Statistical Matching meets Probabilistic Graphical Models

Contributions to Categorical Data Fusion

Eva-Marie Christina Endres

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

Eingereicht am 23. April 2019

Statistical Matching
meets
Probabilistic Graphical Models

Contributions to Categorical Data Fusion

Eva-Marie Christina Endres

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

Eingereicht am 23. April 2019

Erstberichterstatter: Prof. Dr. Thomas Augustin

Zweitberichterstatter: Prof. Dr. Moritz Grosse-Wentrup

Drittberichterstatterin: Prof. Dr. Helga Wagner

Tag der Disputation: 24. Juni 2019

Zusammenfassung

Die sekundäre Analyse bereits verfügbarer Daten kann Zeit, Kosten oder andere Ressourcen einsparen. Allerdings kann die Beantwortung bestimmter Fragstellungen gemeinsame Information über Variablen erfordern, die nicht gemeinsam beobachtet wurden. Statistisches Matching, das die Integration von zwei (oder mehreren) Datensätzen ermöglicht, bietet in solchen Situation eine Lösung. Eine notwendige Voraussetzung dafür ist, dass neben den Variablen, die spezifisch nur in einem der beiden Datensatz vorhanden sind, auch gemeinsame Variablen existieren, die in beiden Datensätzen beobachtet wurden. Diese gemeinsamen Variablen werden verwendet, um den Zusammenhang zwischen den spezifischen Variablen auf Basis der verfügbaren Daten zu schätzen. Dazu ist wichtig, dass die gemeinsamen Variablen gute Prädiktoren für die spezifischen Variablen sind. Ein populärer Weg, gemeinsame Information über nicht gemeinsam erhobene Variablen zu erhalten, basiert auf der Annahme, dass die spezifischen Variablen –bedingt auf die gemeinsamen Variablen– unabhängig sind.

Im Kontext der ersten drei Beiträge dieser kumulativen Dissertation werden neue Methoden für die kategoriale Datenintegration entwickelt, die auf dieser Annahme beruhen. Alle diese neuen Methoden bedienen sich einer Einbettung von statistischem Matching in die Theorie probabilistischer grafischer Modelle. Dabei bildet die bedingte Unabhängigkeitsannahme die zentrale Schnittstelle zwischen statistischem Matching und probabilistischen grafischen Modellen. Mithilfe gerichteter und ungerichteter Graphen werden Abhängigkeitsstrukturen zwischen Variablen dargestellt und eine geeignete Faktorisierung ihrer gemeinsamen Verteilung ermittelt. Dies ermöglicht die Schätzung einzelner Komponenten der gemeinsamen Verteilung auf unterschiedlichen Teilmengen der gegebenen Datenbasis. Ein weiterer Beitrag dieser Thesis nähert sich dem Problem des statistischen Matchings von kategorialen Daten mit einem vorsichtigeren Lösungsvorschlag, der ohne die Annahme der bedingten Unabhängigkeit auskommt. Es wird ein neues, mengenwertiges Imputationsverfahren vorgeschlagen, das die blockweise fehlenden Beobachtungen der spezifischen Variablen durch Mengen von plausiblen Werten ersetzt.

Beitrag 1 befasst sich mit der Schätzung von gerichteten, nicht-zyklischen Graphen auf Teilmengen der vorhandenen Daten. Es werden verschiedene Vorgehensweisen vorgeschlagen, wie diese Subgraphen miteinander zu einem gemeinsamen Bayesnetz kombiniert werden können. Basierend auf dem gemeinsamen, gerichteten Graphen werden diejenigen Faktoren über die Kettenregel für Bayesnetze bestimmt, die die gemeinsame Verteilung aller Variablen bestimmen. Dabei stellt die Annahme der bedingten Unabhängigkeit der spezifischen Variablen gegeben der gemeinsamen Variablen sicher, dass alle Faktoren aus den vorhandenen Daten geschätzt werden können.

Beitrag 2 entwickelt einen Ansatz zum statistischen Matching von kategorialen Daten, der auf einem ungerichteten probabilistischen grafischen Modell basiert. Mithilfe der log-linearen Entwicklung der Multinomialverteilung und der Interpretation des ungerichteten Graphen als Interaktionsgraph, wird ein Markovnetz mit log-linearer Parametrisierung für das statistische Matching hergeleitet. Wiederum gewährleistet die bedingte Unabhängigkeitsannahme, dass alle Komponenten der gemeinsamen Verteilung auf den vorhandenen Daten schätzbar sind.

Beitrag 3 befasst sich mit einem Spezialfall von *Beitrag 2*, nämlich der Integration von binären Daten mithilfe des Ising-Modells. Hierbei handelt sich um ein paarweises Markovnetz, das Interaktionen bis zur maximalen Ordnung zwei zulässt. Die Schätzung der gemeinsamen Verteilung kann für diesen Spezialfall deutlich vereinfacht werden.

Beitrag 4 interpretiert die Datensituation des statistischen Matchings als Problem fehlender Daten. Fehlende Beobachtungen der spezifischen Variablen werden bei der neu vorgeschlagenen unpräzisen Imputation durch Mengen von plausiblen Werten ersetzt. Auf Basis dieser –zum Teil mengenwertigen– Beobachtungen werden untere und obere Schranken für die Wahrscheinlichkeitskomponenten der gemeinsamen Verteilung von gemeinsamen und spezifischen Variablen berechnet. Als Basis für diese Schätzung dient die Theorie der Random Sets.

Abstract

The secondary analysis of already available data can save time, money or other resources. However, answering certain research questions may require joint information about variables that were not observed together. Statistical matching, which allows the integration of two (or more) data files, provides a solution for these situations. The prerequisite for this is that in addition to the variables that are only present in one of the two files, there are also common variables that were observed in both files. These common variables are used to estimate the relation between the specific variables based on the available database. For this purpose, it is important that the common variables are good predictors of the specific variables. A popular way of obtaining joint information about not jointly observed variables is premised on the assumption that the specific variables are conditionally independent given the common variables.

Based on this assumption, new methods for the integration of categorical data are developed in the context of the first three contributions of this cumulative dissertation. All of these new methods use an embedding of statistical matching into the theory of probabilistic graphical models. The conditional independence assumption provides the central interface between statistical matching and probabilistic graphical models. Using directed and undirected graphs, dependence structures between variables are represented and an appropriate factorization of their joint distribution is determined. This factorization allows the estimation of all components of the joint distribution of the common and specific variables on different subsets of the given data. A further contribution to this thesis approaches the problem of statistically matching categorical data with a more cautious solution that works without the assumption of conditional independence. A new, set-valued imputation method is proposed which replaces the block-wise missing observations of the specific variables with sets of plausible values.

Contribution 1 deals with the estimation of directed acyclic graphs on subsets of the available data, and proposes different ways of combining these subgraphs into a joint Bayesian network. On basis of this joint graph, the factors determining the joint distribution of all variables are obtained by the chain rule for Bayesian networks. The assumption of conditional independence of the specific variables given the common variables ensures that all factors are estimable from the available data.

Contribution 2 develops an approach for statistical matching of categorical data based on an undirected probabilistic graphical model. Using the log-linear expansion of the multinomial distribution and the interpretation of the undirected graph as an interaction graph, a Markov network with log-linear parameterization is derived. Again, the conditional independence assumption ensures that all components of the joint distribution are estimable on the existing data.

Contribution 3 deals with a special case of *Contribution 2*, namely the integration of binary data using the Ising model. This is a pairwise Markov network that allows only interactions up to the maximum order of two. The estimation of the joint distribution can be markedly simplified for this special case.

Contribution 4 interprets the data situation of statistical matching as a missing data problem. The newly developed imprecise imputation replaces the missing observations of specific variables by sets of plausible values. On the basis of these partially set-valued observations, lower and upper bounds are calculated for the probability components of the joint distribution of the common and specific variables. As basis for this estimation, we use the theory of random sets.

Acknowledgements

This work could only arise through the valuable support of several people. My special thanks go to . . .

- *. . . Thomas Augustin for giving me the opportunity to discover new research areas and writing this thesis, for taking all the time to discuss research questions, and for the pleasant collaboration. Thank you, Thomas, for your supervision, your motivation and your trust.*
- *. . . Moritz Grosse-Wentrup and Helga Wagner who kindly agreed to examine this dissertation.*
- *. . . Christian Heumann and Volker Schmid for being members of the examination committee.*
- *. . . Francesca Biagini, my mentor in the LMUMentoring program. Thank you for financial support, giving me the opportunity to visit conferences and workshops, and for very valuable conversations that have helped me to always remain confident.*
- *. . . Aziz Omar, Christoph Jansen, Cornelia Fütterer, Georg Schollmeyer, Johanna Brandt, Julia Plaß, Malte Nalenz, Patrick Schwaferts, Paul Fink, my former and current colleagues in the working group. Thank you for many helpful research discussions, for all chats, shared lunches, energizing coffee breaks, joyful Radlerrunden and for being such great team players in every way.*
- *. . . Katrin Newger, one of my co-authors, for a pleasant collaboration and nice chats with coffee and cake.*
- *. . . Melissa Schmoll and Katrin Newger for being the most reliable research assistants. It has always been a pleasure to work with you.*
- *. . . the TutorPlus-Team of the LMU München, especially Alice Murschall who supervised the sixth cohort, for giving me the opportunity to join the TutorPlus program. Thank you for allowing me to learn so much from you.*
- *. . . all members of the Department of Statistics for the great atmosphere and the choir.*
- *. . . Verena Bauer for sharing the office for a long and very nice time.*
- *. . . Jürgen Landes, Barbara Osimani and Roland Pöllinger from the Munich Center of Mathematical Philosophy for stimulating discussions and exciting slides in the parabolic slide in Garching.*
- *. . . my parents for all the encouragement, backing and love. Without you, I would never have made it that far.*
- *. . . Stefan for all the support, for years of patiently enduring all moods, and for giving me the biggest gift of all.*

Author's contributions

The present cumulative dissertation is composed of the following four contributions, each of which aims at the development of a new statistical matching procedure for categorical data. They are arranged according to the order they appear in this thesis:

- *Contribution 1:*

  Endres, E. and Augustin, T. (2016). Statistical matching of discrete data by Bayesian networks, in A. Antonucci, G. Corani and C. P. de Campos (eds), Proceedings of the Eighth International Conference on Probabilistic Graphical Models, Vol. 52 of Proceedings of Machine Learning Research, PMLR, Lugano, Switzerland, pp. 159–170.

  The embedment of statistical matching in the theory of probabilistic graphical models was created by Eva Endres. Furthermore, the estimation of parts of the graph structure on different subsets of the data and their subsequent combination was developed by her. Thomas Augustin helped with the notation of the synthetic joint distribution. The procedure for the generation of a complete synthetic data set as well as the example application on data of the German General Social Survey comes from Eva Endres. Both authors contributed to the revision.

- *Contribution 2*:

  Endres, E. and Augustin, T. (2019). Utilizing log-linear Markov networks to integrate categorical data files, Technical Report 222, Department of Statistics, LMU Munich. Submitted to *Statistica Neerlandica*, date of submission: *January 11th, 2019.*

  The article is a natural continuation of *Contribution 1.* The content was developed and written by Eva Endres. Thomas Augustin aided with revisions and valuable discussions.

- *Contribution 3*:

  Endres, E., Newger, K. and Augustin, T. (2019). Binary data fusion using undirected probabilistic graphical models: Combining statistical matching and the Ising model, Technical Report 223, Department of Statistics, LMU Munich.

  The introduction to statistical matching and graphical models were written by Eva Endres. The chapters on data integration with the Ising model was mainly written, and the simulation study was designed and conducted by her. The idea of looking at the special case of binary data using the Ising model, goes back to Katrin Newger. *Contribution 3* is an extension of her master's thesis (Newger, 2018) and covers a special case of the model in *Contribution 2.* All authors contributed to the revision.

- *Contribution 4*:

  Endres, E., Fink, P. and Augustin, T. (2018). Imprecise imputation: A nonparametric micro approach reflecting the natural uncertainty of statistical matching with categorical data.
  Accepted for publication in the *Journal of Official Statistics.*

  The basic idea of the article, to replace sets of plausible values for missing data to obtain sets of estimates, comes from Eva Endres. The same applies to the three different imputation approaches, which represent different levels of inclusion of dependence structures. Furthermore, the simulation was designed, implemented into code and carried out by her. Paul Fink helped to speed up the calculation and wrote the code for the evaluation of the simulation. The embedding of imprecise imputation into the theory of random sets was accomplished by Paul Fink and aided by Thomas Augustin. All authors contributed to revisions.

Contents

1 Statistical matching

Today, with advancing technologies, it is possible to collect and store more and more data. If collecting new data is very costly, time-consuming, or, for example, medically invasive, it can be of great benefit if existing data can be used for a secondary analysis rather than collecting new data. However, it may happen that there is no existing dataset that contains information about all the variables needed for a particular analysis. In these situations, *statistical matching* might be the solution how the available data could be used anyhow. It aims at the integration of two or more data files sharing a common core of variables. Consider the following two prototypical examples[1] from official statistics and biostatistics.

1.1 Motivating examples

Eurostat is the statistical office of the European Union (Eurostat, 2018). The members of Eurostat are inter alia analysing data collected by the National Statistical Institutes of the EU member states for Europe-wide comparisons. Amongst others, they are also analysing the poverty in the member states. One of the *Europe 2020* (e.g. Eurostat, 2019a) aims is to lift “at least 20 million people out of the risk of poverty or social exclusion” (Eurostat, 2019b). For the investigation and comparison of the actual poverty among different countries of the EU, Serafino and Tonkin (2017b) compare *income*, *expenditure*, and *material deprivation* as convenient operationalizations of poverty. However, since there exists no single data source containing information on all of these variables, they statistically match the *EU-SILC* (*European Union Statistics on Income and Living Conditions*) data with information on material deprivation and income with the *HBS* (*Household Budget Survey*), including information on the expenditure and income. Figure 1.1 shows the data situation for this application (adapted from Serafino and Tonkin, 2017a). Subsequently, they analyse and compare the different poverty measures for Austria, Belgium, Finland, Germany, Spain and the United Kingdom on basis of the synthetic data file derived by statistical matching. Details and results can be found in Serafino and Tonkin (2017b) and Serafino and Tonkin (2017a).

[1] More examples on applications of statistical matching can be found, for instance, in D’Orazio et al. (Chap. 7 2006b).

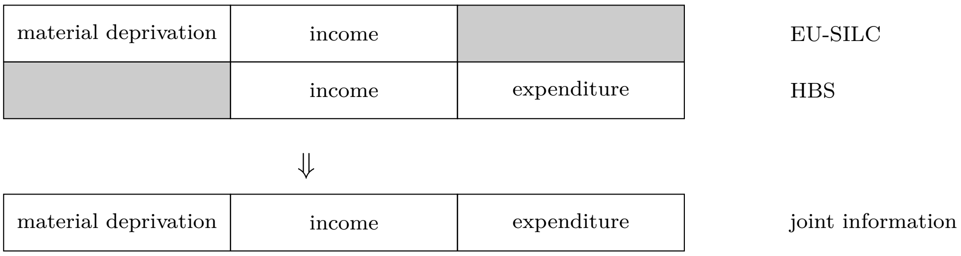


Figure 1.1: Simplified sketch of the Eurostat application for statistical matching (Serafino and Tonkin, 2017b).

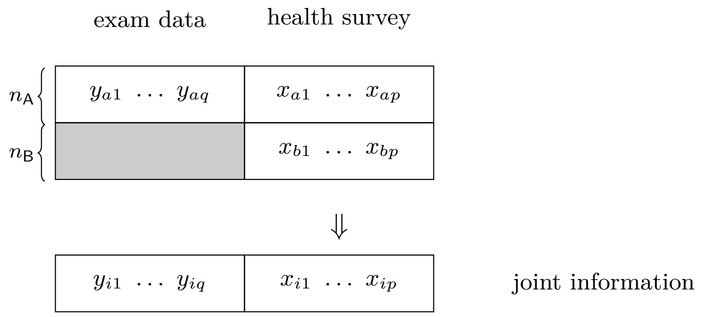


Figure 1.2: Schematic representations of the special data situation in Aluja-Banet et al. (2015).

A second example comes from an application in biostatistics. Aluja-Banet et al. (2015) aim at the estimation of the prevalences of cardiovascular diseases and diabetes. For this purpose, they want to use a large-scale survey (*2006 Health Survey of Catalonia*) which is based on interviews and self-reported answers. However, this kind of data is expected to be inaccurate due to misreportings during the data collecting process, leading to biased results. For this reason, they investigate whether an additional exploitation of detailed clinical information coming from a *health exam data*, which is available for a subsample of this survey, leads to more accurate estimates for the prevalences. The basic data situation of this example is slightly different to the previous. It is depicted in Figure 1.2. Aluja-Banet et al. (2013) claim that this data situation is the simplest in the context of data fusion; they call it *unilateral* fusion. Although this data situation is by definition not directly a statistical matching problem, statistical matching methods can be –and are actually– applied to impute the block of missing entries in the health exam data with the aim to obtain the desired information in this context.

A precise definition of what is understood by the term *statistical matching* in the context of this thesis is given in the following section. Furthermore, notations are clarified and an overview about already existing approaches is given.

An overview on general data situation where statistical matching techniques can be applied, like *database enrichment* corresponding to the previous example by Aluja-Banet et al. (2015), a general missing data situation with *variables missing in groups*, or *split questionnaire survey designs* can be found in Rässler (2004).

1.2 The general data situation, aims and types of already existing approaches

Statistical matching, which is also known under the terms *data fusion*[2], *data merging*, *data matching*, *mass imputation*, *file concatenation* (Rässler, 2002, p. 2), *file matching*, (Little and Rubin, 2002, p. 7), or *synthetical matching* (D'Orazio et al., 2006b, pp. 1–2) means, according to D'Orazio et al. (2006b, p. 2), the integration of two or more data files

(i) that share a set of *common variables* contained in both files;

(ii) each of which contains a set of *specific variables* only observed in one of these files;

(iii) whose sets of observation units are disjoint from each other.

[2] Due to Rässler (2002, p. 2), this term in mainly used in Europe.

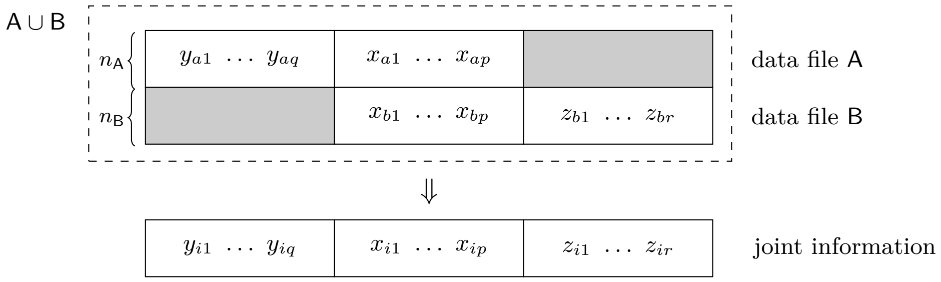


Figure 1.3: Schematic representations of the typical data situation for statistical matching (adapted from D'Orazio et al., 2006b, p. 5).

Figure 1.3 illustrates this data situation visually. It can be formalised as follows.

Throughout this thesis, I consider two[3] data files, A and B, which are composed of n_{A} or n_{B} i.i.d. observations on three sets of categorical[4] variables, $\boldsymbol{X} = \{X_1, \ldots, X_p\}$, $\boldsymbol{Y} = \{Y_1, \ldots, Y_q\}$, $\boldsymbol{Z} = \{Z_1, \ldots, Z_r\}$, coming from a joint probability mass distribution

$$\pi(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) := \mathbb{P}(X_1 = x_1, \ldots, X_p = x_p, Y_1 = y_1, \ldots, Y_q = y_q, Z_1 = z_1, \ldots, Z_r = z_r). \tag{1.1}$$

To distinguish different observation units, they are indexed by $a \in \mathcal{I}_{\mathsf{A}}$ if they hail from file A, and they are indexed by $b \in \mathcal{I}_{\mathsf{B}}$ if they hail from file B. As previously mentioned, $\mathcal{I}_{\mathsf{A}} \cap \mathcal{I}_{\mathsf{B}} = \emptyset$. Within this thesis I will deal exclusively with categorical data. For statistical matching with continuous data, see, for instance D'Orazio et al. (2006b) or Rässler (2002).

As Figure 1.3 indicates, statistical matching can be interpreted as a missing data problem with a special characteristic. The complete block of observations for $\boldsymbol{Y}$ is missing in B, and the block of observations for $\boldsymbol{Z}$ is missing in A. Consequently, the variables $\boldsymbol{Y}$ are called the *specific variables* of A, and $\boldsymbol{Z}$ are the *specific variables* of B. The *common variables* $\boldsymbol{X}$ are observed in both, A and B. Thus, there exists no observation, which simultaneously yields information on all specific variables. Within the context of statistical matching, it can justifiably be assumed that the missing data are at least[5] missing at random (Rässler, 2002, p. 7). Following Little and Rubin (2002, p. 12) this means, that the missingness of a data entry is only dependent on the observed data. Since, the missingness is induced by the sample design and thus deterministic, D'Orazio et al. (2006b, p. 6) even argue that the missing mechanism is MCAR (missing completely at random). It means that the missingness of a data entry is independent from all observed and unobserved values.

As already stated, in this special missingness pattern, there exists no single observation with joint information on $\boldsymbol{Y}$ and $\boldsymbol{Z}$. This fact inevitably leads to a natural uncertainty underlying the task of statistical matching. Since the term *uncertainty* plays a central role in connection with statistical matching, its use will be explained explicitly in the following.

Basically, two sources of uncertainty[6] can be distinguished, depicted in Figure 1.4: *sampling uncertainty* and *identification uncertainty* (e.g. Conti et al., 2016, D'Orazio et al., 2006a). On

[3] The integration of more than two data files is straightforward. For simplicity, I restrict myself to two files in this thesis.

[4] The basic situation is the same for other types of data. However, as mentioned above, in this thesis, I solely consider categorical data fusion.

[5] Concretely, this means that the mechanism is either *missing completely at random* or *missing at random.*

[6] Koopmans (1949), who is also cited by Manski (1995, p. 6), coined the term *identification* and also considers the difference between the uncertainty coming from a finite number of observations and the uncertainty arising from identification problems. Manski (1995, p. 4) separates the statistical inference into a *statistical component* and *identification component.*

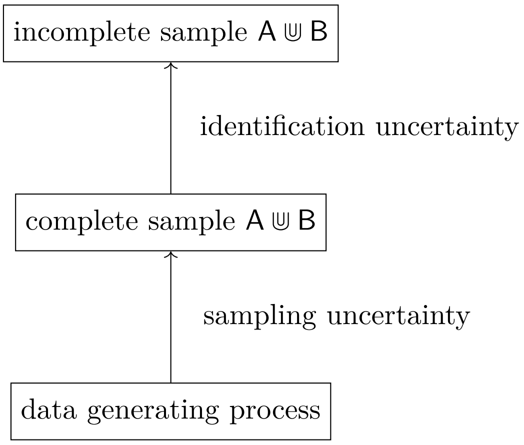

incomplete sample $\mathsf{A} \uplus \mathsf{B}$
identification uncertainty
complete sample $\mathsf{A} \uplus \mathsf{B}$
sampling uncertainty
data generating process

Figure 1.4: Layers of uncertainty in the statistical matching context.

the one hand, as statisticians we usually deal with a finite number of observations in a sample from which we want to estimate, for instance, the true underlying probability distribution. In this context, the source of uncertainty comes from the random variability in our limited sample data. However, this kind of sampling uncertainty is –as usual in the standard literature on statistical matching– not considered further in this thesis. From now on and for the rest of this thesis, I assume that the sampling process is error-free and that the samples A and B yield perfect information on the marginal distributions. On the other hand, since statistical matching is based on the fact that there is no observation with joint information on $\boldsymbol{Y}$ and $\boldsymbol{Z}$, the joint distribution of $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ is not identifiable. Even infinite samples A and B with complete information on the marginals $(\boldsymbol{X}, \boldsymbol{Y})$ and $(\boldsymbol{X}, \boldsymbol{Z})$, would not yield an (point)-identifiable model for $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ (e.g. D'Orazio et al., 2006a). For this purpose, we would need to make further assumptions or use auxiliary information regarding the connection between the specific variables.

D'Orazio et al. (2006b) give an overview about already existing statistical matching approaches that can be used to obtain joint information on variables that have not been jointly observed. These approaches can be divided into the following three categories:

(i) approaches, which build on the conditional independence of the specific variables given the common variables;

(ii) approaches that need auxiliary information about the connection between $\boldsymbol{Y}$ and $\boldsymbol{Z}$ in terms of a third (in)complete file or information about parameters concerning the relation between the specific variables;

(iii) approaches, which approve the natural uncertainty of statistical matching, and which can be summarized under the umbrella term *partial identification*.

Independently of the category of approaches, statistical matching distinguishes between two basic aims (e.g. D'Orazio et al., 2006b, p. 2). On the one hand, the goal might be a complete (but synthetic[7]) data file containing observations on all variables of interest. This aim can be achieved by imputation techniques[8] that fill in every missing entry in the data file by a plaus-

[7] A complete file created by statistical matching is not a real data file with entries coming from observations, but it contains (partly) synthetic observations derived by imputation. As formulated, for instance, by Aluja-Banet et al. (2013, p. 124), these observations are contaminated by uncertainty.

[8] To account for the uncertainty coming from the replacement of missing data entries, Little and Rubin (2002, Chap. 5) and Rässler (2002, Chap. 4) recommend to use, for example, multiple imputation techniques, also in the context of statistical matching.

ible[9], synthetic value (e.g. Little and Rubin, 2002, p. 20). The resulting file could subsequently be used with standard statistical analyses. This goal of statistical matching is summarized under the term *micro approach.* On the other hand, statistical matching might directly aim at the estimation of the joint distribution of $(\boldsymbol{Y}, \boldsymbol{Z})$ or $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$. This objective is termed *macro approach* and it involves also, for instance, the estimation of parameters describing the relationship between the specific variables. Considering categorical data, these parameters can, for example, be the probability components of the joint probability mass distribution concerning the specific variables. In the context of continuous data, this includes, for example, the estimation of the correlation between the specific variables.

In the following, I will give a brief overview about already existing approaches and their allocation into the three categories listed by D'Orazio et al. (2006b).

The first group of approaches is based on an assumption, namely the conditional independence of $\boldsymbol{Y}$ and $\boldsymbol{Z}$ given $\boldsymbol{X}$, in short $\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Z}|\boldsymbol{X}$. The assumption is helpful in the context of statistical matching since it yields an identifiable model for $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ on $\mathsf{A} \uplus \mathsf{B}$ (e.g. D'Orazio et al., 2006b, p.13). Within the macro approach, applying the chain rule and the assumption of conditional independence, the joint probability distribution for sets of categorical variables simplifies as following:

$$\begin{aligned}\pi(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) &= \pi(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{x}) \cdot \pi(\boldsymbol{z}|\boldsymbol{x}) \cdot \pi(\boldsymbol{x}) \\ &= \pi(\boldsymbol{y}|\boldsymbol{x}) \cdot \pi(\boldsymbol{z}|\boldsymbol{x}) \cdot \pi(\boldsymbol{x}). \end{aligned} \tag{1.2}$$

The resulting factors can be estimated from $\mathsf{A} \uplus \mathsf{B}$[10] since now there is no term which is simultaneously dependent on any components in $\boldsymbol{Y}$ and $\boldsymbol{Z}$. Concretely, $\pi(\boldsymbol{y}|\boldsymbol{x})$ is estimated from A, $\pi(\boldsymbol{z}|\boldsymbol{x})$ is estimated from B, and $\pi(\boldsymbol{x})$ is estimated from $\mathsf{A} \uplus \mathsf{B}$. For instance, D'Orazio et al. (2006b, Chap. 2.1; Chap. 2.3) show the application of the parametric macro approach for the multinomial distribution and the multivariate normal distribution, and the nonparametric macro approach using the empirical cumulative distribution function for categorical data and a kernel density estimator for continuous data. More prominent than the macro approaches, are the micro approaches that assume the conditional independence of the specific variables given the common variables. This is because all imputation approaches, which exploit the common variables to generate a connection between $\boldsymbol{Y}$ and $\boldsymbol{Z}$, fall into this category. Thus, all early applications that aimed at the construction of a complete synthetic data file, were based on this assumption (D'Orazio et al., 2006b, p. 13; p. 34). The main disadvantage of approaches belonging to the first group is that the assumption of conditional independence cannot be tested on the available data $\mathsf{A} \uplus \mathsf{B}$. This would require joint information about $\boldsymbol{Y}$ and $\boldsymbol{Z}$. An external, additional data source could help. Moreover, this kind of imputation approaches even establish a conditional independence in the synthetic file (Rässler, 2002, p.4). More detailed information and examples for this type of statistical matching approaches, can be found in D'Orazio et al. (2006b, Chap. 2). Three new statistical matching approaches that fit into this category were developed within the scope of this dissertation. They are presented in *Contributions 1–3.*

Since the assumption of conditional independence is not testable on the data at hand, the second group of approaches that additionally needs external information on the relationships among the specific variables to derive estimates based on point-identification for the joint probability

[9] In this context, I use the term 'plausible' value to refer to a value that could have been the true (unknown) value.

[10] Taking the missing mechanism into account which is either missing at random or missing completely at random, the information in the incomplete sample $\mathsf{A} \uplus \mathsf{B}$ is representative for the unknown complete sample (Pigott, 2001).

distribution (e.g. D'Orazio et al., 2006b, p. 67), might be an alternative. Auxiliary information may, for example, be available as another complete (containing observations on $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$) or incomplete data file (containing observations on $(\boldsymbol{Y}, \boldsymbol{Z})$) decoding the relationship between the specific variables, or there may be information about parameters describing the relationship between the specific variables. For example, D'Orazio et al. (2006b, Chap. 3) show different parametric and nonparametric micro and macro approaches, which are based on the usage of auxiliary information. An exemplary application is found in Singh et al. (1993).

However, as stated above, the conditional independence assumption cannot be tested and auxiliary information might not be available for a certain statistical matching task. For this reason, also a third group of approaches is considered. It is different compared to the previous two because it does not aim at estimates based on point-identifiability. D'Orazio et al. (2004) describe this type of approaches as "assessing all *the possible worlds*, i.e. all parameters' values consistent with the available information". I summarize the approaches within this group under the term *partial identification*[11]. In the context of the macro approach, this means that although the parameters are not point-identifiable on $\mathsf{A} \Cup \mathsf{B}$, the available information can be used to find at least meaningful lower and upper bounds for these parameters, thus resulting in sets of parameter estimates[12] where each element is compatible with the available information. For instance, D'Orazio et al. (2006b, Chap. 4) and Rässler (2002) consider this type of statistical matching procedures in detail. In the notion of the statistical matching micro approach, the aim is the creation of sets of complete synthetic data files (e.g. D'Orazio et al., 2006b, p. 97). *Contribution 4* of this thesis is part of this group of approaches and the first statistical matching micro approach for categorical data in this context.

Especially the approaches of the last type have experienced some developments in the recent years. In the following section I will give a brief overview of current literature on statistical matching.

1.3 Short overview on current literature

Within the context of statistical matching, there are two monographs which give a detailed insight into the data situation, theory and already existing statistical matching approaches. The first one is written by Rässler (2002) and includes but is not limited to measurable objectives of statistical matching. Another central aspect of this monograph is the introduction of new statistical matching approaches based on multiple imputation techniques. The second monograph from D'Orazio et al. (2006b) is a guide for theoretical and practical applications of statistical matching. It gives an overview on different statistical matching techniques, practical issues and applications of statistical matching. Below I will give a short overview of recent articles on the topic of statistical matching.

For instance, Rässler (2002, pp. 1–2, Chap. 3) and Conti et al. (2017) give overviews on literature regarding statistical matching between the early 1970's and the beginning of the 2000's. While

[11] As stated by Di Zio and Vantaggi (2017), this type of approaches is also named *uncertainty analysis* in official statistics.

[12] A characteristic of these sets is that no element of it can be preferred since all of the elements in the set are equally plausible. The term 'equally plausible' should not suggest a (uniform) distribution over the elements of these sets but simply express that every element could have potentially generated the incomplete file $\mathsf{A} \Cup \mathsf{B}$. Every parameter which lies between the lower and upper bound produces a candidate for the joint distribution compatible with the available marginals in A and B. The solution under the conditional independence assumption is also contained in these sets.

Anderson (1957) seems to be the first who considered the case of maximum likelihood estimation with missing data for the multivariate normal distribution, the first applications which can be assigned into the context of statistical matching, were based on the imputation of missing values. Okner (1972) used imputation based on *donation classes*[13], to achieve information on demographics, income and income tax, for both low and high income individuals who have only been separately observed. More recent applications are mainly available from Eurostat, where data from the *EU-SILC* and the *HBS* are statistically matched to compare people's exposure to poverty as described in Section 1.1 (Serafino and Tonkin, 2017a,b, Webber and Tonkin, 2013). Furthermore, they matched the *EU-SILC* with the *European Quality of Life Study* to achieve information on different dimensions on quality of life (Leulescu and Agafitei, 2013, Chap. 2), and integrated the *EU-SILC* data with the *Labour Force Survey* to jointly analyse labour market information and employment-related income (Leulescu and Agafitei, 2013, Chap. 3).

Besides applications, there have been a few recent publications on the methodology of statistical matching, which mainly deal with uncertainty and how it can be measured, and practical issues like selecting appropriate *matching variables*[14]. These approaches can be allocated into the above mentioned group of partial identification approaches.

