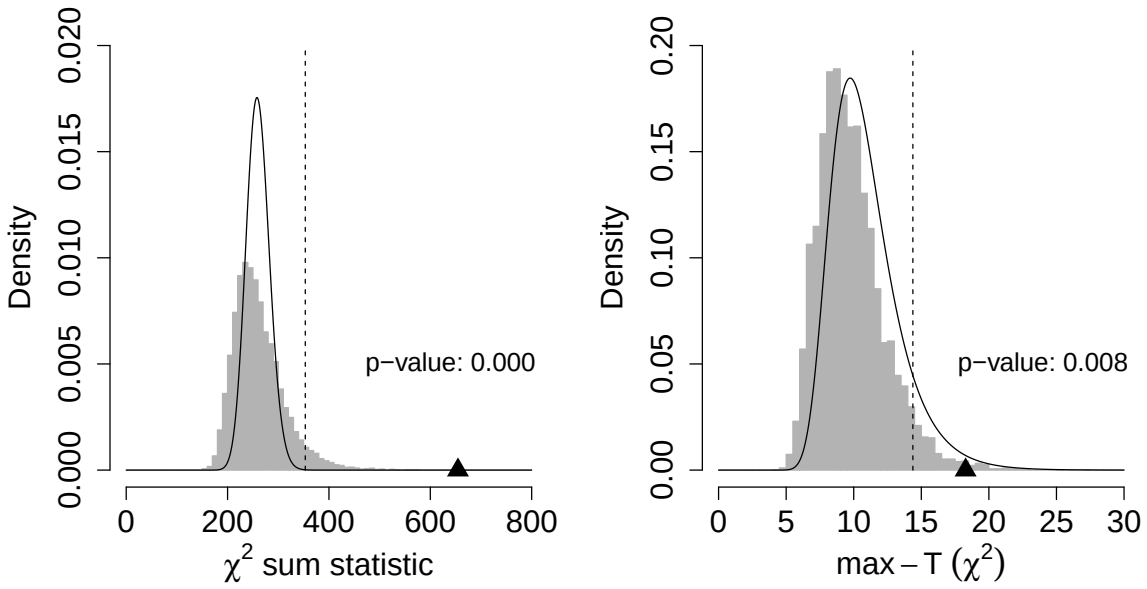

Figure 3.13.: Grey areas show the permutation null distributions of the χ^2 sum statistic Q_{χ^2} and the max-T based on χ^2 test statistics for the complete ICF core set, approximated on the basis of 10000 resamples. Superimposed black curves show the respective analytically derived asymptotic null distributions under the assumption of independence. (For Q_{χ^2}, this analytically derived asymptotic null distribution is the χ^2 distribution with df $= 260$. For max-T (χ^2), the pdf equals $130F(x)^{129}f(x)$, with here $F(x)$ denoting the cumulative distribution function (cdf) and $f(x)$ the pdf of the χ^2 distribution with df $= 2$.) Dashed lines indicate critical values (0.95-quantiles) of the permutation distributions. Filled triangles indicate observed values of Q_{χ^2} and max-T (χ^2).

Table 3.1.: Multiplicity-adjusted P-values for the ICF components, chapters and items that have been identified as significant by at least one of the approaches A1–A5 (see further above for detailed explanations), with $\alpha = 0.05$. Adjusted P-values > 0.05 are indicated by 'ns', which stands for non-significant.

	Multiplicity-adjusted P-values obtained with				
	A1:	A2:	A3:	A4:	A5:
Body functions (b)	0.017	ns	ns	ns	ns
Neuromusculoskeletal and movement-related functions (b7)	0.004	ns	ns	ns	ns
Body structures (s)	0.000	0.005	0.014	0.013	0.008
Structures related to movement (s7)	0.000	0.009	0.014	0.013	0.008
Structure of shoulder region (s720)	0.032	0.032	ns	0.031	ns
Structure of upper extremity (s730)	0.013	0.013	0.014	0.013	0.008
Structure of lower extremity (s750)	0.046	0.046	ns	0.043	0.038
Activities and participation (d)	0.028	0.013	0.020	0.000	0.011
Domestic life (d6)	ns	0.035	ns	0.000	0.029
Doing housework (d640)	ns	0.000	ns	0.000	0.029
Major life areas (d8)	0.003	0.010	0.020	0.025	0.011
Acquiring, keeping and terminating a job (d845)	0.026	0.026	0.020	0.025	0.011
Environmental factors (e)	0.000	0.011	0.022	0.013	0.011
Products and technology (e1)	0.002	0.015	0.022	0.013	0.011
Products and technology for personal use in daily living (e115)	ns	ns	ns	ns	0.028
Products and technology for personal indoor and outdoor mobility and transportation (e120)	0.013	0.013	0.022	0.013	0.011
Support and relationships (e3)	0.017	ns	ns	ns	ns
Services, systems and policies (e5)	0.004	0.012	0.034	0.019	0.016
Architecture and construction services, systems and policies (e515)	0.039	0.039	0.034	0.037	0.016
Housing services, systems and policies (e525)	0.020	0.020	ns	0.019	0.030
Associations and organizational services, systems and policies (e555)	0.026	0.026	ns	0.025	0.027

Table 3.2.: List of the 130 ICF items that have been considered in the stroke study, together with information on which ICF component and ICF chapter each item belongs to. The respective ICF code is given in brackets.

Body functions (b) (comprises 41 ICF items)

Mental functions (b1)
- Consciousness functions (b110)
- Orientation functions (b114)
- Intellectual functions (b117)
- Temperament and personality functions (b126)
- Energy and drive functions (b130)
- Sleep functions (b134)
- Attention functions (b140)
- Memory functions (b144)
- Emotional functions (b152)
- Perceptual functions (b156)
- Higher-level cognitive functions (b164)
- Mental functions of language (b167)
- Calculation functions (b172)
- Mental function of sequencing complex movements (b176)
- Experience of self and time functions (b180)

Sensory functions and pain (b2)
- Seeing functions (b210)
- Functions of structures adjoining the eye (b215)
- Proprioceptive function (b260)
- Touch function (b265)
- Sensory functions related to temperature and other stimuli (b270)
- Sensation of pain (b280)

Voice and speech functions (b3)
- Voice functions (b310)
- Articulation functions (b320)
- Fluency and rhythm of speech functions (b330)

Functions of the cardiovascular, haematological, immunological and respiratory systems (b4)
- Heart functions (b410)
- Blood vessel functions (b415)
- Blood pressure functions (b420)
- Exercise tolerance functions (b455)

Functions of the digestive, metabolic and endocrine systems (b5)
- Ingestion functions (b510)

Defecation functions (b525)

Genitourinary and reproductive functions (b6)
Urination functions (b620)
Sexual functions (b640)

Neuromusculoskeletal and movement-related functions (b7)
Mobility of joint functions(b710)
Stability of joint functions (b715)
Muscle power functions (b730)
Muscle tone functions (b735)
Muscle endurance functions (b740)
Motor reflex functions (b750)
Involuntary movement reaction functions (b755)
Control of voluntary movement functions (b760)
Gait pattern functions (b770)

Body structures (s) (comprises 5 ICF items)

Structures of the nervous system (s1)
Structure of brain (s110)

Structures of the cardiovascular, immunological and respiratory systems (s4)
Structure of cardiovascular system (s410)

Structures related to movement (s7)
Structure of shoulder region (s720)
Structure of upper extremity (s730)
Structure of lower extremity (s750)

Activities and participation (d) (comprises 51 ICF items)

Learning and applying knowledge (d1)
Listening (d115)
Acquiring skills (d155)
Focusing attention (d160)
Reading (d166)
Writing (d170)
Calculating (d172)
Solving problems (d175)

General tasks and demands (d2)
Undertaking a single task (d210)
Undertaking multiple tasks (d220)
Carrying out daily routine (d230)
Handling stress and other psychological demands (d240)

Communication (d3)
- Communicating (receiving) with spoken messages (d310)
- Communicating (receiving) with non-verbal messages (d315)
- Communicating (receiving) with written messages (d325)
- Speaking (d330)
- Producing non-verbal messages (d335)
- Writing messages (d345)
- Conversation (d350)
- Using communication devices and techniques (d360)

Mobility (d4)
- Changing basic body position (d410)
- Maintaining a body position (d415)
- Transferring oneself (d420)
- Lifting and carrying objects (d430)
- Fine hand use (d440)
- Hand and arm use (d445)
- Walking (d450)
- Moving around (d455)
- Moving around in different locations (d460)
- Moving around using equipment (d465)
- Using transportation (d470)
- Driving (d475)

Self-care (d5)
- Washing oneself (d510)
- Caring for body parts (d520)
- Toileting (d530)
- Dressing (d540)
- Eating (d550)
- Looking after one's health (d570)

Domestic life (d6)
- Acquisition of goods and services (d620)
- Preparing meals (d630)
- Doing housework (d640)

Interpersonal interactions and relationships (d7)
- Basic interpersonal interactions (d710)
- Informal social relationships (d750)
- Family relationships (d760)
- Intimate relationships (d770)

Major life areas (d8)
- Acquiring, keeping and terminating a job (d845)
- Remunerative employment (d850)
- Non-remunerative employment (d855)

Basic economic transactions (d860)
Economic self-sufficiency (d870)

Community, social and civic life (d9)
Community life (d910)
Recreation and leisure (d920)

Environmental factors (e) (comprises 33 ICF items)

Products and technology (e1)
Products or substances for personal consumption (e110)
Products and technology for personal use in daily living (e115)
Products and technology for personal indoor and outdoor mobility and transportation (e120)
Products and technology for communication (e125)
Products and technology for employment (e135)
Design, construction and building products and technology of buildings for public use (e150)
Design, construction and building products and technology of buildings for private use (e155)
Assets (e165)

Natural environment and human-made changes to environment (e2)
Physical geography (e210)

Support and relationships (e3)
Immediate family (e310)
Extended family (e315)
Friends (e320)
Acquaintances, peers, colleagues, neighbours and community members (e325)
Personal care providers and personal assistants (e340)
Health professionals (e355)
Health-related professionals (e360)

Attitudes (e4)
Individual attitudes of immediate family members (e410)
Individual attitudes of friends (e420)
Individual attitudes of acquaintances, peers, colleagues, neighbours and community members (e425)
Individual attitudes of personal care providers and personal assistants (e440)
Individual attitudes of health professionals (e450)
Individual attitudes of health-related professionals (e455)
Societal attitudes (e460)

Services, systems and policies (e5)
Architecture and construction services, systems and policies (e515)
Housing services, systems and policies (e525)

Communication services, systems and policies (e535)
Transportation services, systems and policies (e540)
Legal services, systems and policies (e550)
Associations and organizational services, systems and policies (e555)
Social security services, systems and policies (e570)
General social support services, systems and policies (e575)
Health services, systems and policies (e580)
Labour and employment services, systems and policies (e590)

3.8. Discussion

In this chapter we have discussed two-sample global permutation tests for sets of multivariate ordinal variables in potentially high-dimensional set-ups, primarily motivated by the need for statistical tools to analyze data that have been collected by means of the WHO's ICF. Specifically, we have addressed the closely related problems 'SMH against MI' ((3.1) against (3.2)) and 'SMH against MO' ((3.1) against (3.3)). While, under SMH, the ordinal variables' marginal distributions are identical between the two groups to be compared, at least one of these marginal distributions is inhomogeneous under MI and, as a special case thereof, stochastically ordered under MO.

To capture MI and MO, we have proposed sum statistics (see (3.7) and (3.8)), derived as multivariate test statistics under the working assumption that the variables in the set to be tested are independent. Under this assumption, we have found that the test statistic of Klingenberg et al. (2009), which our test statistic for MO is based on, is equivalent to the sum of univariate one-sided CA test statistics. Given that the working independence assumption will be inevitable in most practical situations, this equivalence argues for broader exploration of tests based on simple sum statistics constructed from other traditional univariate test statistics for ordinal data. Compared with tests based on max-T-statistics, such tests usually have more power against alternatives with many weak individual effects, which is an important class of alternatives in ICF-based applications and beyond. This is well known and has been reinforced in our power studies (which have not been shown). Regarding the tests that are proposed in this chapter, there is an additional intuitive explanation why they are expected to be powerful against this class of alternatives: both the χ^2 and CA test statistic, from which our sum statistics are constructed, are score test statistics (Chen, 1993; Agresti, 2002; Smyth, 2003), and it is well known that score test statistics lead to optimal power against alternatives that are close to the null hypothesis.

By means of simulations, we have explored the behaviour of the proposed permutation tests and their max-T-counterparts under SMH, i.e. in null scenarios where the multivariate observations may be non-exchangeable across groups (Figures 3.1–3.6).

The motivation behind has been that despite the theoretically well-founded criticism towards permutation-based inference in such data scenarios, researchers commonly face the problem that no superior (e.g. bootstrap-based) inference methods exist. Of the bootstrap procedures that we have considered (one of which has been shown), none has proved to be a promising alternative to the permutation procedure. We have called this common situation null dilemma. As expected, our simulations have suggested that how deficient the permutation procedure can become depends on the difference in the group-specific covariance matrices, the proportion between the group sizes and the number of variables in the set to be tested. It has come as an initially unexpected observation, however, that the choice of the test statistic and the number of categories per variable seem to play a crucial role as well. For instance, max-T permutation tests have shown remarkable robustness properties under SMH, especially when the max-T based on χ^2 test statistics is used and the number of categories is not too small (Figure 3.3). Subject to our simulations which, admittedly, focus on scenarios with a non-negative uniform correlation structure, it can thus be concluded that theoretical invalidity does not necessarily imply practical invalidity. It is unrealistic to expect simple and generally valid guidelines, but we believe that systematic studies such as ours can help to establish some useful practical recommendations regarding the use of permutation tests under SMH.

The tests presented in this chapter are useful by themselves and, in addition, can be fruitfully combined with multiplicity adjustment procedures for, for example, hypotheses that can be structured in a directed acyclic graph (Goeman and Mansmann, 2008) or in a tree by some prior knowledge (Meinshausen, 2008; Goeman and Solari, 2010; Goeman and Finos, 2012). This is particularly relevant for the analysis of ICF-based data because here such prior knowledge is on hand. However, the tests discussed have their limitations and can be improved in several directions. Firstly, they may be extended to scenarios in which the variables in the set of interest are considered of different importance and researchers would like to account for this in their analyses, rather than to treat all variable-specific test statistics on the same footing. In the ICF context, such scenarios are not unrealistic: the ICF item 'heart functions' (b410) might be considered more important than the ICF item 'voice functions' (b310), for example. Secondly, they may be extended to scenarios in which not all variables are measured on the same ordinal scale. This seems unproblematic if the marginal test statistics maintain the same number of degrees of freedom, as is the case for the CA test statistic. When the χ^2 test statistic is used, however, some standardization will be needed. Thirdly, it is desirable to extend them to scenarios in which two groups are to be compared after adjustment for covariates. In non-randomized ICF studies, for instance, the two groups to be compared often differ substantially with respect to age and BMI, which are the major confounders in studies on human functioning and disability. To

avoid false positive results due to such confounders, it is of utmost importance to be able to adjust for them in the analysis. A simple way to achieve this is to apply the proposed unadjusted tests in covariate-defined strata and subsequently correct for multiplicity over the strata. However, such an approach usually becomes infeasible when there are several potential confounders to adjust for, since the typical sample sizes are too small to construct multivariate strata. Alternatively, the comprehensive theory on GLMs may be exploited to form relevant sum statistics, yet their permutation null distribution will require more assumptions than in the unadjusted case to be valid. We come back to this issue in Chapter 4 where we shall discuss GLM-based global tests which allow for adjustment for covariates.

4. Testing global hypotheses in the generalized linear model

This chapter is devoted to research questions that can be formulated within the context of GLMs; in particular, it is concerned with global tests that can be used to assess, in possibly high-dimensional set-ups, the presence of an association between a set of ordinally scaled covariates and an outcome variable within the range of GLMs, if desired after adjustment for the effect of other covariates. A frequent question from the ICF context that can be answered by means of such tests is whether individuals' profile of functional limitations and disabilities is associated with some subjective quality-of-life score, after adjustment for certain socio-demographic aspects. Section 4.1 provides an overview of the particular contents of this chapter. The chapter, apart from Sections 4.3.5 and 4.4.3, and is mainly based on Jelizarow et al. (2014b).

4.1. Guideline through the chapter

To begin with, let us briefly recall some important points from Chapter 3. Besides discussing a Hotelling-type test along the lines of Agresti and Klingenberg (2005) which treats ordinal data as nominal, in Chapter 3 we have generalized the two-sample permutation test of stochastic order of Klingenberg et al. (2009) from one-sided to two-sided problems. Furthermore, we have shown that, under working independence between the variables in the set to be tested, the test statistic of Klingenberg et al. (2009) is equivalent to the sum of variable-specific one-sided CA test statistics over the whole set. Our own test statistic, Q_{CA}, equals the sum of variable-specific two-sided CA test statistics. The test of Klingenberg et al. (2009) and that from Chapter 3 can thus be seen as permutation-based generalizations of the CA test to higher dimensions. This fact renders them an intuitive choice for set-based analyses of ordinal data, yet they have their limitations. Firstly, they are confined to problems that can be framed as two-group comparisons, such as when the set of interest is to be tested for association with some binary variable. This leaves many possible set relationships with non-binary variables unexplored. Secondly, they do not allow for adjustment for potential confounders. In practice where observational studies are common, however, the possi-

bility of making such adjustments is of utmost importance, in order that false positive findings can be prevented. The present chapter develops two global tests for multivariate ordinal data which overcome the above limitations. The tests are based on different assumptions regarding the distances between the variables' ordered categories, rendering them useful in different practical situations.

The structure of this chapter is as follows. Section 4.2 introduces the statistical framework within which both tests are being constructed. In particular, this is the framework of the 'global test' of Goeman et al. (2004, 2006) which was originally proposed for the analysis of sets of genes or, formulated in statistical terms, sets of metrically scaled variables. Within the broad context of GLMs (McCullagh and Nelder, 1989), the global test exploits the duality between association and prediction: if the set of interest is associated with some other variable, it will improve prediction of that variable. Adopting the terminology of prediction models, the considered null hypothesis is that none of the covariates in the set is associated with the outcome variable, and the alternative hypothesis is that at least one of the covariates in the set shows such an association. Adjustment for other covariates is feasible, provided that their number is smaller than the sample size, which is the standard case in practice. Section 4.3 then elaborates and discusses the two tests proposed. The first test is simply the original global test for metric data applied to scores that need to be assigned a priori to the covariates' categories. In ICF-based applications, one can for example assign 1 to 'no problem', 2 to 'mild to moderate problem' and 4 to 'severe to complete problem' if one believes that the distance between 'mild to moderate problem' and 'severe to complete problem' is twice the distance between 'no problem' and 'mild to moderate problem'. We shall refer to this test as CA-type test, since the CA test is also based on prespecified scores. It turns out that, with data standardized to unit variance, this test is a natural generalization of the traditional two-sided CA test to higher dimensions, covariate-adjusted scenarios and all types of outcome variables that are within the range of GLMs. Immediate connections with the methods from Chapter 3 are pointed out. While the CA-type test expects the user to explicitly choose scores, and making a choice of scores implies making assumptions on the distances between the covariates' categories, the second test which we shall refer to as score-free test is unpredjudiced regarding these distances. As such, it is ideally suited for ordinal covariates because, by definition, the distances between their categories are generally unknown. The unpredjudicedness is achieved through an appropriate dummy-based coding scheme for the ordinal observations which uses only the ordering of the categories. An appealing property of this test is that the test result does not depend on any reference category in the coding scheme. While Section 4.4 examines the behaviour of the two tests by means of simulations, Section 4.5 illustrates their application with data from rehabilitation medicine, and provides practical recommendations on when to favour one or the other. Finally,

Section 4.6 closes the chapter with a short summary and discussion of its contents.

4.2. The 'global test' framework

4.2.1. Hypotheses, test statistic and significance assessment

For a sample of n independent subjects, suppose that we have an $n \times 1$ outcome vector $\boldsymbol{y}$, an $n \times q$ design matrix $\boldsymbol{Z}$ which contains realizations of the covariates we would like to adjust for (e.g. typical potential confounders such as age and sex), and an $n \times p$ design matrix $\boldsymbol{X}$ which contains realizations of the covariates we would like to make inferences about. Suppose further that q is smaller than n, whereas p may exceed n. The data situation may thus be high-dimensional. Under the assumption that the covariates and the outcome variable relate to each other via the GLM, we have

$$g(E(\boldsymbol{y})) = \mathbf{1}\gamma_0 + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{X}\boldsymbol{\beta}, \tag{4.1}$$

where $g(\cdot)$ is the canonical link function for the exponential family distribution of the components of $\boldsymbol{y}$, for example the identity function when the outcome variable is continuous (e.g. some blood parameter) or the logit function when the outcome variable is binary (e.g. some disease subtype). $\mathbf{1}$ is an $n \times 1$ vector of ones, γ_0 denotes an intercept term, $\boldsymbol{\gamma}$ is an unknown $q \times 1$ vector of regression coefficients for the covariates in $\boldsymbol{Z}$, and $\boldsymbol{\beta}$ is an unknown $p \times 1$ vector of regression coefficients for the covariates in $\boldsymbol{X}$. Based on the observed data, we are interested in whether the set of covariates in $\boldsymbol{X}$ as a whole is associated with the outcome $\boldsymbol{y}$, after adjustment for the effect of the covariates in $\boldsymbol{Z}$. This problem can be expressed through the hypotheses

$$H_0 : \boldsymbol{\beta} = \mathbf{0} \quad \text{against} \quad H_A : \boldsymbol{\beta} \neq \mathbf{0}. \tag{4.2}$$

Problem (4.2) is that for which Goeman et al. (2004, 2006) developed the 'global test', based on ideas of le Cessie and van Houwelingen (1995). In particular, they derived a score test statistic that can be employed whatever the dimensionality of the alternative hypothesis is, provided that the respective null hypothesis is low-dimensional. This is in contrast to the classical score, Wald or likelihood ratio (LR) test statistic: they all break down when the number of model parameters under the alternative of interest exceeds the number of subjects in the sample. In explicit terms, the test statistic of Goeman et al. (2004, 2006) has the form

$$S = (\boldsymbol{y} - \boldsymbol{\mu})^\top \boldsymbol{X}\boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{\mu}),$$

where $\boldsymbol{\mu}$ is the expectation of $\boldsymbol{y}$ under the null hypothesis. Because $\boldsymbol{\mu}$ is unknown, its maximum likelihood estimate

$$\hat{\boldsymbol{\mu}} = g^{-1}(\mathbf{1}\hat{\gamma}_0 + \boldsymbol{Z}\hat{\boldsymbol{\gamma}})$$

is plugged in, with $\hat{\gamma}_0$ and $\hat{\boldsymbol{\gamma}}$ being the null model coefficients estimated via an iteratively reweighted least squares algorithm. The resultant test statistic

$$\hat{S} = (\boldsymbol{y} - \hat{\boldsymbol{\mu}})^\top \boldsymbol{X}\boldsymbol{X}^\top (\boldsymbol{y} - \hat{\boldsymbol{\mu}}) \tag{4.3}$$

is thus a quadratic form in the residuals of the null model. For this quadratic form, Goeman et al. (2011) analytically derived an approximate null distribution which is conditional on $\boldsymbol{X}$ and thus remains valid for any correlation between the covariates in the set considered. By means of simulations, this null distribution was shown to perform well with respect to type I error rate control even when the sample size is moderate to small. Alternatively, the test statistic's exact null distribution may be obtained via permutation, yet this procedure is computationally more demanding and, more importantly, it is only valid for problem (4.2) if the null covariates and the covariates in the set to be tested are independent of each other. For significance assessment, the test statistic's permutation null distribution should therefore only come into question if such an independence assumption seems plausible or, trivially, if no covariates are present under the null hypothesis. Here we shall use the approximate null distribution of Goeman et al. (2011) throughout.