First introduced by Rubin (1978), multiple imputation was incorporated into the context of statistical matching for multivariate normally distributed data in Rässler (2002) and Rässler (2004). The multiple replacements of the missing entries are used to achieve lower and upper bounds for the parameters of interest and a measure of uncertainty is defined. D'Orazio et al. (2004) formalize the concept of uncertainty for categorical variables, which is elaborated in D'Orazio et al. (2006a). Here, the basic idea is to find all possible joint distributions which are compatible with the available conditional and unconditional marginals. Furthermore, it is shown how logical constraints can be used to reduce uncertainty. However, as analysed by Vantaggi (2008), the introduction of logical constraints may lead to incoherences[15]. Brozzi et al. (2012) show how to overcome these incoherences by, amongst others, a minimization approach which aims at finding the coherent estimates which are closest to the available information in A and B. Extensions of the ideas raised by D'Orazio et al. (2006a) can be found in Conti et al. (2012), who provide a concept of uncertainty and how to measure it in a parametric and nonparametric context. Conti et al. (2017) study the concept of uncertainty in a nonparametric context. A measure of uncertainty for ordered categorical data is defined by Marella et al. (2012). The data situation of ordered categorical data is studied in detail in Conti et al. (2013) by defining the concept of uncertainty, giving a definition of a measure of uncertainty for ordered categorical data and by explaining how uncertainty can be reduced by logical constraints.

In recent years, there have also been some approaches to solving practical problems in statistical matching, such as the measurement of the *matching noise*[16], for example, in Conti et al. (2008). Another practical issue is the actual choice of the matching variables. For instance, D'Orazio et al. (2017) deal with this problem by selecting those common variables which reduce the uncertainty between the specific variables. A very specific issue that may arise in statistical matching applications have been addressed in Di Zio and Vantaggi (2017), who consider the case of misclassified common variables. Furthermore, the practical issue of the incorporation of sampling weights into the procedure of statistical matching is considered in Conti et al. (2016). Other publications deal with the handling of auxiliary information, which may be present in

[13] A donation class is a subset of homogeneous observations with similar realizations in the matching variables.

[14] Matching variables are those common variables which are actually chosen to be used for statistical matching purposes, i.e. for establishing a connection between the specific variables (e.g. D'Orazio et al., 2017).

[15] In short, it means that it is possible that there is not a single distribution that is compatible with the given marginals while satisfying the logical constraints.

[16] Matching noise is defined as the discrepancy the data generating process and the joint distribution in the complete synthetic data file (Marella et al., 2008).

some cases. In this category is Ahfock et al. (2016), who include auxiliary information for (high-dimensional) multivariate normally distributed data to find positive-definite completions of the partially determined covariance matrix. This type of publication also includes Zhang (2015), who considers the special case of proxy variables being available for the specific variables. Further practical issues as the harmonization of the two data files that should be matched, the choice whether A, B, or $\mathsf{A} \uplus \mathsf{B}$ should be imputed applying the micro approach, or the assessment of the quality of a statistical matching method are, for instance, considered in D'Orazio et al. (2006b, Chap. 6, Chap. 7; Chap. 2.4; Chap. 1.4).

1.4 Dissociation of selected other topics: record linkage, propensity score matching, sensor data fusion, and missing data

Since the terms 'matching' and 'fusion' also appear in other contexts in statistics, this section should give brief differentiations from selected other topics.

The most similar topic to statistical matching is *record linkage*, also known under the terms *data matching*, *data linkage*, *entity resolution*, *object identification* or *field matching* (Christen, 2012, p. ix). Although, as mentioned by Rässler (2002, p. 6) the roots of record linkage and statistical matching are historically related, in the context of record linkage, the files A and B contain identical individuals, i.e. $\mathcal{I}_{\mathsf{A}} \cap \mathcal{I}_{\mathsf{B}} \neq \emptyset$. Thus, one main task of record linkage is the *de-duplication* or *duplicate detection* of observations in A and B, sometimes under the aid of (unique) identifiers. The main consequence of A and B containing the same individuals is that the resulting file $\mathsf{A} \uplus \mathsf{B}$ is not necessarily synthetic. If the aim of matching the same individuals was reached, the file $\mathsf{A} \uplus \mathsf{B}$ could have been observed in reality.

Another area which can easily be mixed with statistical matching is *propensity score matching*, which is especially popular in biostatistics. Following, for instance, D'Agostino, Jr. (1998), propensity score matching can be used to estimate unbiased treatment effects in non-randomized observational studies[17]. Originating from Rosenbaum and Rubin (1983), propensity score matching was embedded into the context of statistical matching by Rässler (2002). Since it aims at finding *statistical twins*, one in the treatment group and one in the control group, the propensity score can also be used to find a suitable[18] donor record in B whose entries for the specific variables are used to replace missing entries[19] of $\boldsymbol{Z}$ in the corresponding recipient record in A.

In statistics, we can also find another subject, which is sometimes called 'data fusion' (e.g. Elmenreich, 2002, p. 8): *(multi)sensor fusion*. It aims at the integration of information arising from multiple (different) sensors (or more general, information sources) to achieve 'better' overall information forming a unified picture (Khaleghi et al., 2013). In this context, 'better' can mean, for instance, more precise measurements or a reduction of missing information (e.g. Khaleghi et al., 2013).

Last but not least, statistical matching should also be dissociated from the more general area of missing data. As mentioned above, statistical matching can be interpreted as a missing data problem (for details, see, for instance D'Orazio et al., 2006b, Chap. 1.3). However, it is a special,

[17] Of course, unbiased estimates can only be achieved for known and observed confounding covariates (e.g. Kuss et al., 2016). Otherwise, only randomization can guarantee equal distributions in the different groups that are to be compared.

[18] The term 'suitable' means that the donor and recipient records are similar regarding their common variables.

[19] More details on imputation in the context of statistical matching can be found in Section 3.4.

and rather extreme form[20] of missing data since we additionally face the identification problem. In the context of statistical matching, missing data methods as *complete-case analysis*[21] or *available-case analysis*[22] are not applicable since we have no single observation with joint information on the specific variables, as discussed above. This is the reason why we additionally have to assume the conditional independence of $\boldsymbol{Y}$ and $\boldsymbol{Z}$ given $\boldsymbol{X}$, use auxiliary information, or partial identification approaches to establish a relationship between $\boldsymbol{Y}$ and $\boldsymbol{Z}$.

1.5 Contribution of this thesis to statistical matching: statistical matching meets probabilistic graphical models

This thesis deals exclusively with the development of new methods for the integration of categorical data. *Contributions 1–3* are based on the assumption of conditional independence, which is represented by probabilistic graphical models. Although it can be argued that this assumption is unjustified, there certainly exist data situations in which there are enough common variables that are good predictors for the specific variables supporting the conditional independence. As already mentioned by Tsamardinos et al. (2012), if "the number of common variables is large it is unlikely that Y provides additional information for Z, than what $\boldsymbol{X}$ already provides". For example, in a survey context, demographic variables might be a good choice for the common variables; or, in a medical context, clinical variables as age, different blood values, gender, or pre-existing diseases.

Probabilistic graphical models are able to visualize the conditional and unconditional dependencies among the sets of common and specific variables. This allows an intuitive access to the relevant field of knowledge. In addition, already existing expert knowledge about possible dependence structures or parameter values can be directly included. The same holds for auxiliary information in terms of a third complete or incomplete file, or in terms of prior knowledge on certain parameters. If no prior knowledge is available, it is possible to resort to existing and well-researched structure estimation algorithms to learn the dependencies between the variables of interest in a data-driven manner. Another advantage of these models is that –although not explicitly used in this work– they are also suitable for continuous variables and for data files with mixed categorical and continuous variables. Furthermore, structure learning algorithms for probabilistic graphical models are able to automatically select the most appropriate matching variables for each specific variable. Not every common variable is necessarily connected to every specific variables as it happens to be in standard statistical matching approaches. Certain common variables are only used as matching variable if the structure learning algorithm for the graphical model has found a direct dependence between them and at least one specific variable.

Contribution 4 of this thesis introduces a very general imputation approach, which can also be applied in situations where the conditional independence assumption is unjustified and no auxiliary information is available. It respects the identification problem of statistical matching without forcing estimates based on point-identifiability. It is the first statistical matching procedure of this type which aims at the construction of sets of synthetic data files in the context of the micro approach. The main advantages of this approach are that it is easy to apply, the components of any conditional or unconditional distribution are straightforwardly estimated by relative frequencies, and auxiliary information in terms of *logical constraints* (see, for instance,

[20] For instance, Little and Rubin (2002, p. 5) show different types of missing data patterns.

[21] Complete-case analysis excludes all observations with missing entries from the analysis (Little and Rubin, 2002, Chap. 3).

[22] Available-case analysis uses all individuals for the analysis for which the variables of interest have been observed (Little and Rubin, 2002, Chap. 3).

D'Orazio et al., 2006a, for a detailed a definition and example of the term logical constraints) can be incorporated. Additionally, an R-package called `impimp` was implemented by Fink and Endres (2019) to make the *imprecise imputation* approach of *Contribution 4* accessible for all potential users.

In the following chapter, I will first of all recapitulate all necessary basic concepts of probabilistic graphical models before establishing the link to statistical matching.

2 Probabilistic graphical models

In short, probabilistic graphical models are used to represent joint distributions in a simple and compact way, exploiting existing (conditional) independences among the set of considered random variables. This type of models is applied in various contexts. Pourret et al. (2008), for instance, demonstrate the versatility of applications. To set the course for this chapter, I will first clarify some basic concepts and introduce the corresponding notation.

2.1 Basic concepts of probabilistic graphical models

In general, a graph consists of a set of *nodes* and a set of *edges*. To be consistent with most of the literature on graphical models, I denote the set of nodes with $\dot{\boldsymbol{X}} = \{\dot{X}_1, \ldots, \dot{X}_p\}$[23]. Two elements of $\dot{\boldsymbol{X}}$, for instance $\dot{X}_j$ and $\dot{X}_{j'}$, can be connected by an *undirected edge* or a *directed edge*, as displayed in Figure 2.1. The set of edges, in the following denoted by $\boldsymbol{E}$, corresponding to a graphical model, is a subset of the Cartesian product $\dot{\boldsymbol{X}} \times \dot{\boldsymbol{X}}$. A directed edge $\dot{X}_j \rightarrow \dot{X}_{j'}$ is denoted as the pair $(\dot{X}_j, \dot{X}_{j'})$ being an element of $\boldsymbol{E}$. Analogously, for an oppositely directed edge, $(\dot{X}_{j'}, \dot{X}_j) \in \boldsymbol{E}$. For an undirected edge $\dot{X}_j - \dot{X}_{j'}$, either the pair $(\dot{X}_j, \dot{X}_{j'})$ or the pair $(\dot{X}_{j'}, \dot{X}_j)$ is an element of the corresponding edge set $\boldsymbol{E}$. Throughout this thesis, I solely consider graphical models that contain only one type of edges, either directed or undirected. Undirected graphs are in the following denoted by $\mathcal{H}$, while directed graphs are denoted by $\mathcal{G}$.

Before moving to probabilistic graphical models, some basic terms have to be clarified. All of the following definitions can, for instance, be found in Koller and Friedman (2009), Murphy (2012), or Lauritzen (1996). Considering a directed edge $\dot{X}_j \rightarrow \dot{X}_{j'}$, we distinguish between the *parent* node $\dot{X}_j$ and the *child* node $\dot{X}_{j'}$. The arrow, representing the edge, points from the parent to the child. All previous nodes with direct or indirect connections pointing to a certain node, i.e. all parents, the parents of the parents, and so on, are summarized as *ancestors* of this node. Vice versa, all children, grand-children etc., are called *descendants* of this node. For undirected edges, we just differentiate whether nodes are *adjacent* or not, i.e. $\dot{X}_j$ and $\dot{X}_{j'}$ are *neighbours* if there is a direct connection $\dot{X}_j - \dot{X}_{j'}$ between them.

A set of directed or undirected edges builds a *path* between $\dot{X}_1$ and $\dot{X}_p$ via $\dot{X}_2, \dot{X}_3, \ldots, \dot{X}_{p-1}$ if, for every $j = 1, \ldots, p-1$, there exits a directed edge $\dot{X}_j \rightarrow \dot{X}_{j+1}$ or an undirected edge $\dot{X}_j - \dot{X}_{j+1}$, respectively. The concept of a *trail* is similar but the directionality is neglected there. This means that there is a trail between $\dot{X}_1$ and $\dot{X}_p$ via $\dot{X}_2, \dot{X}_3, \ldots, \dot{X}_{p-1}$ if, for every $j = 1, \ldots, p-1$, there exits either a directed edge $\dot{X}_j \rightarrow \dot{X}_{j+1}$, or $\dot{X}_j \leftarrow \dot{X}_{j+1}$, or an undirected edge $\dot{X}_j - \dot{X}_{j+1}$. For example, $\dot{X}_1 \rightarrow \dot{X}_2 \rightarrow \dot{X}_3$ is a path between $\dot{X}_1$ and $\dot{X}_3$, while $\dot{X}_1 \rightarrow \dot{X}_2 \leftarrow \dot{X}_3$ describes a trail from $\dot{X}_1$ to $\dot{X}_3$.

After clarifying these basic concepts, we can now describe what characterizes a probabilistic graphical model. In general, a probabilistic graphical model consists of two components:

[23] This notation is so far unrelated to the notation of common variables used in the first chapter. The connection is first made in Section 2.4.

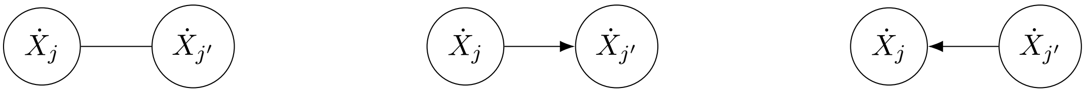


Figure 2.1: Possible undirected and directed connections between nodes in a graphical model.

1. a graph, whose nodes $\dot{\boldsymbol{X}}$ represent random variables $\boldsymbol{X}$ and whose structure represents information about (conditional) independence relations among these variables, and
2. a joint probability distribution $\mathbb{P}$ over $\boldsymbol{X}$.

The information about independencies provided by the graph, is used to find a compact representation of the joint probability distribution which simplifies computations considerably by factorizing $\mathbb{P}$ into smaller and easy calculable pieces.

In the following, I will focus on two types of probabilistic graphical models, starting with *Bayesian networks*, which build on a *directed acyclic*[24] *graph* (*DAG*). After recapitulating how the factorization in the context of Bayesian networks proceeds and why it is feasible, I will continue with *Markov networks* (also known as *Markov random fields*) which are based on an undirected graph structure.

2.2 Bayesian networks

Summarizing, a Bayesian network is a pair $(\mathcal{G}, \mathbb{P})$, where $\mathcal{G}$ encodes a set of conditional probability assertions which forms the basis of a factorization of the *global (joint) probability distribution* $\mathbb{P}$ over $\boldsymbol{X}$ into *local probability models*. These local probability models are in fact conditional probability distributions, one for each node depending on its parent's realizations[25], according to a graph $\mathcal{G}$ (e.g. Koller and Friedman, 2009, Chap. 3). Thus, every node in the graph is linked to a conditional probability table reflecting the conditional distribution of this node given the configuration of its parents. Assuming that every node is conditionally independent of its non-descendants given its parents (which is the so-called (local) *Markov assumption* (e.g. Murphy, 2012, p. 310)), the well-known chain rule for probabilities can be transferred into the *chain rule of Bayesian networks*. It factorizes the global probability distribution of a Bayesian network with graph $\mathcal{G} = (\dot{\boldsymbol{X}}, \boldsymbol{E})$ as

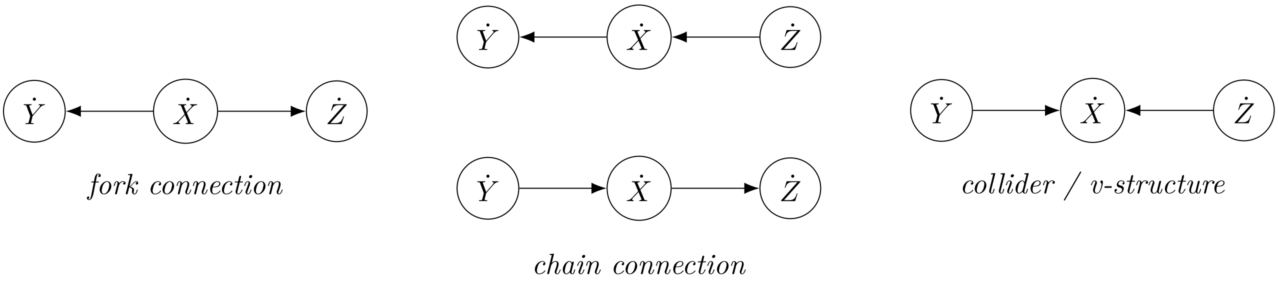
$$\pi(x_1, \ldots, x_p) := \prod_{j=1}^{p} \pi(x_j | \boldsymbol{pa}(X_j)), \tag{2.1}$$

where $\pi(x_j|\boldsymbol{pa}(X_j))$ denotes the conditional probability of $X_j = x_j$ given the realizations $\boldsymbol{pa}(X_j)$ of its set of parent nodes $\boldsymbol{Pa}(X_j)$.

This factorization of the global distribution using the local conditional probabilities of every node given its parent's instantiations can be justified by the *representation theorems* of Bayesian networks (see, for instance, Koller and Friedman, 2009, pp. 62–63). To understand what the representation theorems say, we first have to recapitulate what is called an *I-map*. A graph $\mathcal{G}$ is an independence-map, short I-map, for $\mathbb{P}$ if all conditional independencies that can be read from $\mathcal{G}$, denoted by $\mathrm{Ind}(\mathcal{G})$, also hold in $\mathbb{P}$. This means that the set of conditional independence assertions associated with $\mathcal{G}$ must be a subset of the conditional independence assertions that

[24] As the name suggests, no cycles must occur in a directed acyclic graph. This means that there exists no path from any node X_j which leads back to X_j (after an arbitrary number of steps).

[25] The terms *realization*, *instantiation*, and *configuration* will be used synonymously in this thesis and also in the contributions.



fork connection

chain connection

collider / v-structure

Figure 2.2: Graphical representation of possible (indirect) connections between nodes in a Bayesian network (e.g. Koller and Friedman, 2009, Chap. 3).

hold in $\mathbb{P}$[26], i.e. $\text{Ind}(\mathcal{G}) \subseteq \text{Ind}(\mathbb{P})$. The first theorem says that if $\mathcal{G}$ is an I-map for $\mathbb{P}$, then $\mathbb{P}$ factorizes according to $\mathcal{G}$, which means that $\mathbb{P}$ satisfies Equation (2.1) and it can be represented by a Bayesian network. The second theorem covers the opposite case, namely if $\mathbb{P}$ factorizes according to $\mathcal{G}$, then $\mathcal{G}$ is an I-map for $\mathbb{P}$. This means that $\mathcal{G}$ can be used to find the independencies that hold in $\mathbb{P}$. Thus, as mentioned by Koller and Friedman (2009, pp. 51–52), a graph $\mathcal{G}$ accompanied with a Bayesian network, provides information for a compact representation of a joint distribution and it represents a set of conditional independence assumptions corresponding to this joint distribution. Both notions are equivalent (Koller and Friedman, 2009, pp. 51–52). Up to now, we know that if $\mathcal{G}$ is an I-map for P, all local independencies that we can read from $\mathcal{G}$, hold in $\mathbb{P}$ (Koller and Friedman, 2009, pp. 68–69).

Using the concept of *d-separation*, we can also read global independencies from $\mathcal{G}$, which are beyond the local Markov property (see, e.g. Lauritzen, 1996, Chap. 3.2.2). This applies in particular to indirect connections between nodes that run over several other nodes.

For a graph with node set $\{\dot{Y}, \dot{Z}, \dot{\boldsymbol{X}}\}$, d-separation distinguishes between three possible indirect connections –the *fork connection*, the *chain connection*, and the *collider* or *v-structure*– between two nodes $\dot{Y}$ and $\dot{Z}$. These connections are displayed in Figure 2.2 for three nodes $\{\dot{Y}, \dot{Z}, \dot{X}\}$. D-separation is used to investigate whether the trail between $\dot{Y}$ and $\dot{Z}$ is *active* or not, given that the variables $\boldsymbol{X}$ assigned to $\dot{\boldsymbol{X}}$ are observed[27]. Generally, as, for instance, defined in Koller and Friedman (2009, p. 71), a trail between $\dot{Y}$ and $\dot{Z}$ is active given $\dot{\boldsymbol{X}}$ if

(i) $\dot{X}$ is in $\dot{\boldsymbol{X}}$, i.e. X is observed, for every v-structure in the trail,

(ii) no other node –beyond those belonging to a v-structure as mentioned in (i)– is in $\dot{\boldsymbol{X}}$.

Thus, within the *fork connection* and the *chain connection*, the trail between $\dot{Y}$ and $\dot{Z}$ is active if $\dot{X} \notin \dot{\boldsymbol{X}}$. The trail between $\dot{Y}$ and $\dot{Z}$ is active in the *collider connection* if $\dot{X} \in \dot{\boldsymbol{X}}$, or if at least one descendant of $\dot{X}$ is in $\dot{\boldsymbol{X}}$. Whenever there is no active trail between any $\dot{Y} \in \dot{\boldsymbol{Y}}$ and $\dot{Z} \in \dot{\boldsymbol{Z}}$ given $\dot{\boldsymbol{X}}$, the nodes $\dot{Y}$ and $\dot{Z}$ are d-separated given $\dot{\boldsymbol{X}}$. And, moreover, whenever two nodes $\dot{Y}$ and $\dot{Z}$ are d-separated in the graph given $\dot{\boldsymbol{X}}$, the corresponding random variables Y and Z are conditionally independent given $\boldsymbol{X}$. The concept of d-separation fulfils the properties of *soundness*[28] and *completeness*[29], and it leads to a set of global independencies that hold

[26] A distribution can thus have more than one I-map. The I-maps differ in their complexity and the number of parameters.

[27] To keep the notations simple, I will just write 'given $\dot{\boldsymbol{X}}$' meaning that the corresponding variables are observed.

[28] The soundness of d-separation states that if two nodes in $\mathcal{G}$ are d-separated, the corresponding random variables are indeed conditionally independent in $\mathbb{P}$ (e.g. Koller and Friedman, 2009, p. 72).

[29] Completeness means that if two nodes in $\mathcal{G}$ are not d-separated, the corresponding random variables are conditionally dependent in some distribution $\mathbb{P}$ factorizing according to $\mathcal{G}$ (for details, see Koller and Friedman, 2009, Chap. 3.3.2).

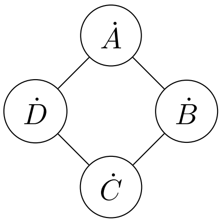


Figure 2.3: Graph structure of the 'Misconception example' in Koller and Friedman (2009, pp. 82–83).

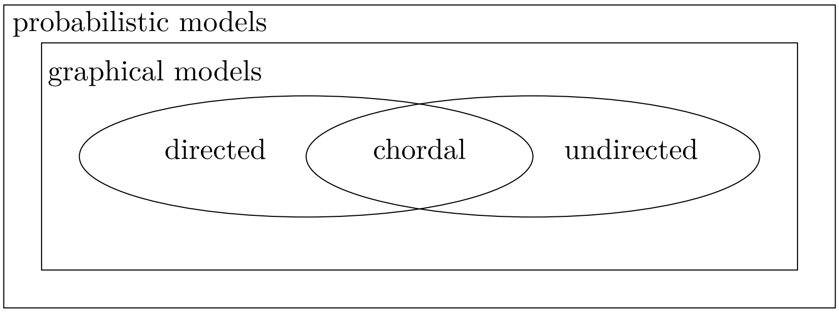


Figure 2.4: Overview on types of probabilistic graphical models, see, for instance, Murphy (2012, p. 664).

in $\mathbb{P}$ and that are equivalent to the local independencies (e.g. Koller and Friedman, 2009, p. 117).

A graph $\mathcal{G}$ is a *perfect map* (*P-map*) for a probability distribution $\mathbb{P}$, if $\text{Ind}(\mathbb{P}) = \text{Ind}(\mathcal{G})$ for all global independencies (e.g. Barber, 2012, p. 72). However, not for all independencies exists a directed acyclic graph that is a P-map for $\mathbb{P}$[30]. This means that even if $\mathcal{G}$ is an I-map for $\mathbb{P}$, it is not necessarily also a *dependence-map*[31] (*D-map*) and vice versa.

A famous example, called the 'Misconception conception' example, for a set of conditional independence assertions which cannot be represented perfectly by a directed acyclic graph is given by Koller and Friedman (2009, pp. 82–83; Chap. 4). The set of conditional independencies $\{(A \perp\!\!\!\perp C|\{B, D\}), (B \perp\!\!\!\perp D|\{A, C\})\}$ should apply to a distribution $\mathbb{P}$. As Koller and Friedman (2009) show, this set of independencies can only be represented by a Markov network, as displayed in Figure 2.3. How this exactly works and how conditional independence assertions are encoded in an undirected graph, will be clarified in the following section.

2.3 Markov networks

Besides Bayesian networks, undirected Markov networks make up the second big part of probabilistic graphical models[32], see Figure 2.4. In contrast to Bayesian networks, which are factorized according to conditional probability distributions, Markov networks are parameterized and factorized using so-called *factors*. Very general, a factor f over a set of random variables $\boldsymbol{X}$, also

[30] And of course, the same also holds for undirected graphs. Markov networks are, for instance, inappropriate to represent the independencies in a v-structure. For instance, the collider in Figure 2.2 encodes the set $\{(Y \perp\!\!\!\perp Z), (Y \not\perp\!\!\!\perp Z|X)\}$ which cannot be represented by an undirected graph.

[31] For a D-map holds that $\text{Ind}(\mathbb{P}) \subseteq \text{Ind}(\mathcal{G})$ (e.g. Barber, 2012, p. 72).

[32] The intersection between directed graphs and undirected graphs are so-called *chordal* or *decomposable* graphs. For details, see, for instance, Lauritzen (1996, Chap. 4.4). Moreover, for example, Koller and Friedman (2009, Chap. 4.5), show how Bayesian networks can be transferred into Markov networks by *moralization* and vice versa by *triangulation*.

known as *affinity function*, *compatibility function*, or *potential function*, is a mapping $f : \mathcal{X} \to \mathbb{R}^+$ (e.g. Koller and Friedman, 2009, p. 104), where $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_p$ denotes the set of possible realizations of $\boldsymbol{X}$. We restrict the codomain of the factor mapping to the positive real numbers to achieve the *Gibbs distribution* (e.g. Barber, 2012, p. 59) as the normalized factor product[33]

$$\pi(\boldsymbol{x}) := \frac{1}{N} \cdot \prod_{o=1}^{m} f(\boldsymbol{C}_o), \tag{2.2}$$

where

$$N := \sum_{\boldsymbol{x} \in \mathcal{X}} \prod_{o=1}^{m} f(\boldsymbol{C}_o) \tag{2.3}$$

is a normalizing constant, known as *partition function*, ensuring that $\pi(\boldsymbol{x})$ is a probability mass distribution. The product runs over m factors, one for each *(maximal) clique*[34] $\boldsymbol{C}_o \subseteq \dot{\boldsymbol{X}}$, $o = 1, \ldots, m$, in the graph $\mathcal{H}$ (e.g. Barber, 2012, p. 59). This implies that the corresponding Markov network has an edge $\dot{X}_j - \dot{X}_{j'}$, whenever $\{\dot{X}_j, \dot{X}_{j'}\}$ is an element of at least one maximal clique.

Similar as in the context of Bayesian networks, conditional independence is closely connected to *active*[35] trails in the network structure. Whenever a path from $\dot{Y}$ to $\dot{Z}$ is *blocked* by $\dot{\boldsymbol{X}}$, meaning that $\boldsymbol{X}$ is observed, Y and Z are conditionally independent given X. The nodes $\dot{Y}$ and $\dot{Z}$ are said to be separated by $\dot{\boldsymbol{X}}$ (e.g. Koller and Friedman, 2009, p. 115). Thus, the concept of separation yields a set of global conditional independencies.

The local independencies in Markov networks are not as clear as in Bayesian networks. For an undirected graphs $\mathcal{H} = (\dot{\boldsymbol{X}}, \boldsymbol{E})$, we have to consider two different concepts of local independence (e.g. Lauritzen, 1996, p. 32):

(i) the *pairwise Markov property*:
for all pairs of non-adjacent nodes $\dot{X}_j \in \dot{\boldsymbol{X}}$ and $\dot{X}_{j'} \in \dot{\boldsymbol{X}}$ holds $X_j \perp\!\!\!\perp X_{j'} | \boldsymbol{X} \setminus \{X_j, X_{j'}\}$, i.e. all nodes, which are no direct neighbours in the graph, are conditionally independent given all other nodes in the graph;

(ii) the *local Markov property*:
for any node $\dot{X}_j \in \dot{\boldsymbol{X}}$ holds $X_j \perp\!\!\!\perp \boldsymbol{X} \setminus \{X_j, \mathrm{MB}(X_j)\} | \mathrm{MB}(X_j)$, where $\mathrm{MB}(X_j)$ denotes the *Markov blanket* of X_j, which is the set of all direct neighbours of X_j. The local Markov property states that a node is conditionally independent of every other node except its neighbours, given its Markov blanket.

These concepts of local independence are equivalent only for *positive distributions*[36]. Moreover, they are also both equivalent with the global independencies for positive distributions. For details, see, for instance, Koller and Friedman (2009, Chap. 4.3) or Lauritzen (1996, Chap. 3.2). For proofs that separation in Markov networks is also sound and complete and that analogue

[33] A product over two factors $f(\boldsymbol{V}, \boldsymbol{X})$ and $f(\boldsymbol{X}, \boldsymbol{W})$, for three disjoint sets $\boldsymbol{X}$, $\boldsymbol{V}$, $\boldsymbol{W}$, is again a factor, mapping from $\mathcal{V} \times \mathcal{X} \times \mathcal{W}$ to the positive real numbers, where $f(\boldsymbol{V}, \boldsymbol{X}, \boldsymbol{W}) \mapsto f(\boldsymbol{V}, \boldsymbol{X}) \cdot f(\boldsymbol{X}, \boldsymbol{W})$, (e.g. Koller and Friedman, 2009, p. 107).

[34] A clique $\boldsymbol{C}$ is a subset of $\dot{\boldsymbol{X}}$, where every pair of nodes associated with $\boldsymbol{C}$, is connected by an edge. The clique is maximal if $\boldsymbol{C}^+$ is not a clique, for any $\boldsymbol{C}^+ \supset \boldsymbol{C}$. See, for instance, Koller and Friedman (2009, p. 35).

[35] A path or trail is active if it is not blocked by at least one observed node. In order keep the explanations and notations simple, I do not explicitly distinguish between nodes and random variables in this paragraph.

[36] In a positive distribution, every probability component of the probability mass distributions is strictly greater than zero.

representation theorems[37] as in Bayesian networks hold for positive distributions, see, for instance, also Koller and Friedman (2009, Chap. 4.3).

A particular property of Markov networks is that the exact factorization in terms of the Gibbs distribution cannot be uniquely be determined by the graph (e.g. Koller and Friedman, 2009, p. 123). It means that different factor products induce the same graph. For example, the single factor $f(A, B, C)$ yields the same graph as the product $f(A, B)f(B, C)f(C, A)$. Both describe the Markov network $\mathcal{H} = (\{\dot{A}, \dot{B}, \dot{C}\}, \{(\dot{A}, \dot{B}), (\dot{B}, \dot{C}), (\dot{C}, \dot{A})\})$. The difference is that the first factorization is based on maximal cliques, while the second uses subsets of them which are 'just' cliques.

There are certainly other ways to parameterize a Markov network. One of them is to use *log-linear models*. They transform factors into so-called *energy functions*:

$$e(\boldsymbol{C}) := -\log(f(\boldsymbol{C})). \tag{2.4}$$

This leads to the joint probability

$$\begin{aligned}\pi(\boldsymbol{x}) &= \frac{1}{N} \cdot \prod_{o=1}^{m} \exp(-e(\boldsymbol{C}_o)) \\ &= \frac{1}{N} \cdot \exp\big\{-\sum_{o=1}^{m} e(\boldsymbol{C}_o)\big\}. \end{aligned} \tag{2.5}$$

This parameterization is used in *Contributions 2* and *3*, where the negative (overall) energy corresponds to the linear predictor of a regression model. In *Contribution 2*, the parameterization is furthermore based on a *log-linear expansion*[38] of the multinomial distribution. In this context, the factor potentials are indeed distributions. Thus, normalization is already satisfied[39] leading to $N = 1$.

2.4 Building a bridge from probabilistic graphical models to statistical matching

After the recapitulation of all necessary terms and definitions, the bridge from probabilistic graphical models to statistical matching can be built. The conditional independence assumption plays the central role and connects the two areas.

As we know, statistical matching uses the assumption that the sets of specific variables $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are conditionally independent given the set of common variables $\boldsymbol{X}$. To reflect this central assumption in a Bayesian network, we have to ensure that every node $\dot{Y} \in \dot{\boldsymbol{Y}}$ is d-separated from every node $\dot{Z} \in \dot{\boldsymbol{Z}}$ by at least one node $\dot{X} \in \dot{\boldsymbol{X}}$. This means that we restrict the graph structure of the Bayesian network to fork connections and chain connections as depicted in Figure 2.2. A collider connection cannot be incorporated since the parameterization would be based on the conditional probability of X given Y and Z, which cannot be computed from the marginals

[37] If $\mathbb{P}$ factorizes over $\mathcal{H}$, $\mathcal{H}$ is an I-map for $\mathbb{P}$. If $\mathcal{H}$ is an I-map for $\mathbb{P}$ and $\mathbb{P}$ is a positive distribution, $\mathbb{P}$ factorizes according to Equation (2.2). The latter statement is known as *Hammersley-Clifford theorem* (e.g. Koller and Friedman, 2009, p. 116).

[38] For an introduction to log-linear models based on log-linear expansions, see, for instance, Whittaker (1990, Chap. 7).

[39] By the way, the same applies for Bayesian networks whose chain rule is a special case of a Gibbs distribution where the factors are equal to the conditional distributions.

available in the data files A and B. Analogously, in a Markov network, the nodes $\dot{Y} \in \dot{\boldsymbol{Y}}$ and $\dot{Z} \in \dot{\boldsymbol{Z}}$ have to be separated by at least one $\dot{X} \in \dot{\boldsymbol{X}}$. The basic form of an undirected graph that should be used for statistical matching is $\dot{Y} - \dot{X} - \dot{Z}$. Further (conditional) independencies can be incorporated.

Probabilistic graphical models have already been used by Landes and Williamson (2016) to solve the statistical matching problem. They show how to find a joint distribution that has maximum entropy from all distributions that are compatible with the available data. To reduce the complexity of this problem, they employ Bayesian networks and exploit the estimated set of conditional independence assertions to find the distribution having maximum entropy.

In the following chapter, I will give brief summaries on the contributions of this thesis. *Contributions 1–3* are devoted to the task of integrating the statistical matching of categorical data into the context of probabilistic graphical models. *Contribution 4* treats the statistical matching micro approach as an imputation task. However, as we will see in Section 3.4, imprecise imputation can also be incorporated into the framework of probabilistic graphical models, namely an imprecise version of Bayesian networks which are known as *credal networks* (Cozman, 2000).

3 Contributing material: summaries, comments, and perspectives

In this chapter I will summarize and discuss each of the four contributions of this cumulative thesis. Overall concluding remarks on this thesis will be given in Chapter 4. *Contribution 1* addresses statistical matching of categorical data with Bayesian networks, while *Contribution 2* treats statistical matching with Markov networks for arbitrary categorical data. *Contribution 3* addresses a special case of *Contribution 2*, namely statistical matching of binary data using the pairwise undirected Ising model. *Contribution 4* uses a new imputation procedure for statistical matching aiming at the construction of sets of complete synthetic files which are used to estimate lower and upper bounds for the parameters of interest.

3.1 *Contribution 1:* Statistical matching of discrete data by Bayesian networks

3.1.1 Summary

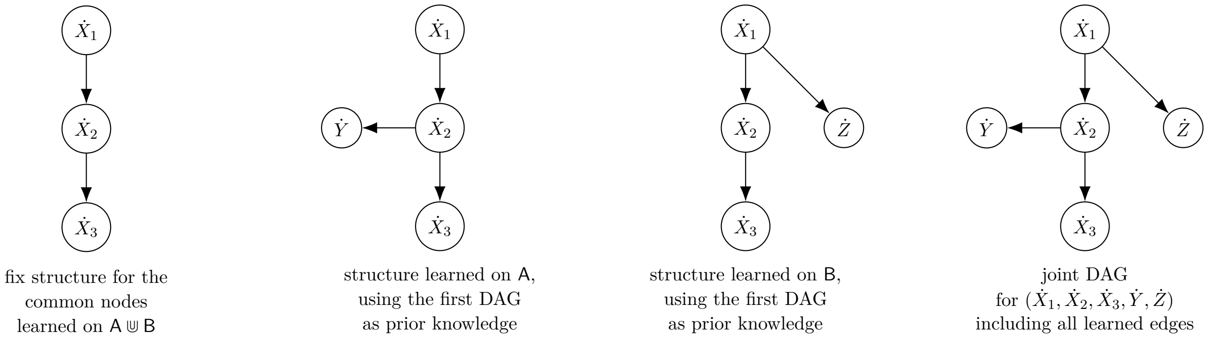
In *Contribution 1*, we use the relation between d-separation of nodes in a graphical model and the conditional independence of their corresponding random variables. Again, we consider two data files A and B, with n_{A} i.i.d observations of $\boldsymbol{Y}$ and $\boldsymbol{X}$, and n_B i.i.d. observations of $\boldsymbol{X}$ and $\boldsymbol{Z}$. Our aim is to estimate a Bayesian network structure from the available data using already available structure learning algorithms under the constraint that the resulting structure reflects the conditional independence assumption that is necessary for statistical matching. Concretely, this means that the graph structure is restricted to a chain connection or fork connection as depicted in Figure 2.2. Both structures ensure that any $\dot{Y} \in \dot{\boldsymbol{Y}}$ is d-separated from any $\dot{Z} \in \dot{\boldsymbol{Z}}$ given at least one $\dot{X} \in \dot{\boldsymbol{X}}$ and thus, $Y \perp\!\!\!\perp Z|X$, for all $Y \in \boldsymbol{Y}$, $Z \in \boldsymbol{Z}$ given at least one[40] $X \in \boldsymbol{X}$. Without loss of generality, we restrict the graph structure to a fork connection[41] in *Contribution 1*.

In *Contribution 1*, we introduce two basically different ways to obtain a *joint graph* structure for $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ from $\mathsf{A} \uplus \mathsf{B}$. In short terms, the first procedure generates a *subgraph*[42] for the common variables, which is used as prior knowledge. Afterwards, two further subgraphs are learned and added that include the edges among the specific nodes and connect the common nodes with the specific nodes. In the second procedure, we learn two distinct subgraphs, one for $(\dot{\boldsymbol{X}}, \dot{\boldsymbol{Y}})$ on A and one for $(\dot{\boldsymbol{X}}, \dot{\boldsymbol{Z}})$ on B. Since the parts concerning the common variables

[40] It is of course also possible that two specific nodes are d-separated by a set of common nodes. A marginal independence between the specific nodes is admittedly conceivable but not reasonable in the context of statistical matching.

[41] The fork connection and chain connection are indeed equivalent. Simple transformations proof that $\pi(x, y, z) = \pi(y|x)\pi(z|x)\pi(x) = \pi(z|x)\pi(x|y)\pi(y) = \pi(y|x)\pi(x|z)\pi(z)$.

[42] With the term subgraph I refer to a part of the graph that considers only a subset of all nodes and their connecting edges.



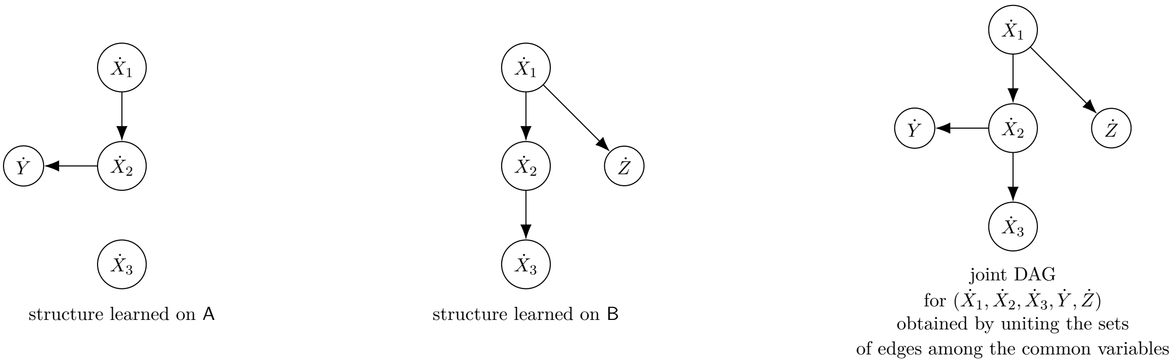
fix structure for the common nodes learned on $\mathsf{A} \Cup \mathsf{B}$

structure learned on A, using the first DAG as prior knowledge

structure learned on B, using the first DAG as prior knowledge

joint DAG for $(\dot{X}_1, \dot{X}_2, \dot{X}_3, \dot{Y}, \dot{Z})$ including all learned edges

Figure 3.1: Toy example to visualize the first procedure with a fix structure for the common variables to obtain a joint graph for $(\dot{X}_1, \dot{X}_2, \dot{X}_3, \dot{Y}, \dot{Z})$.



structure learned on A

structure learned on B

joint DAG for $(\dot{X}_1, \dot{X}_2, \dot{X}_3, \dot{Y}, \dot{Z})$ obtained by uniting the sets of edges among the common variables

Figure 3.2: Toy example to visualize the second procedure with edge union to obtain a joint graph for $(\dot{X}_1, \dot{X}_2, \dot{X}_3, \dot{Y}, \dot{Z})$.

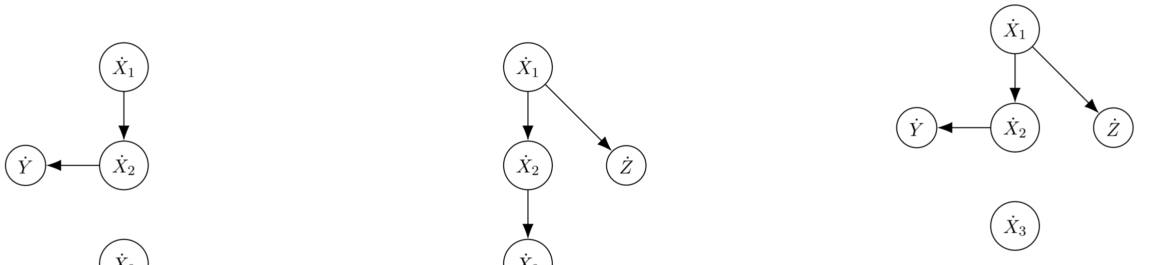
might differ, we unite or intersect the corresponding edge sets.

The generation of a joint graph for the common and specific variables will also be important in *Contributions 2* and *3*. Thus, I will explain it in more detail in the following. All procedures are visualized for a toy example of five variables (X_1, X_2, X_3, Y, Z) in Figures 3.1, 3.2, and 3.3.

The first procedure estimates a fix graph structure for $\boldsymbol{X}$ based on all available $n_{\mathsf{A}} + n_{\mathsf{B}}$ observations using an arbitrary structure learning algorithm for Bayesian networks. This results in a set of estimated edges $\hat{\boldsymbol{E}}^{\mathsf{A}\Cup\mathsf{B}}_{\dot{\boldsymbol{X}}} \subseteq \dot{\boldsymbol{X}} \times \dot{\boldsymbol{X}}$, which exclusively includes edges between the common nodes[43]. In the following step, we use the set $\hat{\boldsymbol{E}}^{\mathsf{A}\Cup\mathsf{B}}_{\dot{\boldsymbol{X}}}$ as prior knowledge for the estimation of the remaining structure of the joint graph. This means that under the constraint that $\hat{\boldsymbol{E}}^{\mathsf{A}\Cup\mathsf{B}}_{\dot{\boldsymbol{X}}}$ is fix and cannot be changed, we add all remaining edges among the sets of specific nodes and also the edges that connect the specific nodes with the common nodes.

We use data file A to estimate the edges among the set of specific nodes $\dot{\boldsymbol{Y}}$, and to estimate the edges that connect the specific nodes in $\dot{\boldsymbol{Y}}$ with the common nodes in $\dot{\boldsymbol{X}}$. As previously mentioned, we use $\hat{\boldsymbol{E}}^{\mathsf{A}\Cup\mathsf{B}}_{\dot{\boldsymbol{X}}}$ as prior knowledge during this estimation process meaning that no edge are added, deleted or reverted. The resulting set of estimated edges $\hat{\boldsymbol{E}}^{\mathsf{A}}_{\dot{\boldsymbol{X}},\dot{\boldsymbol{Y}}}$ is then a subset of $\{\dot{\boldsymbol{Y}} \times \dot{\boldsymbol{Y}}\} \cup \{\dot{\boldsymbol{X}} \times \dot{\boldsymbol{Y}}\} \cup \hat{\boldsymbol{E}}^{\mathsf{A}\Cup\mathsf{B}}_{\dot{\boldsymbol{X}}}$, under the constraint that the previously estimated structure for the common nodes is not changed during the estimation process. The DAG on A is now given as $\hat{\mathcal{G}}^{\mathsf{A}}_{\dot{\boldsymbol{X}},\dot{\boldsymbol{Y}}} = (\dot{\boldsymbol{X}} \cup \dot{\boldsymbol{Y}}, \hat{\boldsymbol{E}}^{\mathsf{A}}_{\dot{\boldsymbol{X}},\dot{\boldsymbol{Y}}})$. We repeat the analogue procedure for file B to obtain a second DAG

[43] With *common nodes* and *specific nodes*, I refer to those nodes representing the common variables or specific variables, respectively.



structure learned on A

structure learned on B

joint DAG
for $(\dot{X}_1, \dot{X}_2, \dot{X}_3, \dot{Y}, \dot{Z})$
obtained by intersecting the sets
of edges among the common variables

Figure 3.3: Toy example to visualize the second procedure with edge intersection to obtain a joint graph for $(\dot{X}_1, \dot{X}_2, \dot{X}_3, \dot{Y}, \dot{Z})$.

$\hat{\mathcal{G}}^{\mathsf{B}}_{\dot{\boldsymbol{X}},\dot{\boldsymbol{Z}}} = (\dot{\boldsymbol{X}} \cup \dot{\boldsymbol{Z}}, \hat{\boldsymbol{E}}^{\mathsf{B}}_{\dot{\boldsymbol{X}},\dot{\boldsymbol{Z}}})$. Since we used the constraint that $\hat{\boldsymbol{E}}^{\mathsf{A} \Cup \mathsf{B}}_{\dot{\boldsymbol{X}}}$ is fix, both DAGs, for A and B, contain the same structure for the common nodes. This fact allows us to easily integrate $\hat{\mathcal{G}}^{\mathsf{A}}_{\dot{\boldsymbol{X}},\dot{\boldsymbol{Y}}}$ and $\hat{\mathcal{G}}^{\mathsf{B}}_{\dot{\boldsymbol{X}},\dot{\boldsymbol{Z}}}$ to obtain a joint graph for $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$:

$$\hat{\mathcal{G}}^{\mathsf{A} \Cup \mathsf{B}}_{\dot{\boldsymbol{X}},\dot{\boldsymbol{Y}},\dot{\boldsymbol{Z}}} := (\dot{\boldsymbol{X}} \cup \dot{\boldsymbol{Y}} \cup \dot{\boldsymbol{Z}}, \hat{\boldsymbol{E}}^{\mathsf{A}}_{\dot{\boldsymbol{X}},\dot{\boldsymbol{Y}}} \cup \hat{\boldsymbol{E}}^{\mathsf{B}}_{\dot{\boldsymbol{X}},\dot{\boldsymbol{Z}}}). \tag{3.1}$$

The second procedure to achieve a joint graph for all variables of interest is to independently estimate two DAGs $\hat{\mathcal{G}}^{\mathsf{A}}_{\dot{\boldsymbol{X}},\dot{\boldsymbol{Y}}} = (\dot{\boldsymbol{X}} \cup \dot{\boldsymbol{Y}}, \hat{\boldsymbol{E}}^{\mathsf{A}}_{\dot{\boldsymbol{X}},\dot{\boldsymbol{Y}}})$ and $\hat{\mathcal{G}}^{\mathsf{B}}_{\dot{\boldsymbol{X}},\dot{\boldsymbol{Z}}} = (\dot{\boldsymbol{X}} \cup \dot{\boldsymbol{Z}}, \hat{\boldsymbol{E}}^{\mathsf{B}}_{\dot{\boldsymbol{X}},\dot{\boldsymbol{Z}}})$, for A and for B, respectively. Within this second procedure, the graph structures for the common variables are not necessarily the same for both DAGs due to random variations in the data. However, both DAGs are again restricted to our basic assumption $\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Z} | \boldsymbol{X}$, i.e. we only allow the basic structures $\dot{X} \to \dot{Y}$ for the DAG on A and $\dot{X} \to \dot{Z}$ for the DAG on B. A joint DAG for $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ can now be obtained by edge union[44]

$$\hat{\mathcal{G}}^{\mathsf{A} \Cup \mathsf{B}}_{\dot{\boldsymbol{X}},\dot{\boldsymbol{Y}},\dot{\boldsymbol{Z}}} := (\dot{\boldsymbol{X}} \cup \dot{\boldsymbol{Y}} \cup \dot{\boldsymbol{Z}}, \hat{\boldsymbol{E}}^{\mathsf{A}}_{\dot{\boldsymbol{X}},\dot{\boldsymbol{Y}}} \cup \hat{\boldsymbol{E}}^{\mathsf{B}}_{\dot{\boldsymbol{X}},\dot{\boldsymbol{Z}}}) \tag{3.2}$$

or edge intersection

$$\hat{\mathcal{G}}^{\mathsf{A} \Cup \mathsf{B}}_{\dot{\boldsymbol{X}},\dot{\boldsymbol{Y}},\dot{\boldsymbol{Z}}} := (\dot{\boldsymbol{X}} \cup \dot{\boldsymbol{Y}} \cup \dot{\boldsymbol{Z}}, (\hat{\boldsymbol{E}}^{\mathsf{A}}_{\dot{\boldsymbol{X}},\dot{\boldsymbol{Y}}} \cap \hat{\boldsymbol{E}}^{\mathsf{B}}_{\dot{\boldsymbol{X}},\dot{\boldsymbol{Z}}}) \cup \hat{\boldsymbol{E}}^{\mathsf{A}}_{\dot{\boldsymbol{Y}},\dot{\boldsymbol{Y}}-\dot{\boldsymbol{X}}} \cup \hat{\boldsymbol{E}}^{\mathsf{B}}_{\dot{\boldsymbol{Z}},\dot{\boldsymbol{Z}}-\dot{\boldsymbol{X}}}), \tag{3.3}$$

where $\hat{\boldsymbol{E}}^{\mathsf{A}}_{\dot{\boldsymbol{Y}},\dot{\boldsymbol{Y}}-\dot{\boldsymbol{X}}}$ denotes the set of edges that was estimated on A and exclusively contains edges among the specific nodes $\boldsymbol{Y}$ and connections from the common variables $\boldsymbol{X}$ to the specific nodes $\boldsymbol{Y}$, i.e. $\hat{\boldsymbol{E}}^{\mathsf{A}}_{\dot{\boldsymbol{Y}},\dot{\boldsymbol{Y}}-\dot{\boldsymbol{X}}} = \hat{\boldsymbol{E}}^{\mathsf{A}}_{\dot{\boldsymbol{Y}},\dot{\boldsymbol{X}}} \setminus \hat{\boldsymbol{E}}^{\mathsf{A}}_{\dot{\boldsymbol{X}}}$. The set $\hat{\boldsymbol{E}}^{\mathsf{B}}_{\dot{\boldsymbol{Z}},\dot{\boldsymbol{Z}}-\dot{\boldsymbol{X}}}$ is defined analogously.

Given the structure of the joint DAG for all variables $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$, we can derive the joint distribution with the aid of the chain rule for Bayesian networks which factorizes in the context of statistical matching as

$$\pi(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) := \prod_{j=1}^{p} \pi(x_j | \boldsymbol{pa}(X_j)) \cdot \prod_{k=1}^{q} \pi(y_k | \boldsymbol{pa}(Y_k)) \cdot \prod_{\ell=1}^{r} \pi(z_\ell | \boldsymbol{pa}(Z_\ell)). \tag{3.4}$$

[44] The union of the two sets of edges may lead to a cyclic graph structure. To solve this problem, the *feeback arc set*, which is defined as the set of edges which must be removed from a cyclic graph to receive an acyclic graph (e.g. Bastert and Matuszewski, 2001), must be found. The respective edges can now be removed or reverted in the joint DAG.

Since the assumption of the conditional independence of the specific variables given the common variables in Equation (1.2) is already incorporated in the graph structure and thus also in this factorization, all factors can be estimated from the available data files. The factors $\pi(x_j|\boldsymbol{pa}(X_j))$ are only dependent on the common variables and can be estimated from $\mathsf{A} \uplus \mathsf{B}$. The factors $\pi(y_k|\boldsymbol{pa}(Y_k))$ are dependent on the specific variables in $\boldsymbol{Y}$ and their parents which are either also in $\boldsymbol{Y}$ or element of the common variables $\boldsymbol{X}$ and can thus be estimated from A. Analogously, $\pi(z_\ell|\boldsymbol{pa}(Z_\ell))$ with $Z_\ell \in \boldsymbol{Z}$ and $\boldsymbol{pa}(Z_\ell) \in \boldsymbol{X} \cup \boldsymbol{Z}$ can be estimated from B. With the estimation of these factors, which we combine to an estimation of the joint probability distribution, the statistical matching macro approach is finished. For the micro approach, we would have to go one step further and impute the missing entries in $\mathsf{A} \uplus \mathsf{B}$ with random draws from the posterior distributions[45]. It means that the missing entries of the a-th observation $\boldsymbol{z}_a = (z_{a1}, \ldots, z_{ar})$, $a \in \mathcal{I}_\mathsf{A}$, are replaced by a random draw from the estimated posterior distribution $\hat{\pi}(\boldsymbol{z}|\boldsymbol{x}_a)$ given the observations of the common variables $\boldsymbol{x}_a$, and the missing entries of the b-th observation $\boldsymbol{y}_b = (y_{b1}, \ldots, y_{bq})$, $b \in \mathcal{I}_\mathsf{B}$, are replaced by draws from $\hat{\pi}(\boldsymbol{y}|\boldsymbol{x}_a)$. Single or multiple imputation techniques can be applied for this purpose.

In an application based on data of the German General Social Survey (GESIS – Leibniz Institute for the Social Sciences, 2013), we showed in Section 3 of *Contribution 1* that our new statistical matching procedure based on Bayesian networks is applicable with already existing software. The quality of the resulting synthetic distribution and data was investigated using Rässler's quality levels for statistical matching (Rässler, 2002, Chap. 2.5).

3.1.2 Comments and Perspectives

In future work on statistical matching with Bayesian networks, the different effects of edge union and edge intersection (see Equations (3.2) and (3.3)) should be properly investigated. Although there was no big difference in the application based on the German General Social Survey, it could be expected that edge intersection removes dependencies from the graph structure. If an edge between two nodes has only been found in one of the available data files, it would not appear in the joint graph. The factorization of the joint probability distribution would then be faulty. With edge union, this problem does not occur. Here, at most additional dependencies could have been found by chance, which are then included in the factorization. However, this does not make the factorization erroneous, but at most computationally expensive.

The estimation of the graph structure depends on the choice of the structure learning algorithm. In the application, we used the score-based *hill climbing* algorithm in combination with the *Bayesian information criterion* (for details, see, for instance Margaritis, 2003, Chap. 2.7.2). The hill climbing algorithm is a heuristic *greedy search algorithm* that works by adding, removing and changing edges, starting with a random or empty graph structure (Nagarajan et al., 2013, p. 19). It chooses the structure that achieves the highest score (which is in our application based on the Bayesian information criterion). It is possible that this structure learning algorithm runs into a local maximum and that the structure, which is globally the best, cannot be found. It should be investigated whether a hybrid[46] structure learning as, for example, the *max-min hill climbing* (Tsamardinos et al., 2006) which is, according to Gasse et al. (2014), one of the most "powerful state-of-the-art algorithms for Bayesian network structure learning", leads to better statistical matching results.

[45] Answering probability queries in a Bayesian network for a variable given the realizations of some other observed variables yields the posterior distributions (e.g. Koller and Friedman, 2009, pp. 5–6).

[46] The term hybrid refers to structure learning algorithms which combine constrained-based and score-based algorithms to get the benefits of both strategies (e.g. Nagarajan et al., 2013, p. 20).

Although there is nothing inherently causal about a Bayesian network (e.g. Murphy, 2012, p. 312), it would also be possible to apply a causal learning algorithm to estimate the graph structure. For instance, Tsamardinos et al. (2012) describe methods for statistical matching, summarized under the term *Integrative Causal Analysis (INCA)*, which are based on finding one or all causal models which are simultaneously consistent with the data at hand and all available prior knowledge. Especially for data situation where the number of common variables is low, these approaches seem to perform better than standard statistical matching approaches (Tsamardinos et al., 2012).

Our statistical matching procedure can also be further developed for exclusively continuous data, where the joint density is assumed to follow a multivariate normal distribution. This type of Bayesian networks is known as *Gaussian networks* (e.g. Koller and Friedman, 2009, Chap. 7). The local distribution for each continuous variable is a normal distribution with a mean which is linearly dependent on its parents. Probably even more interesting would be the further development of our statistical matching method with Bayesian networks for mixed categorical and continuous data. For this purpose, we could use so-called *conditional linear Gaussian Bayesian networks* (e.g. Kjræulff and Madsen, 2013, Chap. 4.1.2). Within this class of probabilistic graphical models, the local distributions for a discrete variable $X^d \in \boldsymbol{X}^d$ is a conditional probability table $\pi(x^d|\boldsymbol{pa}(X^d))$ as in usual Bayesian networks, and a continuous variable $X^c \in \boldsymbol{X}^c$ is assumed to follow a conditional normal distribution $f(x^c|\boldsymbol{pa}(X^c))$ that depends on the instantiation of its discrete parents. The mean of this conditional Gaussian distribution furthermore depends linearly on the continuous parents, the variance is independent of them. It must be noted that discrete nodes may necessarily only have discrete parents considering this type of graphical model. Then, the joint distribution is a multivariate normal mixture density over $\boldsymbol{X} = \boldsymbol{X}^d \cup \boldsymbol{X}^c$ (Madsen, 2008). In the context of statistical matching, under the conditional independence assumption, we would receive the following joint density for $\boldsymbol{X} = \boldsymbol{X}^d \cup \boldsymbol{X}^c$, $\boldsymbol{Y} = \boldsymbol{Y}^d \cup \boldsymbol{Y}^c$, and $\boldsymbol{Y} = \boldsymbol{Y}^d \cup \boldsymbol{Y}^c$:

$$\begin{aligned} f(\boldsymbol{x}^d, \boldsymbol{x}^c, \boldsymbol{y}^d, \boldsymbol{y}^c, \boldsymbol{z}^d, \boldsymbol{z}^c,) = & \prod_{x^d \in \boldsymbol{x}^d} \pi(x^d|\boldsymbol{pa}(X^d)) \cdot \prod_{x^c \in \boldsymbol{x}^c} f(x^c|\boldsymbol{pa}(X^c)) \\ & \cdot \prod_{y^d \in \boldsymbol{y}^d} \pi(y^d|\boldsymbol{pa}(Y^d)) \cdot \prod_{y^c \in \boldsymbol{y}^c} f(y^c|\boldsymbol{pa}(Y^c)) \\ & \cdot \prod_{z^d \in \boldsymbol{z}^d} \pi(z^d|\boldsymbol{pa}(Z^d)) \cdot \prod_{z^c \in \boldsymbol{z}^c} f(z^c|\boldsymbol{pa}(Z^c)) \end{aligned} \tag{3.5}$$

Moreover, the incorporation of auxiliary information into statistical matching with Bayesian networks can be realized straightforwardly. The parameters, i.e. (the probability components of) the local models, or only a subset of them can externally be fixed by an expert. If auxiliary information is available in terms of a third file containing joint information on $(\boldsymbol{X}, \boldsymbol{Y})$ or $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$, the assumption of conditional independence can be repealed and directed edges between the specific variables are allowed. Also a collider connection as in Figure 2.2 with or without an edge between $\dot{Y}$ and $\dot{Z}$ would be conceivable.

Not a repeal but an attenuation of the conditional independence assumption can be reached by using *credal networks* which are an imprecise version of Bayesian networks. In simple terms, a credal network is composed of a directed acyclic graph and a collection of conditional *credal sets*[47] which correspond to our local probability models. Concretely, a credal network can

[47] Following Antonucci et al. (2014, p. 208), I will define a credal set for a categorical variable as a closed convex set of probability mass distributions over this variable. More details on this will be given in Subsection 3.4.2.

be interpreted as a set of Bayesian networks which share the same graph but have different conditional and unconditional parameter values (e.g. de Cooman et al., 2010). They are, for instance, used in contexts where the parameter values cannot be precisely determined due to disagreements of field experts (e.g. de Cooman et al., 2010). In future research it should be determined to what extent this notion can be transferred to the context of statistical matching, where there are two data files, which are likely to yield different parameter values for the common variables resulting from sampling uncertainty.

Using sets of conditional probabilities as local models allows to apply different concepts of independence. Although *strong independence*, which is a straightforward generalization of stochastic independence, can be used to combine the local models to a set of joint distributions, also, for instance, the weaker and asymmetric concept of *epistemic irrelevance* would be conceivable. These two independence concepts coincide in the precise case. For details and definitions of these two independence concepts for credal sets, see, for instance, Antonucci et al. (2014), Cozman and Walley (2005) or de Bock (2015, Chap. 4.4). A more detailed discussion on the usage of credal networks in the context of statistical matching and imprecise imputation can be found in Subsection 3.4.2.

3.2 *Contribution 2:* Utilizing log-linear Markov networks to integrate categorical data files

3.2.1 Summary

Contribution 2 is the undirected analogue to *Contribution 1* and the basic data situation is exactly the same. We again assume that the specific variables are conditionally independent given the common variables and relate this assumption to the basic structure $\dot{Y} - \dot{X} - \dot{Z}$ for our undirected graph. This structure ensures that any specific node $\dot{Y} \in \dot{\boldsymbol{Y}}$ is graphically separated from any specific node $\dot{Z} \in \dot{\boldsymbol{Z}}$ given at least one node $\dot{X} \in \dot{\boldsymbol{X}}$. Thus, all paths from $\dot{\boldsymbol{Y}}$ to $\dot{\boldsymbol{Z}}$ are blocked by $\dot{\boldsymbol{X}}$.

If the graph structure of the Markov network for $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ is not determined by hand, we have to use a structure learning algorithm (e.g. Koller and Friedman, 2009, Chap. 20.7) to estimate it from the available data A and B. Again we have the problem that we have no complete file available which gives us information about the dependence between $\boldsymbol{Y}$ and $\boldsymbol{Z}$. Thus, for structure learning on A and B, we can proceed as described in *Contribution 1*. One way would be to learn again a fix structure for the common variables $\boldsymbol{X}$ based on the $n_{\mathsf{A}} + n_{\mathsf{B}}$ available observations and use the resulting undirected edges as prior knowledge. The subgraphs representing the structure among the specific variables and the connections between the specific variables and the common variables, can be learned under the constraint of the fix structure for $\boldsymbol{X}$ and added as described in Subsection 3.1.1. Alternatively, we can follow the second structure learning procedure of *Contribution 1* and independently estimate two graphs $\hat{\mathcal{H}}^{\mathsf{A}}_{\dot{\boldsymbol{X}},\dot{\boldsymbol{Y}}} = (\dot{\boldsymbol{X}} \cup \dot{\boldsymbol{Y}}, \hat{\boldsymbol{E}}^{\mathsf{A}}_{\dot{\boldsymbol{X}},\dot{\boldsymbol{Y}}})$ and $\hat{\mathcal{H}}^{\mathsf{B}}_{\dot{\boldsymbol{X}},\dot{\boldsymbol{Z}}} = (\dot{\boldsymbol{X}} \cup \dot{\boldsymbol{Z}}, \hat{\boldsymbol{E}}^{\mathsf{B}}_{\dot{\boldsymbol{X}},\dot{\boldsymbol{Z}}})$ receptively on A and B. Again, the sets of edges can be combined by edge union or edge intersection, analogously to Equations (3.2) and (3.3).

As already indicated in Section 2.3, we use a log-linear model to parameterize a Markov network. More concretely, we employ the *log-linar expansion* (e.g. Whittaker, 1990, p. 206) of the multinomial distribution. Thus, the factor potentials are already distributions and restricted to the unit interval, and we can set the normalization constant to 1. Equation (2.5), defined for a set of nodes $\dot{\boldsymbol{X}}$ and a set of edges $\boldsymbol{E} = \dot{\boldsymbol{X}} \times \dot{\boldsymbol{X}}$, then simplifies to

$$\pi(\boldsymbol{x}) = \exp\big\{ -\sum_{o=1}^{m} e(\boldsymbol{C}_o)\big\} = \exp\big\{ \sum_{\boldsymbol{C}^* \subseteq \boldsymbol{X}} u_{\boldsymbol{C}^*}(\boldsymbol{x})\big\}, \tag{3.6}$$

where the negative energy is represented by the sum over the so-called u-terms which is equivalent to a linear predictor of a regression model, applying dummy coding for the categorical covariates. In contrast to former parameterizations in Equation (2.2) using the Gibbs distribution, we now sum over all subsets $\boldsymbol{C}^*$ of $\boldsymbol{X}$, i.e. all elements of the power set $\mathcal{P}(\mathcal{X})$ of $\boldsymbol{X}$ which means that our linear predictor contains all possible interaction terms, all main effects, and an intercept term for the empty set which is also an element of the power set.

A u-term $u_{\boldsymbol{C}^*}(\boldsymbol{x})$ corresponds to a log-odds and equals 0 in this representation if we either consider the reference category of one of the respective categorical variables in $\boldsymbol{C}^*$, or if two variables $X_j \in \boldsymbol{X}$ and $X_{j'} \in \boldsymbol{X}$, both element of $\boldsymbol{C}^*$, are (conditionally) independent. Concrete examples for the representation of a log-linear model with the aid of the log-linear expansion of the multinomial distribution, are given in Appendix A of *Contribution 2*.

However, this type of log-linear model is still very general and we have to restrict it to fit the needs of Markov networks. To ensure that our model is indeed a *graphical model* (e.g. Tutz, 2011, p. 341, p. 346),

i all variables which appear together in a higher-order u-term also have to appear in all combinations in lower-order terms (restriction for *hierarchical*[48] models);

ii the higher-order u-terms of variables which already appear together in lower-order terms must be included in the model (restriction for *graphical*[49] models).

In a graphical log-linear model, the *interaction graph*[50] corresponds to the independence graph and the highest-order u-terms (interaction terms) equal the maximal cliques (Whittaker, 1990, p. 209). Thus, all terms that we need to determine the joint distribution for our Markov networks according to Equation (3.6), can be read off the Markov network structure.