4.2.2. Properties of tests from the 'global test' family

The global test exhibits several properties (P1–P6) making it amenable to broad and efficient use in practice. As previously mentioned in Section 4.2.1,

(a) it is applicable both in the case of low-dimensional and high-dimensional alternatives (P1),

(b) it allows for covariate adjustment without further assumptions (P2),

(c) it is valid even under correlation (P3), and

(d) it can be performed at low computational costs, since an analytical approximation of the test statistic's null distribution is at hand (P4).

Besides that,

(e) it possesses an optimality property, which follows from the fact that it has been constructed as a score test. In particular, it has optimal average power to detect alternatives uniformly distributed on the p-dimensional ball $\left\|\boldsymbol{\beta}\right\| \leq \epsilon$, for $\epsilon \downarrow 0$. In less technical terms, among all possible tests, the global test maximizes the average power against alternatives that are in a neighbourhood of the null hypothesis (P5). On average, it is thus the best test to use if it is expected that all or most covariates in the set are only weakly associated with the outcome variable. It is important to note, however, that this optimality property is meant in terms of the chosen parametrization of the covariates under the alternative; changing

the parametrization means changing the shape of the neighbourhood of the null hypothesis where the test is optimal.

Finally,

(f) the test statistic (4.3) can be written as

$$\hat{S} = \sum_{k=1}^{p} \left[\boldsymbol{x}_k^\top (\boldsymbol{y} - \hat{\boldsymbol{\mu}}) \right]^2,$$

that is, the sum of covariate-specific test statistics over the whole set, where $\boldsymbol{x}_k$ is the kth column of $\boldsymbol{X}$ (P6). We shall see later on in Sections 4.3 and 4.5 that this property proves to be useful in various respects. Noting that, at convergence of the null model, it holds

$$(\boldsymbol{y} - \hat{\boldsymbol{\mu}}) = (\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{y} - \hat{\boldsymbol{\mu}}),$$

where $\boldsymbol{I}$ denotes the n-dimensional identity matrix,

$$\boldsymbol{H} = \tilde{\boldsymbol{Z}} \left(\tilde{\boldsymbol{Z}}^\top \boldsymbol{W} \tilde{\boldsymbol{Z}} \right)^{-1} \tilde{\boldsymbol{Z}}^\top \boldsymbol{W}$$

with

$$\tilde{\boldsymbol{Z}} = (\mathbf{1} | \boldsymbol{Z})$$

is the asymmetric hat matrix of the null model, and

$$\boldsymbol{W} = \mathrm{diag}(\phi v(\hat{\boldsymbol{\mu}}))$$

is the covariance matrix of $\boldsymbol{y}$ under the null hypothesis, with ϕ being the dispersion parameter and $v(\cdot)$ the variance function of the distribution of the components of $\boldsymbol{y}$, the kth covariate-specific test statistic can in turn be written as

$$\hat{S}_k = \left[\boldsymbol{x}_k^\top (\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{y} - \hat{\boldsymbol{\mu}}) \right]^2.$$

From this representation we can immediately see that the contribution of each covariate to the overall test statistic is determined by its residual variance, adjusted for the null covariates. Whether this implicit weighting is appropriate or not depends on the application, such that some standardization might become necessary. We come back to this issue in Section 4.3.4. For further interpretations of the test statistic (4.3) we refer to Goeman et al. (2004, 2006), and to Goeman et al. (2004) and Solari et al. (2012) for connections with penalized likelihood and random effects methods.

Essentially, the framework of the global test is defined by (4.1)–(4.3), and all tests constructed within it enjoy the properties P1–P6. For sets of metrically scaled covariates,

several such tests have already been implemented, each of which is suited for a different outcome type: a global test for the linear model (for continuous outcomes) (Goeman et al., 2004), the logit model (for binary outcomes) (Goeman et al., 2004), the multinomial logit model (for multi-class outcomes), the Poisson model (for count outcomes), and an extended global test for the Cox proportional hazards model (for survival outcomes) (Goeman et al., 2005). In Section 4.3 we discuss how, within the above framework, this versatile methodology can be made applicable to sets of ordinally scaled covariates.

4.3. Handling ordinal covariates within the 'global test' framework

4.3.1. Preliminaries

In what follows, suppose that the covariates in the set of interest are ordinal, and let c_k denote the number of categories of the kth covariate. For convenience of notation, let the ordered categories of unknown distance be labelled with numbers 1 to c_k. (In the ICF-based application considered in Section 4.5, for example, the numbers 1 to 3 stand for the categories 'no problem', 'mild to moderate problem' and 'severe to complete problem' or for the categories 'facilitator', 'neither barrier nor facilitator' and 'barrier', as has already been noted in Section 1.1.) For x_{ik}, the ith realization of the kth covariate, we thus have:

$$x_{ik} \in \{1, \ldots, c_k\}.$$

Technically, the ordinal covariates' special character manifests itself in the fact that their realizations typically need to be recoded in order to enable proper specification of the model under the alternative. Direct use of the labels would imply the assumption that the covariates' categories are equally-spaced. Given that the numbers 1 to c_k are arbitrary and merely meant to indicate which of the categories have been observed, this may not always be desirable. Hence, if we want to render the global test methodology sensitive towards the covariates' ordinal nature, we need to recode the x_{ik}s appropriately. Two approaches to do so are presented in Sections 4.3.2 and 4.3.3, resulting in two different tests for sets of ordinal data which both enjoy the properties P1–P6 described in Section 4.2.2.

4.3.2. Cochran-Armitage-type approach with prespecified scores

The approach

The first approach codes observations on an ordinal scale in the same fashion as does the CA test for trend, hence the name **CA-type approach**. Essentially, this means that the numbers 1 to c_k are transformed into scores that need to be assigned a priori to the ordinal covariates' categories, and the observed scores are then treated as if they were metric observations. Our motivation to consider such a score-dependent approach within the global test framework stems from the wide popularity of the CA test in statistical practice and particularly in medical applications. Formally, the transformation rule that characterizes the CA-type approach can be expressed by

$$\tilde{x}_{ik} = u_k(v) \text{ if } x_{ik} = v, \tag{4.4}$$

where $v = 1, \ldots, c_k$ indexes the ordered categories and $u_k(v)$ denotes the score assigned to the vth category of the kth covariate. It is easy to see that direct use of the numbers 1 to c_k is a special case of (4.4). In particular, direct use of the numbers 1 to c_k corresponds to (4.4) with $u_k(v) = v$. The CA-type test statistic then is

$$\hat{S}_{\text{CA}} = (\boldsymbol{y} - \hat{\boldsymbol{\mu}})^\top \tilde{\boldsymbol{X}} \tilde{\boldsymbol{X}}^\top (\boldsymbol{y} - \hat{\boldsymbol{\mu}}), \tag{4.5}$$

where $\tilde{\boldsymbol{X}}$ is the score-transform of the design matrix $\boldsymbol{X}$ in terms of (4.4). Thus, the test statistic (4.5) is the original test statistic (4.3) applied to prespecified scores. We shall refer to the resultant statistical hypothesis test as **CA-type test**.

A special case: from the Cochran-Armitage-type test to the generalized Cochran-Armitage test

A special variant of the test statistic (4.5) arises when the outcome variable is binary, the null model contains only an intercept and the columns $\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_p$ of $\tilde{\boldsymbol{X}}$ are standardized to have unit variance. In particular, under these conditions, the test statistic (4.5) is equivalent to the sum of covariate-specific two-sided CA test statistics. The proof is a straightforward calculation and is given further below. We can immediately conclude from this relationship that, with $\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_p$ standardized to unit variance, the resultant CA-type test is a proper generalization of the traditional two-sided CA test in three important directions: to higher dimensions, to covariate-adjusted scenarios and to all types of outcome variables that are within the range of GLMs. As such, it can likewise be seen as a generalization of the CA sum statistic-based test that has been presented in Chapter 3. We shall refer to this special variant of the CA-type test as **generalized CA test**. The standardization of the columns of $\tilde{\boldsymbol{X}}$, and its implications, will be further discussed in Section 4.3.4.

We now prove the above-mentioned relationship between the test statistic (4.5) and the traditional two-sided CA test statistic. Because the test statistic (4.5) can be written as the sum of covariate-specific test statistics over the whole set, we can examine the univariate case without loss of generality. Let the outcome variable be binary and 0/1-coded. With $\tilde{x}_{ik}$ and y_i the ith component of $\tilde{\boldsymbol{x}}_k$ and $\boldsymbol{y}$, respectively, be

$$\sum_{i=1}^{n} \delta_{y_i 1} =: n_2$$

the number of subjects with outcome 1 ('cases'),

$$n - n_2 =: n_1$$

the number of subjects with outcome 0 ('controls'),

$$\sum_{i=1}^{n} \delta_{\tilde{x}_{ik} u_k(v)} y_i =: n_{2kv}$$

the number of cases with $\tilde{x}_{ik} = u_k(v)$ and

$$\sum_{i=1}^{n} \delta_{\tilde{x}_{ik} u_k(v)} - n_{2kv} =: n_{1kv}$$

the number of controls with $\tilde{x}_{ik} = u_k(v)$. For the logit model without null covariates and with the columns of $\tilde{\boldsymbol{X}}$ standardized to have unit variance, the test statistic (4.5) can be written as the weighted sum

$$\hat{S}' = \sum_{k=1}^{p} w_k \left[\tilde{\boldsymbol{x}}_k^\top (\boldsymbol{y} - \mathbf{1} \frac{n_2}{n}) \right]^2,$$

where

$$w_k = \left\{ \frac{1}{n} \left[\tilde{\boldsymbol{x}}_k^\top \tilde{\boldsymbol{x}}_k - \frac{1}{n} \left(\tilde{\boldsymbol{x}}_k^\top \mathbf{1} \right)^2 \right] \right\}^{-1}.$$

We can write the kth covariate-specific test statistic $\hat{S}'_k = \left[\tilde{\boldsymbol{x}}_k^\top (\boldsymbol{y} - \mathbf{1} \frac{n_2}{n}) \right]^2$ as

$$\begin{aligned}
\hat{S}'_k &= \left[\sum_{i=1}^{n} \tilde{x}_{ik} \left(y_i - \frac{n_2}{n} \right) \right]^2 \\
&= \left[\sum_{i=1}^{n} \sum_{v=1}^{c_k} \delta_{\tilde{x}_{ik} u_k(v)} u_k(v) \left(y_i - \frac{n_2}{n} \right) \right]^2 \\
&= \left[\sum_{v=1}^{c_k} u_k(v) \left(\sum_{i=1}^{n} \delta_{\tilde{x}_{ik} u_k(v)} y_i - \sum_{i=1}^{n} \delta_{\tilde{x}_{ik} u_k(v)} \frac{n_2}{n} \right) \right]^2 \\
&= \left[\sum_{v=1}^{c_k} u_k(v) \left(n_{2kv} - \frac{n_2}{n} n_{1kv} - \frac{n_2}{n} n_{2kv} \right) \right]^2 \\
&= \left[\sum_{v=1}^{c_k} u_k(v) \left(\frac{n_1 n_{2kv}}{n} - \frac{n_2 n_{1kv}}{n} \right) \right]^2.
\end{aligned}$$

Likewise, we can write the kth covariate-specific weight w_k as

$$\begin{aligned} w_k &= \left\{ \frac{1}{n} \left[\sum_{i=1}^{n} \tilde{x}_{ik}^2 - \frac{1}{n} \left(\sum_{i=1}^{n} \tilde{x}_{ik} \right)^2 \right] \right\}^{-1} \\ &= \left\{ \frac{1}{n^2} \left[n \sum_{i=1}^{n} \sum_{v=1}^{c_k} \delta_{\tilde{x}_{ik} u_k(v)} u_k^2(v) - \left(\sum_{i=1}^{n} \sum_{v=1}^{c_k} \delta_{\tilde{x}_{ik} u_k(v)} u_k(v) \right)^2 \right] \right\}^{-1} \\ &= \left\{ \frac{1}{n^2} \left[n \sum_{v=1}^{c_k} u_k^2(v)(n_{1kv} + n_{2kv}) - \left(\sum_{v=1}^{c_k} u_k(v)(n_{1kv} + n_{2kv}) \right)^2 \right] \right\}^{-1}. \end{aligned}$$

It is now easy to see that, up to a constant factor, $w_k \hat{S}'_k$ is equivalent to the square of the one-sided CA test statistic (see for example Freidlin et al. (2002) for this most frequently used formulation of the latter), which in turn is the two-sided CA test statistic. Thus, $\hat{S}'$ is equivalent to the sum of traditional two-sided covariate-specific CA test statistics. (The constant factor corresponds exactly to that by means of which $\hat{S}'$ is rescaled in order to be able to compute its approximate null distribution (Goeman et al., 2011).)

Choice of scores

For the validity of the CA-type test, the concrete choice of scores is not relevant, provided that this choice has been made without inspection of the data observed. When it comes to the test's power, however, the choice of scores is crucial. The crux is that the scores reflect the suspected relationship between the covariates in the set to be tested and the outcome variable. For example, choosing equally-spaced scores for all covariates in the set reflects the suspicion that the relationship is linear, that is, that the outcome changes linearly between two adjacent categories of at least one covariate in the set. If the suspicion is correct, the CA-type test will be powerful. If it is not correct, that is, if the choice of scores is poor, it may happen that the test has no power at all. We shall illustrate this point by means of simulations in Section 4.4.

In connection with the choice of scores, two issues deserve particular emphasis. Firstly, the CA-type test has the desirable property that two sets of scores

$$\{(u_k(1), \ldots, u_k(c_k))\}_{k=1}^{p}$$

and

$$\left\{(u'_k(1), \ldots, u'_k(c_k))\right\}_{k=1}^{p}$$

lead to the same test result if constants $s, t \in \mathbb{R}$ exist such that

$$u'_k(v) = s \cdot u_k(v) + t$$

for all v and k. The outcome of the test is thus the same for scores that are linear transforms of each other, such as (1, 2, 4) and (3, 5, 9) or (10, 20, 40). Practically speaking, this means that the test result solely depends on the kind of the suspected relationship and not on the — to some extent subjective — numerical scale that has been chosen to reflect it. This property may come as a surprise because, obviously, the test statistic (4.5) is not invariant to every linear transformation of the scores used. The reason why the outcome of the test nevertheless is so lies in the way in which the test statistic needs to be rescaled before its approximate null distribution can be derived analytically (Goeman et al., 2011). For details on the rescaling we refer to Goeman et al. (2011), and here limit ourselves to just mentioning its welcome consequences. Secondly, because the CA-type test is a two-sided test and as such does not depend on the sign of the true regression coefficients for the covariates in the set of interest, it will not be sensitive towards the direction of the suspected relationship of each covariate with the outcome variable. For illustration, for some set that only contains ordinal covariates with three categories (e.g. ICF items after the five-level ordinal scale originally used in the ICF components b, s and d and the nine-level ordinal scale originally used in the ICF component e have been coarsened), any of the 2^p possible mixtures of the strictly monotonically increasing scores (1, 2, 4) and the strictly monotonically decreasing scores (-1, -2, -4) will lead to the same test result. This should be kept in mind in order to prevent false inferential conclusions.

The CA-type test is useful whenever the research interest focuses on the detection of relatively specific alternatives. In such situations, the fact that the test requires specification of scores for all covariates in the set to be tested, and that making a choice of scores means making assumptions on the distances between the covariates' categories, will seldom pose considerable problems. Rather, it can be taken advantage of in order to direct the power of the test towards the desired alternative. When the research interest is broader in the sense that many different alternatives are considered equally important, however, '[...] scientists may feel that the assignment of scores is slightly unscrupulous, or at least they are uncomfortable about it. [...]' (Cochran, 1954). A test that is useful in such situations is discussed in the next section.

4.3.3. Score-free approach

The approach

The second approach to handling ordinality dispenses with scores altogether, hence the name **score-free approach**. It codes ordinal observations by using the dummy-based coding scheme of Walter et al. (1987), sometimes called split coding (Gertheiss et al., 2011). This means that the numbers 1 to c_k are transformed such that the or-

dinal covariates are no more represented one-dimensionally but multi-dimensionally by groups of dummies, with each group corresponding to one ordinal covariate. As opposed to classical dummies, the dummies used here contain information on the ordering of the covariates' categories. In explicit terms, the transformation rule that characterizes the score-free approach is

$$d_{ik\tilde{v}} = \begin{cases} 1 & \text{if } x_{ik} > \tilde{v} \\ 0 & \text{otherwise,} \end{cases} \tag{4.6}$$

where $d_{ik\tilde{v}}$ is the ith component of $\boldsymbol{d}_{k\tilde{v}}$, which is the $\tilde{v}$th dummy vector for the kth covariate, and $\tilde{v} = 1, \ldots, \tilde{c}_k$ with $\tilde{c}_k := c_k - 1$. The score-free test statistic then is

$$\hat{S}_{\text{SF}} = (\boldsymbol{y} - \hat{\boldsymbol{\mu}})^\top \boldsymbol{D}\boldsymbol{D}^\top (\boldsymbol{y} - \hat{\boldsymbol{\mu}}), \tag{4.7}$$

where

$$\boldsymbol{D} = \left(\boldsymbol{D}_1 | \ldots | \boldsymbol{D}_p\right)$$

is the dummy-transform of the design matrix $\boldsymbol{X}$ in terms of (4.6), with

$$\boldsymbol{D}_k = \left(\boldsymbol{d}_{k1} | \ldots | \boldsymbol{d}_{k\tilde{c}_k}\right)$$

denoting the kth group of dummy vectors. We shall refer to the resultant statistical hypothesis test as **score-free test**. Because the $\boldsymbol{D}_k$s are $n \times \tilde{c}_k$ matrices, we have $\tilde{c}_k$ (rather than one) model parameters for the kth covariate, so that the dimension of the alternative in (4.2) increases from $1+q+p$ to $1+q+\sum_{k=1}^{p} \tilde{c}_k$. We may thus encounter an alternative that is high-dimensional even when the data situation in itself is low-dimensional. As pointed out in Section 4.2.1, however, test statistics constructed within the global test framework can be used whatever the dimensionality of the alternative hypothesis is, and therefore no problems occur from that.

The $\boldsymbol{D}_k$s obtained through (4.6) are easy to interpret: the first dummy vector tells us whether the sample has been classified higher than into the first category, the second dummy vector tells us whether the sample has been classified higher than into the second category, and so on. The respective model parameters are similarly easy to interpret: $\beta_{k\tilde{v}}$, the $\tilde{v}$th regression coefficient for the kth covariate, describes the distance between category $\tilde{v}$ and $\tilde{v}+1$, that is, the difference between the effects of category $\tilde{v}$ and $\tilde{v}+1$. Effectively, this means that the first category is taken to be the reference category, and that the effects of the first and the second category are assumed to be more similar than the effects of the first and the third category, which in turn are assumed to be more similar than the effects of the first and the fourth category, and so on. Stated differently, it is expected that the outcome changes rather smoothly than jaggedly across the categories, which is intuitively plausible for covariates measured

on an ordinal scale. It is important to emphasize, however, that no assumptions are made on the particular size of the $\beta_{k\tilde{v}}$s. The resultant score-free test is therefore ideally tailored to ordinal data: it incorporates the ordering of the covariates' categories, but at the same time it is unprejudiced regarding the distances between them. This is well reflected in the power properties of the test, as simulations in Section 4.4 will confirm: the range of alternatives it can detect varies from linear to umbrella-like relationships between the covariates in the set of interest and the outcome variable, with monotonic relationships being more likely to be detected than non-monotonic relationships.

Robustness against the choice of the reference category

It has just been said that the transformation rule (4.6) effectively takes the first category to be the reference category, and that it defines dummies under the assumption of 'smoothness'. Analogous transformation rules or coding schemes may be written up with any other of the categories as reference category. A general formulation is

$$d_{ik\tilde{v}}^{(r)} = \begin{cases} -1 & \text{if } x_{ik} \leq \tilde{v} \wedge \tilde{v} < r \\ 1 & \text{if } x_{ik} > \tilde{v} \wedge \tilde{v} \geq r \\ 0 & \text{otherwise,} \end{cases} \tag{4.8}$$

where $r \in \{1, \ldots, c_k\}$ is the chosen reference category. It is easy to see that (4.8) reduces to (4.6) when $r = 1$. The interpretation of the respective $\boldsymbol{D}_k^{(r)}$s is slightly more intricate than above. For example, for $c_k = 3$ and $r = 2$, the first dummy vector tells us whether the sample has been classified lower than into the second category, and the second dummy vector tells us whether the sample has been classified higher than into the second category. For $c_k = 3$ and $r = 3$, in contrast, the first dummy vector tells us whether the sample has been classified lower than into the second category, and the second dummy vector tells us whether the sample has been classified lower than into the third category. At first sight, this may suggest that different choices of the reference category lead to different test statistics and hence to potentially different inferential conclusions. This, however, is not the case, which is convenient because the choice of the reference category is often arbitrary. In particular, it is readily verified that the different nature of the $\boldsymbol{D}_k^{(r)}$s does not affect the interpretation of the $\tilde{v}$th regression coefficient as the distance between category $\tilde{v}$ and $\tilde{v}+1$. We thus have $\beta_{k\tilde{v}}^{(r)} = \beta_{k\tilde{v}}$ for all r, meaning that the parametrization of the model under the alternative does not depend on the choice of the reference category. Intuitively, it is therefore clear that any score-free test statistic $\hat{S}_{\text{SF}}^{(r)}$ which is derived based on (4.8) must be equivalent to the test statistic (4.7), provided that the null model includes at least an intercept. A formal proof of this valuable invariance property is provided below in the next paragraph. The score-free test may thus be regarded as a test that randomly picks one category on the

ordinal scale and, starting from there, parametrizes the distances between adjacent categories, thereby keeping them flexible.