The joint distribution of $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ in the context of statistical matching and under the conditional independence assumption[51] is determined as

$$\begin{aligned}\pi(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) &= \exp\{\log\pi(\boldsymbol{x},\boldsymbol{y}) + \log\pi(\boldsymbol{x},\boldsymbol{z}) - \log\pi(\boldsymbol{x})\} \\ &:= \exp\left\{ \sum_{\boldsymbol{C}^*\in\mathcal{P}(\boldsymbol{X}\cup\boldsymbol{Y})} u_{\boldsymbol{C}^*}(\boldsymbol{x},\boldsymbol{y}) + \sum_{\boldsymbol{C}^*\in\mathcal{P}(\boldsymbol{X}\cup\boldsymbol{Z})} u_{\boldsymbol{C}^*}(\boldsymbol{x},\boldsymbol{z}) - \sum_{\boldsymbol{C}^*\in\mathcal{P}(\boldsymbol{X})} u_{\boldsymbol{C}^*}(\boldsymbol{x}) \right\}. \end{aligned} \tag{3.7}$$

Again, none of the terms in the joint distribution is simultaneously dependent on any $Y \in \boldsymbol{Y}$ and $Z \in \boldsymbol{Z}$. This means that we can estimate all parameters, needed to specify this distribution, from the available data A and B.

[48] For instance, a model over three variables X, Y, Z that contains as maximum order interaction term for $\{X,Y,Z\}$, gets hierarchical, if we add (interaction) terms for each element in $\{\{X,Y\},\{X,Z\},\{Y,Z\},X,Y,Z\}$.

[49] For instance, a model over three variables X, Y, Z that contains the (interaction) terms $\{\{X,Y\},\{X,Z\},\{Y,Z\},X,Y,Z\}$, gets graphical, if we add an interaction term for $\{X,Y,Z\}$.

[50] An interaction graph represents variables as nodes and interactions between variables are, as the name suggests, represented by edges between the corresponding variables (Whittaker, 1990, p. 209).

[51] In contrast to *Contribution 1*, we use the equality $\pi(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) = \dfrac{\pi(\boldsymbol{x},\boldsymbol{y})\cdot\pi(\boldsymbol{x},\boldsymbol{z})}{\pi(\boldsymbol{x})}$ to express the conditional independence of the specific variables given the common variables.

The applicability of this statistical matching procedure is again shown using the data of the German General Social Survey (GESIS – Leibniz Institute for the Social Sciences, 2013). The quality of the statistical matching results is determined based on the quality levels proposed by Rässler (2002, Chap. 2.5).

3.2.2 Comments and Perspectives

Since both statistical matching with Bayesian networks in *Contribution 1* and with Markov networks in *Contribution 2* have just been investigated in an example data file, the next step for future research should be conduction of extensive simulation studies.

Simulating data from a Bayesian network is straightforward and in accordance with the *topological ordering*[52] of the nodes in the graph. A synthetic observation is simulated from a DAG $\mathcal{G}$ as follows. First, we have to sample values from the unconditional marginal distributions for all nodes that have no parents, i.e. the nodes which are in the first places of the topological ordering. Given these sampled values, we randomly draw from the conditional distributions of the children of these nodes. This step is repeated until all nodes of the ordering are passed. Given a certain number of simulated observations, we can allocate them into two files and delete the blocks of observations of the specific variables from the respective file. This leads to the typical data situation of statistical matching.

The simulation of data based on a log-linear Markov network as considered in *Contribution 2* is, at least theoretically, also easy to perform. Practically, we have to pre-define the values of all probability components of the joint distribution. Subsequently, the corresponding u-terms can be computed and we can simulate observations from the log-linear model. More difficult would be the simulation of data from a Markov network parameterized by factors as in Equation (2.2). Except the positivity restriction, the concept of 'compatibility' in terms of a factor is arbitrary and would have to be determined first.

Independently of considering Bayesian networks or Markov networks for simulation, a further difficulty –besides the more or less arbitrary choice of parameter values– is that we have to pre-define the structure of the graph we want to sample from. In order to cover a large amount of possible data situations, which ensures that the performance of our statistical matching procedure is appropriately investigated, we have to consider graph structures with different numbers of nodes and different independence assertions. Besides the analysis of the statistical matching method in situations where the conditional independence assumption holds, we should also consider graph structures that admit active paths between the specific variables.

Moreover, a simulation study should cover different structure learning algorithms. The results of the exemplary application of statistical matching with Markov networks in *Contribution 2* give reason to believe that a more reliable graph structure would have led to better statistical matching results.

As for statistical matching with Bayesian networks, further steps for future research might be the extension of the method in *Contribution 2* for exclusively continuous and mixed continuous and categorical data. For the former purpose, there exists the class of *Gaussian Markov random fields*, also known as *Gaussian graphical models*, which are, for instance, addressed in Lauritzen (1996, Chap. 5) or Whittaker (1990, Chap. 6). This type of probabilistic graphical models is based on a multivariate normal distribution. Information on packages for the statistical software R (R Core Team, 2019) that can be used for the application of Gaussian graphical models can be

[52] The nodes $\dot{X}_1, \ldots, \dot{X}_p$ are topologically ordered with respect to a directed graph if $j < j'$ for $\dot{X}_j \rightarrow \dot{X}_{j'}$, for all $j, j' \in \{1, \ldots, p\}$ (e.g. Koller and Friedman, 2009, p. 36).

found in Højsgaard et al. (2012, Chap. 4). For mixed continuous and categorical data, there also already exists a class of undirected probabilistic graphical models, the so-called *mixed interaction models* which combine Gaussian graphical models with the log-linear models for categorical data (e.g. Lauritzen, 1996, Chap. 6). This results in a joint distribution over the set of continuous and categorical variables that follows a conditional Gaussian distribution, where the local models of the continuous variables are assumed to follow a normal distribution whose mean may depend on the parents. The variance, as in conditional linear Gaussian Bayesian networks, is independent of the parent's realizations (e.g. Højsgaard et al., 2012, p. 119).

3.3 *Contribution 3:* Binary data fusion using undirected probabilistic graphical models

3.3.1 Summary

Continuing the research on statistical matching with the aid of probabilistic graphical models, inevitably leads to the consideration of special cases. *Contribution 3* is such a project and deals with the integration of exclusively binary data utilizing the Ising model, arising from statistical physics. In two ways, the Ising model is a special case of the more general log-linear Markov model considered in *Contribution 2*: firstly, as already mentioned, the Ising model exclusively considers binary variables, and secondly, the Ising model is a pairwise Markov network which means that the maximum order of interaction terms is two. This restrictions make the estimation of the joint distribution for binary variables computationally very efficient since all high-order interactions terms of the log-linear model are neglected.

For *Contribution 3*, we again consider the data situation as displayed in Figure 1.3 and a Markov network parameterization as in Equation (2.5). The Ising model, developed by Ising (1925), was initially used to explain *ferromagnetism*[53], where the nodes of the corresponding graph represent atoms or particles (e.g. Björnberg, 2009, p. 3). However, it is also suitable for other physical or biological systems, for cell structures, or sociological applications which are based on a set of binary random variables (Kindermann and Snell, 1980, pp. 4–5). Originally, the two-dimensional Ising model consisted of elements that described the *states*, also called *spins* (e.g. Björnberg, 2009, p. 3), of the atoms of a ferromagnetic field which are arranged in a grid or lattice (e.g. McCoy and Wu, 1973, p. 2). A spin can be positive or negative. Its direction is influenced by an external magnetic field, and the directions of its adjacent spins (e.g. Kindermann and Snell, 1980, p. 3). The latter characteristic of the Ising model restricts the corresponding Markov network to a pairwise Markov network, exclusively modelling interactions between pairs of neighbouring spins, while the former characteristic describes what a statistician would name a *main effect* for each spin. The main effects and the interaction effects are expressed in terms of an *energy*. The magnetic field for p binary random variables $X_1, \dots, X_p$ tends to stay in the configuration $\boldsymbol{x} = (x_1, \dots, x_p) \in \mathcal{X} = \mathcal{X}_1 \times \dots \times \mathcal{X}_p$, with $\mathcal{X}_j = \{0, 1\}$ for $j \in \{1, \dots, p\}$, that has the highest probability to occur (van Borkulo et al., 2014, p. 2 of the supplementary information) and that requires the least energy. In a ferromagnetic field, low energy magnets tend to have spins which are mainly pointing into one direction (Kindermann and Snell, 1980, p. 7). The overall energy of the Ising model for an undirected graph $\mathcal{H} = (\dot{\boldsymbol{X}}, \boldsymbol{E})$ can be expressed in terms of the

[53] In a ferromagnet, particles tend to point in the same direction (e.g. Murphy, 2012, p. 668).

Hamiltonian function (e.g. van Borkulo et al., 2014, p. 2 of the supplementary information)

$$H(\boldsymbol{x}) := \sum_{\dot{X}_j \in \dot{\boldsymbol{X}}} e(x_j) + \sum_{(\dot{X}_j, \dot{X}_{j'}) \in \boldsymbol{E}} e(x_j, x_{j'}) \tag{3.8}$$

$$:= - \sum_{\dot{X}_j \in \dot{\boldsymbol{X}}} \tau_j \; x_j - \sum_{(\dot{X}_j, \dot{X}_{j'}) \in \boldsymbol{E}} \beta_{j,j'} \; x_j \; x_{j'}. \tag{3.9}$$

The overall energy of the Ising model results from summing over the energies related to the main effects and the interaction effects, where

$$e(x_j) := -\log\{f(x_j)\} = -\tau_j \; x_j \tag{3.10}$$

and

$$e(x_j, x_{j'}) := -\log\{f(x_j, x_{j'})\} = -\beta_{j,j'} \; x_j \; x_{j'}. \tag{3.11}$$

Equation (3.10) contains the main effects τ_j for every node $\dot{X}_j$ in the graph, and Equation (3.10) contains the pairwise interaction effects $\beta_{j,j'}$ for all adjacent nodes $\dot{X}_j, \dot{X}_{j'}$, for $j \neq j'$, $j, j' \in \{1, \dots, p\}$. Thus, the negative energy equals the linear predictor of this log-linear model.

Incorporating the Hamiltonian function into Equation (2.5) yields the joint distribution of the Ising model as

$$\pi(\boldsymbol{x}) = \frac{1}{N} \cdot \exp\Big\{ -H(\boldsymbol{x}) \Big\} = \frac{1}{N} \cdot \exp\Big\{ \sum_{\dot{X}_j \in \dot{\boldsymbol{X}}} \tau_j \; x_j + \sum_{(\dot{X}_j, \dot{X}_{j'}) \in \boldsymbol{E}} \beta_{j,j'} \; x_j \; x_{j'} \Big\}, \tag{3.12}$$

$$\text{with the normalizing constant} \quad N = \sum_{\boldsymbol{x} \in \mathcal{X}} \exp\Big\{ \sum_{\dot{X}_j \in \dot{\boldsymbol{X}}} \tau_j \; x_j + \sum_{(\dot{X}_j, \dot{X}_{j'}) \in \boldsymbol{E}} \beta_{j,j'} \; x_j \; x_{j'} \Big\}, \tag{3.13}$$

which is needed since the factors of this log-linear Markov network are no distributions and thus not per se normalized.

Embedding statistical matching into the framework of the Ising model is based on the same idea as in *Contribution 2*. We again ensure that all paths from any node $\dot{Y} \in \dot{\boldsymbol{Y}}$ to any node $\dot{Z} \in \dot{\boldsymbol{Z}}$ is blocked by at least one $\dot{X} \in \dot{\boldsymbol{X}}$. Given a graph $\mathcal{H} = (\dot{\boldsymbol{X}} \cup \dot{\boldsymbol{Y}} \cup \dot{\boldsymbol{Z}}, \boldsymbol{E})$, which can either be determined by an expert of the corresponding domain or estimated as described in Subsection 3.2.1, the Hamiltonian for the statistical matching data situation is given as

$$\begin{aligned} H(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) := & - \sum_{\dot{X}_j \in \dot{\boldsymbol{X}}} \tau_j \; x_j - \sum_{\dot{Y}_k \in \dot{\boldsymbol{Y}}} \upsilon_k \; y_k \\ & - \sum_{\dot{Z}_\ell \in \dot{\boldsymbol{Z}}} \phi_\ell \; z_\ell - \sum_{(\dot{X}_j, \dot{X}_{j'}) \in \boldsymbol{E}} \beta_{j,j'} \; x_j \; x_{j'} \\ & - \sum_{(\dot{Y}_k, \dot{Y}_{k'}) \in \boldsymbol{E}} \gamma_{k,k'} \; y_k \; y_{k'} - \sum_{(\dot{Z}_\ell, \dot{Z}_{\ell'}) \in \boldsymbol{E}} \delta_{\ell,\ell'} \; z_\ell \; z_{\ell'} \\ & - \sum_{(\dot{X}_j, \dot{Y}_k) \in \boldsymbol{E}} \epsilon_{j,k} \; x_j \; y_k - \sum_{(\dot{X}_j, \dot{Z}_\ell) \in \boldsymbol{E}} \zeta_{j,\ell} \; x_j \; z_\ell. \end{aligned} \tag{3.14}$$

The main effects τ_j and the interaction effects $\beta_{j,j'}$ can be estimated from $\mathsf{A} \uplus \mathsf{B}$, while the main effects υ_k and the interaction effects $\gamma_{k,k'}$ and $\epsilon_{j,k}$ can be estimated from A, and the main effects ϕ_ℓ and the interaction effects $\delta_{\ell,\ell'}$ and $\zeta_{j,\ell}$ can be estimated from B, for all $j \neq j'$, $j, j' \in \{1, \dots, p\}$, $k \neq k'$, $k, k' \in \{1, \dots, q\}$, $\ell \neq \ell'$, $\ell, \ell' \in \{1, \dots, r\}$.

In *Contribution 3*, we also conducted a simulation study to assess the quality of the statistical matching results obtained by our proposed procedure. We simulated random structures for the Ising model and furthermore varied the number of nodes in the graph, the number of observations, and the sizes of the interaction coefficients. Most important, we also analysed the influence of the conditional independence assumption on the statistical matching results. For this purpose, we simulated data where the conditional independence of any $Y \in \boldsymbol{Y}$ and any $Z \in \boldsymbol{Z}$ given at least one $X \in \boldsymbol{X}$ holds, where the assumption is violated for some $Y \in \boldsymbol{Y}$ and $Z \in \boldsymbol{Z}$, and where the assumption is violated for all specific variables. Although we found that the statistical matching results are best in situations where this central assumption indeed applies, we could also find combinations of our simulation parameters where the synthetic distribution derived by statistical matching was very close to the distribution in the complete simulated data even though the assumption was violated.

3.3.2 Comments and Perspectives

Now that we have investigated statistical matching by Bayesian networks, Markov networks and the Ising model, we can ask ourselves which procedure would be the best to use in an application. However, this question is very hard to answer. As we already know, if a distribution has a v-structure in its dependence structure, its *P-map* is restricted to a directed acyclic graph. If it includes a diamond shaped structure as displayed in Figure 2.3, we need to use an undirected graph to achieve the most appropriate factorization based on the corresponding *P-map*. From the available data it cannot be seen or tested which structure the *P-map* of the corresponding joint distribution has. This task can at best be solved by an expert that is familiar with the substantial context of the application. If we can only find an *I-map* which is not also a *P-map* for the regarded joint distribution, the estimation of the factors and especially the normalizing constant can computationally be very expensive. This is due to the fact that an *I-map* can include more dependencies than actually needed.

If we focus on the interpretability and the understandability of a probabilistic graphical model for a potential user, probably Bayesian networks should be preferred. The arbitrary factors of a Markov network, which are a measure for the compatibility of random variables and their realizations, have no direct interpretation as the local models of a Bayesian network which are conditional probabilities. Considering the availability of software for the estimation of Bayesian networks and Markov networks, again Bayesian networks should be preferred. However, as we show in *Contribution 2*, log-linear Markov networks can be estimated as generalized linear regression (e.g. Tutz, 2011, Fahrmeir et al., 2013) models which are available in most statistical software. And, especially, for the statistical programming software R (R Core Team, 2019), there exists the additional package IsingFit (van Borkulo et al., 2016) for the estimation of the Ising model based on *lasso* regularization in combination with the *extended Bayesian information criterion* (e.g. van Borkulo et al., 2014, Barber and Drton, 2015). The implemented learning algorithms penalize the number of nodes and the number of neighbours in a pseudolikelihood approach, which makes the computation of the normalizing constant tractable also for larger numbers of nodes as described in Koller and Friedman (e.g. 2009, Chap. 20.6.1) or van Borkulo et al. (2014, supplementary information).

It would be a fruitful area for further work to investigate other special forms of probabilistic graphical models for their suitability to statistical matching purposes. Maximum order interactions can be incorporated by, for example, considering a *Hopfield network* (e.g. Murphy, 2012, p. 669). Although it may not seem to be appropriate to consider a fully connected graph for statistical matching, the interaction coefficients regarding any combination of $Y \in \boldsymbol{Y}$ and $Z \in \boldsymbol{Z}$ would be set to zero due to the assumption of conditional independence. The *Potts model* is another special case which generalizes the pairwise Ising model to categorical variables with more than two states. Details for this model are, for instance, given in Wainwright and Jordan (2008, pp. 43–44) or Wu (1982).

3.4 *Contribution 4:* Imprecise imputation: a nonparametric micro approach reflecting the natural uncertainty of statistical matching with categorical data

3.4.1 Summary

Contribution 4 of this thesis also deals with the problem of statistical matching, whereas this time the focus is not on a probabilistic graphical model but on a newly developed imputation method. The special thing about this method is that it does not rely on the conditional independence assumption as the former contributions. Our method, which we named *imprecise imputation*, can be allocated to the group of *partial identification approaches* as listed on page 4. It is a set-valued imputation method, which aims at the construction of a complete synthetic data file. Therefore, it is a statistical matching micro approach and moreover, the first micro approach[54] that reflects the uncertainty inherent to statistical matching.

Within this contribution, we interpret statistical matching as a missing data problem with a special block-wise missingness pattern. To deal with data that contains incomplete observations, there are various procedures. One frequently used approach, besides *complete-case analyses* or *available-case analyses* (e.g. Little and Rubin, 2002, Chap. 3), is imputation, i.e. the substitution of the missing values with suitable real or artificial values to derive complete (but at least partly synthetic) data. To find suitable values (*donor* values) for the replacement of the missing values (*recipient* values), one can use parametric or nonparametric approaches. If imputation is based on *donation classes*, all observations are divided into homogeneous classes which are constructed on basis of the realizations of the common variables. Potential donors to replace a recipient's value are only chosen within the same donation class.

A well-known and often used method is *hot deck imputation*, where similar records from the same sample are used to substitute the missing values (e.g. Little and Rubin, 2002, p. 62). In *Contribution 4*, we deduce *imprecise imputation* as a generalization of hot deck imputation which imputes sets of values instead of a single value for every missing entry in $\mathsf{A} \uplus \mathsf{B}$.

We propose three different imputation approaches which are all based on the data situation described in Section 1.2. They differ in the way they choose donor values from B to substitute missing entries in A[55]:

(i) *domain imputation* replaces every missing value $z_{a\ell}$ by its domain $\mathcal{Z}_\ell$, i.e.

$$\tilde{\mathfrak{z}}_{a\ell} := \mathcal{Z}_\ell, \quad \forall\, a \in \mathcal{I}_{\mathsf{A}}, \ell \in \{1, \ldots, r\}. \tag{3.15}$$

(ii) *variable-wise imputation* based on D donation classes replaces every missing value $z_{a\ell}$ by the complete set of live[56] (donor) values, i.e.

$$\tilde{\mathfrak{z}}_{a\ell} := \{z_\ell | b \in \mathcal{I}^d_{\mathsf{B}}\}, \quad \forall\, a \in \mathcal{I}^d_{\mathsf{A}}, d \in \{1, \ldots, D\}, \ell \in \{1, \ldots, r\}, \tag{3.16}$$

[54] The partial identification approaches listed in the overview on current literature in Section 1.3 are all statistical matching macro approaches.

[55] The imputation of the missing values of $\boldsymbol{Y}$ in B works analogously.

[56] In accordance with D'Orazio et al. (2006b), we use the term 'live values' to refer to values which have actually been observed in $\mathsf{A} \uplus \mathsf{B}$.

where $\mathcal{I}_{\mathsf{A}}^{d} \subseteq \mathcal{I}_{\mathsf{A}}$ denotes the subset of indices which are in A and members of the d-th donation class. The analogue applies to $\mathcal{I}_{\mathsf{B}}^{d}$.

(iii) *case-wise imputation* based on D donation classes replaces all missing entries $\boldsymbol{z}_a = (z_{a1}, \ldots, z_{ar})$ of the a-th observation simultaneously as

$$\tilde{\boldsymbol{\mathfrak{z}}}_a := \{(z_{b1}, \ldots, z_{br}) | b \in \mathcal{I}_{\mathsf{B}}^{d}\}, \quad \forall\, a \in \mathcal{I}_{\mathsf{A}}^{d}, d \in \{1, \ldots, D\}, \ell \in \{1, \ldots, r\}. \tag{3.17}$$

We use the symbols $\boldsymbol{\mathfrak{y}}$ and $\boldsymbol{\mathfrak{z}}$ to denote imprecise, set-valued realizations. The tilde expresses that we refer to an observation in the data file which is synthetic and generated by imputation. The partially synthetic observations in a complete file where all missing observations of $\boldsymbol{Y}$ in B and $\boldsymbol{Z}$ in A have been replaced, are either $(\boldsymbol{x}_a, \boldsymbol{y}_a, \tilde{\boldsymbol{\mathfrak{z}}}_a)$ for $a \in \mathcal{I}_{\mathsf{A}}$ or $(\boldsymbol{x}_b, \tilde{\boldsymbol{\mathfrak{y}}}_b, \boldsymbol{z}_b)$ for $b \in \mathcal{I}_{\mathsf{B}}$.

The results of a statistical matching micro approach and a statistical matching macro approach can usually be straightforwardly converted into each other. In the context of imprecise imputation, this relationship is not as obvious, so I will now go into more detail on how the partially set-valued data obtained by our approach can be used to find probability assertions for an event of interest. In contrast to, for instance, *fractional hot deck imputation*[57], we directly process the partially set-valued data by employing so-called *random sets* to obtain lower and upper bounds for the probabilities of the events of interest[58].

In order to use the partially set-valued observations in the resulting complete synthetic data file to obtain probability statements about some events of interest, we have to replace the random variables $\boldsymbol{Y}$ and $\boldsymbol{Z}$ by the *disjunctive*[59] (or *epistemic*) *random sets* $\boldsymbol{\mathfrak{Y}}$ and $\boldsymbol{\mathfrak{Z}}$. All realizations of $\boldsymbol{\mathfrak{Y}}$ and $\boldsymbol{\mathfrak{Z}}$ are now interpreted as set-valued. Thus, intrinsically precise observations $\boldsymbol{\mathfrak{y}}_a$ and $\boldsymbol{\mathfrak{z}}_b$ are treated as singletons. In simple terms, a random set is a set-valued generalization of a random variable and it is defined as a multi-valued mapping from the sample space Ω to the corresponding power set of the Cartesian product of the sets of possible values of the random variables with non-empty images. Following Couso et al. (e.g. 2014, p. 3), we define the random set in the context of imprecise imputation as

$$(\boldsymbol{X}, \boldsymbol{\mathfrak{Y}}, \boldsymbol{\mathfrak{Z}}) : \Omega \mapsto \mathcal{P}(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}) \backslash \emptyset. \tag{3.18}$$

Using this definition, we are able to derive lower and upper probabilities, π_* and π^*, for an event $\mathcal{E} \subseteq \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ of interest under *strong measurability*[60] as

$$\pi_*(\mathcal{E}) := \pi(\mathcal{E}_*), \text{ and } \pi^*(\mathcal{E}) := \pi(\mathcal{E}^*) \tag{3.19}$$

where π denotes the precise probability measure. The term $\mathcal{E}_* := \{\omega \in \Omega | (\boldsymbol{X}, \boldsymbol{\mathfrak{Y}}, \boldsymbol{\mathfrak{Z}})(\omega) \subseteq \mathcal{E}\}$ defines the *lower inverse* which is the set of elements in Ω contained in the event of interest whose images are unequal to the empty set (Couso et al., 2014, p. 12), and $\mathcal{E}^* := \{\omega \in \Omega | (\boldsymbol{X}, \boldsymbol{\mathfrak{Y}}, \boldsymbol{\mathfrak{Z}})(\omega) \cap \mathcal{E} \neq \emptyset\}$ defines the *upper inverse* which is the set of elements in Ω 'hitting' the

[57] Fractional hot deck imputation also imputes sets of plausible values for a missing entry in a data file. However, its aim is to obtain a precise synthetic observation from this set of imputed values using a special weighting scheme. Thus, it does not directly process the set-valued imputations but the resulting precise synthetic observations. Details can be found in Kim and Fuller (2004).

[58] The bounds obtained by the approach suggested in *Contribution 4* 'envelop' the results of fractional hot deck imputation in the sense that they are a superset.

[59] The *disjunctive* interpretation refers to partial knowledge about a precise quantity. It is in contrast to the *conjunctive* or *ontic* interpretation which refers to the perfect knowledge about an imprecise quantity. For details, see, for instance, Couso and Dubois (2014).

[60] A mapping is strongly measurable if the *upper inverse* is an element of the sigma algebra of the probability space for all elements in the sigma algebra of the corresponding measurable space. See Couso et al. (2014, Chap. 2) for a concrete definition and further details. Since we are exclusively dealing with categorical data, we can employ the power set of $\mathcal{P}(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}) \backslash \emptyset$ as suitable sigma algebra.

event (Couso et al., 2014, p. 12). As shown by Couso et al. (2014, Chap. 2.3), these bounds enclose the true probability of an underlying *ill-observed*[61] random variable. The lower and upper probabilities $\pi_*(\mathcal{E})$ and $\pi^*(\mathcal{E})$ induce all probability measures contained in the corresponding credal set. Conditioning within the context of disjunctive random sets is recapitulated in Section 4.2 in *Contribution 4*. The estimation of the lower and upper probabilities is described in Section 4.3 in *Contribution 4*. The estimators are derived by the relative frequencies within the partially set-valued data, where the lower probability is computed using all observations which are a subset or equal the event of interest, while the upper probability is computed using all observations that do not contradict the event of interest.

In a simulation study, we investigated the performance and usefulness of imprecise imputation for statistical matching. For this purpose, we developed a new simulation procedure to generate categorical data (with two or three possible categories) following a pre-defined dependence structure. The results of this simulation study show that indeed the lower and upper probabilities almost always envelope the true parameters of the underlying marginal and joint distributions. Furthermore, we could show that the intervals bounded by the lower and upper probabilities are small enough to provide useful information.

To make imprecise imputation available for users, we wrote a package called ‘impimp’ for the statistical programming software R (Fink and Endres, 2019). Domain imputation, variable-wise imputation and case-wise imputation are implemented. Furthermore, the package provides functions to estimate the lower and upper bounds for the parameters of interest, i.e. conditional and unconditional probability components, from imprecisely imputed data files.

3.4.2 Comments and Perspectives

Although we have introduced imprecise imputation in the framework of statistical matching, it is a general imputation method that can be used for any kind of missing data. Especially imprecise domain imputation, which imputes all possible values for a missing data entry, does not rely on the assumption that the missing data mechanism is missing (completely) at random. Thus, it is also applicable for data, which are *missing not at random*[62].

Furthermore, as I will explain in more detail in the concluding remarks in Chapter 4, the data file generated by imprecise imputation can be analysed by standard statistical methods for complete data if we sample precise data files[63] from it. If we sample only one complete synthetic data file from it, this approach corresponds to a single imputation approach. Repeating the sampling of precise data files several times, we can apply a multiple imputation pooling technique to bring the results obtained from the different files together. Although this procedure would yield precise results, it is less cautious than using the random set approach described in detail in *Contribution 4* to derive bounds for the probability estimates of certain events.

We could also take it to the extreme and use the most cautious imaginable approach. Assuming that we are interested in the probability mass distributions according to a collection of random variables $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$, the only statement that we can make –without making further assumptions

[61] This means that we interpret the random set in the disjunctive notion as representing imprecise outcomes of a precise random variable (Couso et al., 2014, pp. 17–18).

[62] Following Little and Rubin (2002, p. 12), missing not at random means that the missingness is dependent on the missing values in the data itself.

[63] With ‘precise data file’, I refer to a data file containing precise observations only. Interpreted in the context of *Contribution 4*, the precise observations are singletons.

or using additional information– is that $\pi(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})$ is indeed a probability distribution. Thus, there exists a *set* of possible distributions for $(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z})$, which is the so-called *vacuous* credal set[64] and which can be defined as

$$\begin{aligned}\mathcal{M}_0(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}) := \{\pi | \forall(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) \in \mathcal{X}\times\mathcal{Y}\times\mathcal{Z} : \pi(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) \geq 0,\\ \sum_{(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}} \pi(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) = 1\}.\end{aligned} \tag{3.20}$$

In the context of statistical matching there is additional information given in the incomplete file $\mathsf{A} \uplus \mathsf{B}$ which we can use to reduce the vacuous credal set. Imprecise domain imputation, which is the most cautious imputation approach in *Contribution 4*, reduces the vacuous credal set using the dependencies given in $\mathsf{A} \uplus \mathsf{B}$. A further reduction can be derived by variable-wise imputation and case-wise imputation, each of which strengthens the existing dependence relations even more. As we have already shown in *Contribution 4*, the credal sets $\mathcal{M}^D$, $\mathcal{M}^{VW}$, and $\mathcal{M}^{CW}$ of probability distributions for $(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z})$ obtained by domain imputation, variable-wise imputation and case-wise imputation, respectively, are nested

$$\mathcal{M}^{CW}(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}) \subseteq \mathcal{M}^{VW}(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}) \subseteq \mathcal{M}^{D}(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}) \subseteq \mathcal{M}_0(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}). \tag{3.21}$$

These credal sets can be further reduced by the incorporation of *logical constraints.* If the logical constraints are integrated by removing impossible combinations of observations from the imprecisely imputed data file *before* the estimation of the probability distributions of interest takes place, the resulting marginal distributions and the joint distribution are always compatible. However, this is different when a statistical matching macro approach is used. With this kind of approaches, the estimation takes place first and then the credal sets are further reduced. Now it may happen that incoherences appear. As already noted in the overview of literature in Section 1.3, Vantaggi (2008), Brozzi et al. (2012), and Di Zio and Vantaggi (2017) consider the issue of incoherence in the context of statistical matching.

A further reduction of these credal set can be achieved by using a conditional independence concept. As already indicated in Subsection 3.1.2, within the context of imprecise probabilities, we are provided with different independence concepts as, for example, *strong independence* or *epistemic irrelevance*, which coincide in the precise case.

For a conditional credal set $\mathcal{M}(Y,Z|X)$ of three categorical random variables, the specific variables Y and Z are strongly independent given X if[65]

$$\pi(y,z|x) = \pi(y|x)\cdot\pi(z|x) \tag{3.22}$$

holds for every $\pi(y,z|x) \in \text{ext}[\mathcal{M}(Y,Z|X)]$, and every $x \in \mathcal{X}$ (e.g. Antonucci et al., 2014, p. 213). The term $\text{ext}[\mathcal{M}(Y,Z|X)]$ denotes the extreme points[66] of the conditional credal set.

This conditional independence concept can be used to formulate an amended version of the Markov property, which is the basis to define the chain rule for credal networks. The following

[64] To stay consistent with the usually used notation in the context of imprecise probabilistic graphical models, the notation in this thesis differs to the notation in *Contribution 4*. Usually, a credal set is a set of probability measures on the corresponding measurable space, for the same event (e.g. Cozman and Walley, 2005). Since I am exclusively dealing with categorical variables, I use a set of probability mass distributions to represent this credal set in accordance with the respective literature. The vacuous credal set $\mathcal{M}_0(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z})$ describes the set of *all* possible probability distributions for $(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z})$ (e.g. Antonucci et al., 2014, p. 208).

[65] This definition of conditional stochastic independence can be derived by dividing Equation (1.2) by $\pi(\boldsymbol{x})$.

[66] Extreme points of a credal set cannot be expressed as convex combinations of other elements in this set (e.g. Antonucci et al., 2014, p. 208). Graphically, they specify the extreme points of the polytope in the corresponding probability simplex.

form of the chain rule results from a fork connection as depicted in Figure 2.2:

$$\begin{aligned}\mathcal{M}(X,Y,Z) &= \mathcal{M}(Y|\boldsymbol{Pa}(Y)) \otimes \mathcal{M}(Z|\boldsymbol{Pa}(Z)) \otimes \mathcal{M}(X|\boldsymbol{Pa}(X)) && (3.23)\\ &= \mathcal{M}(Y|X) \otimes \mathcal{M}(Z|X) \otimes \mathcal{M}(X) && (3.24)\\ &:= \mathrm{CH}\,\{\pi|\pi(x,y,z) = \pi(y|x)\cdot\pi(z|x)\cdot\pi(x), \forall(x,y,z) \in \mathcal{X}\times\mathcal{Y}\times\mathcal{Z},\\ &\qquad \forall\pi(y|x) \in \mathrm{ext}[\mathcal{M}(Y|X)], \pi(z|x) \in \mathrm{ext}[\mathcal{M}(Z|X)],\\ &\qquad \pi(x) \in \mathrm{ext}[\mathcal{M}(X)]\}\,. && (3.25)\end{aligned}$$

The three factors on the right hand side of Equation (3.23) represent the local imprecise probability models for the nodes $\dot{X}$, $\dot{Y}$, and $\dot{Z}$ in the corresponding DAG which can be simplified to Equation (3.24) in the context of three nodes. In Equation (3.25), CH denotes the convex hull of the set. The joint credal set $\mathcal{M}(X,Y,Z)$ is also called *strong extension* (Cozman, 2000) and it includes all precise probability distributions that can be derived by the elements of the conditional and unconditional credal sets satisfying the chain rule of Bayesian networks as defined in Equation (2.1).

This general approach can also be reconciled with imprecise imputation. As, for example, shown by (Antonucci et al., 2014, Chap. 9.4.1), the local models for a *non-separately specified* credal network can also be learned from incomplete data. The local models are then restricted to sets of probability tables that are in accordance with the available data, reflecting the dependencies that are present in the data. Given a pre-defined graph structure like the formerly used fork connection, the local models corresponding to every node given its parents, are learned from data by considering all possible complete data files that can be achieved by imputing all possible realizations for a missing entry. Since we are considering all possible complete data files that are compatible with the already available data, we end up with sets of conditional or unconditional local models for every node. This procedure is very similar to imprecise domain imputation with the exception that domain imputation does not predetermine any conditional or unconditional dependencies by default. However, using, for instance, strong independence in combination with domain imputation would –in most cases– yield smaller intervals for the estimates of the parameters of interest.

Moreover, this procedure resembles the *robust Bayesian estimator* introduced by Ramoni and Sebastiani (2001), which bounds conditional probability estimates in the context of Bayesian networks –obtained in a Bayesian estimation framework using so-called virtual frequencies– by considering all possible consistent completions of an incomplete data file, without making assumptions about the underlying missing data mechanism.

A weaker and asymmetric independence concept in the context of imprecise probabilities is *epistemic irrelevance*[67]. Following, for instance, Cozman and Walley (2005), we say that Z is epistemically irrelevant to Y given X if the knowledge about the value of Z does not reduce our uncertainty about Y if we already know the value of X. In the context of credal sets this epistemic irrelevance can be expressed as the equality of the convex hull of $\mathcal{M}(Y|X,Z)$ and the convex hull of $\mathcal{M}(Y|X)$ for all $z \in \mathcal{Z}$ (Cozman and Walley, 2005).

Strong independence implies epistemic irrelevance. Thus, the strong extension is a subset or equal to the *extension based on epistemic irrelevance*[68], sometimes also called the *epistemic extension* (e.g. Mauá et al., 2014). The set of parameter estimates in the context of statistical

[67] As stated by de Cooman et al. (2010), this independence concept comes into play when the *sensitivity analysis interpretation* of credal networks is not sustainable due to inherent imprecision. The sensitivity analysis interpretation means that for a given graph structure, the parameters are precise but unknown and a set of possible parameters is considered to account for this uncertainty.

[68] As stated by de Campos and Cozman (2007), it is the largest joint credal sets satisfying the Markov property under epistemic irrelevance.

matching can be reduced, although not as much as with strong independence. See, for instance, de Cooman et al. (2010), de Campos and Cozman (2007) or de Bock and de Cooman (2015) for discussions and the usage of epistemic irrelevance as independence concept in the context of credal networks.

4 Overall concluding remarks

This cumulative thesis considers the problem of statistical matching with categorical data, which arises if we need joint information about variables that have not been jointly observed. The contributions of the thesis aim at the development of new statistical matching techniques. All presented methods are suitable for the inclusion of expert knowledge. In *Contributions 1–3*, which make use of the theory of probabilistic graphical models, additional knowledge might be considered in the form of dependence structures among sets of variables, or the parameters of the model can be determined –partially or completely– by hand. The basic link between statistical matching and probabilistic graphical models is built by the assumption of conditional independence between the specific variables given the common variables. As this assumption is sometimes considered critical, in *Contribution 4*, I also regard a statistical matching procedure in a partial identification context. Imprecise imputation, which aims at the replacement of missing values in a data file by sets of plausible values, does not rely on unjustified assumptions and supports the inclusion of logical constraints in the imputation step. In contrast to standard imputation techniques, it does not aim at the construction of precise data files but it yields partial set-valued observations. These observations are subsequently processed to obtain lower and upper bounds for the estimates of the parameters of interest.

Any of the approaches presented in the contributions of this thesis can be used to build a synthetic data basis for the analysis with standard statistical methods. In *Contributions 1–3* it is straightforward to sample observations from the joint distribution obtained by statistical matching, as mentioned in Section 3.2.2. Furthermore, we can decide whether the resulting data file should be entirely synthetic meaning that a complete observations is randomly drawn or if we just want to fill the missing values in A, B, or $\mathsf{A} \Cup \mathsf{B}$ as sketched in Section 3.1.1 for the statistical matching approach with Bayesian networks. These synthetic data obtained from random draws based on the statistically matched joint distribution also serves as basis for multiple imputation. Several data sets can be simulated and separately analysed with a subsequent pooling of the results.

Although not as straightforward as in the context of probabilistic graphical models, also the partially set-valued synthetic file obtained by imprecise imputation can be used to generate complete synthetic and precise data files which can be analysed by standard statistical methods. This aim is achieved by sampling a precise donor value from the set of imputed values. This sampling can be performed with or without a former weighting of the elements of the imprecisely imputed sets. A multiple imputation approach is then realised by repeating these random draws several times and pooling the results. Moreover, the partially set-valued data file obtained by imprecise imputation can be interpreted as a hull of multiple imputation. It means that the set-valued data already encapsulates all possible precise data files that can be generated with single or multiple imputation.

However, besides the mentioned advantages, the newly introduced procedures are –as other statistical matching methods– facing some limitations. All procedures that have been introduced in this thesis are based on two basic assumptions:

(i) the missing data mechanism is ignorable;

(ii) the data in A and B are independently and identically distributed, originating from a joint probability distribution $\pi(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$.

In practice, the first assumption seems to be justified as being “a consequence of the data generating process” (D’Orazio et al., 2006b, p. 99). However, the second assumptions cannot be verified and it may not be valid in real applications. Further research might explore whether the *knowledge-based bias correction* approach proposed by Krak and van der Gaag (2014) can be used in the context of statistical matching for non i.i.d. data. They correct the estimates for an early-warning system for classical swine fever in pigs obtained from a Bayesian network approach on data originating from different sources by incorporating expert knowledge.

Other questions that require further investigation concern, for instance, a systematic comparison of different statistical matching procedures by simulation studies. In addition, up to now there seems to be no statistical matching method that addresses the uncertainty arising from the identifiability problem as well as the uncertainty inherent in sampling. A further natural progression of this work is to develop a statistical matching procedure suitable for ordinal data. Furthermore, a new project could address the question whether correction methods for misclassification improve the estimates obtained from the synthetic data by interpreting the synthetic observations as true observations occupied with a measurement error.

Summing up, statistical matching of different data sources requires a lot of former preparation which has not directly been considered in this thesis. This includes the comparison of different definitions and operationalizations of the involved variables, the selection of the matching variables, and the selection of the matching procedure. Even if all these steps are carefully conducted, we must not forget that the results obtained from synthetic, statistically matched data typically reflect the models and assumptions used for the integration process (e.g. Drechsler, 2010, p. 109).

Bibliography

Ahfock, D., S. Pyne, S. X. Lee, and G. J. McLachlan (2016). Partial identification in the statistical matching problem. *Computational Statistics & Data Analysis 104*, 79–90. DOI: https://doi.org/10.1016/j.csda.2016.06.005.

Aluja-Banet, T., J. Daunis-i-Estadella, N. Brunsó, and A. Mompart-Penina (2015). Improving prevalence estimation through data fusion: Methods and validation. *BMC Medical Informatics and Decision Making 15*, 49. DOI: https://doi.org/10.1186/s12911-015-0169-z.

Aluja-Banet, T., J. D. i Estadella, and Y. Chen (2013). Enriching a large-scale survey from a representative sample by data fusion: Models and validation. In C. Davino and L. Fabbris (Eds.), *Survey Data Collection and Integration*, pp. 121–138. Berlin: Springer.

Anderson, T. W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association 52*, 200–203.

Antonucci, A., C. P. de Campos, and M. Zaffalon (2014). Probabilistic graphical models. In T. Augustin, F. P. Coolen, G. de Cooman, and M. C. Troffaes (Eds.), *Introduction to Imprecise Probabilities*, pp. 207–229. Chichester: Wiley. DOI: https://doi.org/10.1002/9781118763117.

Barber, D. (2012). *Bayesian Reasoning and Machine Learning.* Cambridge: Cambridge University Press.

Barber, R. F. and M. Drton (2015). High-dimensional Ising model selection with Bayesian information criteria. *Electronic Journal of Statistics 9*, 567–607. DOI: https://doi.org/10.1214/15-EJS1012.

Bastert, O. and C. Matuszewski (2001). Layered drawings of digraphs. In M. Kaufmann and D. Wagner (Eds.), *Drawing Graphs: Methods and Models*, pp. 87–120. Berlin: Springer.

Björnberg, J. E. (2009). *Graphical Representations of Ising and Potts Models: Stochastic Geometry of the Quantum Ising Model and the Space–Time Potts Model.* Ph. D. thesis, University of Cambridge.

Brozzi, A., A. Capotorti, and B. Vantaggi (2012). Incoherence correction strategies in statistical matching. *International Journal of Approximate Reasoning 53*, 1124–1136. DOI: https://doi.org/10.1016/j.ijar.2012.06.009.

Christen, P. (2012). *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection.* Berlin: Springer.

Conti, P. L., D. Marella, and M. Scanu (2008). Evaluation of matching noise for imputation techniques based on nonparametric local linear regression estimators. *Computational Statistics & Data Analysis 53*, 354–365. DOI: https://doi.org/10.1016/j.csda.2008.07.041.

Conti, P. L., D. Marella, and M. Scanu (2012). Uncertainty analysis in statistical matching. *Journal of Official Statistics 28*, 69–88.

Conti, P. L., D. Marella, and M. Scanu (2013). Uncertainty analysis for statistical matching of ordered categorical variables. *Computational Statistics & Data Analysis 68*, 311–325. DOI: https://doi.org/10.1016/j.csda.2013.07.004.

Conti, P. L., D. Marella, and M. Scanu (2016). Statistical matching analysis for complex survey data with applications. *Journal of the American Statistical Association 111*, 1715–1725. DOI: https://doi.org/10.1080/01621459.2015.1112803.

Conti, P. L., D. Marella, and M. Scanu (2017). How far from identifiability? A systematic overview of the statistical matching problem in a non parametric framework. *Communications in Statistics – Theory and Methods 46*, 967–994. DOI: https://doi.org/10.1080/03610926.2015.1010005.

Couso, I. and D. Dubois (2014). Statistical reasoning with set-valued information: Ontic vs. epistemic views. *International Journal of Approximate Reasoning 55*, 1502–1518. DOI: https://doi.org/10.1016/j.ijar.2013.07.002.

Couso, I., D. Dubois, and L. Sánchez (2014). *Random Sets and Random Fuzzy Sets as Ill-Perceived Random Variables.* Cham: Springer. DOI: https://doi.org/10.1007/978-3-319-08611-8.

Cozman, F. G. (2000). Credal networks. *Artificial Intelligence 120*, 199–233.

Cozman, F. G. and P. Walley (2005). Graphiod properties of epistemic irrelevance and independence. *Annals of Mathematics and Artificial Intelligence 45*, 173–195. DOI: https://doi.org/10.1007/s10472-005-9004-z.

D'Agostino, Jr., R. B. (1998). Tutorial in biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine 17*, 2265–2281.

de Bock, J. (2015). *Credal Networks under Epistemic Irrelevance: Theory and Algorithms.* Ph. D. thesis, Universiteit Gent.

de Bock, J. and G. de Cooman (2015). Credal networks under epistemic irrelevance: The sets of desirable gambles approach. *International Journal of Approximate Reasoning 56*, 178–207. DOI: https://doi.org/10.1016/j.ijar.2014.07.002.

de Campos, C. P. and F. G. Cozman (2007). Computing lower and upper expectations under epistemic independence. *International Journal of Approximate Reasoning 44*, 244–260. DOI: https://doi.org/10.1016/j.ijar.2006.07.013.

de Cooman, G., F. Hermans, A. Antonucci, and M. Zaffalon (2010). Epistemic irrelevance in credal nets: The case of imprecise Markov trees. *International Journal of Approximate Reasoning 51*, 1029–1052. DOI: https://doi.org/10.1016/j.ijar.2010.08.011.

Di Zio, M. and B. Vantaggi (2017). Partial identification in statistical matching with misclassification. *International Journal of Approximate Reasoning 82*, 227–241. DOI: https://doi.org/10.1016/j.ijar.2016.12.015.

D'Orazio, M., M. Di Zio, and M. Scanu (2004). Statistical matching and the likelihood principle: Uncertainty and logical constraints. Technical Report Contributi 2004/1, Italian Statistical Ististute.

D'Orazio, M., M. Di Zio, and M. Scanu (2006a). Statistical matching for categorical data: Displaying uncertainty and using logical constraints. *Journal of Official Statistics 22*, 137–157.

D'Orazio, M., M. Di Zio, and M. Scanu (2006b). *Statistical Matching: Theory and Practice.* Chichester, United Kingdom: Wiley. DOI: https://doi.org/10.1002/0470023554.

D'Orazio, M., M. Di Zio, and M. Scanu (2017). The use of uncertainty to choose matching variables in statistical matching. *International Journal of Approximate Reasoning 90*, 433–440. DOI: https://doi.org/10.1016/j.ijar.2017.08.015.

Drechsler, J. (2010). *Generating Multiply Imputed Synthetic Datasets: Theory and Implementation.* Ph. D. thesis, Otto-Friedrich-Universität Bamberg.

Elmenreich, W. (2002). *Sensor Fusion in Time-Triggered Systems.* Ph. D. thesis, Technische Universität Wien.

Eurostat (2018). About Eurostat: Overview. `https://ec.europa.eu/eurostat/en/about/overview`, [Accessed 29.01.2019].

Eurostat (2019a). Glossary: EU 2020 strategy. `https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:EU_2020_Strategy`, [Accessed: 09.04.2019].

Eurostat (2019b). People at risk of poverty or social exclusion. `https://ec.europa.eu/eurostat/statistics-explained/index.php/People_at_risk_of_poverty_or_social_exclusion`, [Accessed: 19.01.2019].

Fahrmeir, L., T. Kneib, S. Lang, and B. Marx (2013). *Regression: Models, Methods and Applications.* Heidelberg: Springer. DOI: https://doi.org/10.1007/978-3-642-34333-9.

Fink, P. and E. Endres (2019). *impimp: Imprecise Imputation for Statistical Matching.* URL: https://CRAN.R-project.org/package=impimp. R package version 0.3.1.

Gasse, M., A. Aussem, and H. Elghazel (2014). A hybrid algorithm for Bayesian network structure learning with application to multi-label learning. *Expert Systems with Applications 41*, 6755–6772. DOI: https://doi.org/10.1016/j.eswa.2014.04.032.

GESIS – Leibniz Institute for the Social Sciences (2013). Allgemeine Bevölkerungsumfrage der Sozialwissenschaften ALLBUS 2012/German General Social Survey GGSS 2012. GESIS Data Archive, Cologne. ZA4614 Data file Version 1.1.1, DOI: https://doi.org/10.4232/1.12209.

Højsgaard, S., D. Edwards, and S. Lauritzen (2012). *Graphical Models with R.* New York: Springer. DOI: https://doi.org/10.1007/978-1-4614-2299-0.

Ising, E. (1925). Beitrag zur Theorie des Ferromagnetismus. *Zeitschrift für Physik 31*, 253–258.

Khaleghi, B., A. Khamis, F. O. Karray, and S. N. Razavi (2013). Multisensor data fusion: A review of the state-of-the-art. *Information Fusion 14*, 28 – 44. DOI: https://doi.org/10.1016/j.inffus.2011.08.001.

Kim, J. K. and W. Fuller (2004). Fractional hot deck imputation. *Biometrika 91*, 559–578. DOI: https://doi.org/10.1093/biomet/91.3.559.

Kindermann, R. and J. L. Snell (1980). *Markov Random Fields and Their Applications.* Providence: American Mathematical Society.

Kjræulff, U. B. and A. L. Madsen (2013). *Bayesian Networks and Influence Diagrams: A Guide to Construction and Analysis* (2nd ed.). New York: Springer. DOI: https://doi.org/10.1007/978-1-4614-5104-4.

Koller, D. and N. Friedman (2009). *Probabilistic Graphical Models: Principles and Techniques.* Cambridge, MA: MIT Press.

Koopmans, T. (1949). Identification problems in economic model cosntruction. *Econometrica 17*, 125–144.

Krak, T. E. and L. C. van der Gaag (2014). Knowledge-based bias correction - a case study in veterinary decision support. In T. Schaub, G. Friedrich, and B. O'Sullivan (Eds.), *ECAI 2014: 21st European Conference on Artificial Intelligence*, Amsterdam, pp. 453–460. IOS Press. DOI: https://doi.org/10.3233/978-1-61499-419-0-489.

Kuss, O., M. Blettner, and J. Bögermann (2016). Propensity score: An alternative method of analyzing treatment effects. *Deutsches Ärzteblatt International 113*, 597–603.

Landes, J. and J. Williamson (2016). Objective Bayesian nets from consistent datasets. In A. Giffin and K. H. Knuth (Eds.), *AIP Conference Proceedings*, Volume 1757, Potsdam, NY, pp. 020007–1 – 020007–8. DOI: https://doi.org/10.1063/1.4959048.

Lauritzen, S. L. (1996). *Graphical Models.* Oxford: Oxford University Press. Reprinted version with corrections.

Leulescu, A. and M. Agafitei (2013). Statistical matching: A model based approach for data integration. Luxembourg: Publications Office of the European Union. Methodologies and Working Papers, DOI: https://doi.org/10.2785/44822.

Little, R. J. and D. B. Rubin (2002). *Statistical Analysis with Missing Data* (2nd ed.). Hoboken: Wiley. DOI: https://doi.org/10.1002/9781119013563.

Madsen, A. (2008). Belief update in CLG Bayesian networks with lazy propagation. *International Journal of Approximate Reasoning 49*, 503–521. DOI: https://doi.org/10.1016/j.ijar.2008.05.001.

Manski, C. F. (1995). *Identification Problems in the Social Sciences.* Cambridge: Harvard University Press.

Marella, D., P. L. Conti, and M. Scanu (2012). Uncertainty in statistical matching for discrete categorical variables. `https://www.sis-statistica.it/old_upload/contenuti/2013/09/RS12-Uncertainty-in-statistical-matching-for-discrete.pdf` (46th Scientific Meeting of the Italian Statistical Society), [Accessed: 19.02.2019].

Marella, D., M. Scanu, and P. L. Conti (2008). On the matching noise of some nonparametric imputation procedures. *Statistics & Probability Letters 78*, 1593–1600. DOI: https://doi.org/10.1016/j.spl.2008.01.020.

Margaritis, D. (2003). *Learning Bayesian Network Model Structure from Data.* Ph. D. thesis, School of Computer Science, Carnegie Mellon University.

Mauá, D., C. de Campos, A. Benavoli, and A. Antonucci (2014). Probabilistic inference in credal networks: New complexity results. *Journal of Artificial Intelligence Research 50*, 603–637.

McCoy, B. M. and T. T. Wu (1973). *The Two-Dimensional Ising Model.* Cambridge: Harvard University Press.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective.* Cambridge: MIT Press.

Nagarajan, R., M. Scutari, and S. Lèbre (2013). *Bayesian Networks in R with Applications in Systems Biology.* New York: Springer.

Newger, K. (2018). Statistical matching of categorical data with Markov networks. Master's thesis, Ludwig-Maximilians-Universität München. DOI: https://doi.org/10.5282/ubm/epub.59116.

Okner, B. A. (1972). Constructing a new data base from existing microdata sets: The 1966 merge file. *Annals of Economic and Social Measurement 1*, 325–342.

Pigott, T. D. (2001). A review of methods for missing data. *Educational Research and Evaluation 7*, 353–383. DOI: https://doi.org/10.1076/edre.7.4.353.8937.

Pourret, O., P. Naim, and B. Marcot (Eds.) (2008). *Bayesian Networks: A Practical Guide to Applications*. Chichester: Wiley.

R Core Team (2019). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. URL: https://www.R-project.org. [Accessed 08.03.2019].

Ramoni, M. and P. Sebastiani (2001). Robust learning with missing data. *Machine Learning 45*, 147–170.

Rässler, S. (2002). *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*. New York: Springer. DOI: https://doi.org/10.1007/978-1-4613-0053-3.

Rässler, S. (2004). Data fusion: Identification problems, validity, and multiple imputation. *Austrian Journal of Statistics 33*, 153–171.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*, 41–55.

Rubin, D. B. (1978). Multiple imputations in sample surveys – A phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 20–28.

Serafino, P. and R. Tonkin (2017a). Comparing poverty estimates using income, expenditure and material deprivation. In A. B. Atkinson, A.-C. Guio, and E. Marlier (Eds.), *Monitoring social inclusion in Europe*, pp. 241–258. Luxembourg: Publications Office of the European Union. `http://ec.europa.eu/eurostat/documents/3217494/8031566/KS-05-14-075-EN-N.pdf/c3a33007-6cf2-4d86-9b9e-d39fd3e5420c`, [Accessed: 28.01.2019].

Serafino, P. and R. Tonkin (2017b). Statistical matching of European Union statistics on income and living conditions (EU-SILC) and the household budget survey. Luxembourg: Publications Office of the European Union. Collection: Statistical working papers, DOI: https://doi.org/10.2785/933460.

Singh, A. C., H. J. Mantel, M. D. Kinack, and G. Rowe (1993). Statistical matching: Use of auxiliary information as an alternative to the conditional independence assumption. *Survey Methodology 19*, 59–79.

Tsamardinos, I., L. E. Brown, and C. F. Aliferis (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning 65*, 31–78. DOI: https://doi.org/10.1007/s10994-006-6889-7.

Tsamardinos, I., S. Triantafillou, and V. Lagani (2012). Towards integrative causal analysis of heterogeneous data sets and studies. *Journal of Machine Learning Research 13*, 1097–1157.

Tutz, G. (2011). *Regression for Categorical Data.* Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511842061.

van Borkulo, C., D. Borsboom, S. Epskamp, T. Blanken, L. Boschloo, R. Schoevers, and L. Waldorp (2014). A new method for constructing networks from binary data. *Scientific Reports 4*, 1–10. DOI: https://doi.org/10.1038/srep05918.

van Borkulo, C., S. Epskamp, and with contributions from Alexander Robitzsch (2016). *IsingFit: Fitting Ising Models Using the ELasso Method.* URL: https://CRAN.R-project.org/package=IsingFit. R package version 0.3.1.

Vantaggi, B. (2008). Statistical matching of multiple sources: A look through coherence. *International Journal of Approximate Reasoning 49*, 701–711. DOI: https://doi.org/10.1016/j.ijar.2008.07.005.

Wainwright, M. J. and M. I. Jordan (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning 1*, 1–305. DOI: https://doi.org/10.1561/2200000001.

Webber, D. and R. Tonkin (2013). Statistical matching of EU-SILC and the Household Budget Survey to compare poverty estimates using income, expenditures and material deprivation. Luxembourg: Publications Office of the European Union. Methodologies and Working Papers, DOI: https://doi.org/10.2785/4151.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics.* Chichester: Wiley.

Wu, F. Y. (1982). The Potts model. *Reviews of Modern Physics 54*, 235–268.

Zhang, L.-C. (2015). On proxy variables and categorical data fusion. *Journal of Official Statistics 31*, 783–807. DOI: https://doi.org/10.1515/jos-2015-0045.

Attached contributions

Contribution 1: pp. 46–57

Endres, E. and Augustin, T. (2016). Statistical matching of discrete data by Bayesian networks, in A. Antonucci, G. Corani and C. P. de Campos (eds), Proceedings of the Eighth International Conference on Probabilistic Graphical Models, Vol. 52 of Proceedings of Machine Learning Research, PMLR, Lugano, Switzerland, pp. 159–170.

The original publication is available at

`http://proceedings.mlr.press/v52/endres16.html`

JMLR: Workshop and Conference Proceedings vol 52, 159-170, 2016 PGM 2016

Statistical Matching of Discrete Data by Bayesian Networks

Eva Endres EVA.ENDRES@STAT.UNI-MUENCHEN.DE
Thomas Augustin AUGUSTIN@STAT.UNI-MUENCHEN.DE
Department of Statistics, Ludwig-Maximilians-Universität München
Munich (Germany)

Abstract

Statistical matching (also known as data fusion, data merging, or data integration) is the umbrella term for a collection of methods which serve to combine different data sources. The objective is to obtain joint information about variables which have not jointly been collected in one survey, but on two (or more) surveys with disjoint sets of observation units. Besides specific variables for the different data files, it is indispensable to have common variables which are observed in both data sets and on basis of which the matching can be performed. Several existing statistical matching approaches are based on the assumption of conditional independence of the specific variables given the common variables. Relying on the well-known fact that d-separation is related to conditional independence for a probability distribution which factorizes along a directed acyclic graph, we suggest to use probabilistic graphical models as a powerful tool for statistical matching. In this paper, we describe and discuss first attempts for statistical matching of discrete data by Bayesian networks. The approach is exemplarily applied to data collected within the scope of the German General Social Survey.

Keywords: Statistical matching; data fusion; data merging; data integration; probabilistic graphical models; Bayesian networks; conditional independence.

1. Introduction

Nowadays data is omnipresent and is constantly being collected, for example, by authorities, companies, or by surveillance systems. Thus, an immense amount of qualitative and quantitative data is already available for researchers. To save time and costs, it is much more effective to use already existing data sources for statistical analysis instead of planning and carrying out new surveys. However, single data sources are barely adequate to answer varying research questions, particularly in the case when we need joint information about variables that have not jointly been observed but in two (or more) different surveys. Let us assume that information about a specific set of variables is available in the first of the two data set, and in the second data set we have information about a disjoint set of variables. Given that there is also a set of partially overlapping variables in both data sets, we are able to fuse these data sources to achieve joint information. This procedure is commonly known as *statistical matching* (*data fusion*, *data merging*, or *data integration*). For example, Rässler (2002) or D'Orazio et al. (2006a) described different methods for statistical matching. Several of these methods are mainly based on a certain kind of conditional independence (CI, throughout the paper) assumption. This assumption is strongly related to d-separation which is a basic concept of probabilistic graphical models, where the (in)dependence structure among a set of variables is naturally represented by a graph. For this reason, we suggest to utilize probabilistic graphical models for statistical matching. In this paper, we focus on discrete data and Bayesian networks.

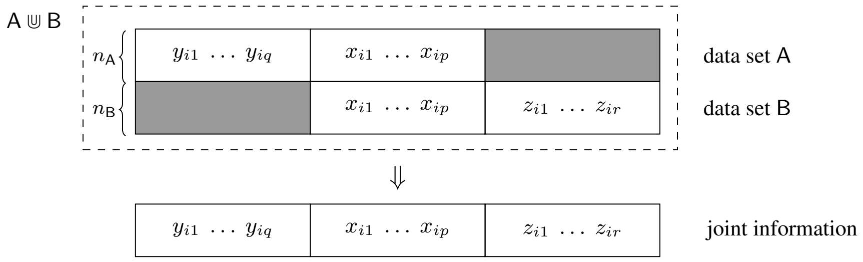


Figure 1: Schematic representation of the statistical matching problem (adapted from D'Orazio et al., 2006a, p. 5).

The paper is structured as follows. Section 1.1 outlines the framework and basic problem of statistical matching. It introduces the basic idea of statistical matching under the assumption of CI. Subsequently, Section 1.2 briefly recalls the definition of Bayesian networks and clarifies the used notations. Section 2.1 describes our basic idea for statistical matching of discrete data by Bayesian networks. Afterwards, the procedure is elucidated in three steps. In Section 3, we illustrate the proposed matching approaches by an application in the context of the German General Social Survey. Section 3.1 gives an introductory summary of the data, while Section 3.2 shows the actual application example. The corresponding results are presented in Section 3.3, which is followed by a conclusion and discussion in Section 4.

1.1 The Framework of Statistical Matching

Following, for instance, D'Orazio et al. (2006a), statistical matching aims at the combination of two (or more) data sources to gain joint information about not jointly observed variables. The data sources characteristically have a partially overlapping set of variables and disjoint sets of observations. Throughout the paper, let us assume that we have two data sets A and B, indexed by $\mathcal{I}_\mathsf{A}$ and $\mathcal{I}_\mathsf{B}$, respectively, with n_A and n_B i.i.d. observations following a common discrete distribution P. Both data sets contain information on the vector of *common variables* $\mathbf{X} = (X_1, \dots, X_p)'$, as well as vectors of *specific variables* $\mathbf{Y} = (Y_1, \dots, Y_q)'$ in A and $\mathbf{Z} = (Z_1, \dots, Z_r)'$ in B, respectively. To cut the matter short: we do not have joint information about $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$, but on $(\mathbf{X}, \mathbf{Y})$ and $(\mathbf{X}, \mathbf{Z})$. The schematic representation of this situation in Figure 1 illustrates this general framework, and shows that the statistical matching problem can also be interpreted as a missing data problem, where the shaded areas reflect the missing values. D'Orazio et al. (2006a) state that the missing values are missing completely at random (MCAR) in most of the standard applications. They give a brief justification for this statement and explain its consequences in their first chapter.

Basically, we can distinguish two main approaches of statistical matching: the *macro approach* and the *micro approach*. The main objective of the macro approach is to estimate the joint probability distribution of $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$, (or any characteristic of it), while the micro approach is geared to additionally generate a synthetic data set that contains all variables of interest (e.g. D'Orazio et al., 2006a, pp. 13). To reach these aims, it is common practice to use procedures that are premised on the assumption of CI of $\mathbf{Y}$ and $\mathbf{Z}$ given $\mathbf{X}$. This technical assumption guarantees that the joint distribution of $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ is identifiable and thus estimable on the incomplete i.i.d. sample A ⊎ B, i.e. the union of the two data sets A and B with missing joint observations of $\mathbf{Y}$ and $\mathbf{Z}$. Hence, its joint probability distribution is fully described by its probability mass distribution

$$p_{\mathbf{X},\mathbf{Y},\mathbf{Z}}(\mathbf{x},\mathbf{y},\mathbf{z}) = P(X_1 = x_1,\dots,X_p = x_p, Y_1 = y_1,\dots,Y_q = y_q, Z_1 = z_1,\dots,Z_r = z_r),$$

$x_j \in \mathcal{X}_j$, $y_k \in \mathcal{Y}_k$, $z_\ell \in \mathcal{Z}_\ell$, $j = 1,\dots,p$, $k = 1,\dots,q$, $\ell = 1,\dots,r$, where $\mathcal{X} = \mathcal{X}_1 \times \dots \mathcal{X}_p$, $\mathcal{Y} = \mathcal{Y}_1 \times \dots \times \mathcal{Y}_q$, and $\mathcal{Z} = \mathcal{Z}_1 \times \dots \times \mathcal{Z}_r$ denote the domains of $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$, and $(\mathbf{x},\mathbf{y},\mathbf{z}) := (x_1,\dots,x_p,y_1,\dots,y_q,z_1,\dots,z_r)$. Collecting all probability components of $\mathbf{p}_{\mathbf{X},\mathbf{Y},\mathbf{Z}}(\mathbf{x},\mathbf{y},\mathbf{z})$ yields a vector $\mathbf{p}$, whose $|\mathcal{X}| \cdot |\mathcal{Y}| \cdot |\mathcal{Z}|$ entries can be considered as parameters, representing the probability entries of the multidimensional contingency table of $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$. The parameters of the corresponding multinomial distribution directly follow from $\mathbf{p}$ by taking trivial restrictions on the probability components into account. Under the assumption of CI of $\mathbf{Y}$ and $\mathbf{Z}$ given $\mathbf{X}$, the joint distribution is fully determined by the conditional distributions of $\mathbf{Y}$ given $\mathbf{X}$, and $\mathbf{Z}$ given $\mathbf{X}$, together with the marginal distribution of $\mathbf{X}$. Therefore, under the assumption of CI, the parameter vector $\mathbf{p}$ simplifies to $\mathbf{p}^{\mathsf{A} \Cup \mathsf{B}} := (\mathbf{p}_{\mathbf{Y}|\mathbf{X}}, \mathbf{p}_{\mathbf{Z}|\mathbf{X}}, \mathbf{p}_{\mathbf{X}})'$ whose components are either estimable from observations $(\mathbf{x}_i, \mathbf{y}_i)$, $i \in \mathcal{I}_\mathsf{A}$, or $(\mathbf{x}_i, \mathbf{z}_i)$, $i \in \mathcal{I}_\mathsf{B}$, or $\mathbf{x}_i$, $i \in \{\mathcal{I}_\mathsf{A} \cup \mathcal{I}_\mathsf{B}\}$, and whose likelihood given $\mathsf{A} \Cup \mathsf{B}$ becomes

$$L(\mathbf{p}^{\mathsf{A} \Cup \mathsf{B}} | \mathsf{A} \Cup \mathsf{B}) = \prod_{i \in \mathcal{I}_\mathsf{A}} p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}_i|\mathbf{x}_i) \prod_{i \in \mathcal{I}_\mathsf{B}} p_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}_i|\mathbf{x}_i) \prod_{i \in \{\mathcal{I}_\mathsf{A} \cup \mathcal{I}_\mathsf{B}\}} p_{\mathbf{X}}(\mathbf{x}_i), \tag{1}$$

by selecting the appropriate component of $\mathbf{p}^{\mathsf{A} \Cup \mathsf{B}}$ for every observation.

Conditional independence, which is the crucial basis of this approach, is strongly related to the d-separation-criterion in probabilistic graphical models (e.g. Kjræulff and Madsen, 2013, pp. 32). In a probabilistic graphical model, random variables are represented by nodes. If two nodes are d-separated by a third node, it follows that the two random variables corresponding to the former two nodes are conditionally independent given the third random variable corresponding to the latter node. Based on this, we suggest to use probabilistic graphical models, more precisely, Bayesian networks for statistical matching. We clarify the notations for Bayesian networks based on discrete data used throughout this paper, hereinafter.