We now prove the invariance of the score-free test statistic to the choice of the reference category. Consider the transformation rule (4.8), and let

$$\hat{S}_{\mathrm{SF}}^{(r)} = (\boldsymbol{y} - \hat{\boldsymbol{\mu}})^{\top} \boldsymbol{D}^{(r)} \boldsymbol{D}^{(r)\top} (\boldsymbol{y} - \hat{\boldsymbol{\mu}})$$

be the respective score-free test statistic, for some reference category $r \in \{1, \ldots, c_k\}$. To prove that $\hat{S}_{\mathrm{SF}}^{(r)}$ is invariant to the choice of the reference category, we must first rewrite it. Let $\boldsymbol{I}$ and $\boldsymbol{H}$ be defined as in Section 4.2.2, and let

$$\tilde{\boldsymbol{H}} = \tilde{\boldsymbol{Z}} \left(\tilde{\boldsymbol{Z}}^{\top} \tilde{\boldsymbol{Z}} \right)^{-1} \tilde{\boldsymbol{Z}}^{\top}$$

denote the projection matrix that $\boldsymbol{H}$ becomes in the case of the linear model with normally distributed errors. Using that

$$(\boldsymbol{y} - \hat{\boldsymbol{\mu}}) = (\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{y} - \hat{\boldsymbol{\mu}}),$$

and noting that

$$\boldsymbol{H}\tilde{\boldsymbol{H}} = \tilde{\boldsymbol{H}}$$

and therefore

$$(\boldsymbol{I} - \boldsymbol{H}) = (\boldsymbol{I} - \boldsymbol{H}) \left(\boldsymbol{I} - \tilde{\boldsymbol{H}} \right),$$

we can write $\hat{S}_{\mathrm{SF}}^{(r)}$ in the more cumberstone form

$$\hat{S}_{\mathrm{SF}}^{(r)} = (\boldsymbol{y} - \hat{\boldsymbol{\mu}})^{\top} (\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \tilde{\boldsymbol{H}}) \boldsymbol{D}^{(r)} \boldsymbol{D}^{(r)\top} (\boldsymbol{I} - \tilde{\boldsymbol{H}})(\boldsymbol{I} - \boldsymbol{H})^{\top} (\boldsymbol{y} - \hat{\boldsymbol{\mu}}).$$

Let $\boldsymbol{d}_{k\tilde{v}}^{(r)}$ be the $\tilde{v}$th dummy vector for the kth covariate. We notice that all that happens when we go from $\boldsymbol{d}_{k\tilde{v}}^{(r)}$ to $\boldsymbol{d}_{k\tilde{v}}^{(r+1)}$ is that the entries of the rth dummy vector are subtracted by 1. In equations, this means

$$\boldsymbol{d}_{k\tilde{v}}^{(r+1)} = \boldsymbol{d}_{k\tilde{v}}^{(r)} - \mathbf{1}\delta_{r\tilde{v}},$$

where $\delta_{r\tilde{v}} = 1$ if $r = \tilde{v}$ and $\delta_{r\tilde{v}} = 0$ otherwise. The different $\boldsymbol{d}_{k\tilde{v}}^{(r)}$s can thus be transformed into each other by shifts by the vector of ones. Because the vector of ones is in the null space of the projection defined by $(\boldsymbol{I} - \tilde{\boldsymbol{H}})$, it follows immediately that $(\boldsymbol{I} - \tilde{\boldsymbol{H}})\boldsymbol{D}^{(r)}$ is invariant to the choice of the reference category, provided that the null model is non-empty (i.e. it includes at least an intercept). Consequently, any choice of the reference category will lead to the same test statistic, which completes the proof.

4.3.4. Ordinal covariates on different scales

In practice, the most frequently encountered situation is that where all covariates in the set to be tested are measured on the same ordinal scale, that is, where $c_k = c$ for all k. This section briefly discusses practical solutions to potential issues that may arise in situations where the covariates are measured on different ordinal scales.

An important property of the test statistics (4.5) and (4.7) is that they can be decomposed into covariate-specific contributions. For the former, the contribution of each covariate to the overall test statistic is determined by its residual variance, adjusted for the null covariates. For the latter, the covariate-specific contribution is determined by the summed residual variances of the respective dummies, likewise adjusted for the null covariates. This becomes apparent from the fact that the test statistics can be written as

$$\hat{S}_{\text{CA}} = \sum_{k=1}^{p} \left[\tilde{\boldsymbol{x}}_k^\top (\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{y} - \hat{\boldsymbol{\mu}}) \right]^2$$

and

$$\hat{S}_{\text{SF}} = \sum_{k=1}^{p} \sum_{\tilde{v}}^{\tilde{c}_k} \left[\boldsymbol{d}_{k\tilde{v}}^\top (\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{y} - \hat{\boldsymbol{\mu}}) \right]^2,$$

respectively, where $\boldsymbol{H}$ is the hat matrix of the null model (see Section 4.2.1). In general, this implicit weighting of the covariates is desirable: covariates with high residual variance usually carry more potentially important information than those with low residual variance, so they should have more influence on the test result. However, when the covariates are measured on different ordinal scales, this weighting will in some way be distorted by the fact that covariates with many categories are more likely to lead to high residual variance than covariates with few categories. Given that the metric level of measurement is more informative than the ordinal one, and that the finer the ordinal scale the closer it is to the metric scale, one may argue that it is only intuitive to give more weight to covariates with many categories than to covariates with few categories. In some instances, however, one might want to correct for the imbalance between the ordinal scales used. This can be accomplished by standardizing each of the covariates to unit variance before the CA-type or the score-free test is being performed. For the CA-type test statistic $\hat{S}_{\text{CA}}$, this means that we need to replace $\tilde{\boldsymbol{x}}_k$ by

$$\tilde{\boldsymbol{x}}'_k = \frac{\tilde{\boldsymbol{x}}_k}{\sqrt{n^{-1} \tilde{\boldsymbol{x}}_k^\top (\boldsymbol{I} - \boldsymbol{H}') \tilde{\boldsymbol{x}}_k}},$$

where $\boldsymbol{H}' = n^{-1}\mathbf{1}\mathbf{1}^\top$. For the score-free test statistic $\hat{S}_{\text{SF}}$, standardization of each covariate to unit variance means that we need to replace $\boldsymbol{d}_{k\tilde{v}}$ by

$$\boldsymbol{d}'_{k\tilde{v}} = \frac{\boldsymbol{d}_{k\tilde{v}}}{\sqrt{n^{-1} \text{trace}\left[\boldsymbol{D}_k^\top (\boldsymbol{I} - \boldsymbol{H}') \boldsymbol{D}_k \right]}}.$$

For the CA-type test, the standardization procedure just described leads directly to the generalized CA test discussed in the second paragraph of Section 4.3.2. Unlike the ordinary CA-type test with unstandardized $\tilde{\boldsymbol{x}}_k$s, the generalized CA test is thus 'scale-corrected' by construction. One should be aware, however, that any correction for the imbalance caused by covariates measured on different ordinal scales comes at the price of a reweighting of covariates measured on the same ordinal scale, both in case of the CA-type and the score-free test. Whether it is sensible or not to pay this price depends on the application. When most of the covariates in the set to be tested are measured on the same ordinal scale and their residual variances, and therefore their implicit weights, are expected to vary considerably, it seems preferable to keep the original weighting, that is, not to standardize each of the covariates to unit variance. In contrast, when the covariates in the set of interest differ greatly in the number of categories and covariates with the same number of categories are expected to contribute similarly to the test result, it may be reasonable to perform the standardization. Such scenarios are, however, rather uncommon in practice, since covariates making up a set often describe similar aspects, and covariates describing similar aspects are typically measured on similar scales.

4.3.5. Practical realization in R

Both the CA-type and the score-free test can be performed by using the function `gt()` from the R package `globaltest` of Goeman and Oosting (2012). For illustration, suppose that, in some ICF study with sample size ten, we would like to test the two ICF items `b114` and `b134` jointly for association with a binary and 0/1-coded outcome `outc`, and that we would like to adjust for some potential confounder `conf`. Suppose further that the data observed are as follows.

```
outc <- c(0,0,0,0,0,1,1,1,1,1)
b114 <- c(2,1,3,2,3,2,1,1,2,3)
b134 <- c(1,1,2,1,3,3,1,2,2,2)
conf <- rnorm(10,mean=20,sd=10)
```

Then

```
library(globaltest)
gt(response=outc,null=~conf,alternative=~b114+b134,model="logistic")
```

performs the CA-type test with the equally-spaced scores (1, 2, 3), and the respective R output has the form

```
p-value Statistic Expected Std.dev #Cov
  0.614       8.7     15.3    14.8    2 .
```

To perform the CA-type test with other scores, we first need to recode the numbers 1 to 3 in the vectors `b114` and `b134` in terms of the transformation rule (4.4). If we wish to perform the generalized CA test rather than the CA-type test (i.e. if we wish to standardize the covariates to unit variance), this can be realized by means of the argument `standardized`. The command that we concretely need then is

```
library(globaltest)
gt(response=outc,null=~conf,alternative=~b114+b134,model="logistic",
   standardize=TRUE).
```

The score-free test can be performed simply via

```
library(globaltest)
gt(response=outc,null=~conf,alternative=~ordered(b114)+ordered(b134),
   model="logistic").
```

4.4. Cochran-Armitage-type versus score-free test: a simulation study for binary outcomes

4.4.1. Simulation set-up

In Section 4.3 we have stated that the CA-type test would be useful in situations where the research interest focuses on the detection of relatively specific alternatives, and that the score-free test in turn would be useful in situations where many different alternatives are considered equally important, that is, where the research interest is rather broad. In this section we present a small simulation study which we conducted with the primary objective of illustrating and further clarifying these statements. For this purpose, we examined the performance of the CA-type and the score-free test for different set-outcome relationships. The CA-type test was based on the equally-spaced scores 1 to c_k throughout, and both tests were used in their ordinary form with unstandardized covariates.

Throughout the study, the outcome variable was binary and 0/1-coded, the set to be tested comprised

$$p = 100$$

independent ordinally scaled covariates with the same number $c = 3$ of categories, and there were no covariates to be adjusted for. The sample sizes considered were

$$n = \{20, 40, 60, 80, 100\}.$$

Our major interest thus lay in high-dimensional data scenarios. Within this general set-up, we studied five different set-outcome relationships: linear, non-strictly monotonic, asymmetric umbrella, umbrella and mixed. For completeness, we further studied the null case of no relationship, even though, in principle, good type I error rate control can be expected due to the fact that our tests have been constructed within the global test framework. To obtain data sets for which the different relationships can be found, we used that, in the set-up considered, the set-outcome relationship is determined by the trend in the binomial proportions of sample units with outcome 1 (and 0, respectively) across the categories of each of the 100 covariates. With

$$(b^1_{k1}, b^1_{k2}, b^1_{k3}) =: \boldsymbol{b}^1_k$$

and

$$(b^0_{k1}, b^0_{k2}, b^0_{k3}) =: \boldsymbol{b}^0_k$$

denoting the kth covariate's binomial proportions of sample units with outcome 1 and 0, respectively, where

$$b^1_{kv}, b^0_{kv} \in (0, 1)$$

and

$$b^1_{kv} + b^0_{kv} = 1$$

for $v = 1, 2, 3$, the particular patterns of binomial proportions that we examined in our study were

(a) S0 (null case): $\boldsymbol{b}^1_1 = \cdots = \boldsymbol{b}^1_{100} = (0.5, 0.5, 0.5)$,

(b) S1 (linear): $\boldsymbol{b}^1_1 = \cdots = \boldsymbol{b}^1_{24} = (0.4, 0.5, 0.6)$;
$\boldsymbol{b}^1_{25} = \cdots = \boldsymbol{b}^1_{100} = (0.5, 0.5, 0.5)$,

(c) S2 (non-strictly monotonic): $\boldsymbol{b}^1_1 = \cdots = \boldsymbol{b}^1_{24} = (0.35, 0.55, 0.55)$;
$\boldsymbol{b}^1_{25} = \cdots = \boldsymbol{b}^1_{100} = (0.5, 0.5, 0.5)$,

(d) S3 (asymmetric umbrella): $\boldsymbol{b}^1_1 = \cdots = \boldsymbol{b}^1_{24} = (0.4, 0.6, 0.5)$;
$\boldsymbol{b}^1_{25} = \cdots = \boldsymbol{b}^1_{100} = (0.5, 0.5, 0.5)$,

(e) S4 (umbrella): $\boldsymbol{b}_1^1 = \cdots = \boldsymbol{b}_{24}^1 = (0.45, 0.65, 0.45)$;
$\boldsymbol{b}_{25}^1 = \cdots = \boldsymbol{b}_{100}^1 = (0.5, 0.5, 0.5)$ and

(f) S5 (mixed): $\boldsymbol{b}_1^1 = \cdots = \boldsymbol{b}_6^1 = (0.4, 0.5, 0.6)$;
$\boldsymbol{b}_7^1 = \cdots = \boldsymbol{b}_{12}^1 = (0.35, 0.55, 0.55)$;
$\boldsymbol{b}_{13}^1 = \cdots = \boldsymbol{b}_{18}^1 = (0.4, 0.6, 0.5)$;
$\boldsymbol{b}_{19}^1 = \cdots = \boldsymbol{b}_{24}^1 = (0.45, 0.65, 0.45)$;
$\boldsymbol{b}_{25}^1 = \cdots = \boldsymbol{b}_{100}^1 = (0.5, 0.5, 0.5)$.

For S1–S5, the number of informative covariates in the set was thus chosen as 24. For each desired pattern, random data sets were generated as follows. Firstly, the binary outcome was drawn from a Bernoulli distribution with probability of success equal to 0.5. Secondly, conditionally on the outcome, realizations of each of the 100 ordinal covariates were drawn from covariate-specific independent multinomial distributions such that the desired pattern of binomial proportions resulted; multinomial distributions that satisfy this condition were determined based on Bayes' theorem. For a more detailed description of the data generation technique see Section 4.4.3. The power (type I error rate) was then estimated from 10000 random data sets as the average rejection rate of false (true) null hypotheses, and the desired significance level was $\alpha = 0.05$. The simulation margin of error thus amounted to $\pm 2\{0.05(1-0.05)/10000\}^{1/2} \approx \pm 0.0044$. The results from our simulation experiments are reported in Section 4.4.2.

4.4.2. Simulation results

Table 4.2 summarizes the average rejection rates obtained with our two tests for the simulation scenarios S0–S5 described in the previous section. Under the null hypothesis of no association between the set and the outcome variable (scenario S0), both tests offer good type I error rate control: nearly all deviations of the actual type I error rate from the nominal one lie within the simulation margin of error of approximately $\pm 0.44\%$, which confirms the general usability of the approximate null distribution of Goeman et al. (2011). Under the different alternative hypotheses of interest (scenarios S1–S5), we find for the CA-type test that its power increases the better the prespecified scores reflect the true set-outcome relationship. This is intuitively plausible and typical for score-dependent methods for ordinal data, such as for the traditional univariate CA test. The dependence of the CA-type test's power properties on the choice of scores becomes particularly evident when we contrast the results for S1 and S4 with each other. To recap, we have chosen the scores (1, 2, 3) throughout the set, which reflects linearity of the suspected relationship between the covariates in the set and the outcome variable. This is exactly the kind of set-outcome relationship that is true for S1. As Table 4.2 shows, this accurate match between the prespecified scores and the true set-outcome

relationship renders the CA-type test powerful, even slightly more powerful than the score-free test. For S4, in contrast, the CA-type test has basically no power at all. Apparently, this is owing to the fact that here the degree of misspecification of scores is fairly large, since S4 represents an umbrella-like set-outcome relationship. (A side remark: to have power to detect umbrella-like set-outcome relationships, we would have had to choose umbrella-shaped scores such as, for example, $(1, 2, 1)$.) As can be further seen from Table 4.2, an entirely different picture than for the CA-type test is obtained for the score-free test. In particular, our results indicate that the latter has power irrespective of what kind of relationship the covariates in the set exhibit with the outcome variable, and that the power to detect monotonic set-outcome relationships (scenarios S1 and S2) exceeds the power to detect non-monotonic ones (scenarios S3 and S4). This specific behaviour of the score-free test has been confirmed by various further simulation experiments that we conducted on this issue (not shown here).

The simplistic character of the scenarios S1–S4 has helped to illustrate the power properties of the CA-type and the score-free test. Scenarios of this kind are, however, unlikely to be encountered in practice. Especially when the set to be tested comprises many covariates, it appears unrealistic that each of these covariates exhibits the same kind of relationship with the outcome variable. A more realistic scenario is represented by S5 where some of the covariates in the set are monotonically related to the outcome variable, whereas others show a non-monotonic relationship. As could have been expected from the results for S1–S4, here the score-free test has more power than the CA-type test. Nevertheless, the score-free test will not *per se* be the better choice. In particular, the fact that the CA-type test requires correctly (or close to correctly) specified scores to be powerful makes it useful in applications where only a specific type of set-outcome relationship is considered important.

Table 4.2.: Average rejection rates for the simulation scenarios S0–S5 (see Section 4.4.1 for detailed descriptions).

	CA-type test*					Score-free test				
Sample size n	20	40	60	80	100	20	40	60	80	100
S0	0.053	0.051	0.056	0.054	0.049	0.055	0.054	0.056	0.052	0.052
S1	0.194	0.435	0.689	0.861	0.948	0.185	0.415	0.662	0.838	0.937
S2	0.180	0.388	0.632	0.812	0.916	0.185	0.402	0.651	0.828	0.925
S3	0.087	0.147	0.224	0.311	0.393	0.136	0.280	0.476	0.648	0.792
S4	0.056	0.060	0.064	0.073	0.075	0.093	0.143	0.217	0.303	0.408
S5	0.119	0.230	0.380	0.520	0.663	0.147	0.302	0.506	0.683	0.827

* based on the equally-spaced scores $(1, 2, 3)$ throughout the set

4.4.3. *Excursus:* simulating a desired set-outcome relationship

The data generation technique that was used for the simulation study presented above has already been described briefly in Section 4.4.1. This excursus extends the descriptions from Section 4.4.1 by more detailed explanations. To recap, the problem that this excursus particularly addresses is the generation of data sets with a desired set-outcome relationship when the outcome variable, say Y, is binary (i.e. $Y \in \{0,1\}$) and the set to be tested comprises independent ordinally scaled covariates, say $X_1,\ldots,X_p$, each of which has three distinct categories (i.e. $X_1,\ldots,X_p \in \{1,2,3\}$).

As has previously been noted in Section 4.4.1, in the set-up considered, the set-outcome relationship is determined by the binomial proportions

$$b^1_{kv} = \Pr(Y = 1 | X_k = v)$$

and

$$b^0_{kv} = \Pr(Y = 0 | X_k = v),$$

respectively, where $b^1_{kv}, b^0_{kv} \in (0,1)$ and $b^1_{kv} + b^0_{kv} = 1$ for $v = 1,2,3$ and $k = 1,\ldots,p$. For chosen b^1_{kv}s, we are now interested to find the conditional distributions of the X_ks given Y, from which data can be drawn easily. By Bayes' theorem, and under the assumption that $P(Y = 0) = P(Y = 1) = 0.5$, we can write

$$\frac{\Pr(X_k = v | Y = 0)}{\Pr(X_k = v | Y = 1)} = \frac{b^0_{kv}}{b^1_{kv}} = \frac{1 - b^1_{kv}}{b^1_{kv}}.$$

More specifically, we have

$$\frac{\Pr(X_k = 1 | Y = 0)}{\Pr(X_k = 1 | Y = 1)} = \frac{1 - b^1_{k1}}{b^1_{k1}}, \tag{4.9}$$

$$\frac{\Pr(X_k = 2 | Y = 0)}{\Pr(X_k = 2 | Y = 1)} = \frac{1 - b^2_{k2}}{b^1_{k2}}, \tag{4.10}$$

$$\frac{\Pr(X_k = 3 | Y = 0)}{\Pr(X_k = 3 | Y = 1)} = \frac{1 - b^3_{k3}}{b^1_{k3}}, \tag{4.11}$$

and we furthermore know that

$$\sum_{v=1}^{3} \Pr(X_k = v | Y = 0) = 1, \tag{4.12}$$

$$\sum_{v=1}^{3} \Pr(X_k = v | Y = 1) = 1. \tag{4.13}$$

We can see immediately that, based on the information provided by (4.9)–(4.13), it will be impossible to solve the problem, since we have five independent equations but six

unknown parameters, namely $\Pr(X_k = 1|Y = 1)$, $\Pr(X_k = 2|Y = 1)$, $\Pr(X_k = 3|Y = 1)$, $\Pr(X_k = 1|Y = 0)$, $\Pr(X_k = 2|Y = 0)$ and $\Pr(X_k = 3|Y = 0)$. Without loss of generality, let us therefore assume that $\Pr(X_k = 1|Y = 1)$ is fix, and we write $\Pr(X_k = 1|Y = 1) =: f^1_{k1}$. This enables us to solve the system of equations, and we eventually arrive at

$$\Pr(X_k = 2|Y = 1) = \frac{1 - \frac{\left(f^1_{k1} - b^1_{k1} f^1_{k1}\right)}{b^1_{k1}} - \frac{\left(1 - b^1_{k3}\right)}{b^1_{k3}} + \frac{\left(f^1_{k1} - b^1_{k3} f^1_{k1}\right)}{b^1_{k3}}}{\frac{\left(1 - b^1_{k2}\right)}{b^1_{k2}} - \frac{\left(1 - b^1_{k3}\right)}{b^1_{k3}}},$$

$$\Pr(X_k = 3|Y = 1) = 1 - f^1_{k1} - \Pr(X_k = 2|Y = 1),$$

$$\Pr(X_k = 1|Y = 0) = \frac{1 - b^1_{k1}}{b^1_{k1}} f^1_{k1},$$

$$\Pr(X_k = 2|Y = 0) = \frac{1 - b^1_{k2}}{b^1_{k2}} \Pr(X_k = 2|Y = 1),$$

$$\Pr(X_k = 3|Y = 0) = \frac{1 - b^1_{k3}}{b^1_{k3}} \Pr(X_k = 3|Y = 1).$$

For $b^1_{k1} = 0.4$, $b^1_{k2} = 0.5$, $b^1_{k3} = 0.6$ and $f^1_{k1} = 0.\bar{3}$, for example, we obtain $\Pr(X_k = 2|Y = 1) = 0.1\bar{6}$, $\Pr(X_k = 3|Y = 1) = 0.5$, $\Pr(X_k = 1|Y = 0) = 0.5$, $\Pr(X_k = 2|Y = 0) = 0.1\bar{6}$ and $\Pr(X_k = 3|Y = 0) = 0.\bar{3}$. This information can now be used to simulate data sets for which the desired set-outcome relationship can be found. It is important to note, however, that the technique just described has its limitations, since some constellations of b^1_{k1}, b^1_{k2}, b^1_{k3} and f^1_{k1} will yield probabilities outside the interval $(0, 1)$.