1.2 Bayesian Networks - Basic Concepts and Notation

A Bayesian network over the discrete random variables $\mathbf{W} = (W_1,\dots,W_s)'$ is composed of a global probability distribution $P(\mathbf{W} = \mathbf{w}) = P(W_1 = w_1,\dots,W_s = w_s)$ and a directed acyclic graph (DAG) $\mathcal{G}_\mathbf{W}$, where each random variable W_m, $m = 1,\dots,s$ is represented by an eponymous node. The graph $\mathcal{G}_\mathbf{W}$ is furthermore defined by a set $\mathfrak{E}_\mathbf{W}$ of directed edges between pairs of nodes which represents the dependencies among the random variables (e.g. Koller and Friedman, 2009, pp. 51). According to the graph $\mathcal{G}_\mathbf{W}$, the joint probability distribution $P(\mathbf{W} = \mathbf{w})$ can be factorized into smaller local probability distributions by applying the so-called *chain rule* of Bayesian networks (e.g. Koller and Friedman, 2009, p. 62)

$$P(\mathbf{W} = \mathbf{w}) = \prod_{m=1}^{s} P(W_m = w_m | \mathbf{Pa}(W_m) = \mathbf{pa}(W_m)) =: \prod_{m=1}^{s} p(w_m|\mathbf{pa}(W_m)), \tag{2}$$

where $\mathbf{Pa}(W_m)$ denotes the vector of parents of node W_m and $\mathbf{pa}(W_m)$ its realizations. This factorization of the global probability distribution exploits the Markov assumption which states that each node is conditionally independent of its non-descendants given its parents (e.g. Kjræulff and Madsen, 2013, p. 8, pp. 32).

In order to estimate a Bayesian network, i.e. a probability distribution and a DAG from data, we have to carry out two steps: structure learning and parameter learning. Structure learning means that we estimate the directed acyclic graph from the available data with the aid of score based, constraint based or hybrid learning algorithms (e.g. Koski and Noble, 2012; Koller and Friedman, 2009, chap. 17). Given the learned structure, we estimate the parameters of the local probability distributions. For this purpose, we can apply maximum likelihood estimation or Bayesian inference (e.g. Koller and Friedman, 2009, chap. 17). These local probability distributions can then be composed to the global probability distribution by means of the above-mentioned chain rule.

2. Statistical Matching by Bayesian Networks

2.1 Basic Idea

The main idea of statistical matching by Bayesian networks is the representation of the (assumed) CI of $\mathbf{Y}$ and $\mathbf{Z}$ given $\mathbf{X}$ by a directed acyclic graph and its extension by incorporating further CI assumptions determined by the Bayesian network approach. To ensure that we derive a Bayesian network which reflects the CI assumptions necessary for statistical matching, we restrict the graph to the basic structure[1] $\mathbf{Y} \leftarrow \mathbf{X} \rightarrow \mathbf{Z}$, where the common variables are the parents of the specific variables, hereinafter. This structure is known as fork connection (e.g. Koski and Noble, 2012).

Unless the joint graph structure is determined by an expert, we estimate two DAGs, one on A, and one on B, and combine them to derive a joint DAG containing all common and specific variables. On the basis of this structure, we learn the parameters of the local probability distributions on the available observations given in the incomplete sample $\mathsf{A} \uplus \mathsf{B}$ either by maximum likelihood estimation or by Bayesian inference. In a more algorithmic way, our proposed (micro) matching approach consists of three steps: estimating and combining the (directed acyclic) graphs for data sets A and B, estimating the corresponding local parameters and combining them to the joint probability distribution, and imputing the missing values in A and B to derive a complete synthetic data set. In the following, the three steps will be explained in detail.

2.2 Step 1: Estimation and Combination of the Graph Structures

For the estimation and combination of the DAGs for A and B, the following two different procedures are conceivable.

Procedure 1 (fix graph structure $\mathcal{G}_{\mathbf{X}}$ for the common variables in A and B): Initially, we estimate the Bayesian network structure $\mathcal{G}_{\mathbf{X}}$, i.e. a directed acyclic graph, for the common variables $\mathbf{X}$ on basis of all observations $\mathbf{x}_i \in \{\mathcal{I}_{\mathsf{A}} \cup \mathcal{I}_{\mathsf{B}}\}$ utilizing common structure learning algorithms for Bayesian networks. The resulting graph is denoted by $\hat{\mathcal{G}}_{\mathbf{X}}^{\mathsf{A} \uplus \mathsf{B}}$. Subsequently, we use the estimated set of directed edges $\hat{\mathfrak{E}}_{\mathbf{X}}^{\mathsf{A} \uplus \mathsf{B}}$ corresponding to $\hat{\mathcal{G}}_{\mathbf{X}}^{\mathsf{A} \uplus \mathsf{B}}$ as prior knowledge and pass it to the structure learning procedures on A and B as prior knowledge to retain the currently estimated graph structure for $\mathbf{X}$. Subject to the condition that the graph structure of the common variables is fixed, we estimate two separate DAGs $\hat{\mathcal{G}}_{\mathbf{X},\mathbf{Y}}^{\mathsf{A}}$ and $\hat{\mathcal{G}}_{\mathbf{X},\mathbf{Z}}^{\mathsf{B}}$ on the data sets A and B, respectively. This procedure ensures that the graph structures regarding to the common variables on A and B are identical and can be

[1] The basic structures $\mathbf{Y} \rightarrow \mathbf{X} \rightarrow \mathbf{Z}$ and $\mathbf{Y} \leftarrow \mathbf{X} \leftarrow \mathbf{Z}$ would be equivalent.

matched without any difficulties. The resulting joint DAG contains nodes for all variables $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ and its set of edges is composed by the union[2] $\hat{\mathfrak{E}}_{\mathbf{X}}^{\mathsf{A}\Cup\mathsf{B}} \cup \hat{\mathfrak{E}}_{\mathbf{X},\mathbf{Y}}^{\mathsf{A}} \cup \hat{\mathfrak{E}}_{\mathbf{X},\mathbf{Z}}^{\mathsf{B}}$.

Procedure 2 (individual graph structures $\mathcal{G}_{\mathbf{X}}^{\mathsf{A}}$ and $\mathcal{G}_{\mathbf{X}}^{\mathsf{B}}$ for the common variables in A and B): For the second procedure for estimating and combining the graph structures, we separately estimate two DAGs $\hat{\mathcal{G}}_{\mathbf{X},\mathbf{Y}}^{\mathsf{A}}$ and $\hat{\mathcal{G}}_{\mathbf{X},\mathbf{Z}}^{\mathsf{B}}$ with sets of edges $\hat{\mathfrak{E}}_{\mathbf{X},\mathbf{Y}}^{\mathsf{A}}$ and $\hat{\mathfrak{E}}_{\mathbf{X},\mathbf{Z}}^{\mathsf{B}}$ independently of one another on A, and on B. Since the observation units in A and B are disjoint, it cannot be ruled out that we derive different graph structures for the common variables on the two different data sets A and B, i.e. $\hat{\mathfrak{E}}_{\mathbf{X}}^{\mathsf{A}} \neq \hat{\mathfrak{E}}_{\mathbf{X}}^{\mathsf{B}}$. To obtain one joint graph structure for all common variables, we suggest to union $\hat{\mathfrak{E}}^{\mathsf{A}\Cup\mathsf{B}} = \hat{\mathfrak{E}}_{\mathbf{X},\mathbf{Y}}^{\mathsf{A}} \cup \hat{\mathfrak{E}}_{\mathbf{X},\mathbf{Z}}^{\mathsf{B}}$ or intersect $\hat{\mathfrak{E}}^{\mathsf{A}\Cup\mathsf{B}} = (\hat{\mathfrak{E}}_{\mathbf{X}}^{\mathsf{A}} \cap \hat{\mathfrak{E}}_{\mathbf{X}}^{\mathsf{B}}) \cup \hat{\mathfrak{E}}_{\mathbf{Y},\mathbf{Y}-\mathbf{X}}^{\mathsf{A}} \cup \hat{\mathfrak{E}}_{\mathbf{Z},\mathbf{Z}-\mathbf{X}}^{\mathsf{B}}$ the sets of edges within the common variables of $\hat{\mathcal{G}}^{\mathsf{A}}$ and $\hat{\mathcal{G}}^{\mathsf{B}}$, where, for example, $\hat{\mathfrak{E}}_{\mathbf{Y},\mathbf{Y}-\mathbf{X}}^{\mathsf{A}}$ denotes the set of edges among the specific variables $\mathbf{Y}$ and the connecting edges between these specific variables and the common variables. Since the sets of edges $\hat{\mathfrak{E}}_{\mathbf{X}}^{\mathsf{A}}$ and $\hat{\mathfrak{E}}_{\mathbf{X}}^{\mathsf{B}}$ correspond to two directed acyclic graphs, for the intersection of both holds that $\hat{\mathfrak{E}}_{\mathbf{X}}^{\mathsf{A}\Cup\mathsf{B}} \subseteq \hat{\mathfrak{E}}_{\mathbf{X}}^{\mathsf{A}}$ and $\hat{\mathfrak{E}}_{\mathbf{X}}^{\mathsf{A}\Cup\mathsf{B}} \subseteq \hat{\mathfrak{E}}_{\mathbf{X}}^{\mathsf{B}}$ and therefore, $\hat{\mathfrak{E}}_{\mathbf{X}}^{\mathsf{A}\Cup\mathsf{B}}$ is also free of cycles. However, the union of these two sets of directed edges may contain cycles. In this case, we search for the *feedback arc set* and revert its elements, so that the resulting graph is free of cycles (e.g. Bastert and Matuszewski, 2001). This procedure yields a common graph structure $\hat{\mathcal{G}}^{\mathsf{A}\Cup\mathsf{B}}$ for the matched Bayesian network. The edges among the specific variables, and between the specific variables and the common variables are preserved in the matched Bayesian network. As a result, the matched DAG contains all variables $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ and its set of directed edges is given by $\hat{\mathfrak{E}}^{\mathsf{A}\Cup\mathsf{B}}$.

2.3 Step 2: Estimation of the Local Parameters and the Joint Probability Distribution

In the second step, we need to estimate and combine the local probability distributions of all variables in the Bayesian network. As described above, the global probability distribution represented by a Bayesian network is the product over the local (conditional) probability distributions. Applying the chain rule from Equation (2) for Bayesian networks in the statistical matching context yields

$$\begin{aligned}\hat{P}^{\mathsf{A}\Cup\mathsf{B}}(\mathbf{X}=\mathbf{x},\mathbf{Y}=\mathbf{y},\mathbf{Z}=\mathbf{z}) = {} & \prod_{k=1}^{q} \hat{p}_{\hat{\mathcal{G}}_{\mathbf{X},\mathbf{Y}}^{\mathsf{A}}}(y_k|\mathbf{pa}(Y_k)) \cdot \prod_{\ell=1}^{r} \hat{p}_{\hat{\mathcal{G}}_{\mathbf{X},\mathbf{Z}}^{\mathsf{B}}}(z_\ell|\mathbf{pa}(Z_\ell)) \\ & \cdot \prod_{j=1}^{p} \hat{p}_{\hat{\mathcal{G}}_{\mathbf{X}}^{\mathsf{A}\Cup\mathsf{B}}}(x_j|\mathbf{pa}(X_j)) \end{aligned} \tag{3}$$

as an estimator for the joint probability distribution. Just as in the likelihood function in Equation (1), the different terms of this joint probability distribution are estimable on A, B, or $\mathsf{A} \Cup \mathsf{B}$, respectively. In the event that our original concern was macro statistical matching, we are now finished. Otherwise, we additionally need to perform Step 3.

2.4 Step 3: Imputation of the Missing Values

In the last optional step, our aim is to construct a synthetic data file containing observations of all common and specific variables. The most obvious approach is the imputation of the missing values

[2] The union of these three sets of directed edges contains no cycles. The argument is based on the following three facts: 1. None of the sets $\hat{\mathfrak{E}}_{\mathbf{X}}^{\mathsf{A}\Cup\mathsf{B}}$, $\hat{\mathfrak{E}}_{\mathbf{X},\mathbf{Y}}^{\mathsf{A}}$, and $\hat{\mathfrak{E}}_{\mathbf{X},\mathbf{Z}}^{\mathsf{B}}$ corresponding to the three DAGs contains cycles. 2. The subsets only concerning the common variables are identical in all of the three sets. 3. The two subsets concerning the specific variables $\mathbf{Y}$ and $\mathbf{Z}$ are disjoint and can therefore not produce cycles.

in $\mathsf{A} \uplus \mathsf{B}$. Specifically, this means that we impute values for $\mathbf{Z}$ in A and $\mathbf{Y}$ in B. The values of $\mathbf{X}$ remain unaffected in A as well as in B. This ensures that, in any case, the marginal as well as the joint distributions of the common variables are maintained. For the purpose of imputation, we can directly draw synthetic values for Y_k, $k = 1, \dots, q$ or Z_ℓ, $\ell = 1, \dots, r$, given the realizations of $\mathbf{X}$, for every $i \in \mathcal{I}_A$, or for every $i \in \mathcal{I}_B$, respectively, from the estimated posterior distributions $\hat{P}^{\mathsf{A} \uplus \mathsf{B}}(\mathbf{Z} = \mathbf{z} | \mathbf{X} = \mathbf{x})$ and $\hat{P}^{\mathsf{A} \uplus \mathsf{B}}(\mathbf{Y} = \mathbf{y} | \mathbf{X} = \mathbf{x})$.

3. The German General Social Survey

3.1 The Data Base

To illustrate the proposed approach of statistical matching by utilizing Bayesian networks, we exemplarily apply it to the German General Social Survey (GGSS/ALLBUS) (GESIS – Leibniz Institute for the Social Sciences, 2013). This survey periodically collects information on attitudes, behavior, and the social structure in Germany every two years since 1980 (GESIS – Leibniz Institute for the Social Sciences, 2016). After preparation, these data are available for research and teaching and are therefore frequently used for statistical analysis.

In this paper, we apply our suggested approach to data which has been collected in 2012 and which contains, inter alia, information on demography, religiousness and physical health of the respondents. Originally, this survey covered 3480 observations of 752 variables. (For details see GESIS – Leibniz Institute for the Social Sciences (2013).) For this illustration, we extracted the following 17 variables[3] as common or specific variables:

- common: *sex*, *age*, *graduation*, professional *activity*, *marital* status, and net *income* of the respondents,
- specific in A: *denomination*, frequency of *church*goings, frequency of experiencing the presence of *God* through faith, frequency of experiences that can only be explained through the intervention of *supernatural* powers, any experience with miracle *healers*/spirit healers, and frequency of *pray*ing
- specific in B: frequency of visiting a *doctor*, *hospital* stay in the last 12 month, number of *cigarettes* per day, *alcoholic* beverages per day, and general *health*.

In many statistical matching applications, the common variables include demographic information. This is because in most of the surveys, questions about the demographic background of the respondents are very common. However, this fact does not eo ipso justify to assume CI between the sets of specific variables given demographic information.

The continuous variables *income* and *age* have been discretized by interval discretization into categorical variables with finally six possible, different realizations (levels) for income, and 17 levels for age. Variables levels with less than 20 observations have globally been ignored. After excluding the missing values, we obtain a data set with 800 observation. This data set is randomly split into two subsets, each containing $n_{\mathsf{A}} = n_{\mathsf{B}} = 400$ observations. In the first subset, we remove all observations regarding to the variables $\mathbf{Z}$, and in the second data set, we remove observations of $\mathbf{Y}$. This procedure yields two data sets A ($\in \mathbb{R}^{400 \times 12}$) and B ($\in \mathbb{R}^{400 \times 11}$) which can then be

[3] In the following, we mainly use the abbreviations written in *slanted* font to refer to the variables.

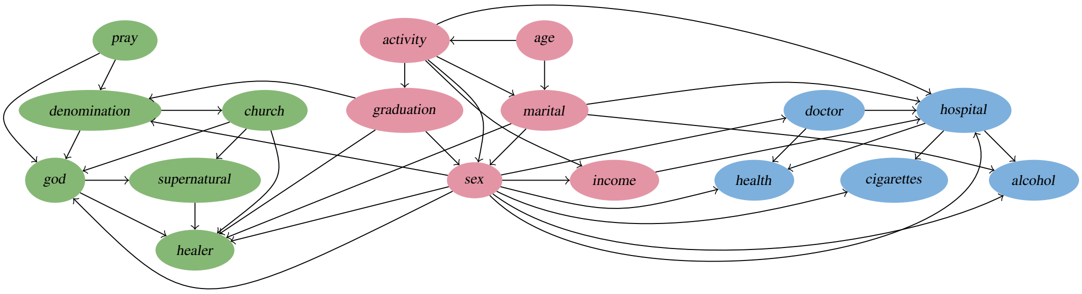


Figure 2: Graph structure of the matched Bayesian network $\hat{\mathcal{G}}^{\mathsf{A}\Cup\mathsf{B}}$ using Procedure 1 in Step 1 of the statistical matching approach.

matched as if they stem from two different surveys. Additionally, the matched synthetic data file can then be compared to the original file.

For the practical implementation of the suggested matching approach, we used the statistical software `R` (version 3.3.0 R Core Team, 2016), and the package `bnlearn` (version 3.9 Scutari, 2010; Nagarajan et al., 2013).

3.2 Statistical Matching of the GGSS Data by Bayesian Networks

For Procedure 1 of Step 1 in our statistical matching approach, we rely on $n = n_{\mathsf{A}} + n_{\mathsf{B}}$ observations of the six common variables to estimate the graph structure $\mathcal{G}_{\mathbf{X}}^{\mathsf{A}\Cup\mathsf{B}}$. For the estimation of the DAG structure, we use a bootstrap approach with model averaging to learn the directed acyclic graph which additionally estimates a measure for the strength of an edge to appear in the final DAG (Scutari and Nagarajan, 2011). In concrete, the structure is learned with the aid of the hill climbing algorithm in combination with the Bayesian information criterion as score which is applied to 500 bootstrap samples of size $\frac{2}{3} \cdot n$ (e.g. Nagarajan et al., 2013; Margaritis, 2003). To achieve a Bayesian network that represents the intended CI assumptions, the algorithm is limited to structures which are in line with the fork connection. Every edge that appeared in one of the bootstrap iterations is incorporated into the final graph, except for cycle-causing arcs. During the bootstrap structure learning, the strength of each edge to appear in the final DAG is computed as its relative frequency of appearance in the bootstrap folds. Starting with the edge with the highest strength, all edges are incorporated into the final DAG. In the event that an edge would cause a cycle, it is ignored and the edges with higher strengths stay incorporated in the final DAG. Since the structures for the common variables are fixed and identical in this procedure, we can merge the two graphs into one single Bayesian network as displayed in Figure 2.

Given the joint Bayesian network structure, we estimate the parameters of the local distributions by maximum likelihood. (A Bayesian estimation approach is also conceivable for this purpose.) Hence, the estimators equal the (conditional) empirical fractions of the variable levels. For nodes with several parent nodes it is likely that there exist combinations of parent instantiation which have not been observed in the present data. We cannot estimate the parameters for this child-node

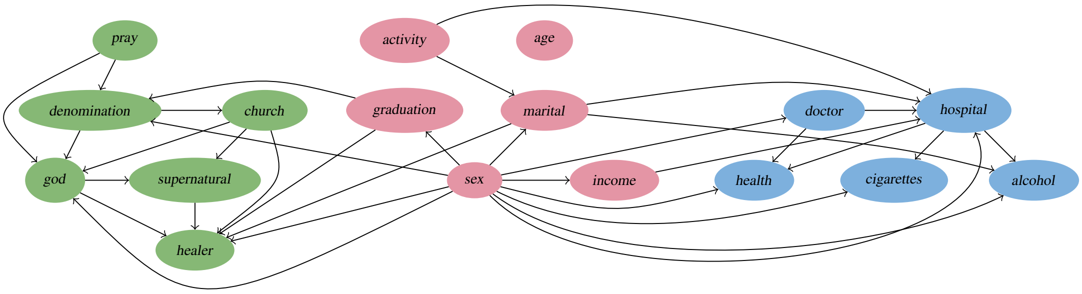


Figure 3: Graph structure of the matched Bayesian network $\hat{\mathcal{G}}^{\mathsf{A} \uplus \mathsf{B}}$ using edge intersection in Procedure 2 in Step 1 of the statistical matching approach.

given unobserved parental characteristics. In these cases, we assume the child-node to be uniformly distributed given its parent instantiations. This ad hoc assumption only slightly influences the micro approach because only rare combinations of instantiations are affected. Nevertheless, the influence of this assumption should be investigated more extensively in future research. Finally we impute the missing values in $\mathsf{A} \uplus \mathsf{B}$ by random draws from the posterior.

Within Procedure 2 for Step 1 of the statistical matching approach, we estimate two different graph structures for A and B, again with the bootstrap approach. To receive graph structures that represent the block-wise CI of the statistical matching framework, we again restrict the graphs $\hat{\mathcal{G}}^{\mathsf{A}}$ and $\hat{\mathcal{G}}^{\mathsf{B}}$ to the fork connection, where the common variables are the parents of the specific variables. The resulting two graph structures only regarding the common variables differ in a few details. Therefore, as mentioned above, the sets of edges of graphs $\hat{\mathcal{G}}^{\mathsf{A}}$ and $\hat{\mathcal{G}}^{\mathsf{B}}$ are combined by either intersection as displayed in Figure 3 or by union as displayed in Figure 4. The combination strategy using edge union leads to the following issue: in $\hat{\mathcal{G}}^{\mathsf{A}}$ we estimated the edge *sex*→*activity*, while in $\hat{\mathcal{G}}^{\mathsf{B}}$ the reverted edge *sex*←*activity* has been estimated. Applying the idea of feedback arc sets to this situation leads to the decision that *sex* is the parent of *activity* in the final matched graph. After the estimation and combination of the local probability distributions, we impute the missing values of $\mathsf{A} \uplus \mathsf{B}$ on the same principle as above. Using the same start value for the random number generator yields the same results for all matched Bayesian networks, derived by Procedure 1 or Procedure 2. This is due to the fact that all specific variables have the same variables as parents in every matched DAG.

3.3 Results

For the assessment of the accuracy of a statistical matching procedure, Rässler (2002) distinguishes four *quality levels* in descending order: (i) preserving the individual values, i.e. the matched values equal the true values, (ii) preserving joint distributions, (iii) preserving correlation structures, which corresponds to association in our case, and (iv) preserving marginal distributions. To exemplarily assess the quality of the derived complete synthetic data files by statistical matching with Bayesian networks, we limit ourselves to the latter two points, i.e. the comparison of the marginal distribu-

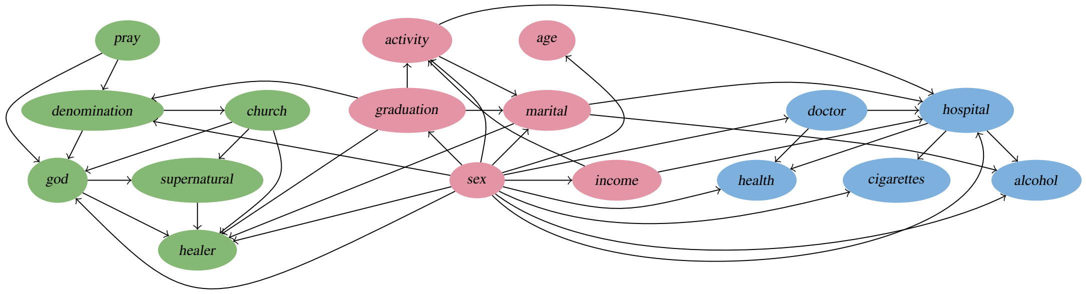


Figure 4: Graph structure of the matched Bayesian network $\hat{\mathcal{G}}^{\mathsf{A}\Cup\mathsf{B}}$ using edge union in Procedure 2 in Step 1 of the statistical matching approach.

tions and the bivariate association structures in the original GGSS data and the matched synthetic data in the following. Quality level (i) is generally very difficult to fulfill and not that important (e.g. D'Orazio et al., 2006a, p. 10). The second quality level should be examined extensively in future simulation studies where the true joint distribution is known.

Since the imputation process does not change the marginal distributions of the common variables, it is sufficient to consider the marginal distributions of the specific variables exclusively. To emphasize the contrast between the original and matched marginal distributions, we compute the Jensen-Shannon divergence between the parameters in the matched synthetic data set and the estimated parameters in the original GGSS data set (e.g. Lin, 1991). Table 1 presents the rounded results for the Jensen-Shannon divergence using the base 2 logarithm. It is apparent from this table

denomination	*church*	*god*	*supernatural*	*healer*	*pray*
0.000	0.001	0.001	0.004	0.021	0.002

doctor	*hospital*	*cigarettes*	*alcohol*	*health*
0.247	0.003	0.304	0.322	0.182

Table 1: Jensen-Shannon divergence of the synthetic marginal distributions and the marginals in the original GGSS data. (new table)

that a part of the marginal distributions of the synthetic and original data are very similar. However, the other part shows that the matching process did not preserve the parameters of the marginal distributions. For example, for the variable *denomination* the Jensen-Shannon divergence has a rounded value of 0. This means that the parameters of his variable are globally very similar in the matched synthetic data and the original GGSS data. However, the Jensen-Shannon divergence for *alcohol* is rather large with a rounded value of 0.322. Taken together, the results shown in Table 1 indicate that the matching process performed rather good with regard to the preservation of the marginal distributions of the variables *denomination*, *church*, *god*, *supernatural*, *healer*, *pray*, and *hospital*. However, the marginal distributions of the remaining variables which concern the physical health, are not so well retained. These results are mostly also confirmed by the p-values of the univariate χ^2-tests with a corresponding null hypothesis which states that the marginal distributions of the original GGSS data and the matched data are equal. Within the set of specific variables regarding religiousness, we recognize stronger associations in average between the single variables than in the specific set regarding to the physical health. There is evidence that this strength of association also affects the preservation of the marginals of the specific variables which should be investigated more closely in future research.

Furthermore, we compare the bivariate associations between the specific variables in the matched synthetic data with the corresponding associations in the GGSS data to receive an impression of the matching quality. To this end, we determine Sakoda's adjusted Pearson's $C \in [0, 1]$ (corrected contingency coefficient) for every bivariate combination of specific variables Y_k and Z_ℓ, $k = 1, \ldots, q$, $\ell = 1, \ldots, r$. This coefficient is independent of the sample size and the dimension of the contingency table. The absolute deviations of the associations in the matched synthetic data and the GGSS data which range from 0.02 to 0.149 indicate that the association structures in both data files are similar. The mean absolute deviation of the association has a rounded value of 0.046 and a standard deviation of 0.035.

4. Conclusion and Discussion

In this paper, we represented first attempts to utilize probabilistic graphical models as a powerful tool for statistical matching. Concretely, an approach for statistical matching of discrete data by Bayesian networks is described which we will further develop end extend in future work.

To the authors' knowledge, there is no statistical matching approach implemented which can deal with discrete data only. For this reason, we have not yet compared our results to a kind of gold standard statistical matching approach. This makes it even more important to stress that the generalizability of the results of the application example is subject to certain limitations. For instance, the choice of the common variables was more or less arbitrary. The association between the common variables and the specific variables should, in practice, be measured (e.g. D'Orazio et al., 2006a, pp. 167). In our GGSS example these associations vary rather stable between 0.13 and 0.20 for the specific variables on religiousness, and with a wide range between 0.02 and 0.44 for the specific variables on physical health. Although demographic variables are often selected as matching variables because they are collected in most of the surveys, it is not ensured that they are convenient to justify CI between the specific variables given the common ones. Note that this assumption can, in general, not be tested on the incomplete sample $\mathsf{A} \uplus \mathsf{B}$ in the statistical matching framework. Further sources of weakness which could have affected the results of the application example are the assumption of uniformly distributed parameters for not observed parent instantiations as mentioned in Section 3.2, and the choice of the structure learning algorithm. Additionally, the representativeness of the synthetic data set should be examined more accurately. In the event that the original data is known, like in the application example above, the assumption of CI should also explicitly be tested.

The approaches introduced in this paper will serve as a base for future research of how probabilistic graphical models can be utilized for statistical matching. A natural progression of this work is to consider not only discrete but also continuous and mixed discrete and continuous variables for statistical matching. Additionally, the natural ordering of ordinal variables should not be ignored. Further research is also required to determine if the use of undirected probabilistic graphical models is more promising. Although directed graphical models are appropriate to map many real-world problems, it is not always reasonable to set a direction between associated variables. In statistical matching it is also common practice to use auxiliary information to estimate the parameters of the joint probability distribution. This may mean the inclusion of a third complete or incomplete file or information about inestimable parameters (e.g. D'Orazio et al., 2006a, chap. 3). In addition, the inclusion of auxiliary information to probabilistic graphical models in terms of predefined graph structures or parameters, would be a fruitful area for future work. More broadly, in future research we should also take advantage of imprecise probabilistic graphical models (see, e.g., Cozman, 2000; Antonucci et al., 2014, for a survey) to robustify the whole modeling process, including a relaxation of the CI assumption by the different concepts of independence for imprecise probabilities (for an overview, see, e.g., Miranda and de Cooman, 2014). The stability of estimates based on very few observations can also be improved by these generalizations. Moreover, we should also consider, in the spirit of Manski (2003) and Vansteelandt et al. (2006), to use partial identified models or systematic sensitivity analysis to avoid the strong assumption of CI (see also D'Orazio et al., 2006b). In many surveys mainly discrete information is collected and a statistical matching approach for this kind of data is surely beneficial. Furthermore, this also allows for surveys which reduce the respondent's burden by not asking a respondent a complete questionnaire but only specific blocks of questions. The resulting data could then be matched.

Acknowledgments

We thank the GESIS - Leibniz Institute for the Social Sciences for providing the GGSS data. We are also very grateful to the LMUMentoring program, supporting young female researchers for providing financial support for the first author to travel to this conference. Last but not least a very special thank is due to three anonymous referees for their stimulating comments. Especially one of them gave very valuable comments and constructive suggestions which helped us to improve our work.

References

A. Antonucci, C. de Campos, and M. Zaffalon. Probabilistic graphical models. In T. Augustin, F. Coolen, G. de Cooman, and M. Troffaes, editors, *Introduction to Imprecise Probabilities*, pages 207–229, Chichester, United Kingdom, 2014. Wiley.

O. Bastert and C. Matuszewski. Layered drawings of digraphs. In M. Kaufmann and D. Wagner, editors, *Drawing Graphs: Methods and Models*, pages 87–120. Springer, Berlin, 2001.

F. Cozman. Credal networks. *Artificial Intelligence*, 120(2):199–233, 2000.

M. D'Orazio, M. Di Zio, and M. Scanu. *Statistical Matching: Theory and Practice*. Wiley, Chichester, United Kingdom, 2006a.

M. D'Orazio, M. Di Zio, and M. Scanu. Statistical matching for categorical data: Displaying uncertainty and using logical constraints. *Journal of Official Statistics*, 22(1):137–157, 2006b.

GESIS – Leibniz Institute for the Social Sciences. Allgemeine Bevölkerungsumfrage der Sozialwissenschaften ALLBUS 2012/German General Social Survey GGSS 2012, 2013. ZA4614 Data file Version 1.1.1, doi:10.4232/1.12209.

GESIS – Leibniz Institute for the Social Sciences. GESIS - ALLBUS: ALLBUS Home, 2016. URL `http://www.gesis.org/en/allbus/allbus-home/`. [Accessed 10.05.2016].

U. Kjræulff and A. Madsen. *Bayesian Networks and Influence Diagrams: A Guide to Construction and Analysis*. Springer, New York, 2nd edition, 2013.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, Cambridge, MA, 2009.

T. Koski and J. Noble. A review of Bayesian networks and structure learning. *Mathematica Applicanda*, 40(1):53–103, 2012.

J. Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

C. Manski. *Partial Identification of Probability Distributions*. Springer, New York, NY, 2003.

D. Margaritis. *Learning Bayesian Network Model Structure from Data*. PhD thesis, Carnegie-Mellon University, Pittsburgh, PA, 2003.

E. Miranda and G. de Cooman. Structural judgements. In T. Augustin, F. Coolen, G. de Cooman, and M. Troffaes, editors, *Introduction to Imprecise Probabilities*, pages 207–229, Chichester, United Kingdom, 2014. Wiley.

R. Nagarajan, M. Scutari, and S. Lèbre. *Bayesian Networks in R: With Applications in Systems Biology*. Springer, New York, NY, 2013.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL `https://www.R-project.org`.

S. Rässler. *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*. Springer, New York, NY, 2002.

M. Scutari. Learning Bayesian networks with the bnlearn R package. *Journal of Statistical Software*, 35(3):1–22, 2010.

M. Scutari and R. Nagarajan. On identifying significant edges in graphical models. In A. Hommersom and P. Lucas, editors, *Proceedings of the Workshop 'Probabilistic Problem Solving in Biomedicine' of the 13th Artificial Intelligence in Medicine (AIME) Conference*, pages 15–27, 2011.

S. Vansteelandt, E. Goetghebeur, M. Kenward, and G. Molenberghs. Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Statistica Sinica*, 16(3):953–979, 2006.

Contribution 2: pp. 60–83

Endres, E. and Augustin, T. (2019). Utilizing log-linear Markov networks to integrate categorical data files, Technical Report 222, Department of Statistics, LMU Munich.

The original publication is available at

`https://doi.org/10.5282/ubm/epub.61678`

Submitted to *Statistica Neerlandica*, date of submission: *January 11th, 2019.*

LMU

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

INSTITUT FÜR STATISTIK

Eva Endres and Thomas Augustin

Utilizing log-linear Markov networks to integrate categorical data files

Technical Report Number 222, 2019
Department of Statistics
University of Munich

http://www.statistik.uni-muenchen.de

Utilizing log-linear Markov networks to integrate categorical data files

Eva Endres[*] Thomas Augustin[†]
Department of Statistics, LMU München

17th April 2019

Abstract

The integration of different data sharing only a subset of variables will become even more relevant in the future. With the aid of data fusion techniques, already existing data can be exploited to carry out new statistical analyses, circumventing the expensive collection of new data. This paper presents a new statistical matching method for categorical data based on a conditional independence assumption. The method uses undirected graphical models to visualize dependencies among variables, and obtains a powerful factorization of their joint distribution. It is used to estimate the probability components of the joint distribution despite the underlying identification problem. We embed the problem of statistical matching into the theory of log-linear Markov networks and show an exemplary application of this new method based on data of the German General Social Survey. The results indicate that the joint distribution can be reconstructed fairly well through the proposed statistical matching method.

Keywords: conditional independence, data fusion, log-linear model, Markov random field, probabilistic graphical model, statistical matching

1 Introduction and description of the problem

Statistical matching, which terms the integration of already existing data, became increasingly important in the last years. On the one hand, the collection of new data is expensive and time-consuming. On the other hand, if data originate from long questionnaires, we must be aware of the inaccuracy resulting from potential non-response. As already stated by D'Orazio et al. (2006a) or Rässler (2002), these are strong arguments against the collection of new data but for performing secondary analysis of already available data sources.

However, we are confronted with a serious challenge in secondary analysis if we need joint information about variables which have not been jointly observed. If we though have data files which share some of their variables, i.e. the intersection of the variable sets is not the empty set, we are able to integrate these files. See, for instance, Serafino and Tonkin (2017) and Aluja-Banet et al. (2015), for applications of statistical matching in the context of official statistics and epidemiology.

Figure 1 shows a schematic representation of the basic scope. In the following, we will call the variables which are contained in a single data file only, the *specific variables*, and the variables which are present in both files the *common variables*. Although we can

[*] eva.endres@stat.uni-muenchen.de
[†] augustin@stat.uni-muenchen.de

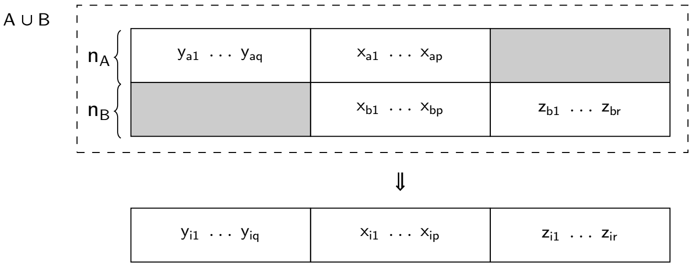


Figure 1: Schematic representation of the statistical matching problem (see D'Orazio et al., 2006a, p.5 (modified)).

justifiably assume that the observations of the specific variables are missing completely at random (e.g. D'Orazio et al., 2006a, p. 6), we are not per se able to find an identifiable model of all variables of interest based on the available data files without further assumptions or information.

Statistical matching yields the solution for this issue. As previously mentioned, with statistical matching we are able to extract joint information about variables which have been collected in different surveys. *Joint information* can either be the joint probability distribution (or any of its characteristics) or a complete (but synthetic) data file which contains all variables of interest and reflects the structure of the true but unknown complete file (e.g. D'Orazio et al., 2006a, p. 2). The former aim describes the so-called statistical matching *macro approach* while the latter refers to the *micro approach*.

In the present paper, we embed the statistical matching task into the framework of undirected probabilistic graphical models and use log-linear Markov networks (e.g. Koller and Friedman, 2009) to obtain estimates for the components of the joint probability distribution. Section 2 introduces the general framework and notations for statistical matching, and discusses the central role of the conditional independence assumption. Section 3 recalls the basic concepts of log-linear Markov networks and links them with the problem of categorical data integration. Section 4 shows the application of the new statistical matching approach based on Markov networks for data of the German General Social Survey. Finally, we give a summary and an outlook in Section 5.

2 Statistical matching

2.1 The basic framework

Statistical matching (or also called data fusion or data integration) refers to a data situation as displayed in Figure 1. Let A be a data file with n_A categorical observations $(x_1, \dots, x_p, y_1, \dots, y_q)$ of the variables in the sets $\boldsymbol{X} = \{X_1, \dots, X_p\}$ and $\boldsymbol{Y} = \{Y_1, \dots, Y_q\}$, and B a data file with n_B categorical observations $(x_1, \dots, x_p, z_1, \dots, z_r)$ of the variables in the sets $\boldsymbol{X}$ and $\boldsymbol{Z} = \{Z_1, \dots, Z_r\}$. The sets of possible realizations of the random variables are denoted by $\mathcal{X}_j$, $\mathcal{Y}_k$, and $\mathcal{Z}_\ell$ for X_j, Y_k, and Z_ℓ, respectively, for $j = 1, \dots, p$, $k = 1, \dots, q$, and $\ell = 1, \dots r$.

If we treat the files A and B as a single data source $A \uplus B$ with $n = n_A + n_B$ observations created from the union of A and B, statistical matching can be interpreted as a missing data problem with a special missingness pattern. The gray areas in Figure 1 display the blocks of missing entries in the combined file $A \uplus B$. As we can see from this visualization, the special task of statistical matching arises from the fact that there is no single observation which gives us information on all variables $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$. This leads to a serious identification problem during the estimation of the parameters of the joint distribution.

Regardless of which concrete approach we use to solve this identification problem, we have to make one basic assumption: the observations in both files A and B have to be independently and identically distributed following a joint probability distribution $\pi(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) := \mathbb{P}(X_1 = x_1, \ldots, X_p = x_p, Y_1 = y_1, \ldots, Y_q = y_q, Z_1 = z_1 \ldots, Z_r = z_r)$. Since we focus on categorical data, this joint distribution can be expressed by a multi-dimensional probability table whose entries are our parameters of interest (under the constraint that the sum over all entries equals 1).

For instance, D'Orazio et al. (2006a) describe different approaches, which can be split into three different groups according to their basic concepts, how statistical matching can be applied in practice:

1. The first group of approaches is based on the assumption of conditional independence of the specific variables given the common variables.

2. The second type of approaches exploits potentially available auxiliary information. This may be a third data file with joint observations on the specific variables or even all variables of interest. In a parametric setting, it would furthermore be conceivable that there exists information about certain parameters, for example, from pilot studies.

3. The last group of approaches directly addresses the identification problem of statistical matching. Instead of relying on additional assumptions or auxiliary information, the uncertainty corresponding to the identification problem is respected and sets of parameter estimates are obtained for the macro approach, or sets of complete synthetic data files are created for the micro approach.

For examples of the second and third type of approaches see, for instance, Singh et al. (1993), Di Zio and Vantaggi (2017), D'Orazio et al. (2006b), or Endres et al. (2018). As mentioned above, we emphasize on approaches based on the conditional independence of the specific variables given the common variables which is closely connected to the concept of *separation* in the context of *probabilistic graphical models*. Some papers in which directed acyclic graphs are examined for the statistical matching task have already been published. For instance, Landes and Williamson (2016) show how to learn a Bayesian network which coincides with the marginal distributions of the present data and whose corresponding joint distribution has maximum entropy. Endres and Augustin (2016) introduce an approach on how to learn a joint Bayesian network for the available (incomplete) data. Already existing available structure learning algorithms are adapted to learn a joint directed acyclic graph of $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ on $A \uplus B$. The network structure is the basis of subsequent parameter estimation using an adapted version of the chain rule for Bayesian networks. Another idea of intersecting the data integration problem with graphical models is described, for instance, by Tsamardinos et al. (2012) and Janzing (2018). These data fusion approaches aim at the detection of causal models which are consistent with the available data.

In this paper, we consider the case when there is no natural directionality regarding the relationship between variables. In this situation, a Bayesian network which is based on a *directed* acyclic graph is not the means of choice. However, there is a class of undirected probabilistic graphical models which also has the potential to meet the aims of statistical matching, namely Markov networks. They are closely related to Bayesian networks, yet they differ in a key aspect: Markov networks build on an *undirected* graph. To prepare for the relationship between statistical matching and Markov networks, we take a closer look at the concept of conditional independence in the following subsection.

2.2 The role of the conditional independence assumption

As mentioned above, due to the identification problem, the parameters of the joint distribution which concern the relationship between the specific variables $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are not directly estimable. This is where the assumption of the conditional independence of $\boldsymbol{Y}$ and $\boldsymbol{Z}$ given $\boldsymbol{X}$ comes in. Applying the chain rule and the definition of conditional independence, the probability distribution of $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ is given by

$$\begin{aligned}\pi(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) &= \pi(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{x}) \cdot \pi(\boldsymbol{z}|\boldsymbol{x}) \cdot \pi(\boldsymbol{x}) \\ &= \pi(\boldsymbol{y}|\boldsymbol{x}) \cdot \pi(\boldsymbol{z}|\boldsymbol{x}) \cdot \pi(\boldsymbol{x}) \\ &= \frac{\pi(\boldsymbol{x}, \boldsymbol{y}) \cdot \pi(\boldsymbol{x}, \boldsymbol{z})}{\pi(\boldsymbol{x})}. \end{aligned} \tag{1}$$

Looking at this factorization, we can easily see that $\pi(\boldsymbol{x}, \boldsymbol{y})$ is only dependent on $\boldsymbol{Y}$ and $\boldsymbol{X}$ and thus is estimable on data file A, whereas $\pi(\boldsymbol{x}, \boldsymbol{z})$ can be estimated from the second file B, and the third term $\pi(\boldsymbol{x})$ is estimable on all n observations. Since we can legitimately assume that we are in a MCAR (missing completely at random) situation (D'Orazio et al., 2006a, p. 6), the blocks of missing entries of $\boldsymbol{Y}$ in B, and $\boldsymbol{Z}$ in A can be ignored within the estimation-step and the available data $A \uplus B$ is representative for the (not available) complete file (e.g. Pigott, 2001).

From this point we can build the bridge to probabilistic graphical models. The graph of a probabilistic graphical models can be viewed as a map which visualizes (in)dependencies. If all independencies which are represented by the graph are also present in the corresponding probability distribution, the graph is said to be an independence map (*I-map*) for this distribution (e.g. Pearl, 1988, p. 92). These I-maps lead to a factorization of the probability distribution according to the cliques of the graph (e.g. Studený, 2010, p. 46). In Endres and Augustin (2016) we also build upon the factorization of probability distributions but in the context of Bayesian networks which are based on directed acyclic graphs (DAG). Since there are situations where variables interact but where there is no natural direction of this connection, we consider the embedding of Markov networks into the context of statistical matching in the present paper. We will explain it in detail in the next section after a short revision of some necessary foundations of Markov networks.

3 Markov networks and statistical matching

3.1 Basic concepts and notations of log-linear Markov networks

As a start, we will briefly recall the definition of log-linear Markov networks and fix our notations. See, for example, Koller and Friedman (2009) or Lauritzen (1996) for detailed explanations of undirected graphical models. For reasons of readability, we only consider one set of discrete random variables $\boldsymbol{X} = \{X_1, \dots, X_p\}$ in this subsection and do not explicitly refer to the statistical matching framework but describe log-linear Markov networks for arbitrary situations. The concrete application of Markov networks for the purpose of statistical matching will be described in the next subsection.

In the subsequent explanations, we refer to a certain variable which is an element of $\boldsymbol{X}$ with the symbol X_j, $j \in \{1, \dots, p\}$, while certain subsets of $\boldsymbol{X}$ are characterized by an index set $\boldsymbol{j} \subseteq \{1,\dots,p\}$ such that $\boldsymbol{X_j} := \{X_j : X_j \in \boldsymbol{X}, j \in \boldsymbol{j}\}$. The corresponding realizations $\boldsymbol{x} = (x_1,\dots,x_p)$ are analogously indexed and $x_j \in \mathcal{X}_j = \{0,1,\dots,d_j - 1\}$ for every $j \in \{1,\dots,p\}$. Referring to the d_j different categories of the j-th variable as $\{0,1,\dots,d_j - 1\}$ does not imply any ordering.

A Markov network over the set of categorical variables $\boldsymbol{X} = \{X_1,\dots,X_p\}$ is represented by an undirected graph $\mathcal{H} = (\mathring{\boldsymbol{V}}, \boldsymbol{E})$. The symbol $\mathring{\boldsymbol{V}}$ denotes the set of p vertices in the graph, representing the p random variables in $\boldsymbol{X}$. To preserve readability, we will set $\mathring{\boldsymbol{V}} \equiv \mathring{\boldsymbol{X}}$ and thus $\mathcal{H} = (\mathring{\boldsymbol{X}}, \boldsymbol{E})$. The circle above a symbol refers to nodes where symbols without circle refer to the corresponding random variables. With the symbol × indicating the Cartesian product, $\boldsymbol{E} \subseteq \mathring{\boldsymbol{X}} \times \mathring{\boldsymbol{X}}$ terms the set of pairwise (undirected) edges in the graph. Interpreting $\mathcal{H}$ as *independence graph*, the pair $(\mathring{X}_i, \mathring{X}_j)$ is not an element of $\boldsymbol{E}$ iff the corresponding and non-adjacent variables X_i and X_j are conditionally independent given $\boldsymbol{X} \setminus \{X_i, X_j\}$ (pairwise *Markov assumption*). In the following, we assume that there exists a (everywhere) positive probability mass distribution $\pi(\boldsymbol{x}) = \mathbb{P}(X_1 = x_1,\dots,X_p = x_p)$ that factorizes over $\mathcal{H}$, and thus the local, the pairwise and the global Markov assumptions coincide (see, e.g. Koller and Friedman, 2009, p. 119). Consequently, the (in)dependencies among the set of variables $\boldsymbol{X}$ are visualized by $\mathcal{H}$ and can be read off the graph. Two sets of variables $\mathbf{X_f}$ and $\mathbf{X_g}$ are conditionally independent given $\mathbf{X_h}$, written $\mathbf{X_f} \perp\!\!\!\perp \mathbf{X_g} | \mathbf{X_h}$, if there is no *active path* between any nodes $\mathring{X}_f \in \mathring{\boldsymbol{X}}_\mathbf{f}$ and $\mathring{X}_g \in \mathring{\boldsymbol{X}}_\mathbf{g}$ given $\mathring{\boldsymbol{X}}_\mathbf{h}$ in $\mathcal{H}$, for disjoint sets $\mathring{\boldsymbol{X}}_\mathbf{f}, \mathring{\boldsymbol{X}}_\mathbf{g}, \mathring{\boldsymbol{X}}_\mathbf{h}$ each of which is a subset of $\boldsymbol{V}$. The node sets $\mathring{\boldsymbol{X}}_\mathbf{f}$ and $\mathring{\boldsymbol{X}}_\mathbf{g}$ are then said to be *separated* by $\mathring{\boldsymbol{X}}_\mathbf{h}$ (see,e.g. Studený, 2010, p. 43).

Since we are dealing with categorical data which can be represented by multi-dimensional contingency tables, we suggest to use the log-linear parameterization of Markov networks. The corresponding joint probability is then given as

$$\pi(\boldsymbol{x}) = \exp\left\{ \sum_{\boldsymbol{C} \subseteq \boldsymbol{X}} u_{\boldsymbol{C}}(\boldsymbol{x}) \right\}, \tag{2}$$

which is also known under the term *log-linear expansion* (of the multinomial distribution) (e.g. Whittaker, 1990, p. 206). In this representation of a log-linear model, we sum over all subsets $\boldsymbol{C}$ of $\boldsymbol{X}$ (i.e. over all elements of the power set $\mathcal{P}(\boldsymbol{X})$ of $\boldsymbol{X}$) under the constraint that $u_{\boldsymbol{C}}(\boldsymbol{x}) = 0$ if $x_j = 0$ for $X_j \in \boldsymbol{C}$. The sum within the curly brackets is equivalent to a linear predictor of a regression model where the u-terms correspond to the regression parameters. In the log-linear expansion of the multinomial distribution, these u-terms are log-odds and can be interpreted as such. Some log-linear representations for selected cases are shown in Appendix A.

Graphical models are a subset of the more general class of log-linear models (see, e.g. Tutz, 2011, p. 346) which

1. include all lower-order terms of variables which appear together in a higher-order term (*hierarchical* model) and
2. include the higher-order terms of variables whose pairwise terms are all also contained in the model (*graphical* model).

A graphical log-linear model can be represented by an *interaction graph* which coincides with the independence graph whenever there exists an interaction term for each *clique* in the graph, and where the *maximal cliques* (e.g. Whittaker, 1990, p. 209) determine the highest-order interaction terms. (The term maximal clique refers to a subset of $\boldsymbol{V}$ where every pair of nodes is connected by an edge (e.g. Koller and Friedman, 2009, p. 35).) Thus, we are able to read the interaction terms off the undirected graph structure. The term $u_{\boldsymbol{C}}(\boldsymbol{x})$ equals zero if $\{\mathring{X}_i, \mathring{X}_j\} \subseteq \mathring{\boldsymbol{X}}$ but $(\mathring{X}_i, \mathring{X}_j) \notin \boldsymbol{E}$. The highest-order interaction terms determine the *generating class* of the log-linear model (see, e.g. Lauritzen, 1996, p. 82).

There is also a close connection between the interpretation of such a log-linear model and the separation in graphs. Whenever the sets of nodes $\mathring{\boldsymbol{X}}_{\mathbf{f}}$ and $\mathring{\boldsymbol{X}}_{\mathbf{g}}$ are separated by another set $\mathring{\boldsymbol{X}}_{\mathbf{h}}$ in $\mathcal{H}$, it holds that $\mathbf{X_f} \perp\!\!\!\perp \mathbf{X_g} | \mathbf{X_h}$ in the corresponding distribution, and all interaction terms over $\boldsymbol{X}_{\mathbf{f}}$ and $\boldsymbol{X}_{\mathbf{g}}$ are zero. It means that $u_{\boldsymbol{C}}(\boldsymbol{x}) = 0$ if $\{X_f, X_g\} \in \boldsymbol{C}$ for any $X_f \in \boldsymbol{X}_{\mathbf{f}}$ and $X_g \in \boldsymbol{X}_{\mathbf{g}}$. Thus, the joint probability distribution factorizes to the product of two functions m_1 and m_2. This factorization is usually referred to as *factorization criterion* (e.g. Højsgaard et al., 2012, p. 11 and p. 32).

Since we are dealing with exclusively categorical data in this paper, we will in the following apply a multinomial distribution. For *decomposable* graphical models, this leads to closed-form maximum likelihood estimators. (In decomposable models, every cycle of minimum length four has a shortcut (e.g. Tutz, 2011, p. 352).) Details can be found, for instance, in Højsgaard et al. (2012, p. 31). For arbitrary graphical models, the maximum likelihood estimators can be determined by iterative methods like, for instance, iterative proportional fitting (see, e.g. Højsgaard et al., 2012, p. 35).

3.2 Utilizing Markov networks for statistical matching

As mentioned above, within the framework of statistical matching, the available observations in $A \Cup B$ are assumed to be i.i.d. realizations of a joint distribution $\pi(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$ with missing (completely at random) values $\boldsymbol{Y}$ in B and $\boldsymbol{Z}$ in A. As consequence, we can imagine that there exists a true underlying file with complete information on all variables $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$. Furthermore, assuming that $\mathbb{P}$ factorizes over a Markov network, there also exists a true underlying Markov network structure. In the following this true network structure, denoted by $\mathcal{H}^{A \Cup B}$, is supposed to be known, or at least that the information in A and B is sufficient to estimate an error-free version $\hat{\mathcal{H}}^{A \Cup B}$ of the true network structure. $\mathcal{H}^{A \Cup B}$ is composed of a set of undirected edges $\boldsymbol{E}^{A \Cup B}$ and a set of nodes $\mathring{\boldsymbol{V}}^{A \Cup B} \equiv \mathring{\boldsymbol{X}} \cup \mathring{\boldsymbol{Y}} \cup \mathring{\boldsymbol{Z}}$ with cardinality $p + q + r$.

To meet the requirements for solving the statistical matching problem, we assume that the specific variables $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are conditionally independent given the common variables $\boldsymbol{X}$. In the graph of the corresponding Markov network, none of the pairs $(\mathring{Y}_k, \mathring{Z}_\ell)$ is in $\boldsymbol{E}^{A \Cup B}$, $k = 1, \ldots, q$, $\ell = 1, \ldots, r$. Hence, there exist no direct paths between any nodes $\mathring{Y}_k \in \mathring{\boldsymbol{Y}}$ and $\mathring{Z}_\ell \in \mathring{\boldsymbol{Z}}$, i.e. the specific variables are separated by at least one $\mathring{X}_j \in \mathring{\boldsymbol{X}}$. This separation ensures that the parameters $u_{\boldsymbol{C}}$ of the log-linear Markov model are zero if $\{Y_k, Z_\ell\} \in \boldsymbol{C}$.

To achieve an estimation equation extended for statistical matching purposes, we need to incorporate the log-linear representation of Equation (2) into the factorization based on the conditional independence assumption in Equation (1). Statistical matching by log-linear Markov networks is then implemented by

$$\begin{aligned}\pi(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) &= \exp\{\log \pi(\boldsymbol{x}, \boldsymbol{y}) + \log \pi(\boldsymbol{x}, \boldsymbol{z}) - \log \pi(\boldsymbol{x})\} \\ &= \exp\left\{ \sum_{\boldsymbol{C} \in \mathcal{P}(\boldsymbol{X} \cup \boldsymbol{Y})} u_{\boldsymbol{C}}(\boldsymbol{x}, \boldsymbol{y}) + \sum_{\boldsymbol{C} \in \mathcal{P}(\boldsymbol{X} \cup \boldsymbol{Z})} u_{\boldsymbol{C}}(\boldsymbol{x}, \boldsymbol{z}) - \sum_{\boldsymbol{C} \in \mathcal{P}(\boldsymbol{X})} u_{\boldsymbol{C}}(\boldsymbol{x}) \right\} \qquad (3)\end{aligned}$$

under analogue constraints as for Equation (2). This means that a summand is zero either if one of the corresponding realizations is zero (i.e. it equals the reference category) or if the corresponding nodes are separated in $\mathcal{H}^{A \Cup B}$ (i.e. the pairwise edges are not in

$\boldsymbol{E}^{A \uplus B}$). As it can easily been seen from the equation, none of the terms is simultaneously dependent on the specific variables $\boldsymbol{Y}$ and $\boldsymbol{Z}$. Thus, all terms are separately estimable on different parts of the data, namely the first term can be estimated from A, the second from B, and the third on $A \uplus B$. This means that we now have an identifiable model for the incomplete sample $A \uplus B$, and have come up with a solution for the statistical matching macro approach.

4 Illustrative application

To show the practical applicability of our statistical matching approach, we use data of the German General Social Survey collected by GESIS – Leibniz Institute for the Social Sciences (2016). After a registration, the data can be downloaded from `www.doi.org/10.4232/1.12209`. All analyses are conducted by the statistical programming software R (R Core Team, 2018, version 3.5.1). The R code for all analyses is available on request.

4.1 The German General Social Survey

The German General Social Survey is a cross-sectional study which has been carried out every two years since 1980. It serves as data source to analyse attitudes and behaviors in the German society. For our application, we use the data of the GGSS 2012 which focuses amongst others on health-related topics. The data are composed of 3480 observations of 752 variables. For our illustration, we extract seven categorical variables, which we split into common and specific variables:

common: the SEX and the AGE of the respondent, and whether the respondent is EMPLOYED,

specific in A: the intensity of smoking (SMOKE) and how much ALCOHOL the respondent drinks,

specific in B: how many times the respondent visited a DOCTOR in the past 12 months, and how often the respondent exercises for at least 20 minutes (SPORT).

Since our focus is not on the missing data problem of the survey itself, we delete the observations with missing entries for our purposes. This results in a data file with 1375 observations. To reduce structural zeros to a minimum, we also summarize some of the categories. Finally, we have five binary variables and two variables with three categories (*age* and *smoke*). The term *structural zero* usually refers to zero entries in the true probability mass distribution. However, in our application, the true underlying probability distribution of the considered (GGSS) population is unknown and we have to use the (estimated) sample distribution as reference. Zero entries in this sample distribution are no 'true' structural zeros, yet we call them so because this sample distribution serves as our reference distribution. To mimic the situation of statistical matching, we randomly split our data file into two files A and B as follows:

- file A has 688 observations of SEX, AGE, EMPLOYED, SMOKE, and ALCOHOL
- file B has 687 observations of SEX, AGE, EMPLOYED, DOCTOR, and SPORT.

An exemplary extract of this data situation is displayed in Table 1.

Using our notation, we consider the following sets of common and specific variables:

$$\boldsymbol{X} = \{\text{SEX}, \text{AGE}, \text{EMPLOYED}\},\ \boldsymbol{Y} = \{\text{SMOKE}, \text{ALCOHOL}\},\ \boldsymbol{Z} = \{\text{SPORT}, \text{DOCTOR}\}.$$

The (aggregated) possible realizations for each variable are listed in Appendix C.1.

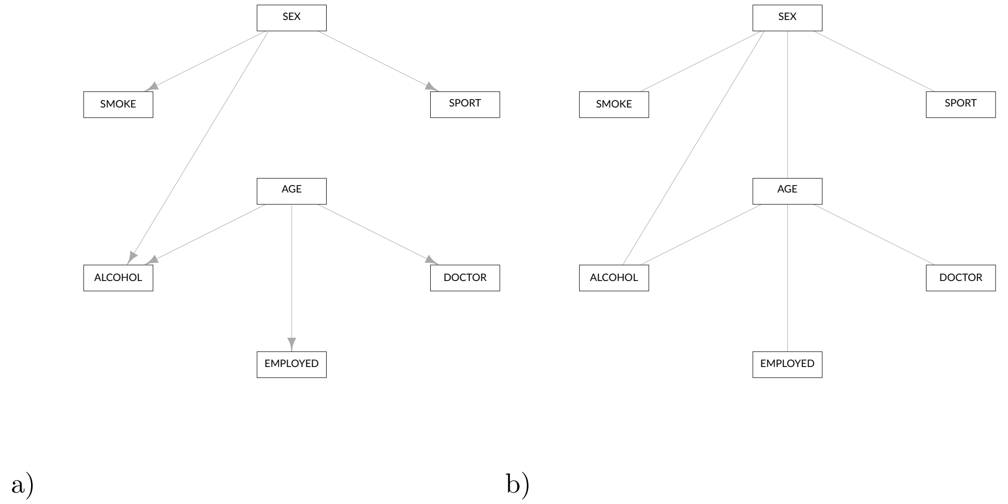


Figure 2: Joint DAG based on $A \uplus B$ on the left side. Joint undirected graph based on $A \uplus B$ on the right side, derived by moralization of the DAG.

4.2 Statistical matching of the GGSS data with log-linear Markov networks

4.2.1 The Markov network structure

The true network structure for the data of the German General Social Survey is unknown and has to be learned from the data. In Endres and Augustin (2016), we introduced a statistical matching technique which is based on Bayesian networks. Different parts of the joint Bayesian network are learned on different parts of the data at hand and subsequently combined. To obtain the structure of the joint Markov network on $A \uplus B$, we also follow this procedure and moralize the resulting DAG. Maybe this looks inconvenient at first, but this procedure has the advantage that we end up with a decomposable graph. Thus, closed-form ML-estimates for the probability components of the joint distribution of $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ are available. Of course, also other structure learning algorithm for Markov networks can be adapted for this step. Figure 2 shows the joint DAG on the left side which was estimated on $A \uplus B$. On the right hand side, we see the moralized graph, i.e. the structure of the joint Markov network on $A \uplus B$. The estimation and moralization was performed in R using the R-package *bnlearn* by Scutari (2010, version 4.3). Specifically, the structure was learned with the aid of the score-based *hill-climbing*-algorithm which was applied on 500 bootstrap samples and combined by model averaging.

4.2.2 Estimation of the parameters of the log-linear Markov network

According to the graph, we eliminate the entries of the powersets of $\boldsymbol{X}$, $\boldsymbol{X} \cup \boldsymbol{Y}$, and $\boldsymbol{X} \cup \boldsymbol{Z}$ whose corresponding u-terms are equal to zero (i.e. the pairwise connections are not element of the set of edges of the graph) and we obtain the following reduced sets $\mathcal{P}^{*}$

containing the remaining relevant entries:

$$\begin{aligned}
\mathcal{P}^*(\boldsymbol{X}) =& \{\varnothing, \text{SEX}, \text{AGE}, \text{EMPLOYED}, \{\text{SEX}, \text{AGE}\}, \{\text{AGE}, \text{EMPLOYED}\}\}\\
\mathcal{P}^*(\boldsymbol{X} \cup \boldsymbol{Y}) =& \{\varnothing, \text{SEX}, \text{AGE}, \text{EMPLOYED}, \text{SMOKE}, \text{ALCOHOL}, \{\text{SEX}, \text{AGE}\},\\
& \{\text{AGE}, \text{EMPLOYED}\}, \{\text{SEX}, \text{SMOKE}\}, \{\text{SEX}, \text{ALCOHOL}\}, \{\text{AGE}, \text{ALCOHOL}\},\\
& \{\text{SEX}, \text{AGE}, \text{ALCOHOL}\}\}\\
\mathcal{P}^*(\boldsymbol{X} \cup \boldsymbol{Z}) =& \{\varnothing, \text{SEX}, \text{AGE}, \text{EMPLOYED}, \text{SPORT}, \text{DOCTOR}, \{\text{SEX}, \text{AGE}\},\\
& \{\text{AGE}, \text{EMPLOYED}\}, \{\text{SEX}, \text{SPORT}\}, \{\text{AGE}, \text{DOCTOR}\}\}.
\end{aligned}$$

Applying Equation (3) leads to the estimation equation

$$\begin{aligned}
&\tilde{\pi}(\text{sex}, \text{age}, \text{employed}, \text{smoke}, \text{alcohol}, \text{sport}, \text{doctor})\\
&\quad = \exp\Bigg\{ \log \hat{\pi}^A(\text{sex}, \text{age}, \text{employed}, \text{smoke}, \text{alcohol})\\
&\qquad + \log \hat{\pi}^B(\text{sex}, \text{age}, \text{employed}, \text{sport}, \text{doctor}) - \log \hat{\pi}^{A \uplus B}(\text{sex}, \text{age}, \text{employed})\Bigg\}\\
&\quad = \exp\Bigg\{ u^A_{\varnothing} + u^A_{\{\text{SEX}\}}(\text{sex}) + u^A_{\{\text{AGE}\}}(\text{age}) + u^A_{\{\text{EMPLOYED}\}}(\text{employed}) + u^A_{\{\text{SMOKE}\}}(\text{smoke})\\
&\qquad + u^A_{\{\text{ALCOHOL}\}}(\text{alcohol}) + u^A_{\{\text{SEX,AGE}\}}(\text{sex}, \text{age}) + u^A_{\{\text{AGE,EMPLOYED}\}}(\text{age}, \text{employed})\\
&\qquad + u^A_{\{\text{SEX,SMOKE}\}}(\text{sex}, \text{smoke}) + u^A_{\{\text{SEX,ALCOHOL}\}}(\text{sex}, \text{alcohol})\\
&\qquad + u^A_{\{\text{AGE,ALCOHOL}\}}(\text{age}, \text{alcohol}) + u^A_{\{\text{SEX,AGE,ALCOHOL}\}}(\text{sex}, \text{age}, \text{alcohol})\\
&\qquad + u^B_{\varnothing} + u^B_{\{\text{SEX}\}}(\text{sex}) + u^B_{\{\text{AGE}\}}(\text{age}) + u^B_{\{\text{EMPLOYED}\}}(\text{employed}) + u^B_{\{\text{SPORT}\}}(\text{sport})\\
&\qquad + u^B_{\{\text{DOCTOR}\}}(\text{doctor}) + u^B_{\{\text{SEX,AGE}\}}(\text{sex}, \text{age}) + u^B_{\{\text{AGE,EMPLOYED}\}}(\text{age}, \text{employed})\\
&\qquad + u^B_{\{\text{SEX,SPORT}\}}(\text{sex}, \text{sport}) + u^B_{\{\text{AGE,DOCTOR}\}}(\text{age}, \text{doctor})\\
&\qquad - u^{A \uplus B}_{\varnothing} - u^{A \uplus B}_{\{\text{SEX}\}}(\text{sex}) - u^{A \uplus B}_{\{\text{AGE}\}}(\text{age}) - u^{A \uplus B}_{\{\text{EMPLOYED}\}}(\text{employed})\\
&\qquad - u^{A \uplus B}_{\{\text{SEX,AGE}\}}(\text{sex}, \text{age}) - u^{A \uplus B}_{\{\text{AGE,EMPLOYED}\}}(\text{age}, \text{employed})\Bigg\}, \qquad (4)
\end{aligned}$$

where the superscripts indicate which data file is used to estimate the corresponding term. To be able to distinguish between the true distributions π, the distributions estimated on the complete GGSS sample $\hat{\pi}$ are marked with a circumflex, and the synthetic distributions $\tilde{\pi}$, estimated with the aid of statistical matching, are from now on marked with a tilde.

For the concrete implementation in R, we use a generalized Poisson regression model with a log-link. With this regression model, we estimate the parameters of the log-linear model and obtain the fitted values. A justification why this procedure is appropriate in our context can be found in Appendix B. Furthermore, Appendices C.2 and C.3 contain an interpretation of the u-terms and their actual estimates regarding to Equation 4.