4.5. Application 2: functioning and disability in multiple sclerosis

Data set and question of interest

To illustrate the application of the methods presented in this chapter, we analyzed data from the multi-centre cross-sectional study on functional limitations and disabilities in MS of Holper et al. (2010). The study was conducted in one rehabilitation centre in Germany and three rehabilitation centres in Switzerland from 2007 to 2008, and it was based on $p = 129$ ICF items from the extended ICF checklist (listed further below in this section in Table 4.4 which, in addition, provides information on the ICF items' particular tree structure). The recruitment of the individuals involved and the data collection were performed by physicians and other health professionals specifically trained for this purpose in ICF workshops. The considered data set includes $n = 93$ individuals of which 33 were diagnosed with the MS form primary progressive MS (PP MS) and

60 with the MS form secondary progressive MS (SP MS). In brief, PP MS patients suffer from a steady increase in functional limitation and disability without clear attacks, whereas SP MS patients suffer from unpredictable attacks of functional limitation and disability followed by periods of remission, and eventually experience a decline without periods of remission. Aside from ICF item-based information, the considered data set provides disease-related and socio-demographic characteristics on the individuals (see Holper et al. (2010) for complete details). The question of interest was now whether there is an association between MS patients' ICF component-specific profiles and the MS form they suffer from, after adjustment for the effect of age, sex and sum score from the Beck Depression Inventory (BDI) II (Beck et al., 1996). Unlike in the ICF-based application previously presented in Section 3.7, here it was thus not of interest to localize effects further (for instance on the level of ICF chapters). The test result of the ICF-based application presented in this section therefore serves as an example of a user-driven compromise in the sense of Section 2.3.

To answer the above question, the ICF-based data were first preprocessed as described in Section 1.1. In particular, this means that both the five-level ordinal scale of ICF items of the ICF components 'body functions' (b), 'body structures' (s) and 'activities and participation' (d) and the nine-level ordinal scale of ICF items of the ICF component 'environmental factors' (e) were coarsened to three levels (see Figures 1.1 and 1.2), and that the additional answer option 9 (not applicable), which has been observed merely once, was recoded into the answer option 0 (no problem/neither barrier nor facilitator). As has already been noted in Section 1.1, the scale coarsening potentially reduces the number of ICF items for which one or more categories have remained unobserved in the sample and, thereby, facilitates data analysis. With the three-level ordinal scale, our data set still comprises 39 ICF items (14 from b, 4 from s, 12 from d, and 9 from e) for which merely two of the three categories could be observed. Given that in the vast majority of cases (35 of 39) it was the highest category (i.e. 'severe to complete problem' or 'barrier') that has remained unobserved, the respective ICF items were treated as if they had been measured on the same two-level scale.

Methods

To test for association between each of the ICF components and the MS form (coded with 0 for PP MS and 1 for SP MS), merely for the purpose of illustration of differences between the CA-type and the score-free test, we applied them both and contrasted the respective results with each other. In practice, of course, one should decide for one test or the other, which is sensibly done based upon power considerations. We used the CA-type test with the equally-spaced scores (1, 2, 3) throughout, noting that score-dependent methods for ordinal data are commonly used together with equally-

spaced scores, and we adjusted our analysis for age, sex and BDI score. Because we were interested in testing the four ICF components simultaneously, it was necessary to adjust the respective P-values for multiplicity. We did so by means of the Bonferroni-Holm procedure (Holm, 1979). Given that, of the total of 129 ICF items, 34 ICF items belong to the ICF component b, 13 to the ICF component s, 51 to the ICF component d and 31 to the ICF component e, one could alternatively use the multiplicity adjustment rule (2.1) which respects the fact that the ICF components are of different size. Here, however, we preferred to treat the ICF components on the same footing. The results obtained are discussed below.

Results

Table 4.3 displays the Bonferroni-Holm adjusted P-values for the ICF components b, s, d and e, obtained with the CA-type and the score-free test for the logit model. At the standard level of significance $\alpha = 0.05$, the CA-type and the score-free test lead to the same inferential conclusions for b, s and d: while b and d are found to be significantly associated with the MS form, no such association can be revealed for s. It can thus be said that PP MS and SP MS patients differ in their overall pattern of restrictions of body functions as well as activities and participation. When it comes to the ICF component e, the CA-type test clearly maintains the null hypothesis of no association with the MS form, whereas the score-free test rejects it. Recalling the simulation results on power from Section 4.4.2, this may indicate that the ICF component e comprises ICF items that exhibit a non-monotonic relationship with the MS form. Figure 4.1 helps to clarify whether this is the case: it shows the ICF item-specific contributions to the test statistics $\hat{S}_{\text{CA}}$ (left panel) and $\hat{S}_{\text{SF}}$ (right panel) for the entire set e. If now an ICF item is non-monotonically related to the MS form, its influence on $\hat{S}_{\text{CA}}$ is likely to be smaller compared to its influence on $\hat{S}_{\text{SF}}$. Among the 31 ICF items included in e, it becomes readily visible from the figure that this is particularly true for the ICF item 'light' (e240). A look into the data in fact suggests the presence of a non-monotonic relationship: the estimated binomial proportions across the categories of the ICF item e240 are 0.21, 0.54 and 0.24 for PP MS patients and, consequently, 0.79, 0.46 and 0.76 for SP MS patients. The fact that this is fairly close to an umbrella-like relationship explains why the influence of e240 on $\hat{S}_{\text{CA}}$ is considerably less pronounced than on $\hat{S}_{\text{SF}}$. Noting that non-monotonic relationships seem to be present for 17 further ICF items in e (see hatched bars in Figure 4.1), and that numerous of these ICF items belong to the most influential ones in the set, it is of little surprise that here our two tests have lead to different inferential conclusions. In this context, the ICF item 'climate' (e225) deserves particular mention. On the one hand, it is the ICF item that contributes most to the CA-type test statistic. On the other hand, we find from Figure 4.1 that it belongs

to those ICF items for which the data suggest a non-monotonic relationship with the MS form. At first sight, this may be somewhat counterintuitive. A closer look into the data provides clarification: the estimated binomial proportions across the categories of the ICF item e225 are 0.41, 0.58 and 0.22 for PP MS patients and, consequently, 0.59, 0.42 and 0.78 for SP MS patients. This reflects a strongly asymmetric umbrella-like relationship, which explains the considerable influence of e225 on $\hat{S}_{\mathrm{CA}}$.

Differences between PP MS and SP MS patients with respect to functional limitations and disabilities in the course of the disease have previously been reported in the medical literature (A. Thompson, 2004; Amato et al., 2006). On the basis of individual ICF items, however, the presence of such differences could so far not be confirmed; merely some descriptive observations in that direction were made (Holper et al., 2010). In contrast to that, our results show that, on the basis of ICF components, proper statistical evidence in favour of the phenomenology communicated in the medical literature can be provided. This well exemplifies the potential practical benefit of the tests developed in this chapter.

As an additional but rather informal step, we performed the CA-type and the score-free test separately for each ICF item, even though the classical univariate scenario is not that by which the tests' development has been motivated. For comparison, we performed ICF item-specific LR tests, based on both the CA-type and the score-free approach to handling ordinality. As with the analysis of the ICF components, we adjusted for age, sex and BDI score. After Bonferroni-Holm correction of the ICF item-specific P-values, we find that for none of the 129 ICF items a statistically significant effect can be detected, irrespective of which of the four tests is being used. Our univariate results are thus in line with the earlier mentioned univariate results of Holper et al. (2010). Furthermore, they tell us that, if we had attempted to exploit the ICF items' tree structure by means of Meinshausen's procedure (see Section 2.3), the tree level of individual ICF items would not have been reached.

Table 4.3.: Multiplicity-adjusted P-values via Bonferroni-Holm for the ICF components b, s, d and e.

	CA-type test	Score-free test
Body functions (b)	0.030	0.021
Body structures (s)	0.345	0.328
Activities and participation (d)	0.049	0.023
Environmental factors (e)	0.145	0.039

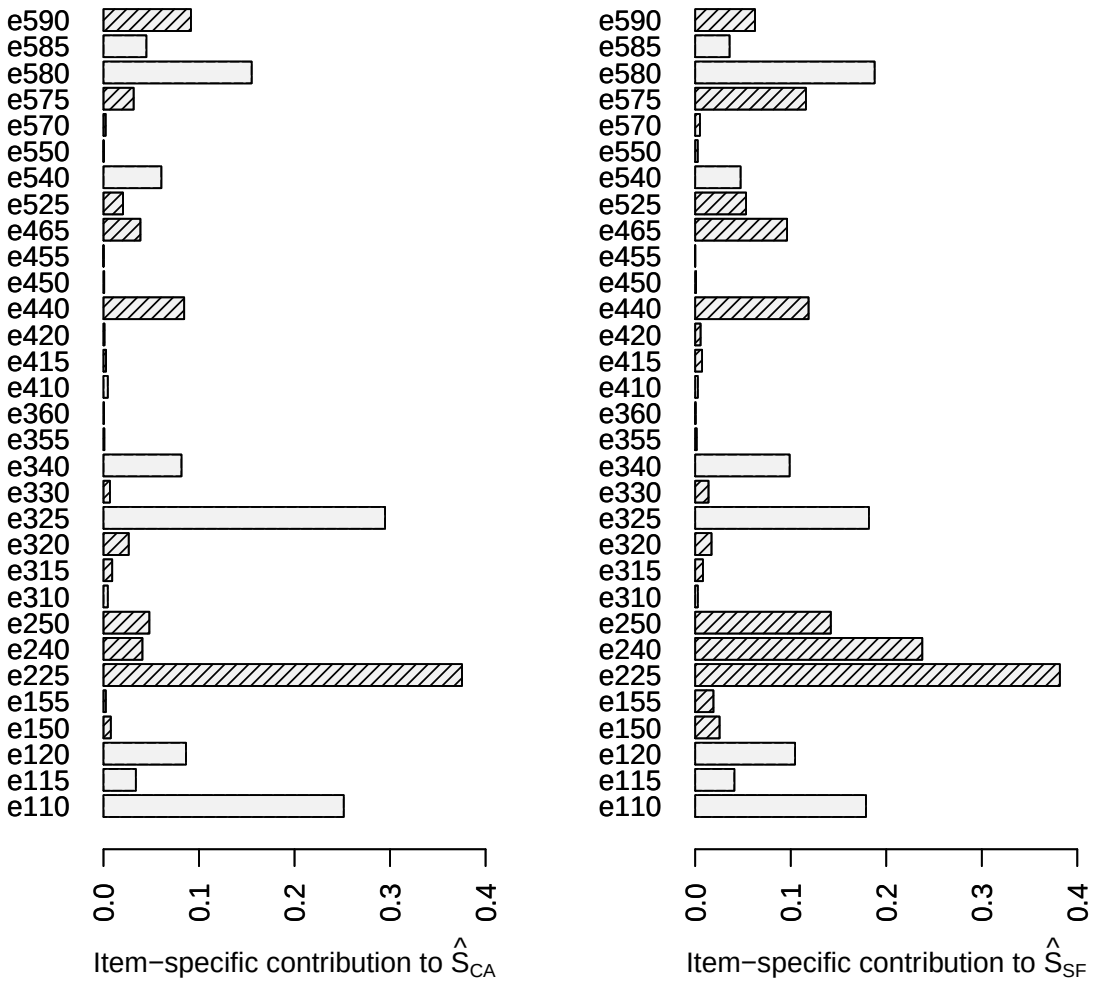

Figure 4.1.: ICF item-specific contributions to the CA-type test statistic $\hat{S}_{CA}$ and the score-free test statistic $\hat{S}_{SF}$ for the ICF component e. Hatched bars belong to those ICF items for which the data suggest a non-monotonic relationship with the MS form.

Table 4.4.: List of the 129 ICF items that have been considered in the MS study, together with information on which ICF component and ICF chapter each item belongs to. The respective ICF code is given in brackets.

Body functions (b) (comprises 34 ICF items)

Mental functions (b1)
- Orientation functions (b114)
- Intellectual functions (b117)
- Temperament and personality functions (b126)
- Energy and drive functions (b130)
- Sleep functions (b134)
- Attention functions (b140)
- Memory functions (b144)
- Psychomotor functions (b147)
- Emotional functions (b152)
- Perceptual functions (b156)
- Thought functions (b160)
- Higher-level cognitive functions (b164)
- Mental functions of language (b167)

Sensory functions and pain (b2)
- Seeing functions (b210)
- Hearing functions (b230)
- Vestibular functions (b235)
- Sensation of pain (b280)

Voice and speech functions (b3)
- Voice functions (b310)
- Articulation functions (b320)
- Fluency and rhythm of speech functions (b330)

Functions of the cardiovascular, haematological, immunological and respiratory systems (b4)
- Exercise tolerance functions (b455)

Functions of the digestive, metabolic and endocrine systems (b5)
- Digestive functions (b515)
- Defecation functions (b525)
- Weight maintenance functions (b530)
- Sensations associated with the digestive system (b535)
- Thermoregulatory functions (b550)

Genitourinary and reproductive functions (b6)
- Urination functions (b620)
- Sexual functions (b640)

Neuromusculoskeletal and movement-related functions (b7)
- Mobility of joint functions(b710)
- Muscle power functions (b730)
- Muscle tone functions (b735)
- Involuntary movement functions (b765)
- Gait pattern functions (b770)
- Sensations related to muscles and movement functions (b780)

Body structures (s) (comprises 13 ICF items)

Structures of the nervous system (s1)
- Structure of brain (s110)
- Spinal cord and related structures (s120)

*Structures of eye, ear and related structures (s2)**

*Structures involved in voice and speech (s3)**

*Structures related to the digestive, metabolic and endocrine system (s5)**

Structures related to the genitourinary and reproductive systems (s6)
- Structure of urinary system (s610)
- Structure of reproductive system (s630)

Structures related to movement (s7)
- Structure of head and neck region (s710)
- Structure of shoulder region (s720)
- Structure of upper extremity (s730)
- Structure of pelvic region (s740)
- Structure of lower extremity (s750)
- Structure of trunk (s760)

Note: For the ICF chapters s2, s3 and s5 (marked with *) merely an overall assessment of the individuals is available, measured on the same five-level ordinal scale that is used for ICF items in the ICF components b, s and d. For simplicity, s2, s3 and s5 are henceforth treated as ICF chapters that include only one item.

Activities and participation (d) (comprises 51 ICF items)

Learning and applying knowledge (d1)
- Watching (d110)
- Acquiring skills (d155)
- Thinking (d163)
- Reading (d166)
- Writing (d170)
- Solving problems (d175)
- Making decisions (d177)

General tasks and demands (d2)
- Undertaking a single task (d210)
- Undertaking multiple tasks (d220)
- Carrying out daily routine (d230)

Communication (d3)
- Speaking (d330)
- Conversation (d350)

Mobility (d4)
- Transferring oneself (d420)
- Lifting and carrying objects (d430)
- Fine hand use (d440)
- Hand and arm use (d445)
- Walking (d450)
- Moving around (d455)
- Moving around in different locations (d460)
- Moving around using equipment (d465)
- Using transportation (d470)
- Driving (d475)

Self-care (d5)
- Washing oneself (d510)
- Caring for body parts (d520)
- Toileting (d530)
- Dressing (d540)
- Eating (d550)
- Drinking (560)
- Looking after one's health (d570)

Domestic life (d6)
- Acquisition of goods and services (d620)
- Preparing meals (d630)
- Doing housework (d640)
- Caring for household objects (d650)
- Assisting others (d660)

Interpersonal interactions and relationships (d7)
- Basic interpersonal interactions (d710)
- Complex interpersonal interactions (d720)
- Relating with strangers (d730)
- Formal relationships (d740)
- Informal social relationships (d750)
- Family relationships (d760)
- Intimate relationships (d770)

Major life areas (d8)
- Vocational training (d825)
- Higher education (d830)
- Remunerative employment (d850)
- Basic economic transactions (d860)
- Complex economic transactions (d865)
- Economic self-sufficiency (d870)

Community, social and civic life (d9)
- Community life (d910)
- Recreation and leisure (d920)
- Religion and spirituality (d930)
- Human rights (d940)

Environmental factors (e) (comprises 31 ICF items)

Products and technology (e1)
- Products or substances for personal consumption (e110)
- Products and technology for personal use in daily living (e115)
- Products and technology for personal indoor and outdoor mobility and transportation (e120)
- Design, construction and building products and technology of buildings for public use (e150)
- Design, construction and building products and technology of buildings for private use (e155)

Natural environment and human-made changes to environment (e2)
- Climate (e225)
- Light (e240)
- Sound (e250)

Support and relationships (e3)
- Immediate family (e310)
- Extended family (e315)
- Friends (e320)
- Acquaintances, peers, colleagues, neighbours and community members (e325)
- People in positions of authority (e330)
- Personal care providers and personal assistants (e340)
- Health professionals (e355)
- Health-related professionals (e360)

Attitudes (e4)
- Individual attitudes of immediate family members (e410)
- Individual attitudes of extended family members (e415)
- Individual attitudes of friends (e420)
- Individual attitudes of personal care providers and personal assistants (e440)
- Individual attitudes of health professionals (e450)

Individual attitudes of health-related professionals (e455)
Social norms, practices and ideologies (e465)

Services, systems and policies (e5)
Housing services, systems and policies (e525)
Transportation services, systems and policies (e540)
Legal services, systems and policies (e550)
Social security services, systems and policies (e570)
General social support services, systems and policies (e575)
Health services, systems and policies (e580)
Education and training services, systems and policies (e585)
Labour and employment services, systems and policies (e590)

4.6. Discussion

In this chapter we have developed two tests that enable researchers to assess the presence of an association between a set of ordinal covariates and an outcome variable within the range of GLMs, mainly motivated by the need for statistical tools to analyze data that have been collected by means of the WHO's ICF. Feasibility independent of the dimensionality of the alternative hypothesis, validity under any correlation and the possibility of covariate adjustment render the tests widely useful in practice. Our first test, the score-based CA-type test, expects the user to make assumptions on the distances between the covariates' categories, and its power is then directed towards the set-outcome relationship that is in line with these assumptions. Under mild conditions, we have shown that this test is a proper generalization of the traditional CA test to higher dimensions, covariate-adjusted scenarios and GLM-specific outcomes. Our second test, the score-free test, respects the ordering of the covariates' categories while dispensing with assumptions on the distances between them, and its power is spread over a wide range of possible set-outcome relationships, with more emphasis put on monotonic than on non-monotonic ones. In practice, whether to employ the CA-type or the score-free test depends on whether some specific alternative or many different alternatives are considered important, such that recommendations can only be made with reference to concrete applications.

One scenario outside the ICF context where the score-free test promises to be more appropriate than the CA-type test is when sets of SNPs in genetic association studies of complex diseases are to be tested. To test individual SNPs in case-control situations, it is common practice to use the CA test for trend, where the scores are chosen such that they reflect the underlying genetic model (Freidlin et al., 2002; Balding, 2006). To test sets of SNPs, the SNP-specific CA test statistics are often combined into one test statistic for the entire set, and critical values are obtained via some resampling proce-

dure (Balding, 2006; Hoh and Ott, 2003). This popularity of the CA test for the analysis of SNP data speaks for the usefulness of the CA-type and, as a special variant, the generalized CA test in this context. For complex diseases, however, the genetic model is typically unknown, and the choice of scores hence unclear. To overcome this issue, one can perform separate tests for each genetic model and then build some weighted average of the respective results. As pointed out by Balding (2006), it will mostly be sensible to choose the weights such that greater plausibility of the additive model is reflected but that the resultant test still has power to detect effects that are far from additive. The fact that this corresponds to the power properties of the score-free test without that subjectivity comes into play, since no weights need to be specified, argues for future explorations of this test in the context of SNP set analyses.

Although standard univariate problems are not those which our tests have originally been intended for, it is important to emphasize that the latter may be valuable in such situations as well. It should be kept in mind, however, that the tests proposed are score tests and as such only have optimal average power when the departure from the null hypothesis is small, that is, when the effect of the covariate considered is weak.

Global tests for sets of nominal covariates have not been considered in this chapter. Besides the CA-type and the score-free test, however, the `R` package `globaltest` likewise implements a global test that is tailored to covariates measured on a nominal scale. Application of this test to sets of ordinal covariates can be sensible, yet only in instances where monotonic and non-monotonic set-outcome relationships are considered equally important. In such instances, provided that the outcome variable is binary, it may be worthwhile to compare the performance of the just mentioned test for sets of nominal covariates with that of the χ^2 sum statistic-based permutation test from Chapter 3, which likewise assumes the data to be nominally scaled.

Finally, the tests proposed are not only useful by themselves but, in addition, can be fruitfully combined with multiplicity adjustment procedures for, for example, hypotheses that can be structured in a directed acyclic graph (Goeman and Mansmann, 2008) or in a tree by some expert knowledge (Meinshausen, 2008; Goeman and Solari, 2010; Goeman and Finos, 2012). This is particularly relevant for the analysis of ICF-based data because here such prior knowledge is on hand. However, both the CA-type and the score-free test are limited with respect to the outcome variables that they can handle. Future research problems, whether posed within or outside the ICF context, will therefore call for extensions of both tests for more complex models, such as for the cumulative logit model for ordinally scaled outcomes.

5. Contributions, limitations and key conclusions of this thesis

This chapter looks at the present thesis from different yet related perspectives: from the statistical and methodical perspective, the historical perspective, the ICF perspective and, eventually, from an overall perspective. Separately for each perspective, the relevant contributions of this thesis to the literature are briefly reviewed. For clarity, it is merely the most important contributions that are pointed out.

Statistical and methodical perspective

Seen from the statistical and methodical perspective, this thesis has contributed various tests of association which can be used for the analysis of sets of multivariate ordinal variables in possibly high-dimensional data situations. The first test may be understood as a permutation-based generalization of the two-sample χ^2-test of homogeneity to higher dimensions, and the second test as a permutation-based generalization of the traditional two-sided CA test to higher dimensions (see Chapter 3). In this context, it has furthermore been shown that the recently proposed two-sample test of Klingenberg et al. (2009), which the second test is based on, is a permutation-based generalization of the one-sided CA test to higher dimensions. This interpretation is useful, since it justifies and motivates the use of the test of Klingenberg et al. (2009) in many situations in which it has not been considered relevant so far. The third and the fourth test both are extensions of the 'global test' for groups of genes of Goeman et al. (2004, 2006, 2011) and can be used to test for association, if desired after adjustment for certain covariates, between a set of ordinally scaled covariates and an outcome variable within the range of GLMs (see Chapter 4). Under mild conditions, the third test has been shown to be a generalization of the traditional CA test to higher dimensions, covariate-adjusted scenarios and GLM-specific outcomes. As such, it includes the second test as a permutation-based special case. This is remarkable, in light of the fact that the two tests have been derived within fundamentally different frameworks. One concern with the above three tests may now be that, in principle, they are not ideally suited for ordinal data. More specifically, while the first test respects the fact that the distances between the ordinal variables' categories are unknown but not the fact that they are

ordered, the opposite is the case for the second and the third test. At first sight, this may not appear satisfactory. From the application viewpoint, however, it may sometimes turn out to be beneficial. The crux is that the way in which ordinality is handled determines the power properties of the resultant test, and because the desired power properties may vary between applications, it is worthwhile to have tests that have been developed under different assumptions. For the second and the third test, there is one further reason why they promise to be useful for set-based analyses of ordinal data: they have been shown to be generalizations of the CA test, which is the most widely used univariate test for ordinal variables. The fourth and last test proposed now sets itself apart from the other ones in that it really takes ordinality into account. From the available literature, we are not aware of any other global test that is ideally tailored to ordinal data. Future research may include extensions of this test for more complex models outside the GLM family, and explorations of its usefulness beyond ICF-based applications by which this work has primarily been motivated.

As one further contribution, this thesis has evaluated, by means of extensive simulations, permutation tests' behaviour under theoretically unfavourable conditions, on the basis of the two-sample permutation tests from Chapter 3. To recap, here theoretically unfavourable means that, under the null hypothesis, the joint distribution of the variables in the set to be tested may differ between both groups; in such scenarios, permutation-based inference will not be valid. The key conclusion that has been drawn from the simulation study is that theoretically invalid permutation tests can still be 'practically valid', and it has been illustrated that the degree of their failure can be considered as a function of numerous parameters, such as for example the proportion between group sizes, the number of variables in the set of interest and, importantly, the test statistic used. From the literature, we are not aware of any other simulation study in this context that has provided comparably insightful information. We believe in fact that systematic studies such as that presented in this thesis can help to establish some useful practical recommendations with respect to the use of permutation tests in scenarios where their theoretical validity is unrealistic but no inference method does exist that is superior to the permutation procedure. Such practical recommendations in turn can help researchers to draw correct conclusions from their permutation-based analyses. One limitation of our study which should be mentioned is that it has merely focused on scenarios with a non-negative uniform correlation structure between variables. This calls for future studies that cover more flexible correlation structures.

Historical perspective

When we adopt the historical perspective on this thesis, there are two points that deserve particular mention. Firstly, the test statistics proposed (i.e. the test statistics

(3.7), (3.8), (4.5) and (4.7)) are all sums of variable-specific test statistics over the whole set to be tested. In contrast to what is widely assumed, however, tests based on such sum statistics, or briefly sum tests, are not a product of the gene set analysis era where practical approaches have been needed to construct tests for sets of genes which are feasible whatever the dimensionality of the research problem is. Rather, they were first introduced far before this era by Chung and Fraser (1958), and several decades later rediscovered for the analysis of microarray-based gene expression data. In the so far hardly explored context of multivariate ordinal data, the tests proposed in this thesis thus connect with an established tradition in hypothesis test construction. Secondly, this thesis has shown initially unexpected connections between some of the tests discussed and the CA test (see 'Statistical and methodical perspective' for more detailed information). The latter was introduced by Cochran in 1954 and, independently thereof, by Armitage in 1955; this thesis has thus connected traditional and contemporary statistical research on methods for ordinally scaled data.

ICF perspective

Seen from the ICF perspective, the present thesis has enabled researchers to perform statistical analyses of ICF-based data in a way that incorporates the available prior knowledge on ICF items' structure. To obtain an impression of the practical benefit of such knowledge-based analyses when compared with standard univariate analyses, which do not make any inferential use of external information, two concrete ICF-based applications have been presented (see Sections 3.7 and 4.5). In both instances the knowledge-based analysis has provided more insight into the phenomenon under study than the conventional univariate analysis. The tests developed in this thesis, combined with appropriate multiplicity adjustment procedures, thus promise to be useful new tools for the statistical analysis of ICF-based data, and in particular in ICF-based applications where the number of ICF items considered exceeds the number of subjects in the sample. Whether these tools can ultimately establish themselves in the ICF community will depend on numerous factors, and certainly require intensive discussion between ICF-oriented researchers and statisticians. Irrespective thereof, as the ICF is more and more used world-wide to collect data on human functioning and disability, the need for statistical tools to answer the research questions that arise from such data will continue to rise.

Overall perspective

The use of prior knowledge, if available, to improve the statistical analysis of high-dimensional data has been amply discussed in the past decade's literature. The de-

velopment of relevant statistical methods, however, has mainly focused on metrically scaled data. Seen from an overall perspective, this thesis has attempted to illustrate that the potential practical benefit of knowledge-based analyses likewise holds for research problems on the ordinal scale and, for this purpose, has developed various tests of association for predefined sets of multivariate ordinal variables in potentially high-dimensional set-ups. While the analysis of ICF-based data has been the primary motivation for this work, the same methods can be used to analyze any other type of possibly high-dimensional multivariate ordinal data that can be structured into sets by external information. Examples include realizations of items in psychodiagnostic tests (e.g. structured into sets by the subdimension that they describe), side or adverse effects in drug safety or toxicity studies (e.g. structured into sets by the body function that they affect) and SNPs in next-generation sequencing studies (e.g. structured into sets by genes). As the inferential exploitment of prior information becomes popular in more and more fields of application, however, and ordinally scaled data arise in almost any of them, future research problems will call for extensions of the tests proposed to more complex scenarios. This includes but is not exhausted by scenarios where it is of interest to test, after adjustment for potential confounders, whether there is an association between a set of ordinal variables and another ordinal variable.

A. Simulation results in detail

This appendix provides the detailed results from the simulation studies described in Sections 3.5 and 3.6. In particular, Tables A.1–A.6 refer to the results obtained for the permutation null distribution (see Section 3.5), whereas Tables A.7–A.12 refer to the results obtained for the bootstrap null distribution (see Section 3.6). Tables A.1–A.4 display the simulation results for the permutation null distribution in the case $c = 4$, and Tables A.5–A.6 display the follow-up simulation results in the case $c = 2$. Tables A.7–A.10 show the simulation results for the bootstrap null distribution in the case $c = 4$, and Tables A.11–A.12 show the follow-up simulation results in the case $c = 2$. Because, in the case $c = 2$, the sum statistics Q_{χ^2} from (3.7) and Q_{CA} from (3.8) as well as their max-T-counterparts are equivalent, the respective results are reported only once.

Table A.1.: Actual type I error rate with the permutation null distribution of the χ^2 sum statistic Q_{χ^2} in the case $\mathbf{c} = \mathbf{4}$. Simulation margin of error for $\alpha = 0.05$: ± 0.0138. See Figure 3.1 in Section 3.5.2 for a visual representation.

	A: $p = 20$											
	(n_1, n_2)											
(ρ_1, ρ_2)	(10,10)	(20,20)	(30,30)	(40,40)	(8,12)	(16,24)	(26,34)	(32,48)	(5,15)	(12,28)	(18,42)	(24,56)
(0,0)	0.056	0.050	0.044	0.052	0.049	0.047	0.042	0.051	0.039	0.042	0.061	0.047
(0.25,0.25)	0.042	0.048	0.053	0.054	0.056	0.048	0.057	0.048	0.051	0.040	0.047	0.052
(0.5,0.5)	0.048	0.065	0.049	0.052	0.054	0.046	0.061	0.053	0.053	0.044	0.048	0.048
(0.75,0.75)	0.049	0.049	0.048	0.043	0.051	0.044	0.063	0.054	0.050	0.054	0.049	0.045
(0,0.25)	0.054	0.051	0.058	0.049	0.044	0.046	0.047	0.048	0.039	0.038	0.037	0.039
(0.25,0.5)	0.050	0.058	0.051	0.050	0.052	0.042	0.054	0.043	0.040	0.030	0.034	0.035
(0.5,0.75)	0.051	0.052	0.054	0.045	0.050	0.039	0.056	0.043	0.046	0.034	0.043	0.034
(0,0.5)	0.059	0.054	0.053	0.054	0.045	0.042	0.043	0.038	0.023	0.032	0.024	0.025
(0.25,0.75)	0.054	0.056	0.056	0.051	0.044	0.030	0.047	0.033	0.038	0.022	0.024	0.028
(0,0.75)	0.068	0.044	0.058	0.060	0.043	0.025	0.037	0.030	0.026	0.018	0.011	0.011
(0.25,0)	0.054	0.051	0.058	0.049	0.056	0.053	0.059	0.051	0.064	0.048	0.062	0.066
(0.5,0.25)	0.050	0.058	0.051	0.050	0.071	0.053	0.059	0.063	0.078	0.061	0.066	0.065
(0.75,0.5)	0.051	0.052	0.054	0.045	0.054	0.053	0.070	0.065	0.069	0.072	0.065	0.075
(0.5,0)	0.059	0.054	0.053	0.054	0.065	0.055	0.064	0.063	0.064	0.079	0.098	0.080
(0.75,0.25)	0.054	0.056	0.056	0.051	0.067	0.074	0.081	0.068	0.095	0.092	0.089	0.088
(0.75,0)	0.068	0.044	0.058	0.060	0.080	0.081	0.078	0.078	0.122	0.118	0.107	0.111
	B: $p = 100$											
	(n_1, n_2)											
(ρ_1, ρ_2)	(10,10)	(20,20)	(30,30)	(40,40)	(8,12)	(16,24)	(26,34)	(32,48)	(5,15)	(12,28)	(18,42)	(24,56)
(0,0)	0.043	0.051	0.044	0.050	0.038	0.046	0.054	0.048	0.051	0.045	0.037	0.038
(0.25,0.25)	0.042	0.064	0.052	0.038	0.053	0.052	0.041	0.043	0.059	0.052	0.051	0.047
(0.5,0.5)	0.046	0.059	0.057	0.050	0.048	0.054	0.050	0.050	0.055	0.051	0.051	0.048
(0.75,0.75)	0.047	0.049	0.055	0.048	0.046	0.048	0.047	0.043	0.059	0.051	0.058	0.048
(0,0.25)	0.051	0.049	0.045	0.042	0.040	0.037	0.034	0.029	0.031	0.019	0.021	0.022
(0.25,0.5)	0.051	0.064	0.057	0.045	0.039	0.040	0.039	0.039	0.040	0.037	0.034	0.029
(0.5,0.75)	0.051	0.050	0.054	0.044	0.040	0.042	0.042	0.045	0.046	0.038	0.034	0.031
(0,0.5)	0.071	0.055	0.061	0.040	0.034	0.031	0.030	0.036	0.026	0.015	0.011	0.011
(0.25,0.75)	0.060	0.054	0.058	0.040	0.034	0.035	0.034	0.035	0.034	0.026	0.018	0.020
(0,0.75)	0.075	0.057	0.059	0.045	0.039	0.022	0.031	0.037	0.028	0.012	0.007	0.005
(0.25,0)	0.051	0.049	0.045	0.042	0.067	0.071	0.060	0.066	0.092	0.081	0.084	0.079
(0.5,0.25)	0.051	0.064	0.057	0.045	0.065	0.073	0.053	0.057	0.089	0.082	0.082	0.066
(0.75,0.5)	0.051	0.050	0.054	0.044	0.059	0.064	0.048	0.053	0.078	0.076	0.075	0.063
(0.5,0)	0.071	0.055	0.061	0.040	0.103	0.103	0.070	0.094	0.157	0.134	0.129	0.126
(0.75,0.25)	0.060	0.054	0.058	0.040	0.072	0.091	0.056	0.066	0.128	0.104	0.110	0.088
(0.75,0)	0.075	0.057	0.059	0.045	0.108	0.114	0.073	0.096	0.200	0.152	0.146	0.133

Table A.2.: Actual type I error rate with the permutation null distribution of the CA sum statistic Q_{CA} in the case $\mathbf{c} = \mathbf{4}$. Simulation margin of error for $\alpha = 0.05$: ± 0.0138. See Figure 3.2 in Section 3.5.2 for a visual representation.

	A: $p = 20$											
	(n_1, n_2)											
(ρ_1, ρ_2)	(10,10)	(20,20)	(30,30)	(40,40)	(8,12)	(16,24)	(26,34)	(32,48)	(5,15)	(12,28)	(18,42)	(24,56)
(0,0)	0.049	0.041	0.051	0.060	0.051	0.049	0.050	0.046	0.040	0.051	0.052	0.043
(0.25,0.25)	0.053	0.049	0.056	0.046	0.048	0.047	0.055	0.045	0.049	0.051	0.058	0.054
(0.5,0.5)	0.051	0.055	0.043	0.052	0.054	0.043	0.059	0.049	0.047	0.046	0.053	0.053
(0.75,0.75)	0.058	0.050	0.047	0.045	0.049	0.046	0.069	0.051	0.040	0.048	0.055	0.057
(0,0.25)	0.050	0.051	0.054	0.047	0.040	0.041	0.052	0.039	0.024	0.035	0.034	0.032
(0.25,0.5)	0.060	0.052	0.048	0.051	0.045	0.037	0.051	0.038	0.028	0.030	0.034	0.037
(0.5,0.75)	0.054	0.048	0.046	0.048	0.045	0.035	0.060	0.045	0.030	0.038	0.037	0.038
(0,0.5)	0.059	0.058	0.054	0.061	0.042	0.029	0.039	0.031	0.016	0.017	0.013	0.017
(0.25,0.75)	0.059	0.053	0.042	0.048	0.038	0.032	0.047	0.030	0.022	0.028	0.020	0.030
(0,0.75)	0.059	0.054	0.054	0.061	0.038	0.030	0.039	0.032	0.014	0.014	0.009	0.015
(0.25,0)	0.050	0.051	0.054	0.047	0.056	0.050	0.060	0.059	0.078	0.076	0.075	0.071
(0.5,0.25)	0.060	0.052	0.048	0.051	0.066	0.052	0.069	0.061	0.079	0.067	0.074	0.067
(0.75,0.5)	0.054	0.048	0.046	0.048	0.055	0.054	0.071	0.061	0.066	0.063	0.071	0.062
(0.5,0)	0.059	0.058	0.054	0.061	0.082	0.067	0.073	0.069	0.118	0.114	0.098	0.090
(0.75,0.25)	0.059	0.053	0.042	0.048	0.070	0.060	0.072	0.067	0.102	0.087	0.083	0.079
(0.75,0)	0.059	0.054	0.054	0.061	0.095	0.082	0.079	0.074	0.153	0.129	0.118	0.109
	B: $p = 100$											
	(n_1, n_2)											
(ρ_1, ρ_2)	(10,10)	(20,20)	(30,30)	(40,40)	(8,12)	(16,24)	(26,34)	(32,48)	(5,15)	(12,28)	(18,42)	(24,56)
(0,0)	0.044	0.063	0.051	0.054	0.046	0.066	0.049	0.054	0.039	0.064	0.046	0.045
(0.25,0.25)	0.049	0.057	0.053	0.050	0.047	0.053	0.046	0.048	0.065	0.048	0.040	0.044
(0.5,0.5)	0.047	0.052	0.061	0.048	0.046	0.054	0.049	0.055	0.062	0.053	0.046	0.049
(0.75,0.75)	0.050	0.054	0.064	0.046	0.046	0.053	0.051	0.056	0.060	0.051	0.046	0.050
(0,0.25)	0.066	0.060	0.057	0.040	0.028	0.032	0.026	0.028	0.023	0.014	0.014	0.012
(0.25,0.5)	0.047	0.061	0.057	0.045	0.037	0.039	0.041	0.039	0.038	0.031	0.030	0.032
(0.5,0.75)	0.049	0.062	0.062	0.046	0.040	0.047	0.048	0.048	0.043	0.036	0.035	0.036
(0,0.5)	0.059	0.057	0.055	0.035	0.027	0.025	0.025	0.033	0.018	0.010	0.005	0.008
(0.25,0.75)	0.055	0.061	0.056	0.046	0.033	0.032	0.035	0.037	0.032	0.023	0.020	0.025
(0,0.75)	0.061	0.059	0.056	0.034	0.026	0.027	0.025	0.036	0.015	0.009	0.006	0.007
(0.25,0)	0.066	0.060	0.057	0.040	0.094	0.093	0.074	0.083	0.138	0.116	0.103	0.115
(0.5,0.25)	0.047	0.061	0.057	0.045	0.062	0.071	0.053	0.063	0.114	0.080	0.080	0.074
(0.75,0.5)	0.049	0.062	0.062	0.046	0.057	0.066	0.051	0.060	0.088	0.068	0.065	0.063
(0.5,0)	0.059	0.057	0.055	0.035	0.106	0.099	0.074	0.096	0.183	0.143	0.131	0.146
(0.75,0.25)	0.055	0.061	0.056	0.046	0.069	0.083	0.058	0.068	0.133	0.096	0.093	0.092
(0.75,0)	0.061	0.059	0.056	0.034	0.103	0.100	0.081	0.093	0.190	0.150	0.135	0.161

Table A.3.: Actual type I error rate with the permutation null distribution of max-T (χ^2) in the case $\mathbf{c} = \mathbf{4}$. Simulation margin of error for $\alpha = 0.05$: ± 0.0138. See Figure 3.3 in Section 3.5.2 for a visual representation.

	A: $p = 20$											
	(n_1, n_2)											
(ρ_1, ρ_2)	(10,10)	(20,20)	(30,30)	(40,40)	(8,12)	(16,24)	(26,34)	(32,48)	(5,15)	(12,28)	(18,42)	(24,56)
(0,0)	0.055	0.058	0.059	0.041	0.044	0.054	0.057	0.052	0.044	0.047	0.054	0.052
(0.25,0.25)	0.049	0.050	0.056	0.045	0.045	0.052	0.052	0.056	0.050	0.049	0.050	0.047
(0.5,0.5)	0.043	0.049	0.052	0.046	0.039	0.043	0.051	0.049	0.048	0.054	0.061	0.063
(0.75,0.75)	0.054	0.047	0.048	0.042	0.043	0.043	0.051	0.049	0.059	0.051	0.049	0.052
(0,0.25)	0.056	0.046	0.053	0.051	0.043	0.043	0.044	0.064	0.036	0.049	0.046	0.058
(0.25,0.5)	0.054	0.053	0.067	0.038	0.056	0.064	0.043	0.049	0.047	0.056	0.057	0.049
(0.5,0.75)	0.063	0.048	0.052	0.043	0.059	0.058	0.058	0.060	0.046	0.053	0.052	0.055
(0,0.5)	0.049	0.048	0.053	0.062	0.047	0.056	0.051	0.059	0.044	0.056	0.060	0.052
(0.25,0.75)	0.053	0.048	0.058	0.049	0.052	0.053	0.046	0.059	0.045	0.050	0.042	0.067
(0,0.75)	0.056	0.048	0.066	0.065	0.040	0.059	0.057	0.054	0.050	0.049	0.053	0.045
(0.25,0)	0.056	0.046	0.053	0.051	0.045	0.060	0.046	0.051	0.046	0.045	0.051	0.049
(0.5,0.25)	0.054	0.053	0.067	0.038	0.040	0.049	0.055	0.043	0.047	0.068	0.048	0.046
(0.75,0.5)	0.063	0.048	0.052	0.043	0.053	0.052	0.050	0.049	0.050	0.058	0.056	0.051
(0.5,0)	0.049	0.048	0.053	0.062	0.051	0.057	0.049	0.051	0.048	0.053	0.053	0.055
(0.75,0.25)	0.053	0.048	0.058	0.049	0.054	0.048	0.048	0.043	0.052	0.058	0.045	0.053
(0.75,0)	0.056	0.048	0.066	0.065	0.049	0.070	0.051	0.052	0.043	0.045	0.049	0.045
	B: $p = 100$											
	(n_1, n_2)											
(ρ_1, ρ_2)	(10,10)	(20,20)	(30,30)	(40,40)	(8,12)	(16,24)	(26,34)	(32,48)	(5,15)	(12,28)	(18,42)	(24,56)
(0,0)	0.050	0.064	0.054	0.064	0.061	0.041	0.045	0.043	0.042	0.053	0.050	0.051
(0.25,0.25)	0.035	0.042	0.058	0.055	0.048	0.051	0.042	0.047	0.046	0.050	0.048	0.052
(0.5,0.5)	0.045	0.051	0.034	0.062	0.045	0.053	0.050	0.043	0.039	0.065	0.055	0.051
(0.75,0.75)	0.047	0.054	0.047	0.051	0.058	0.038	0.054	0.035	0.054	0.051	0.048	0.055
(0,0.25)	0.051	0.052	0.045	0.048	0.058	0.049	0.050	0.052	0.049	0.055	0.048	0.045
(0.25,0.5)	0.055	0.049	0.052	0.061	0.044	0.043	0.036	0.051	0.038	0.060	0.051	0.052
(0.5,0.75)	0.058	0.065	0.040	0.045	0.052	0.054	0.050	0.049	0.072	0.061	0.049	0.055
(0,0.5)	0.055	0.050	0.055	0.049	0.046	0.041	0.053	0.046	0.042	0.049	0.045	0.054
(0.25,0.75)	0.055	0.056	0.052	0.047	0.046	0.043	0.056	0.063	0.068	0.050	0.052	0.056
(0,0.75)	0.058	0.055	0.043	0.041	0.046	0.049	0.054	0.046	0.059	0.055	0.056	0.061
(0.25,0)	0.051	0.052	0.045	0.048	0.054	0.048	0.053	0.060	0.048	0.050	0.047	0.043
(0.5,0.25)	0.055	0.049	0.052	0.061	0.042	0.054	0.054	0.041	0.043	0.060	0.042	0.045
(0.75,0.5),	0.058	0.065	0.040	0.045	0.048	0.050	0.045	0.041	0.038	0.047	0.056	0.046
(0.5,0)	0.055	0.050	0.055	0.049	0.059	0.064	0.052	0.059	0.048	0.059	0.041	0.042
(0.75,0.25)	0.055	0.056	0.052	0.047	0.051	0.048	0.045	0.043	0.040	0.045	0.042	0.048
(0.75,0)	0.058	0.055	0.043	0.041	0.057	0.056	0.058	0.052	0.035	0.043	0.050	0.040

Table A.4.: Actual type I error rate with the permutation null distribution of max-T (CA) in the case $\mathbf{c} = \mathbf{4}$. Simulation margin of error for $\alpha = 0.05$: ± 0.0138. See Figure 3.4 in Section 3.5.2 for a visual representation.

	A: $p = 20$											
	(n_1, n_2)											
(ρ_1, ρ_2)	(10,10)	(20,20)	(30,30)	(40,40)	(8,12)	(16,24)	(26,34)	(32,48)	(5,15)	(12,28)	(18,42)	(24,56)
(0,0)	0.051	0.050	0.052	0.040	0.052	0.053	0.054	0.057	0.046	0.047	0.039	0.041
(0.25,0.25)	0.052	0.044	0.058	0.041	0.053	0.051	0.055	0.050	0.052	0.051	0.052	0.052
(0.5,0.5)	0.043	0.049	0.050	0.050	0.053	0.050	0.053	0.059	0.052	0.052	0.053	0.052
(0.75,0.75)	0.054	0.049	0.055	0.042	0.043	0.050	0.064	0.052	0.046	0.047	0.052	0.055
(0,0.25)	0.060	0.051	0.051	0.040	0.052	0.048	0.054	0.056	0.053	0.058	0.049	0.057
(0.25,0.5)	0.040	0.048	0.054	0.047	0.051	0.058	0.052	0.051	0.054	0.055	0.066	0.065
(0.5,0.75)	0.049	0.046	0.049	0.050	0.052	0.054	0.059	0.058	0.060	0.062	0.054	0.064
(0,0.5)	0.051	0.043	0.057	0.049	0.046	0.049	0.050	0.055	0.059	0.059	0.056	0.054
(0.25,0.75)	0.052	0.061	0.056	0.053	0.045	0.064	0.063	0.061	0.068	0.068	0.061	0.075
(0,0.75)	0.055	0.055	0.055	0.061	0.049	0.067	0.054	0.057	0.069	0.068	0.063	0.066
(0.25,0)	0.060	0.051	0.051	0.040	0.047	0.051	0.058	0.053	0.043	0.049	0.054	0.044
(0.5,0.25)	0.040	0.048	0.054	0.047	0.048	0.036	0.049	0.053	0.044	0.050	0.051	0.044
(0.75,0.5),	0.049	0.046	0.049	0.050	0.039	0.042	0.059	0.053	0.044	0.039	0.047	0.047
(0.5,0)	0.051	0.043	0.057	0.049	0.045	0.043	0.051	0.051	0.050	0.043	0.049	0.040
(0.75,0.25)	0.052	0.061	0.056	0.053	0.042	0.033	0.047	0.052	0.044	0.036	0.046	0.039
(0.75,0)	0.055	0.055	0.055	0.061	0.040	0.034	0.044	0.051	0.038	0.040	0.046	0.039
	B: $p = 100$											
	(n_1, n_2)											
(ρ_1, ρ_2)	(10,10)	(20,20)	(30,30)	(40,40)	(8,12)	(16,24)	(26,34)	(32,48)	(5,15)	(12,28)	(18,42)	(24,56)
(0,0)	0.052	0.057	0.064	0.068	0.046	0.059	0.060	0.047	0.047	0.044	0.042	0.045
(0.25,0.25)	0.045	0.060	0.050	0.053	0.050	0.053	0.055	0.042	0.052	0.049	0.051	0.041
(0.5,0.5)	0.048	0.050	0.049	0.057	0.043	0.063	0.047	0.038	0.058	0.050	0.057	0.044
(0.75,0.75)	0.051	0.054	0.051	0.049	0.052	0.057	0.055	0.048	0.058	0.051	0.047	0.047
(0,0.25)	0.052	0.051	0.047	0.058	0.037	0.055	0.059	0.047	0.046	0.049	0.049	0.049
(0.25,0.5)	0.043	0.050	0.052	0.046	0.054	0.060	0.056	0.055	0.060	0.051	0.058	0.050
(0.5,0.75)	0.052	0.048	0.052	0.053	0.047	0.068	0.053	0.052	0.069	0.072	0.067	0.065
(0,0.5)	0.053	0.048	0.038	0.053	0.040	0.053	0.059	0.052	0.066	0.051	0.061	0.063
(0.25,0.75)	0.055	0.047	0.049	0.047	0.048	0.058	0.059	0.063	0.085	0.080	0.066	0.076
(0,0.75)	0.050	0.045	0.041	0.048	0.048	0.057	0.062	0.065	0.077	0.067	0.065	0.079
(0.25,0)	0.052	0.051	0.047	0.058	0.052	0.049	0.055	0.043	0.052	0.050	0.047	0.043
(0.5,0.25)	0.043	0.050	0.052	0.046	0.048	0.055	0.040	0.031	0.060	0.048	0.041	0.039
(0.75,0.5)	0.052	0.048	0.052	0.053	0.037	0.049	0.044	0.031	0.045	0.038	0.035	0.027
(0.5,0)	0.053	0.048	0.038	0.053	0.058	0.053	0.056	0.033	0.052	0.050	0.036	0.033
(0.75,0.25)	0.055	0.047	0.049	0.047	0.047	0.049	0.046	0.031	0.047	0.043	0.032	0.028
(0.75,0)	0.050	0.045	0.041	0.048	0.053	0.051	0.055	0.039	0.042	0.037	0.031	0.024

Table A.5.: Actual type I error rate with the permutation null distribution of the χ^2 sum statistic Q_{χ^2} in the case $\mathbf{c} = \mathbf{2}$. Simulation margin of error for $\alpha = 0.05$: ± 0.0138. See Figure 3.5 in Section 3.5.2 for a visual representation.

	A: $p = 20$											
	(n_1, n_2)											
(ρ_1, ρ_2)	(10,10)	(20,20)	(30,30)	(40,40)	(8,12)	(16,24)	(26,34)	(32,48)	(5,15)	(12,28)	(18,42)	(24,56)
(0,0)	0.058	0.035	0.063	0.048	0.050	0.048	0.048	0.047	0.043	0.046	0.046	0.061
(0.25,0.25)	0.060	0.046	0.054	0.052	0.051	0.043	0.054	0.046	0.051	0.049	0.053	0.050
(0.5,0.5)	0.057	0.054	0.050	0.047	0.043	0.049	0.062	0.056	0.044	0.051	0.052	0.057
(0.75,0.75)	0.052	0.051	0.048	0.044	0.046	0.052	0.068	0.048	0.049	0.049	0.055	0.054
(0,0.25)	0.057	0.050	0.064	0.051	0.040	0.033	0.047	0.037	0.020	0.034	0.033	0.033
(0.25,0.5)	0.058	0.056	0.046	0.052	0.040	0.039	0.048	0.041	0.029	0.033	0.030	0.041
(0.5,0.75)	0.056	0.051	0.053	0.047	0.038	0.047	0.059	0.046	0.028	0.041	0.040	0.043
(0,0.5)	0.061	0.059	0.055	0.067	0.035	0.032	0.038	0.033	0.014	0.019	0.011	0.014
(0.25,0.75)	0.054	0.056	0.053	0.057	0.043	0.037	0.047	0.038	0.021	0.023	0.021	0.022
(0,0.75)	0.057	0.056	0.050	0.057	0.032	0.029	0.037	0.026	0.013	0.015	0.007	0.015
(0.25,0)	0.057	0.050	0.064	0.051	0.064	0.056	0.059	0.055	0.088	0.068	0.075	0.069
(0.5,0.25)	0.058	0.056	0.046	0.052	0.056	0.065	0.068	0.059	0.081	0.065	0.078	0.072
(0.75,0.5)	0.056	0.051	0.053	0.047	0.047	0.056	0.060	0.058	0.069	0.061	0.075	0.068
(0.5,0)	0.061	0.059	0.055	0.067	0.073	0.076	0.065	0.064	0.132	0.115	0.111	0.094
(0.75,0.25)	0.054	0.056	0.053	0.057	0.063	0.068	0.066	0.063	0.100	0.079	0.102	0.084
(0.75,0)	0.057	0.056	0.050	0.057	0.085	0.086	0.080	0.075	0.134	0.118	0.131	0.110
	B: $p = 100$											
	(n_1, n_2)											
(ρ_1, ρ_2)	(10,10)	(20,20)	(30,30)	(40,40)	(8,12)	(16,24)	(26,34)	(32,48)	(5,15)	(12,28)	(18,42)	(24,56)
(0,0)	0.056	0.050	0.045	0.053	0.046	0.051	0.038	0.060	0.046	0.045	0.041	0.040
(0.25,0.25)	0.050	0.058	0.056	0.047	0.047	0.053	0.051	0.054	0.063	0.054	0.049	0.046
(0.5,0.5)	0.052	0.055	0.064	0.049	0.046	0.058	0.053	0.053	0.066	0.055	0.048	0.054
(0.75,0.75)	0.055	0.049	0.058	0.045	0.043	0.051	0.058	0.051	0.063	0.056	0.055	0.057
(0,0.25)	0.060	0.055	0.055	0.039	0.031	0.032	0.027	0.033	0.021	0.012	0.013	0.015
(0.25,0.5)	0.052	0.061	0.061	0.048	0.038	0.041	0.039	0.040	0.035	0.032	0.032	0.030
(0.5,0.75)	0.050	0.050	0.062	0.045	0.039	0.047	0.051	0.046	0.040	0.036	0.042	0.040
(0,0.5)	0.063	0.056	0.061	0.037	0.025	0.028	0.025	0.032	0.015	0.008	0.006	0.008
(0.25,0.75)	0.053	0.051	0.058	0.043	0.033	0.030	0.036	0.033	0.024	0.023	0.025	0.024
(0,0.75)	0.058	0.057	0.057	0.040	0.033	0.025	0.024	0.032	0.014	0.009	0.004	0.009
(0.25,0)	0.060	0.055	0.055	0.039	0.095	0.087	0.076	0.087	0.146	0.116	0.109	0.123
(0.5,0.25)	0.052	0.061	0.061	0.048	0.060	0.077	0.055	0.062	0.113	0.078	0.083	0.073
(0.75,0.5)	0.050	0.050	0.062	0.045	0.058	0.063	0.062	0.055	0.099	0.070	0.069	0.069
(0.5,0)	0.063	0.056	0.061	0.037	0.107	0.100	0.088	0.089	0.181	0.141	0.137	0.146
(0.75,0.25)	0.053	0.051	0.058	0.043	0.068	0.084	0.065	0.070	0.127	0.098	0.097	0.093
(0.75,0)	0.058	0.057	0.057	0.040	0.097	0.093	0.084	0.093	0.167	0.147	0.133	0.149

Table A.6.: Actual type I error rate with the permutation null distribution of max-T (χ^2) in the case $\mathbf{c} = \mathbf{2}$. Simulation margin of error for $\alpha = 0.05$: ± 0.0138. See Figure 3.6 in Section 3.5.2 for a visual representation.

	A: $p = 20$											
	(n_1, n_2)											
(ρ_1, ρ_2)	(10,10)	(20,20)	(30,30)	(40,40)	(8,12)	(16,24)	(26,34)	(32,48)	(5,15)	(12,28)	(18,42)	(24,56)
(0,0)	0.054	0.055	0.062	0.046	0.046	0.048	0.053	0.045	0.053	0.048	0.052	0.030
(0.25,0.25)	0.058	0.038	0.057	0.054	0.039	0.050	0.061	0.054	0.055	0.056	0.051	0.051
(0.5,0.5)	0.053	0.047	0.057	0.047	0.041	0.040	0.063	0.061	0.048	0.050	0.058	0.059
(0.75,0.75)	0.047	0.050	0.052	0.041	0.047	0.052	0.060	0.052	0.043	0.041	0.053	0.061
(0,0.25)	0.059	0.043	0.068	0.057	0.047	0.046	0.052	0.059	0.054	0.059	0.057	0.047
(0.25,0.5)	0.057	0.049	0.054	0.041	0.050	0.046	0.060	0.053	0.052	0.052	0.058	0.066
(0.5,0.75)	0.048	0.051	0.046	0.046	0.038	0.051	0.069	0.072	0.053	0.059	0.061	0.071
(0,0.5)	0.053	0.053	0.055	0.062	0.040	0.051	0.059	0.054	0.068	0.054	0.059	0.052
(0.25,0.75)	0.048	0.062	0.048	0.049	0.053	0.058	0.060	0.063	0.069	0.061	0.060	0.082
(0,0.75)	0.052	0.058	0.057	0.051	0.048	0.059	0.060	0.055	0.062	0.066	0.061	0.077
(0.25,0)	0.059	0.043	0.068	0.057	0.043	0.044	0.056	0.052	0.049	0.040	0.043	0.041
(0.5,0.25)	0.057	0.049	0.054	0.041	0.038	0.043	0.055	0.047	0.053	0.047	0.049	0.048
(0.75,0.5)	0.048	0.051	0.046	0.046	0.038	0.038	0.063	0.046	0.048	0.040	0.046	0.049
(0.5,0)	0.053	0.053	0.055	0.062	0.047	0.040	0.039	0.056	0.055	0.047	0.050	0.041
(0.75,0.25)	0.048	0.062	0.048	0.049	0.037	0.045	0.056	0.041	0.045	0.038	0.044	0.037
(0.75,0)	0.052	0.058	0.057	0.051	0.042	0.044	0.060	0.040	0.045	0.040	0.040	0.038
	B: $p = 100$											
	(n_1, n_2)											
(ρ_1, ρ_2)	(10,10)	(20,20)	(30,30)	(40,40)	(8,12)	(16,24)	(26,34)	(32,48)	(5,15)	(12,28)	(18,42)	(24,56)
(0,0)	0.039	0.047	0.062	0.060	0.048	0.042	0.051	0.040	0.046	0.054	0.046	0.047
(0.25,0.25)	0.044	0.051	0.050	0.054	0.042	0.052	0.053	0.045	0.044	0.054	0.047	0.050
(0.5,0.5)	0.047	0.050	0.065	0.055	0.049	0.055	0.054	0.053	0.064	0.047	0.049	0.045
(0.75,0.75)	0.052	0.042	0.066	0.047	0.041	0.058	0.046	0.054	0.058	0.048	0.054	0.050
(0,0.25)	0.039	0.045	0.052	0.045	0.043	0.040	0.059	0.038	0.043	0.052	0.046	0.050
(0.25,0.5)	0.047	0.053	0.047	0.048	0.054	0.044	0.058	0.048	0.066	0.045	0.062	0.060
(0.5,0.75)	0.049	0.051	0.065	0.051	0.058	0.066	0.056	0.057	0.066	0.071	0.065	0.074
(0,0.5)	0.040	0.052	0.048	0.064	0.059	0.042	0.061	0.049	0.044	0.038	0.058	0.057
(0.25,0.75)	0.043	0.055	0.056	0.048	0.054	0.061	0.058	0.059	0.082	0.077	0.085	0.087
(0,0.75)	0.045	0.051	0.048	0.044	0.063	0.062	0.053	0.057	0.074	0.071	0.076	0.094
(0.25,0)	0.039	0.045	0.052	0.045	0.042	0.046	0.048	0.046	0.040	0.053	0.047	0.033
(0.5,0.25)	0.047	0.053	0.047	0.048	0.054	0.052	0.048	0.034	0.050	0.046	0.036	0.033
(0.75,0.5)	0.049	0.051	0.065	0.051	0.043	0.044	0.044	0.038	0.048	0.037	0.044	0.030
(0.5,0)	0.040	0.052	0.048	0.064	0.060	0.053	0.052	0.044	0.044	0.058	0.033	0.026
(0.75,0.25)	0.043	0.055	0.056	0.048	0.045	0.051	0.048	0.028	0.050	0.040	0.032	0.027
(0.75,0)	0.045	0.051	0.048	0.044	0.044	0.049	0.053	0.043	0.043	0.040	0.027	0.018

Table A.7.: Actual type I error rate with the bootstrap null distribution of the χ^2 sum statistic Q_{χ^2} in the case $\mathbf{c} = \mathbf{4}$. Simulation margin of error for $\alpha = 0.05$: ± 0.0138. See Figure 3.7 in Section 3.6 for a visual representation.

	A: $p = 20$											
	(n_1, n_2)											
(ρ_1, ρ_2)	(10,10)	(20,20)	(30,30)	(40,40)	(8,12)	(16,24)	(26,34)	(32,48)	(5,15)	(12,28)	(18,42)	(24,56)
(0,0)	0.009	0.010	0.016	0.017	0.003	0.012	0.012	0.019	0.001	0.011	0.021	0.015
(0.25,0.25)	0.014	0.019	0.023	0.031	0.012	0.024	0.025	0.025	0.011	0.015	0.027	0.037
(0.5,0.5)	0.037	0.049	0.043	0.040	0.038	0.039	0.051	0.042	0.033	0.031	0.045	0.040
(0.75,0.75)	0.052	0.045	0.045	0.042	0.051	0.041	0.060	0.047	0.049	0.048	0.049	0.045
(0,0.25)	0.012	0.009	0.022	0.026	0.003	0.012	0.018	0.021	0.004	0.007	0.017	0.018
(0.25,0.5)	0.028	0.037	0.036	0.041	0.025	0.030	0.032	0.031	0.018	0.018	0.025	0.028
(0.5,0.75)	0.048	0.050	0.052	0.039	0.045	0.033	0.053	0.042	0.042	0.028	0.042	0.031
(0,0.5)	0.015	0.026	0.029	0.028	0.009	0.016	0.017	0.029	0.003	0.012	0.010	0.014
(0.25,0.75)	0.040	0.044	0.048	0.045	0.033	0.026	0.040	0.029	0.026	0.017	0.019	0.025
(0,0.75)	0.032	0.032	0.047	0.044	0.017	0.020	0.025	0.028	0.011	0.013	0.008	0.011
(0.25,0)	0.012	0.009	0.022	0.026	0.009	0.018	0.024	0.026	0.010	0.008	0.029	0.034
(0.5,0.25)	0.028	0.037	0.036	0.041	0.034	0.031	0.040	0.047	0.026	0.036	0.051	0.051
(0.75,0.5)	0.048	0.050	0.052	0.039	0.047	0.048	0.063	0.062	0.050	0.059	0.057	0.067
(0.5,0)	0.015	0.026	0.029	0.028	0.024	0.028	0.031	0.043	0.021	0.032	0.044	0.055
(0.75,0.25)	0.040	0.044	0.048	0.045	0.040	0.052	0.065	0.059	0.040	0.068	0.074	0.073
(0.75,0)	0.032	0.032	0.047	0.044	0.049	0.049	0.055	0.061	0.045	0.070	0.078	0.079
	B: $p = 100$											
	(n_1, n_2)											
(ρ_1, ρ_2)	(10,10)	(20,20)	(30,30)	(40,40)	(8,12)	(16,24)	(26,34)	(32,48)	(5,15)	(12,28)	(18,42)	(24,56)
(0,0)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002
(0.25,0.25)	0.011	0.014	0.020	0.020	0.007	0.010	0.018	0.021	0.010	0.012	0.022	0.025
(0.5,0.5)	0.037	0.041	0.048	0.044	0.030	0.041	0.037	0.044	0.037	0.037	0.037	0.042
(0.75,0.75)	0.049	0.049	0.051	0.049	0.049	0.048	0.047	0.040	0.054	0.049	0.055	0.045
(0,0.25)	0.001	0.003	0.003	0.005	0.000	0.001	0.002	0.007	0.000	0.001	0.002	0.004
(0.25,0.5)	0.029	0.034	0.032	0.034	0.018	0.019	0.024	0.026	0.018	0.018	0.022	0.025
(0.5,0.75)	0.047	0.045	0.050	0.042	0.038	0.034	0.039	0.040	0.035	0.036	0.033	0.030
(0,0.5)	0.012	0.019	0.026	0.019	0.007	0.007	0.009	0.017	0.004	0.001	0.003	0.006
(0.25,0.75)	0.037	0.040	0.047	0.037	0.022	0.026	0.030	0.028	0.021	0.020	0.016	0.017
(0,0.75)	0.032	0.034	0.044	0.029	0.019	0.014	0.024	0.029	0.014	0.008	0.005	0.005
(0.25,0)	0.001	0.003	0.003	0.005	0.000	0.004	0.007	0.011	0.003	0.006	0.021	0.008
(0.5,0.25)	0.029	0.034	0.032	0.034	0.027	0.039	0.038	0.040	0.031	0.040	0.050	0.046
(0.75,0.5)	0.047	0.045	0.050	0.042	0.052	0.057	0.047	0.051	0.059	0.062	0.067	0.060
(0.5,0)	0.012	0.019	0.026	0.019	0.013	0.031	0.036	0.040	0.023	0.038	0.058	0.054
(0.75,0.25)	0.037	0.040	0.047	0.037	0.043	0.065	0.047	0.056	0.057	0.072	0.089	0.078
(0.75,0)	0.032	0.034	0.044	0.029	0.034	0.067	0.049	0.073	0.046	0.083	0.102	0.094

Table A.8.: Actual type I error rate with the bootstrap null distribution of the CA sum statistic Q_{CA} in the case $\mathbf{c} = \mathbf{4}$. Simulation margin of error for $\alpha = 0.05$: ± 0.0138. See Figure 3.8 in Section 3.6 for a visual representation.

	A: $p = 20$											
	(n_1, n_2)											
(ρ_1, ρ_2)	(10,10)	(20,20)	(30,30)	(40,40)	(8,12)	(16,24)	(26,34)	(32,48)	(5,15)	(12,28)	(18,42)	(24,56)
(0,0)	0.024	0.025	0.034	0.050	0.010	0.024	0.037	0.037	0.019	0.033	0.039	0.039
(0.25,0.25)	0.045	0.046	0.054	0.046	0.043	0.043	0.052	0.039	0.045	0.043	0.053	0.054
(0.5,0.5)	0.062	0.053	0.043	0.052	0.061	0.047	0.061	0.050	0.058	0.050	0.053	0.054
(0.75,0.75)	0.068	0.056	0.049	0.047	0.055	0.048	0.073	0.055	0.053	0.054	0.050	0.054
(0,0.25)	0.038	0.030	0.045	0.037	0.019	0.028	0.038	0.035	0.011	0.026	0.025	0.027
(0.25,0.5)	0.061	0.056	0.048	0.055	0.050	0.035	0.052	0.036	0.032	0.031	0.035	0.038
(0.5,0.75)	0.065	0.054	0.050	0.051	0.053	0.038	0.060	0.048	0.039	0.042	0.038	0.039
(0,0.5)	0.050	0.051	0.054	0.059	0.040	0.028	0.036	0.033	0.013	0.017	0.012	0.015
(0.25,0.75)	0.067	0.054	0.044	0.051	0.046	0.035	0.048	0.032	0.031	0.028	0.021	0.032
(0,0.75)	0.066	0.052	0.056	0.063	0.042	0.029	0.040	0.035	0.016	0.016	0.009	0.015
(0.25,0)	0.038	0.030	0.045	0.037	0.041	0.039	0.046	0.047	0.044	0.053	0.059	0.060
(0.5,0.25)	0.061	0.056	0.048	0.055	0.072	0.056	0.071	0.061	0.079	0.062	0.075	0.061
(0.75,0.5)	0.065	0.054	0.050	0.051	0.072	0.056	0.070	0.061	0.076	0.061	0.070	0.065
(0.5,0)	0.050	0.051	0.054	0.059	0.075	0.059	0.071	0.068	0.087	0.098	0.091	0.084
(0.75,0.25)	0.067	0.054	0.044	0.051	0.080	0.063	0.071	0.068	0.102	0.087	0.087	0.078
(0.75,0)	0.066	0.052	0.056	0.063	0.091	0.079	0.080	0.071	0.133	0.125	0.114	0.105
	B: $p = 100$											
	(n_1, n_2)											
(ρ_1, ρ_2)	(10,10)	(20,20)	(30,30)	(40,40)	(8,12)	(16,24)	(26,34)	(32,48)	(5,15)	(12,28)	(18,42)	(24,56)
(0,0)	0.000	0.000	0.005	0.011	0.000	0.001	0.007	0.012	0.000	0.002	0.006	0.007
(0.25,0.25)	0.046	0.049	0.050	0.048	0.044	0.050	0.043	0.045	0.055	0.043	0.035	0.045
(0.5,0.5)	0.055	0.054	0.061	0.048	0.058	0.056	0.048	0.060	0.074	0.056	0.048	0.049
(0.75,0.75)	0.055	0.058	0.064	0.047	0.055	0.056	0.050	0.059	0.075	0.053	0.048	0.051
(0,0.25)	0.021	0.027	0.038	0.028	0.012	0.014	0.016	0.022	0.013	0.005	0.008	0.007
(0.25,0.5)	0.052	0.057	0.058	0.047	0.043	0.037	0.040	0.040	0.041	0.032	0.028	0.033
(0.5,0.75)	0.058	0.060	0.064	0.047	0.049	0.049	0.049	0.049	0.057	0.038	0.036	0.039
(0,0.5)	0.051	0.051	0.055	0.035	0.022	0.026	0.025	0.030	0.019	0.009	0.006	0.009
(0.25,0.75)	0.061	0.063	0.058	0.047	0.041	0.033	0.035	0.037	0.037	0.023	0.021	0.025
(0,0.75)	0.068	0.058	0.057	0.039	0.027	0.030	0.024	0.034	0.020	0.010	0.006	0.008
(0.25,0)	0.021	0.027	0.038	0.028	0.033	0.045	0.053	0.060	0.057	0.065	0.071	0.088
(0.5,0.25)	0.052	0.057	0.058	0.047	0.066	0.067	0.055	0.061	0.106	0.076	0.081	0.075
(0.75,0.5)	0.058	0.060	0.064	0.047	0.070	0.070	0.054	0.062	0.104	0.070	0.068	0.065
(0.5,0)	0.051	0.051	0.055	0.035	0.083	0.087	0.072	0.087	0.123	0.111	0.115	0.135
(0.75,0.25)	0.061	0.063	0.058	0.047	0.082	0.083	0.060	0.066	0.136	0.093	0.094	0.093
(0.75,0)	0.068	0.058	0.057	0.039	0.104	0.100	0.080	0.093	0.159	0.135	0.130	0.160

Table A.9.: Actual type I error rate with the boostrap null distribution of max-T (χ^2) in the case $\mathbf{c} = \mathbf{4}$. Simulation margin of error for $\alpha = 0.05$: ± 0.0138. See Figure 3.9 in Section 3.6 for a visual representation.

	A: $p = 20$											
	(n_1, n_2)											
(ρ_1, ρ_2)	(10,10)	(20,20)	(30,30)	(40,40)	(8,12)	(16,24)	(26,34)	(32,48)	(5,15)	(12,28)	(18,42)	(24,56)
(0,0)	0.082	0.063	0.063	0.049	0.068	0.066	0.064	0.061	0.057	0.052	0.055	0.058
(0.25,0.25)	0.072	0.062	0.063	0.050	0.071	0.063	0.065	0.063	0.068	0.055	0.052	0.055
(0.5,0.5)	0.067	0.068	0.060	0.052	0.066	0.058	0.059	0.055	0.057	0.062	0.061	0.070
(0.75,0.75)	0.077	0.058	0.058	0.048	0.080	0.050	0.057	0.052	0.071	0.056	0.052	0.057
(0,0.25)	0.083	0.059	0.061	0.058	0.064	0.050	0.055	0.071	0.047	0.054	0.051	0.060
(0.25,0.5)	0.074	0.066	0.069	0.046	0.074	0.076	0.053	0.053	0.060	0.060	0.061	0.057
(0.5,0.75)	0.078	0.060	0.062	0.047	0.082	0.069	0.063	0.066	0.063	0.061	0.061	0.061
(0,0.5)	0.082	0.066	0.062	0.065	0.068	0.068	0.058	0.063	0.059	0.069	0.065	0.058
(0.25,0.75)	0.075	0.058	0.071	0.055	0.079	0.063	0.057	0.062	0.065	0.061	0.051	0.073
(0,0.75)	0.073	0.062	0.070	0.070	0.060	0.069	0.064	0.058	0.060	0.063	0.056	0.046
(0.25,0)	0.083	0.059	0.061	0.058	0.068	0.070	0.051	0.053	0.060	0.051	0.061	0.050
(0.5,0.25)	0.074	0.066	0.069	0.046	0.067	0.058	0.061	0.050	0.061	0.079	0.052	0.049
(0.75,0.5)	0.078	0.060	0.062	0.047	0.069	0.063	0.055	0.055	0.062	0.063	0.058	0.054
(0.5,0)	0.082	0.066	0.062	0.065	0.076	0.067	0.057	0.055	0.056	0.065	0.057	0.061
(0.75,0.25)	0.075	0.058	0.071	0.055	0.078	0.063	0.054	0.052	0.068	0.066	0.050	0.059
(0.75,0)	0.073	0.062	0.070	0.070	0.069	0.084	0.058	0.053	0.055	0.049	0.054	0.047
	B: $p = 100$											
	(n_1, n_2)											
(ρ_1, ρ_2)	(10,10)	(20,20)	(30,30)	(40,40)	(8,12)	(16,24)	(26,34)	(32,48)	(5,15)	(12,28)	(18,42)	(24,56)
(0,0)	0.087	0.077	0.065	0.066	0.095	0.057	0.054	0.048	0.090	0.063	0.058	0.053
(0.25,0.25)	0.073	0.061	0.068	0.064	0.078	0.067	0.051	0.050	0.106	0.063	0.056	0.057
(0.5,0.5)	0.089	0.070	0.046	0.068	0.083	0.074	0.063	0.051	0.090	0.079	0.067	0.053
(0.75,0.75)	0.100	0.079	0.060	0.057	0.093	0.062	0.066	0.041	0.093	0.064	0.058	0.057
(0,0.25)	0.075	0.076	0.066	0.061	0.082	0.064	0.057	0.061	0.103	0.064	0.058	0.051
(0.25,0.5)	0.090	0.069	0.065	0.066	0.075	0.068	0.049	0.055	0.099	0.075	0.057	0.059
(0.5,0.75)	0.111	0.080	0.057	0.057	0.095	0.079	0.066	0.057	0.110	0.074	0.063	0.057
(0,0.5)	0.091	0.062	0.065	0.059	0.078	0.054	0.067	0.053	0.115	0.061	0.057	0.067
(0.25,0.75)	0.116	0.073	0.069	0.061	0.088	0.069	0.068	0.070	0.116	0.070	0.061	0.060
(0,0.75)	0.100	0.074	0.048	0.050	0.096	0.073	0.063	0.061	0.119	0.072	0.066	0.068
(0.25,0)	0.075	0.076	0.066	0.061	0.077	0.070	0.058	0.067	0.111	0.066	0.053	0.046
(0.5,0.25)	0.090	0.069	0.065	0.066	0.080	0.069	0.063	0.044	0.087	0.070	0.046	0.050
(0.75,0.5)	0.111	0.080	0.057	0.057	0.087	0.069	0.055	0.052	0.087	0.060	0.066	0.053
(0.5,0)	0.091	0.062	0.065	0.059	0.104	0.084	0.060	0.065	0.103	0.077	0.047	0.047
(0.75,0.25)	0.116	0.073	0.069	0.061	0.089	0.068	0.054	0.048	0.087	0.060	0.045	0.054
(0.75,0)	0.100	0.074	0.048	0.050	0.084	0.067	0.068	0.061	0.086	0.057	0.057	0.043

Table A.10.: Actual type I error rate with the bootstrap null distribution of max-T (CA) in the case $\mathbf{c} = \mathbf{4}$. Simulation margin of error for $\alpha = 0.05$: ± 0.0138. See Figure 3.10 in Section 3.6 for a visual representation.

	A: $p = 20$											
	(n_1, n_2)											
(ρ_1, ρ_2)	(10,10)	(20,20)	(30,30)	(40,40)	(8,12)	(16,24)	(26,34)	(32,48)	(5,15)	(12,28)	(18,42)	(24,56)
(0,0)	0.066	0.056	0.054	0.048	0.066	0.058	0.059	0.062	0.056	0.050	0.044	0.048
(0.25,0.25)	0.061	0.054	0.061	0.044	0.067	0.055	0.059	0.057	0.056	0.056	0.059	0.059
(0.5,0.5)	0.056	0.056	0.053	0.053	0.068	0.059	0.055	0.064	0.061	0.055	0.058	0.057
(0.75,0.75)	0.059	0.054	0.056	0.044	0.057	0.054	0.066	0.058	0.054	0.052	0.052	0.056
(0,0.25)	0.069	0.055	0.057	0.043	0.066	0.053	0.060	0.057	0.062	0.063	0.054	0.057
(0.25,0.5)	0.052	0.056	0.065	0.049	0.065	0.063	0.059	0.055	0.068	0.060	0.068	0.068
(0.5,0.75)	0.063	0.053	0.052	0.051	0.063	0.058	0.060	0.066	0.072	0.069	0.058	0.070
(0,0.5)	0.067	0.055	0.061	0.055	0.057	0.059	0.057	0.058	0.072	0.066	0.065	0.054
(0.25,0.75)	0.063	0.067	0.058	0.056	0.065	0.069	0.072	0.064	0.076	0.074	0.068	0.080
(0,0.75)	0.069	0.061	0.062	0.064	0.063	0.073	0.059	0.060	0.081	0.080	0.071	0.071
(0.25,0)	0.069	0.055	0.057	0.043	0.061	0.052	0.064	0.057	0.050	0.053	0.061	0.049
(0.5,0.25)	0.052	0.056	0.065	0.049	0.054	0.041	0.055	0.058	0.051	0.054	0.054	0.047
(0.75,0.5)	0.063	0.053	0.052	0.051	0.047	0.046	0.063	0.054	0.051	0.044	0.050	0.049
(0.5,0)	0.067	0.055	0.061	0.055	0.062	0.048	0.055	0.052	0.055	0.051	0.054	0.041
(0.75,0.25)	0.063	0.067	0.058	0.056	0.053	0.042	0.052	0.056	0.052	0.038	0.050	0.043
(0.75,0)	0.069	0.061	0.062	0.064	0.053	0.037	0.050	0.054	0.039	0.047	0.048	0.041
	B: $p = 100$											
	(n_1, n_2)											
(ρ_1, ρ_2)	(10,10)	(20,20)	(30,30)	(40,40)	(8,12)	(16,24)	(26,34)	(32,48)	(5,15)	(12,28)	(18,42)	(24,56)
(0,0)	0.078	0.064	0.070	0.077	0.073	0.072	0.063	0.052	0.050	0.048	0.047	0.051
(0.25,0.25)	0.078	0.067	0.057	0.060	0.078	0.058	0.064	0.044	0.064	0.058	0.061	0.045
(0.5,0.5)	0.076	0.060	0.062	0.065	0.061	0.071	0.053	0.044	0.071	0.064	0.060	0.051
(0.75,0.75)	0.068	0.057	0.056	0.054	0.059	0.063	0.055	0.050	0.069	0.056	0.054	0.051
(0,0.25)	0.076	0.055	0.054	0.063	0.055	0.067	0.070	0.052	0.057	0.060	0.056	0.055
(0.25,0.5)	0.077	0.061	0.068	0.055	0.072	0.065	0.063	0.063	0.078	0.068	0.070	0.060
(0.5,0.75)	0.077	0.057	0.055	0.057	0.063	0.075	0.059	0.057	0.089	0.083	0.076	0.071
(0,0.5)	0.079	0.060	0.045	0.058	0.058	0.062	0.066	0.054	0.083	0.067	0.068	0.070
(0.25,0.75)	0.082	0.054	0.053	0.060	0.067	0.072	0.069	0.073	0.103	0.091	0.082	0.083
(0,0.75)	.080	0.058	0.052	0.057	0.076	0.079	0.075	0.073	0.104	0.084	0.082	0.091
(0.25,0)	0.076	0.055	0.054	0.063	0.081	0.058	0.071	0.046	0.060	0.063	0.054	0.047
(0.5,0.25)	0.077	0.061	0.068	0.055	0.069	0.067	0.049	0.037	0.066	0.057	0.050	0.042
(0.75,0.5)	0.077	0.057	0.055	0.057	0.059	0.057	0.049	0.037	0.056	0.050	0.035	0.033
(0.5,0)	0.079	0.060	0.045	0.058	0.076	0.064	0.065	0.038	0.058	0.059	0.048	0.037
(0.75,0.25)	0.082	0.054	0.053	0.060	0.067	0.054	0.057	0.034	0.051	0.050	0.037	0.032
(0.75,0)	0.080	0.058	0.052	0.057	0.062	0.059	0.066	0.044	0.049	0.047	0.034	0.029

Table A.11.: Actual type I error rate with the bootstrap null distribution of the χ^2 sum statistic Q_{χ^2} in the case $\mathbf{c} = \mathbf{2}$. Simulation margin of error for $\alpha = 0.05$: ± 0.0138. See Figure 3.11 in Section 3.6 for a visual representation.

	A: $p = 20$											
	(n_1, n_2)											
(ρ_1, ρ_2)	(10,10)	(20,20)	(30,30)	(40,40)	(8,12)	(16,24)	(26,34)	(32,48)	(5,15)	(12,28)	(18,42)	(24,56)
(0,0)	0.018	0.022	0.040	0.042	0.014	0.028	0.037	0.037	0.015	0.027	0.032	0.051
(0.25,0.25)	0.039	0.044	0.053	0.049	0.038	0.041	0.048	0.044	0.041	0.047	0.052	0.048
(0.5,0.5)	0.061	0.052	0.050	0.055	0.047	0.052	0.063	0.056	0.040	0.053	0.054	0.058
(0.75,0.75)	0.056	0.053	0.049	0.047	0.046	0.056	0.070	0.049	0.050	0.052	0.060	0.055
(0,0.25)	0.031	0.035	0.055	0.045	0.023	0.028	0.037	0.034	0.010	0.028	0.028	0.027
(0.25,0.5)	0.056	0.055	0.049	0.053	0.041	0.041	0.049	0.038	0.027	0.033	0.030	0.038
(0.5,0.75)	0.061	0.054	0.052	0.051	0.044	0.049	0.060	0.046	0.026	0.043	0.040	0.042
(0,0.5)	0.049	0.054	0.054	0.065	0.029	0.032	0.038	0.032	0.009	0.016	0.010	0.014
(0.25,0.75)	0.057	0.059	0.053	0.059	0.044	0.037	0.048	0.040	0.021	0.025	0.022	0.025
(0,0.75)	0.057	0.057	0.052	0.057	0.031	0.031	0.039	0.029	0.015	0.015	0.007	0.015
(0.25,0)	0.031	0.035	0.055	0.045	0.040	0.034	0.042	0.049	0.042	0.052	0.057	0.062
(0.5,0.25)	0.056	0.055	0.049	0.053	0.051	0.062	0.069	0.060	0.075	0.065	0.078	0.072
(0.75,0.5)	0.061	0.054	0.052	0.051	0.050	0.058	0.064	0.058	0.069	0.063	0.079	0.065
(0.5,0)	0.049	0.054	0.054	0.065	0.060	0.064	0.061	0.063	0.100	0.097	0.103	0.091
(0.75,0.25)	0.057	0.059	0.053	0.059	0.064	0.071	0.067	0.063	0.096	0.078	0.102	0.085
(0.75,0)	0.057	0.057	0.052	0.057	0.083	0.081	0.079	0.074	0.117	0.114	0.126	0.109
	B: $p = 100$											
	(n_1, n_2)											
(ρ_1, ρ_2)	(10,10)	(20,20)	(30,30)	(40,40)	(8,12)	(16,24)	(26,34)	(32,48)	(5,15)	(12,28)	(18,42)	(24,56)
(0,0)	0.000	0.004	0.005	0.012	0.001	0.004	0.004	0.013	0.000	0.003	0.002	0.010
(0.25,0.25)	0.042	0.046	0.052	0.045	0.039	0.048	0.049	0.052	0.044	0.045	0.045	0.043
(0.5,0.5)	0.051	0.053	0.065	0.050	0.047	0.057	0.054	0.055	0.066	0.060	0.048	0.055
(0.75,0.75)	0.058	0.052	0.061	0.048	0.048	0.056	0.055	0.054	0.071	0.061	0.056	0.061
(0,0.25)	0.018	0.031	0.040	0.031	0.013	0.013	0.019	0.024	0.009	0.005	0.008	0.010
(0.25,0.5)	0.052	0.061	0.064	0.048	0.036	0.039	0.040	0.041	0.032	0.032	0.030	0.032
(0.5,0.75)	0.054	0.049	0.063	0.046	0.043	0.047	0.050	0.048	0.050	0.040	0.042	0.039
(0,0.5)	0.043	0.051	0.057	0.039	0.022	0.027	0.025	0.033	0.013	0.007	0.006	0.007
(0.25,0.75)	0.057	0.054	0.063	0.046	0.035	0.031	0.038	0.033	0.025	0.024	0.025	0.023
(0,0.75)	0.062	0.057	0.056	0.042	0.032	0.025	0.025	0.031	0.014	0.010	0.004	0.009
(0.25,0)	0.018	0.031	0.040	0.031	0.034	0.048	0.052	0.064	0.053	0.061	0.075	0.088
(0.5,0.25)	0.052	0.061	0.064	0.048	0.054	0.074	0.054	0.061	0.095	0.077	0.083	0.070
(0.75,0.5)	0.054	0.049	0.063	0.046	0.061	0.064	0.062	0.062	0.099	0.070	0.073	0.069
(0.5,0)	0.043	0.051	0.057	0.039	0.075	0.085	0.081	0.084	0.128	0.109	0.122	0.140
(0.75,0.25)	0.057	0.054	0.063	0.046	0.067	0.085	0.064	0.071	0.123	0.095	0.097	0.095
(0.75,0)	0.062	0.057	0.056	0.042	0.089	0.091	0.084	0.091	0.142	0.134	0.131	0.145

Table A.12.: Actual type I error rate with the bootstrap null distribution of max-T (χ^2) in the case $\mathbf{c} = \mathbf{2}$. Simulation margin of error for $\alpha = 0.05$: ± 0.0138. See Figure 3.12 in Section 3.6 for a visual representation.

	A: $p = 20$											
	(n_1, n_2)											
(ρ_1, ρ_2)	(10,10)	(20,20)	(30,30)	(40,40)	(8,12)	(16,24)	(26,34)	(32,48)	(5,15)	(12,28)	(18,42)	(24,56)
(0,0)	0.075	0.059	0.063	0.046	0.070	0.064	0.059	0.054	0.096	0.060	0.059	0.031
(0.25,0.25)	0.076	0.040	0.058	0.060	0.075	0.061	0.074	0.060	0.071	0.062	0.063	0.060
(0.5,0.5)	0.102	0.067	0.070	0.059	0.071	0.047	0.073	0.065	0.073	0.062	0.067	0.064
(0.75,0.75)	0.080	0.059	0.054	0.047	0.070	0.061	0.068	0.056	0.075	0.062	0.058	0.067
(0,0.25)	0.077	0.051	0.070	0.059	0.073	0.059	0.063	0.060	0.075	0.072	0.067	0.048
(0.25,0.5)	0.107	0.068	0.057	0.056	0.096	0.055	0.068	0.060	0.070	0.061	0.061	0.076
(0.5,0.75)	0.091	0.063	0.060	0.054	0.080	0.062	0.085	0.077	0.089	0.071	0.065	0.078
(0,0.5)	0.081	0.056	0.062	0.067	0.088	0.062	0.066	0.055	0.079	0.064	0.067	0.056
(0.25,0.75)	0.105	0.077	0.056	0.062	0.091	0.069	0.067	0.069	0.111	0.074	0.067	0.087
(0,0.75)	0.102	0.065	0.062	0.059	0.102	0.067	0.066	0.065	0.119	0.082	0.071	0.080
(0.25,0)	0.077	0.051	0.070	0.059	0.068	0.063	0.062	0.056	0.079	0.054	0.047	0.044
(0.5,0.25)	0.107	0.068	0.057	0.056	0.070	0.052	0.060	0.057	0.065	0.054	0.059	0.052
(0.75,0.5)	0.091	0.063	0.060	0.054	0.063	0.049	0.069	0.052	0.076	0.050	0.057	0.053
(0.5,0)	0.081	0.056	0.062	0.067	0.072	0.054	0.052	0.060	0.077	0.063	0.060	0.041
(0.75,0.25)	0.105	0.077	0.056	0.062	0.078	0.048	0.062	0.047	0.058	0.047	0.054	0.045
(0.75,0)	0.102	0.065	0.062	0.059	0.079	0.054	0.062	0.044	0.062	0.052	0.050	0.042
	B: $p = 100$											
	(n_1, n_2)											
(ρ_1, ρ_2)	(10,10)	(20,20)	(30,30)	(40,40)	(8,12)	(16,24)	(26,34)	(32,48)	(5,15)	(12,28)	(18,42)	(24,56)
(0,0)	0.083	0.072	0.070	0.064	0.090	0.066	0.059	0.046	0.045	0.057	0.060	0.050
(0.25,0.25)	0.103	0.063	0.072	0.066	0.083	0.072	0.058	0.053	0.051	0.063	0.057	0.060
(0.5,0.5)	0.113	0.065	0.070	0.060	0.088	0.070	0.060	0.056	0.100	0.068	0.061	0.051
(0.75,0.75)	0.102	0.059	0.075	0.057	0.080	0.072	0.059	0.059	0.104	0.063	0.065	0.054
(0,0.25)	0.076	0.054	0.065	0.052	0.070	0.062	0.068	0.044	0.044	0.060	0.057	0.060
(0.25,0.5)	0.128	0.068	0.066	0.069	0.110	0.069	0.071	0.063	0.089	0.070	0.071	0.064
(0.5,0.75)	0.102	0.072	0.078	0.062	0.101	0.088	0.060	0.064	0.132	0.087	0.078	0.080
(0,0.5)	0.113	0.059	0.064	0.075	0.096	0.053	0.075	0.055	0.071	0.067	0.073	0.070
(0.25,0.75)	0.108	0.075	0.065	0.050	0.094	0.083	0.068	0.068	0.149	0.108	0.100	0.100
(0,0.75)	0.129	0.069	0.076	0.076	0.106	0.081	0.056	0.080	0.163	0.104	0.096	0.105
(0.25,0)	0.076	0.054	0.065	0.052	0.091	0.064	0.059	0.048	0.039	0.068	0.056	0.041
(0.5,0.25)	0.128	0.068	0.066	0.069	0.097	0.068	0.063	0.042	0.064	0.057	0.046	0.046
(0.75,0.5)	0.102	0.072	0.078	0.062	0.086	0.065	0.055	0.049	0.076	0.049	0.049	0.037
(0.5,0)	0.113	0.059	0.064	0.075	0.084	0.070	0.064	0.054	0.044	0.065	0.043	0.037
(0.75,0.25)	0.108	0.075	0.065	0.050	0.088	0.067	0.064	0.037	0.058	0.050	0.037	0.031
(0.75,0)	0.129	0.069	0.076	0.076	0.066	0.070	0.062	0.051	0.044	0.044	0.037	0.025

Bibliography

A. Thompson (2004). Overview of primary progressive multiple sclerosis (PPMS): similarities and differences from other forms of MS, diagnostic criteria, pros and cons of progressive diagnosis. *Multiple Sclerosis 10*, S2–S7.

Ackermann, M. and K. Strimmer (2009). A general modular framework for gene set enrichment analysis. *BMC Bioinformatics 10*, 47.

Agresti, A. (2001). Exact inference for categorical data: recent advances and continuing controversies. *Statistics in Medicine 20*, 2709–2722.

Agresti, A. (2002). *Categorical Data Analysis*. Hoboken, New Jersey: Wiley.

Agresti, A. and B. Klingenberg (2005). Multivariate tests comparing binomial probabilities, with application to safety studies for drugs. *Journal of the Royal Statistical Society: Series C 54*, 691–706.

Algurén, B., C. Bostan, L. Christensson, B. Fridlund, and A. Cieza (2011). A multidisciplinary cross-cultural measurement of functioning after stroke: Rasch analysis of the brief ICF Core Set for stroke. *Topics in Stroke Rehabilitation 18(suppl 1)*, 573–586.

Amato, M. P., V. Zipoli, and E. Portaccio (2006). Multiple sclerosis-related cognitive changes: a review of cross-sectional and longitudinal studies. *Journal of Neurological Sciences 245*, 41–46.

American Heritage Dictionary (2014). `https://www.ahdictionary.com`. Accessed: 2014-09-15.

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics 11*, 375–386.

Astin, F., K. Atkin, and A. Darr (2008). Family support and cardiac rehabilitation: a comparative study of the experiences of South Asian and White-European patients and their carer's living in the United Kingdom. *European Journal of Cardiovascular Nursing 7(1)*, 43–51.

Balding, D. J. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics 7*, 781–791.

Beck, A. T., R. A. Steer, and G. K. Brown (1996). *Manual for the Beck Depression Inventory-II*. San Antonio: The Psychological Cooperation.

Boonen, A., J. Braun, I. E. van der Horst Bruinsma, F. Huang, W. Maksymowych, N. Kostanjsek, A. Cieza, G. Stucki, and D. van der Heijde (2010). ASAS/WHO ICF Core Sets for ankylosing spondylitis (AS): how to classify the impact of AS on functioning and health. *Annals of the Rheumatic Diseases 69(1)*, 102–107.

Bostan, C., C. Oberhauser, and A. Cieza (2012). Investigating the dimension functioning from a condition-specific perspective and the qualifier scale of the International Classification of Functioning, Disability and Health based on Rasch analyses. *American Journal of Physical Medicine and Rehabilitation 91(suppl)*, S129–S140.

Boulesteix, A.-L. (2005). *Dimension Reduction and Classification with High-Dimensional Microarray Data*. Ph. D. thesis, Faculty of Mathematics, Informatics und Statistics, LMU Munich.

Brown, P. J. (1993). *Measurement, Regression, and Calibration*. Oxford: Oxford University Press.

Chen, J. J. (1993). Trend Test for Overdispersed Proportions. *Biometrical Journal 35(8)*, 949–958.

Chung, J. H. and D. A. S. Fraser (1958). Randomization Tests for a Multivariate Two-Sample Problem. *Journal of the American Statistical Association 53(283)*, 729–735.

Cieza, A., T. Erwert, B. T. Ustün, S. Chatterji, N. Kostanjsekn, and G. Stucki (2004). Development of ICF Core Sets for patients with chronic conditions. *Journal of Rehabiliation Medicine 44(Suppl)*, 9–11.

Cieza, A., S. Geyh, S. Chatterji, N. Kostanjsek, B. T. Ustün, and G. Stucki (2006). Identification of candidate categories of the International Classification of Functioning Disability and Health (ICF) for a Generic ICF Core Set based on regression modelling. *BMC Medical Research Methodology 6*, 36.

Cieza, A., R. Hilfiker, S. Chatterji, N. Kostanjsekn, B. T. Ustün, and G. Stucki (2009). The International Classification of Functioning, Disability, and Health could be used to measure functioning. *Journal of Clinical Epidemiology 62*, 899–911.

Cochran, W. G. (1954). Some methods for strengthening the common chi-squared tests. *Biometrics 10*, 417–451.

Draghici, S., P. Khatri, R. P. Martins, G. C. Ostermeier, and S. Krawetz (2003). Global functional profiling of gene expression. *Genomics 81*, 98–104.

Dudoit, S. and M. J. van der Laan (2008). *Multiple Testing Procedures with Applications to Genomics*. New York: Springer Series in Statistics.

Dudoit, S., M. J. van der Laan, and K. S. Pollard (2004). Multiple Testing. Part I. Single-Step Procedures for Control of General Type I Error Rates. *Statistical Applications in Genetics and Molecular Biology 3(1)*, Article 13.

Efron, B. and R. J. Tibshirani (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Fellinghauer, B. A. G. (2011). *Understanding Human Functioning and Disability Using Graphical Modelsn*. Ph. D. thesis, Department of Mathematics, ETH Zürich.

Fellinghauer, B. A. G., P. Bühlmann, M. Ryffel, M. von Rhein, and J. D. Reinhardt (2013). Stable graphical model estimation with Random Forests for discrete, continuous, and mixed variables. *Computational Statistics & Data Analysis 64*, 132–152.

Fellinghauer, B. A. G., J. D. Reinhardt, and G. Stucki (2010). *In: Research Issues in Physical and Rehabilitation Medicine (F. Franchignoni, editor)*. Pavia: Maugeri Foundation Books.

Fisher, R. A. (1932). *Statistical Methods for Research Workers*. London: Oliver and Boyd.

Frank, I. E. and J. H. Friedman (1993). A statistical view of some chemometrics regression tools. *Technometrics 35*, 109–135.

Freidlin, B., G. Zheng, Z. Li, and J. L. Gastwirth (2002). Trend tests for case-control studies of genetic markers: power, sample size and robustness. *Human Heredity 53(3)*, 146–152.

Fridley, B. L., G. D. Jenkins, and J. M. Biernacka (2010). Self-contained gene-set analysis of expression data: an evaluation of existing and novel methods. *PLoS One 5(9)*, e12693.

Garthwaite, P. H. (1994). An interpretation of partial least squares. *Journal of the American Statistical Association 89*, 122–127.

Gertheiss, J. (2011). *Feature Extraction in Regression and Classification with Structured Predictors*. Ph. D. thesis, Faculty of Mathematics, Informatics und Statistics, LMU Munich.

Gertheiss, J., S. Hogger, C. Oberhauser, and G. Tutz (2011). Selection of ordinally scaled independent variables with applications to international classification of functioning core sets. *Journal of the Royal Statistical Society: Series C 60*, 377–395.

Geyh, S., A. Cieza, J. Schouten, H. Dickson, P. Frommelt, Z. Omar, N. Kostanjsek, H. Ring, and G. Stucki (2004). ICF core sets for stroke. *Journal of Rehabilitation Medicine 36*, 135–141.

Goeman, J. J. and P. Bühlmann (2007). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics 23(8)*, 980–987.

Goeman, J. J. and L. Finos (2012). The inheritance procedure: multiple testing of tree-structured hypotheses. *Statistical Applications in Genetics and Molecular Biology 11(1)*, 1–18.

Goeman, J. J. and U. Mansmann (2008). Multiple testing on the directed acyclic graph of gene ontology. *Bioinformatics 24(4)*, 537–544.

Goeman, J. J. and J. Oosting (2012). Globaltest R package, version 5.18.0. `http://www.bioconductor.org/`.

Goeman, J. J., J. Oosting, A. M. Cleton-Jansen, J. K. Anninga, and H. C. van Houwelingen (2005). Testing association of a pathway with survival using gene expression data. *Bioinformatics 21(9)*, 1950–1957.

Goeman, J. J. and A. Solari (2010). The sequential rejection principle of familywise error control. *Annals of Statistics 38(6)*, 3782–3810.

Goeman, J. J. and A. Solari (2014). Multiple hypothesis testing in genomics. *Statistics in Medicine 33(11)*, 1946–1978.

Goeman, J. J., S. A. van de Geer, F. de Kort, and H. C. van Houwelingen (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics 20(1)*, 93–99.

Goeman, J. J., S. A. van de Geer, and H. C. van Houwelingen (2006). Testing against a high-dimensional alternative. *Journal of the Royal Statistical Society: Series B 68*, 477–493.

Goeman, J. J., H. C. van Houwelingen, and L. Finos (2011). Testing against a high-dimensional alternative in the generalized linear model: asymptotic type I error control. *Biometrika 98*, 381–390.

Good, P. (2002). Extensions of the concept of exchangeability and their applications. *Journal of Modern Applied Statistical Methods 1(2)*, 243–247.

Good, P. (2005). *Permutation, Parametric and Bootstrap Tests of Hypotheses*. New York: Springer Science.

Herrmann, K. H., I. Kirchberger, F. Biering-Sør, and A. Cieza (2011). Differences in functioning of individuals with tetraplegia and paraplegia according to the International Classification of Functioning, Disability and Health (ICF). *Spinal Cord 49(4)*, 534–543.

Hirji, K. F. (1991). A comparison of exact, mid-p, and score tests for matched case-control studies. *Biometrics 47*, 487–496.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika 75*, 800–803.

Hoh, J. and J. Ott (2003). Mathematical multi-locus approaches to localizing complex human trait genes. *Nature Reviews Genetics 4*, 701–709.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics 6*, 65–70.

Holper, L., M. Coenen, A. Weise, G. Stucki, A. Cieza, and J. Kesselring (2010). Characterization of functioning in multiple sclerosis using the ICF. *Journal of Neurology 257(1)*, 103–113.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika 75*, 383–386.

Hotelling, H. (1931). The generalization of Student's ratio. *Annals of Mathematical Statistics 2(3)*, 360–378.

Huang, Y., H. Xu, V. Calian, and J. C. Hsu (2006). To permute or not to permute. *Bioinformatics 22*, 2244–2248.

Hummel, M., R. Meister, and U. Mansmann (2008). GlobalANCOVA: exploration and assessment of gene group effects. *Bioinformatics 24(1)*, 78–85.

Jelizarow, M., A. Cieza, and U. Mansmann (2014a). Global permutation tests for multivariate ordinal data: alternatives, test statistics and the null dilemma. *Journal of the Royal Statistical Society: Series C (Applied Statistics). doi: 10.1111/rssc.12070.*

Jelizarow, M., U. Mansmann, and J. J. Goeman (2014b). A Cochran-Armitage-type and a score-free global test for multivariate ordinal data. *Under revision. Preliminary version: Technical Report 168, Department of Statistics, LMU Munich.*

Kaiser, S. (2011). *Biclustering: Methods, Software and Application.* Ph. D. thesis, Faculty of Mathematics, Informatics und Statistics, LMU Munich.

Kaiser, S. and F. Leisch (2010). orddata: Generation of Artificial Ordinal and Binary Data. *R package, version 0.1*.

Kaiser, S., D. Träger, and F. Leisch (2011). Generating correlated ordinal random values. *Department of Statistics, LMU: Technical Report 94 (Available from http://epub.ub.uni-muenchen.de/12157/)*.

Kaizar, E. E., Y. Li, and J. H. Hsu (2011). Permutation Multiple Tests of Binary Features Do Not Uniformly Control Error Rates. *Journal of the American Statistical Association 106(495)*, 1067–1074.

Kalisch, M., B. A. G. Fellinghauer, E. Grill, M. H. Maathuis, U. Mansmann, P. Bühlmann, and G. Stucki (2010). Understanding human functioning using graphical models. *BMC Medical Research Methodology 10*, 14.

Klingenberg, B., A. Solari, L. Salmaso, and F. Pesarin (2009). Testing Marginal Homogeneity Against Stochastic Order in Multivariate Ordinal Data. *Biometrics 65*, 452–462.

Kong, S. W., W. T. Pu, and P. Park (2006). A multivariate approach for integrating genome-wide expression data and biological knowledge. *Bioinformatics 22*, 2373–2380.

Lancester, H. O. (1961). Significance Tests in Discrete Distributions. *Journal of the American Statistical Association 56 (294)*, 223–234.

le Cessie, S. and H. C. van Houwelingen (1995). Testing the fit of regression models via score tests in random effects models. *Biometrics 51*, 600–614.

Maciejewski, H. (2013). Gene set analysis methods: statistical models andmethodological differences. *Briefings in Bioinformatics. doi:10.1093/bib/bbt002*.

Madden, R., C. Sykes, and T. B. Ustün (2007). *World Health Organization Family of International Classifications: definition, scope and purpose*. Geneva: World Health Organization Press.

Mansmann, U. and R. Meister (2005). Testing differential gene expression in functional groups. *Methods of Information in Medicine 44*, 449–453.

Marcus, R., E. Peritz, and K. R. Gabriel (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika 63*, 655–660.

Martens, H. (2001). Reliable and relevant modelling of real world data: a personal account of the development of PLS regression. *Chemometrics and Intelligent Laboratory Systems 58*, 85–95.

Martens, H. and T. Naes (1989). *Multivariate Calibration*. New York: Wiley.

McCullagh, P. and J. Nelder (1989). *Generalized Linear Models*. Boca Raton: Chapman & Hall.

Meinshausen, N. (2008). Hierarchical testing of variable importance. *Biometrika 95*, 265–278.

Meinshausen, N., M. Maathuis, and P. Bühlmann (2012). Asymptotic optimality of the Westfall-Young permutation procedure for multiple testing under dependence. *The Annals of Statistics 39(6)*, 3369–3391.

Neuhäuser, M. (2010). *Computer-intensive und nichtparametrische statistische Tests*. Oldenburg: Oldenbourg Wissenschaftsverlag.

Oberhauser, C., R. Escorpizo, A. Boonen, G. Stucki, and A. Cieza (2013). Statistical validation of the brief International Classification of Functioning, Disability and Health Core Set for osteoarthritis based on a large international sample of patients with osteoarthritis. *Arthritis Care & Research (Hoboken) 65(2)*, 177–186.

Pesarin, F. (2001). *Multivariate Permutation Tests: With Applications in Biostatistics*. Chichester: Wiley.

Pesarin, F. and L. Salmaso (2010). *Permutation Tests for Complex Data: Theory, Applications and Software*. Chichester: Wiley.

Pollard, K. S. and M. J. van der Laan (2004). Choice of a null distribution in resampling-based multiple testing. *Journal of Statistical Planning and Inference 125*, 85–100.

Prodinger, B., T. Salzberger, G. Stucki, T. Stamm, and A. Cieza (2012). Measuring functioning in people with fibromyalgia (FM) based on the International Classification of Functioning, Disability and Health (ICF) — a psychometric analysis. *Pain Practice 12*, 255–265.

R Development Core Team (2014). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. `http://www.R-project.org/`.

Rauch, A., A. Cieza, and G. Stucki (2008). How to apply the International Classification of Functioning Disability and Health (ICF) for rehabilitation management in clinical practice. *European Journal of Physical and Rehabilitation Medicine 44*, 329–342.

Røe, C., E. Bautz-Holter, and A. Cieza (2013). Low back pain in 17 countries, a Rasch analysis of the ICF core set for low back pain. *International Journal of Rehabilitation Research 36*, 38–47.

Romano, J. (1990). On the behaviour of randomization tests without a groupsymmetry assumption. *Journal of the American Statistical Association 85*, 686–692.

Selb, M., R. Escorpizo, N. Kostanjsek, G. Stucki, B. T. Ustün, and A. Cieza (2014). A guide on how to develop an international classification of functioning, disability and health core set. *European Journal of Physical and Rehabilitation Medicine*.

Shaked, M. and J. G. Shanthikumar (2006). *Stochastic Orders*. New York: Springer Science & Business Media.

Smyth, G. K. (2003). Pearson's Goodness of Fit Statistic as a Score Test Statistic. *Statistics and Science: A Festschrift for Terry Speed. Lecture Notes-Monograph Series 40*, 115–126.

Solari, A., S. le Cessie, and J. J. Goeman (2012). Testing goodness of fit in regression: a general approach for specified alternatives. *Statistics in Medicine 31*, 3656–3666.

Stone, M. and R. J. Brooks (1990). Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal component regression. *Journal of the Royal Statistical Society: Series B 52*, 237–269.

Stucki, G. and G. Grimby (2004). Applying the ICF in medicine. *Journal of Rehabilitation Medicine 44(suppl)*, 5–6.

Stucki, G., N. Kostanjsek, B. T. Ustün, and A. Cieza (2008). ICF-based classification and measurement of functioning. *European Journal of Physical and Rehabilitation Medicine 44*, 315–328.

Troendle, J. F., E. L. Korn, and L. M. McShane (2004). An Example of Slow Convergence of the Bootstrap in High Dimensions. *The American Statistician 58(1)*, 25–29.

Tschiesner, U., C. Oberhauser, and A. Cieza (2011). ICF Core Set for head and neck cancer: do the categories discriminate among clinically relevant subgroups of patients? *International Journal of Rehabilitation Research 34(2)*, 121–130.

U. S. National Library of Medicine (2014). `http://www.pubmed.gov/`. Accessed: 2014-08-10.

Ustün, T. B., S. Chatterji, J. Bickenbach, N. Kostanjsek, and M. Schneider (2003). The International Classification of Functioning, Disability and Health: a new tool for understanding disability and health. *Disability and Rehabilitation 25(11-12)*, 565–571.

Ustün, T. B., S. Chatterji, and N. Kostanjsek (2004). Comments from WHO for the Journal of Rehabilitation Medicine special supplement on ICF core sets. *Journal of Rehabiliation Medicine 44(suppl)*, 7–8.

Walter, S. D., A. R. Feinstein, and C. K. Wells (1987). Coding ordinal independent variables in multiple regression analysis. *American Journal of Epidemiology 125*, 319–323.

Westfall, P., D. Zaykin, and S. Young (2001). *Multiple tests for genetic effects in association studies. In: Biostatistical methods (S. Looney, editor)*. Toloway, New Jersey: Humana Press.

Westfall, P. H. and J. F. Troendle (2008). Multiple Testing with Minimal Assumptions. *Biometrical Journal 50(5)*, 745–755.

Westfall, P. H. and R. D. Wolfinger (1997). Multiple tests with discrete distributions. *The American Statistician 51*, 3–8.

Westfall, P. H. and S. S. Young (1993). *Resampling-based Multiple Testing: Examples and Methods for P-Value Adjustment*. New York: Wiley-Interscience.

World Health Organization (1992). *International Statistical Classification of Diseases and Related Health Problems: ICD-10*. Geneva: World Health Organization Press.

World Health Organization (2001a). *54th World Health Assembly*. Geneva: World Health Organization Press.

World Health Organization (2001b). *International Classification of Functioning, Disability and Health: ICF*. Geneva: World Health Organization Press.

World Health Organization and The World Bank (2011). *World Report on Disability*. Geneva: World Health Organization Press.

Xu, H. and J. C. Hsu (2007). Using the partitioning principle to control the generalized family error rate. *Biometrical Journal 49*, 52–67.

Zheng, G., J. Joo, and Y. Yang (2009). Pearson's test, trend test, and MAX are all trend tests with different types of scores. *Annals of Human Genetics 73(2)*, 133–140.

List of figures

List of tables

Eidesstattliche Versicherung

(gemäß § 8 Abs. 2 Pkt. 5 der Promotionsordnung vom 12.07.2011)

Hiermit versichere ich an Eides statt, dass ich die vorgelegte Dissertation selbständig und ohne unerlaubte Hilfe Dritter verfasst sowie keine anderen als die angegebenen Quellen verwendet habe.

München, den 28.11.2014 .....

Monika Jelizarow