4.2.3 Results

Following the recommendation by Rässler (2002), the performance of our new statistical matching procedure is assessed by investigating the following *quality levels*:

1. the preservation of the marginal distributions,
2. the preservation of the association structure, and

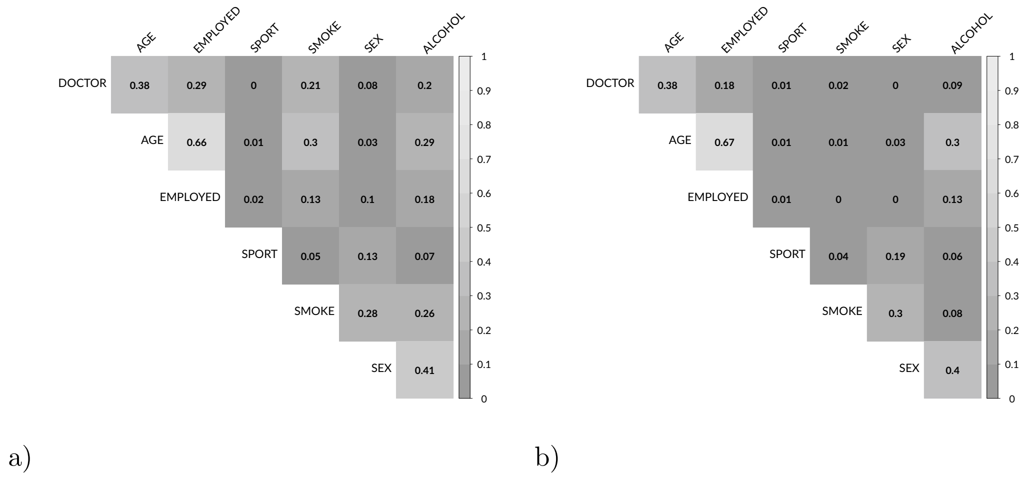
3. the preservation of the joint distribution.

As fourth quality level, Rässler (2002) proposed the preservation of the individual values. It is accomplished if the synthetic values equal the true values. This quality level is not considered in the following since, on the one hand, we have no information about the true values but only on the sample values, and on the other hand, if the joint distribution is well preserved the accordance of the synthetic values with the true values yields no further statistical information.

The first quality level is investigated by computing the Jensen-Shannon divergence (e.g. Lin, 1991) between the univariate marginal distributions $\hat{\pi}(x_j)$, $\hat{\pi}(y_k)$, and $\hat{\pi}(z_\ell)$, estimated on the complete GGSS data sample and the (partly synthetic) univariate marginal distributions $\tilde{\pi}(x_j)$, $\tilde{\pi}(y_k)$, and $\tilde{\pi}(z_\ell)$ determined by statistical matching for every $j = 1, \ldots, p$; $k = 1, \ldots, q$; $\ell = 1, \ldots, r$. The computation of the Jensen-Shannon divergence is problematic if structural zeros appear in the sample distribution. To deal with these cases, we set the structural zero to 10^{-16} which is numerically almost zero. We have also investigated the divergences between all multivariate marginals. The results are not shown here in detail due to their scope, but they are available on request. In summary, the results show that the Jensen-Shannon divergence from the matched distributions to the sample distributions distribution is small (the maximal value is 0.0479) and it becomes larger the more variables are included in the marginals.

The marginals $\tilde{\pi}(x_j)$, $\tilde{\pi}(y_k)$, and $\tilde{\pi}(z_\ell)$ are computed by summarizing over the corresponding components of joint distribution $\tilde{\pi}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$ which is estimated using Equation (4). The estimates $\hat{\pi}(x_j)$, $\hat{\pi}(y_k)$, $\hat{\pi}(z_\ell)$, and $\hat{\pi}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$, all computed on the complete GGSS sample, serve as our references for subsequent comparisons since the true joint distribution over the whole population for the GGSS data is of course unknown. The Jensen-Shannon divergence ($\in [0; 1]$) between the univariate marginals in the GGSS sample and the marginals determined by statistical matching is displayed in Table 2. The divergence is close to zero for all univariate marginals which means that the sample distributions and the statistically matched distributions are very similar. As expected, the smallest differences can be observed between the marginals of the common variables. The largest differences can be observed at the specific variables DOCTOR, SPORT, and SMOKE.

The second quality level is investigated by comparing the corrected contingency coefficient which is also known as Sakoda's adjusted Pearson's C ($\in [0, 1]$). To obtain the values for this association measure for the statistically matched file, we generate a complete synthetic file from $\tilde{\pi}$ by multiplying the number of desired observations with the estimated probability components of $\tilde{\pi}$. Subsequently, we use this synthetic data to compute the corrected contingency coefficients for the statistically matched file. Figure 3 shows the pairwise associations between all variables a) in the GGSS sample and b) the statistically matched data file. As expected, the associations are attenuated in the matched file. Especially the associations of the variable SMOKE with most of the other variables are strongly weakened. The largest difference between the corrected contingency table can be observed between SMOKE and AGE. Although the association is reflected in the file A, the statistical matching procedure was not able to reproduce this connection. The bivariate associations between the other variables, however, seem to be well preserved. Further analyses showed that the weakened associations with the variable SMOKE arise from an error-prone estimation of the graph structure. Especially an additional edge between SMOKE and AGE (which is present in the graph estimated on the complete GGSS sample) markedly improves the results of statistical matching. Another edge between DOCTOR and EMPLOYED improves the results even further. The resulting network structure and the bivariate corrected contingency coefficients are shown in Appendix C.4.



a) b)

Figure 3: Corrected contingency coefficient between pairs of variables on the complete GGSS sample (on the left) and the matched synthetic file (on the right).

The joint distribution of $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ contains 288 ($= 2^5 \cdot 3^2$) probability components, each of which was estimated on the complete GGSS sample. Figure 4 shows the estimates for each probability component of $\pi(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$. It suggests that statistical matching has a tendency to overestimate small probabilities and underestimate large probabilities. The Manhattan distance is 0.455, and 0.416 omitting the structural zeros in the sample distribution. The Jensen-Shannon divergence is 0.073 if we set the structural zeros numerically to zero (10^{-16}), and 0.054 if we ignore them. All in all, the differences move in a rather small range of values, which suggests that our method performs well, at least in this application.

5 Concluding remarks

Within this paper, we presented a new macro approach for statistical matching, based on the assumption of conditional independence of the specific variables given the common variables. This assumption builds a natural bridge to probabilistic graphical models aiming at a graphical representation of the dependencies among a set of variables, which can be used to find a convenient factorization of the joint distribution. For the embedding of statistical matching into the comprehensive theory of probabilistic graphical models, we restrict the graph to a shape that reflects the conditional independence of the specific variables given the common variables. Based on this graph, we estimate the factors that together form the joint distribution with the aid of a log-linear Markov network. Starting with this estimate of the joint distribution, the creation of a complete synthetic data file (micro approach) can easily be realized by drawing samples from it. We showed the applicability of our new approach using data of the German General Social Survey. Our preliminary results have indicated that our approach provides promising results at least for this data file. In particular, the small differences between the sample distribution and the distribution estimated using our statistical matching approach are very positive as we avoided overoptimism by deliberately not selecting the specific and common variables on the basis of previous association analyses. Moreover, all edges were found by a structure learning algorithm and no further substantively justified edges were artificially added. The

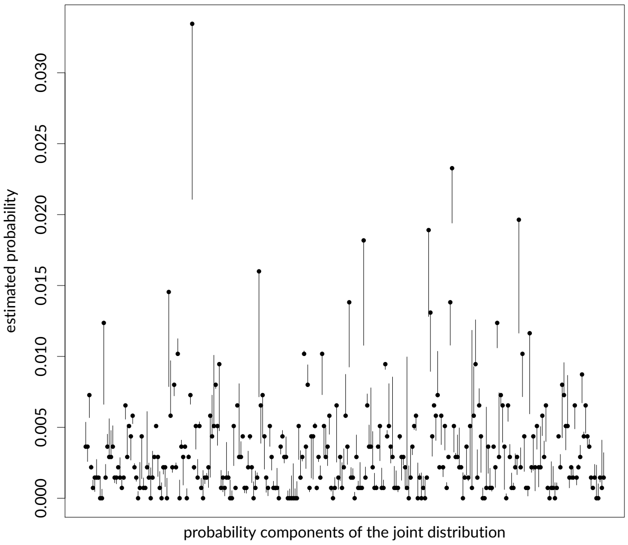


Figure 4: Absolute difference between the sample distribution and the matched distribution in the GGSS data example separately for all probability components of the joint distribution. The black points are the estimates for the components of $\hat{\pi}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$ based on the complete GGSS data. The lines indicate the absolute differences from the sample estimates to the estimates obtained by statistical matching. The endpoints of the lines equal the estimated probability components of $\tilde{\pi}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$.

question raised by these results is whether the statistical matching with Markov networks is equally successful with other data files. For this reason, further data files should be matched with this method and the comparison with other matching methods shall also be carried out. We recommend, as done here, the artificial matching of actually complete data files, where blocks of records are removed by hand because otherwise the results cannot be sufficiently evaluated. Another option would be to carry out simulation studies which would also offer a possibility to investigate how the statistical matching approach performs for situations where this particular conditional independence assumption does not hold. Nevertheless, the simulation of categorical data following a pre-defined dependence structure is associated with rather subtle issues that we have already listed and explained in Endres et al. (2018, App. A). Moreover, more work will need to be done to detect the influence of the structure learning algorithm on statistical matching and also under which conditions a (slightly) misspecified graph structure still leads to sufficiently good statistical matching results. Moreover, a generalization of this macro approach for continuous data or mixed continuous and categorical data would be strongly desirable.

Acknowledgements

We thank the GESIS – Leibniz Institute for Social Sciences for providing data for versatile research purposes. We furthermore like to thank Paul Fink, Cornelia Fütterer, Christoph Jansen, and Georg Schollmeyer for their valuable and constructive suggestions on earlier versions of this work. The first author also thanks the LMUMentoring programme for support. And finally, we thank Melissa Schmoll for excellent research assistance.

References

T. Aluja-Banet, J. Daunis-i-Estadella, N. Brunsó, and A. Mompart-Penina. Improving prevalence estimation through data fusion: Methods and validation. BMC Medical Informatics and Decision Making, 15(1):49, 2015. doi:10.1186/s12911-015-0169-z.

S. G. Baker. The multinomial-Poisson transformation. The Statistician, 43(4):495–504, 1994.

M. Di Zio and B. Vantaggi. Partial identification in statistical matching with misclassification. International Journal of Approximate Reasoning, 82:227–241, 2017. ISSN 0888613X. doi:www.doi.org/10.1016/j.ijar.2016.12.015.

M. D'Orazio, M. Di Zio, and M. Scanu. Statistical Matching: Theory and Practice. Wiley, Chichester, United Kingdom, 2006a. ISBN 9780470023532. doi:www.doi.org/10.1002/0470023554.

M. D'Orazio, M. Di Zio, and M. Scanu. Statistical matching for categorical data: Displaying uncertainty and using logical constraints. Journal of Official Statistics, 22(1): 137–157, 2006b.

E. Endres and T. Augustin. Statistical matching of discrete data by Bayesian networks. In A. Antonucci, G. Corani, and C. P. de Campos, editors, Proceedings of the Eighth International Conference on Probabilistic Graphical Models, volume 52 of Proceedings of Machine Learning Research, pages 159–170, Lugano, Switzerland, 2016. PMLR. URL `http://proceedings.mlr.press/v52/endres16.html`. [Accessed 28.11.2018].

E. Endres, P. Fink, and T. Augustin. Imprecise imputation: A nonparametric micro approach reflecting the natural uncertainty of statistical matching with categorical data. Technical Report 214, Department of Statistics, LMU Munich, 2018.

GESIS – Leibniz Institute for the Social Sciences. Allgemeine Bevölkerungsumfrage der Sozialwissenschaften ALLBUS 2012/German General Social Survey GGSS 2012, 2013. ZA4614 Data file Version 1.1.1.

GESIS – Leibniz Institute for the Social Sciences. GESIS - ALLBUS: ALLBUS Home, 2016. URL `http://www.gesis.org/en/allbus/allbus-home/`. [Accessed 28.11.2018].

S. Højsgaard, D. Edwards, and S. Lauritzen. Graphical Models with R. Use R! Springer, New York, 2012. ISBN 9781461422983. doi:10.1007/978-1-4614-2299-0.

D. Janzing. Merging joint distributions via causal model classes with low VC dimension. ArXiv e-prints, 2018. URL `https://arxiv.org/abs/1804.03206`. [Accessed 28.11.2018].

D. Koller and N. Friedman. Probabilistic Graphical Models: Principles and Techniques. MIT Press, Cambridge, MA, 2009.

J. Landes and J. Williamson. Objective Bayesian nets from consistent datasets. In A. Giffin and K. H. Knuth, editors, AIP Conference Proceedings, volume 1757, pages 020007–1 – 020007–8, Potsdam, NY, USA, 2016. doi:www.doi.org/10.1063/1.4959048.

S. L. Lauritzen. Graphical Models. Oxford University Press, Oxford, 1996. Reprinted version with corrections.

J. Lin. Divergence measures based on the Shannon entropy. IEEE Transactions on Information Theory, 37(1):145–151, 1991. ISSN 00189448. doi:www.doi.org/10.1109/18.61115.

J. Pearl. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Francisco, CA, 1988.

T. D. Pigott. A review of methods for missing data. Educational Research and Evaluation, 7:353–383, 2001. ISSN 13803611. doi:www.doi.org/10.1076/edre.7.4.353.8937.

R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL `https://www.R-project.org`.

S. Rässler. Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches. Springer, New York, NY, 2002. ISBN 9780387955162. doi:www.doi.org/10.1007/978-1-4613-0053-3.

M. Scutari. Learning Bayesian networks with the bnlearn R package. Journal of Statistical Software, 35(3):1–22, 2010. doi:10.18637/jss.v035.i03.

P. Serafino and R. Tonkin. Statistical matching of European Union statistics on income and living conditions (EU-SILC) and the household budget survey, 2017. Collection: Statistical working papers.

A. C. Singh, H. J. Mantel, M. D. Kinack, and G. Rowe. Statistical matching: Use of auxiliary information as an alternative to the conditional independence assumption. Survey Methodology, 19(1):59–79, 1993.

M. Studený. Probabilistic Conditional Independence Structures. Springer, London, 2010.

I. Tsamardinos, S. Triantafillou, and V. Lagani. Towards integrative causal analysis of heterogeneous data sets and studies. Journal of Machine Learning Research, 13:1097–1157, 2012.

G. Tutz. Regression for Categorical Data. Cambridge University Press, Cambridge, 2011. ISBN 9780511842061. doi:10.1017/CBO9780511842061.

J. Whittaker. Graphical Models in Applied Multivariate Statistics. Wiley, Chichester, 1990.

A Log-linear expansions for selected cases

Since, up to our knowledge, it is hard to find some examples for log-linear expansions, we provide some here, in the supporting information. We consider different situations which can easily be extended to higher dimensions. For more information on log-linear expansions we refer, for instance, to Whittaker (1990). Log-linear Markov networks are, for example, described in Lauritzen (1996) or Koller and Friedman (2009).

A.1 One variable with three categories

Let X be a random variable with realizations $x \in \{0,1,2\}$ and let

$$x_1 = \begin{cases} 1 & \text{, if } x = 1 \\ 0 & \text{, otherwise} \end{cases} \quad \text{and} \quad x_2 = \begin{cases} 1 & \text{, if } x = 2 \\ 0 & \text{, otherwise} \end{cases}$$

be dummy variables indicating these realizations. Then the distribution of X can be written as

$$\pi(x) = \pi(0)^{1-x_1-x_2}\pi(1)^{x_1}\pi(2)^{x_2}.$$

Applying the logarithm yields

$$\begin{aligned} \log \pi(x) &= \underbrace{\log \pi(0)}_{u_\varnothing} + x_1 \cdot \underbrace{\log\left(\frac{\pi(1)}{\pi(0)}\right)}_{u_{x1}} + x_2 \cdot \underbrace{\log\left(\frac{\pi(2)}{\pi(0)}\right)}_{u_{x2}} \\ &= u_\varnothing + x_1 \cdot u_{x_1} + x_2 \cdot u_{x_2}. \end{aligned}$$

The u-terms are here constants which we can rewrite as functions $u_X(\cdot)$ of x as follows

$$\log \pi(x) = u_\varnothing + u_{X_1}(x_1) + u_{X_2}(x_2).$$

A.2 2×3-contingency table

Let X and Y be a random variable with realizations $x \in \{0,1\}$ and $y \in \{0,1,2\}$ and let

$$y_1 = \begin{cases} 1 & \text{, if } y = 1 \\ 0 & \text{, otherwise} \end{cases} \quad \text{and} \quad y_2 = \begin{cases} 1 & \text{, if } y = 2 \\ 0 & \text{, otherwise} \end{cases}$$

be dummy variables indicating these realizations. Then the joint distribution of (X,Y) can be written as

$$\pi(x,y) = \pi(0,0)^{(1-x)(1-y_1-y_2)}\pi(1,0)^{x(1-y_1-y_2)}\pi(0,1)^{(1-x)y_1}\pi(1,1)^{xy_1}\pi(0,2)^{(1-x)y_2}\pi(1,2)^{xy_2}.$$

Applying the logarithm yields

$$\begin{aligned} \log \pi(x,y) &= \underbrace{\log \pi(0,0)}_{u_\varnothing} + x \cdot \underbrace{\log\left(\frac{\pi(1,0)}{\pi(0,0)}\right)}_{u_x} + y_1 \cdot \underbrace{\log\left(\frac{\pi(0,1)}{\pi(0,0)}\right)}_{u_{y1}} \\ &+ y_2 \cdot \underbrace{\log\left(\frac{\pi(0,2)}{\pi(0,0)}\right)}_{u_{y2}} + x \cdot y_1 \cdot \underbrace{\log\left(\frac{\pi(0,0)\pi(1,1)}{\pi(1,0)\pi(0,1)}\right)}_{u_{xy1}} + x \cdot y_2 \cdot \underbrace{\log\left(\frac{\pi(0,0)\pi(1,2)}{\pi(1,0)\pi(0,2)}\right)}_{u_{xy2}} \\ &= u_\varnothing + x \cdot u_x + y_1 \cdot u_{y_1} + y_2 \cdot u_{y_2} + x \cdot y_1 \cdot u_{xy_1} + x \cdot y_2 \cdot u_{xy_2}. \end{aligned}$$

Table 1: Linear predictors of the log-linear model in dependence of the realizations of X and Y.

x	y	log-linear model
0	0	$u_\varnothing$
0	1	$u_\varnothing + u_{y_1}$
0	2	$u_\varnothing + u_{y_2}$
1	0	$u_\varnothing + u_x$
1	1	$u_\varnothing + u_x + u_{y_1} + u_{xy_1}$
1	2	$u_\varnothing + u_x + u_{y_2} + u_{xy_2}$

The u-terms are here constants which we can rewrite as functions $u(\cdot)$ of the realizations x and y as

$$\begin{aligned}\log\pi(x,y) &= u_\varnothing + u_X(x) + u_{Y_1}(y_1) + u_{Y_2}(y_2) + u_{\{X,Y_1\}}(x,y_1) + u_{\{X,Y_2\}}(x,y_2)\\ &= u_\varnothing + u_X(x) + u_Y(y) + u_{\{X,Y\}}(x,y) \qquad (5)\end{aligned}$$

with

$$u_X(x) = \begin{cases} u_x & ,x=1\\ 0 & ,x=0,\end{cases} \qquad u_{\{X,Y\}}(x,y) = \begin{cases} u_{xy_2} & ,x=1,y=2\\ u_{xy_1} & ,x=1,y=1\\ 0 & ,x=1,y=0\\ 0 & ,x=0,y=2\\ 0 & ,x=0,y=1\\ 0 & ,x=0,y=0.\end{cases}$$

$$u_Y(y) = \begin{cases} u_{y_2} & ,y=2\\ u_{y_1} & ,y=1\\ 0 & ,y=0,\end{cases}$$

Table 1 shows the linear predictor from the log-linear expansion of $\pi(x,y)$ in dependence of the realizations x and y of X and Y.

A.3 $2\times2\times3$-contingency table

Let X, Y and Z be a random variable with realizations $x \in \{0,1\}$, $y \in \{0,1\}$, and $z \in \{0,1,2\}$. Then the joint distribution of (X,Y,Z) can be written as

$$\begin{aligned}\pi(x,y,z) =&\pi(0,0,0)^{(1-x)(1-y)(1-z)} \cdot \pi(1,0,0)^{x(1-y)(1-z)} \cdot \pi(0,1,0)^{(1-x)y(1-z)}\\ &\cdot\pi(0,0,1)^{(1-x)(1-y)z} \cdot \pi(1,1,0)^{xy(1-z)} \cdot \pi(1,0,1)^{x(1-y)z}\\ &\cdot\pi(0,1,1)^{(1-x)yz} \cdot \pi(1,1,1)^{xyz}.\end{aligned}$$

Applying the logarithm and the assumption that Y and Z are conditionally independent given X yields

$$\begin{aligned}
\log \pi(x,y,z) = \log \pi(x=0,y=0) &+ x \cdot \log\left(\frac{\pi(x=1,y=0)}{\pi(x=0,y=0)}\right) + y \cdot \log\left(\frac{\pi(x=0,y=1)}{\pi(x=0,y=0)}\right) \\
&+ xy \cdot \log\left(\frac{\pi(x=0,y=0)\pi(x=1,y=1)}{\pi(x=1,y=0)\pi(x=0,y=1)}\right) \\
&+ \log \pi(x=0,z=0) + x \cdot \log\left(\frac{\pi(x=1,z=0)}{\pi(x=0,z=0)}\right) + z \cdot \log\left(\frac{\pi(x=0,z=1)}{\pi(x=0,z=0)}\right) \\
&+ xz \cdot \log\left(\frac{\pi(x=0,z=0)\pi(x=1,z=1)}{\pi(x=1,z=0)\pi(x=0,z=1)}\right) \\
&- \log \pi(x=0) - x \cdot \log\left(\frac{\pi(x=1)}{\pi(x=0)}\right) \\
&= u_\varnothing + x \cdot u_x + y \cdot u_y + x \cdot y \cdot u_{xy} + u_\varnothing + x \cdot u_x + z \cdot u_z + x \cdot z \cdot u_{xz} \\
&- u_\varnothing - x \cdot u_x \\
&= u_\varnothing + x \cdot u_x + y \cdot u_y + x \cdot y \cdot u_{xy} + z \cdot u_z + x \cdot z \cdot u_{xz}
\end{aligned}$$

The u-terms are here constants which we can rewrite as functions $u(\cdot)$ of the realizations x, y, and z as

$$\begin{aligned}
\log \pi(x,y,z) &= u_\varnothing + u_X(x) + u_Y(y) + u_{\{X,Y\}}(x,y) + u_\varnothing + u_X(x) + u_Z(z) + u_{\{X,Z\}}(x,z) \\
&- u_\varnothing - u_X(x) \\
&= u_\varnothing + u_X(x) + u_Y(y) + u_{\{X,Y\}}(x,y) + u_Z(z) + u_{\{X,Z\}}(x,z) \\
&= \log(\pi(x,y)) + \log(\pi(x,z)) - \log(\pi(x)).
\end{aligned}$$

B Special features with the estimation in R

In the former sections, we aim at the estimation of the components of the joint probability distribution of the common and the specific variables. For this purpose, we assume that our data follows a multinomial distribution which can be expressed in terms of a log-linear expansion. Thus, the components of Equation (4) can be interpreted as linear predictors of multinomial regression models using a log-link and dummy coding. This also leads to an appropriate log-odds interpretation of the u-terms. However, in R, we use the glm()-function to fit a generalized Poisson regression model. This simplifies the maximization of the likelihood and leads to identical estimates (see Baker, 1994). Furthermore, since the log-linear model based on a Poisson-regression fits the expected cell counts of a multivariate contingency table and we estimate all parameters on different parts of the data, we have to rescale the results we obtain in R.

Let $m(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})$ denote the expected cell counts according to a certain realization $(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})$, and $\hat{m}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})$ the corresponding estimated values. Beginning with the conditional independence assumption (CIA), we obtain

$$\pi(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) \overset{CIA}{=} \frac{\pi(\boldsymbol{x},\boldsymbol{y})\cdot\pi(\boldsymbol{x},\boldsymbol{z})}{\pi(\boldsymbol{x})} = \frac{\frac{m(\boldsymbol{x},\boldsymbol{y})}{n}\cdot\frac{m(\boldsymbol{x},\boldsymbol{z})}{n}}{\frac{m(\boldsymbol{x})}{n}}$$
$$= \frac{m(\boldsymbol{x},\boldsymbol{y})\cdot m(\boldsymbol{x},\boldsymbol{z})}{n\cdot m(\boldsymbol{x})} = \frac{m(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})}{n}.$$

However, since we are facing the statistical matching problem, we cannot estimate neither $\pi(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})$ nor $m(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})$ on basis of all observations but only on basis of a subset of our data. This leads to the problem that the estimated marginals of $\boldsymbol{X}$ differ on A and B, more specifically $\hat{m}^A(\boldsymbol{x}) \neq \hat{m}^B(\boldsymbol{x})$. Thus, we have to take the basis of the estimates into account:

$$\hat{\pi}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) \overset{CIA}{=} \frac{\hat{\pi}^A(\boldsymbol{x},\boldsymbol{y})\cdot\hat{\pi}^B(\boldsymbol{x},\boldsymbol{z})}{\hat{\pi}^{A\uplus B}(\boldsymbol{x})}$$
$$= \frac{\frac{\hat{m}^A(\boldsymbol{x},\boldsymbol{y})}{n_A}\cdot\frac{\hat{m}^B(\boldsymbol{x},\boldsymbol{z})}{n_B}}{\frac{\hat{m}^{A\uplus B}(\boldsymbol{x})}{n}}$$
$$= \frac{n}{n_A\cdot n_B}\cdot\frac{\hat{m}^A(\boldsymbol{x},\boldsymbol{y})\cdot\hat{m}^B(\boldsymbol{x},\boldsymbol{z})}{\hat{m}^{A\uplus B}(\boldsymbol{x})}.$$

In the Poisson regression, the response is connected to the linear predictor η, which is a function of the covariates, by the log-link, i.e. $\log(m(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})) = \eta(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})$. To estimate the joint probability from the model equation, we have to multiply it with a factor that rescales with the number of observations as follows:

$$\hat{\pi}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) = \frac{n}{n_A\cdot n_B}\cdot\exp\left\{\hat{\eta}^A(\boldsymbol{x},\boldsymbol{y}) + \hat{\eta}^B(\boldsymbol{x},\boldsymbol{z}) - \hat{\eta}^{A\uplus B}(\boldsymbol{x})\right\}.$$

The superscripts symbolize which part of the data is used to estimate the corresponding parameters.

Thus, to obtain the estimates for the components of the joint probability distribution from the Poisson regression, the fitted values have to be multiplied by the correction factor $\frac{n}{n_A\cdot n_B}$.

C Further material for the GGSS application

For the application, we use data from the GESIS – Leibniz Institute for the Social Sciences (2016). Specifically, we use the data *ZA4614 (data file Version 1.1.1)* (GESIS – Leibniz Institute for the Social Sciences, 2013). Since we do not want our results to be additionally influenced by the missing data in the data, we remove the observations with missing entries in advance. This guarantees that only the quality of the statistical matching is reflected in the results.

C.1 Summary of possible realizations of the variables in the GGSS data

For the GGSS data, the true joint distribution is unknown and has to be estimated from the data. However, most of the considered variables have a lot of categories which leads to zeros in the estimation because we have much less observations than possible combinations in the categories. To reduce this zeros which are technically no structural zeros in the true distribution but estimated zeros in the empirical distribution, we summary some of the categories to obtain variables with two to three categories. The resulting possible categories are the following, where first is the reference category:

$$\begin{aligned}
&\text{sex} \in \mathcal{X}_{\text{SEX}} = \{\textit{male}, \textit{female}\},\\
&\text{age} \in \mathcal{X}_{\text{AGE}} = \{18-44\ \textit{years}, 45-59\ \textit{years}, \geq 60\ \textit{years}\},\\
&\text{employed} \in \mathcal{X}_{\text{EMPLOYED}} = \{\textit{employed}, \textit{unemployed}\},\\
&\text{smoke} \in \mathcal{Y}_{\text{SMOKE}} = \{\textit{smoker}, \textit{formerly smoked}, \textit{never smoked}\},\\
&\text{alcohol} \in \mathcal{Y}_{\text{ALCOHOL}} = \{\textit{occasionally or often}, \textit{never}\},\\
&\text{sport} \in \mathcal{Z}_{\text{SPORT}} = \{\textit{often}, \textit{seldom or never}\},\\
&\text{doctor} \in \mathcal{Z}_{\text{DOCTOR}} = \{\textit{sometimes or often}, \textit{seldom or never}\}.
\end{aligned}$$

C.2 Interpretation of the u-terms

As mentioned in the paper, the u-terms are interpretable as log-odds. In the following, we will exemplary show for the variables SEX and AGE how the interpretation can be derived from the estimation Equation (4). For better readability, the reference categories of all other variables are coded as 0. The derivation of the interpretation of all other u-terms works analogously.

C.2.1 $u_\varnothing$

$$\begin{aligned}
\pi(0,0,0,0,0,0,0) &= \exp(u_\varnothing)\\
\Leftrightarrow u_\varnothing &= \log(\pi(0,0,0,0,0,0,0))
\end{aligned}$$

C.2.2 $u_{\{\mathbf{SEX}\}}$

$$\begin{aligned}
\pi(\textit{female},0,0,0,0,0,0) &= \exp(u_\varnothing + u_{\{\text{SEX}\}}(\textit{female}))\\
\Leftrightarrow u_{\{\text{SEX}\}}(\textit{female}) &= \log\left(\frac{\pi(\textit{female},0,0,0,0,0,0)}{\pi(\textit{male},0,0,0,0,0,0)}\right)
\end{aligned}$$

C.2.3 $u_{\{\mathbf{AGE}\}}$

$$\begin{aligned}
\pi(0,45-59\ \textit{years},0,0,0,0,0) &= \exp(u_\varnothing + u_{\{\text{AGE}\}}(45-59\ \textit{years}))\\
\Leftrightarrow u_{\{\text{AGE}\}}(45-59\ \textit{years}) &= \log\left(\frac{\pi(0,45-59\ \textit{years},0,0,0,0,0)}{\pi(0,18-44\ \textit{years},0,0,0,0,0)}\right)
\end{aligned}$$

$$\begin{aligned}
\pi(0,\geq 60\ \textit{years},0,0,0,0,0) &= \exp(u_\varnothing + u_{\{\text{AGE}\}}(\geq 60\ \textit{years}))\\
\Leftrightarrow u_{\{\text{AGE}\}}(\geq 60\ \textit{years}) &= \log\left(\frac{\pi(0,\geq 60\ \textit{years},0,0,0,0,0)}{\pi(0,18-44\ \textit{years},0,0,0,0,0)}\right)
\end{aligned}$$

C.2.4 $u_{\{\mathbf{SEX},\mathbf{AGE}\}}$

$$\begin{aligned}
\pi(\textit{female},45-59\ \textit{years},0,0,0,0,0) \quad &= \exp(u_\varnothing + u_{\{\text{SEX}\}}(\textit{female}) + u_{\{\text{AGE}\}}(45-59\ \textit{years})\\
&\qquad + u_{\{\text{SEX,AGE}\}}(\textit{female},45-59\ \textit{years}))
\end{aligned}$$

$$\Leftrightarrow u_{\{\text{SEX,AGE}\}}(female, 45-59\ years))$$
$$= \log\left(\frac{\pi(female, 45-59\ years, 0,0,0,0,0)\cdot\pi(male, 18-44\ years, 0,0,0,0,0)}{\pi(female, 18-44\ years, 0,0,0,0,0)\cdot\pi(male, 45-59\ years, 0,0,0,0,0)}\right)$$

C.3 Estimates for the u-terms

Based on Equation (4), we have computed all estimates for the incorporated u-terms. They are displayed in the following tables, separated on the data used for estimation.

Table 2: Estimated coefficients $\hat{u}^{A\uplus B}$ concerning the common variables $\boldsymbol{X}$, estimated on $A \uplus B$.

variable name(s)	category	$\hat{u}^{A\uplus B}$
∅	*(intercept)*	5.1896
EMPLOYED	*unemployed*	-0.6674
AGE	45 – 59 *years*	-0.0973
AGE	≥ 60 *years*	-1.5395
SEX	*female*	-0.0800
EMPLOYED : AGE	*unemployed* : 45 – 59 *years*	-0.6129
EMPLOYED : AGE	*unemployed* : ≥ 60 *years*	2.3009

Table 3: Estimated coefficients $\hat{u}^{A}$ concerning the common variables $\boldsymbol{X}$ and the specific variables $\boldsymbol{Y}$, estimated on A.

variable name(s)	category	$\hat{u}^{A}$
∅	*(intercept)*	3.2315
SEX	*female*	-1.0673
AGE	45 – 59 *years*	-0.2576
AGE	≥ 60 *years*	-2.1091
ALCOHOL	*never*	-0.9116
SMOKE	*never smoked*	-0.1719
SMOKE	*smoker*	-0.0782
EMPLOYED	*unemployed*	-0.6397
SEX : AGE	*female* : 45 – 59 *years*	-0.1560
SEX : AGE	*female* : ≥ 60 *years*	-0.3784
SEX : ALCOHOL	*female* : *never*	1.1629
AGE : ALCOHOL	45 – 59 *years* : *never*	0.2623
AGE : ALCOHOL	≥ 60 *years* : *never*	1.0845
SEX : SMOKE	*female* : *never smoked*	0.9471
SEX : SMOKE	*female* : *smoker*	0.1169
AGE : EMPLOYED	45 – 59 *years* : *unemployed*	-0.7979
AGE : EMPLOYED	≥ 60 *years* : *unemployed*	2.2951
SEX : AGE : ALCOHOL	*female* : 45 – 59 *years* : *never*	0.2749
SEX : AGE : ALCOHOL	*female* : ≥ 60 *years* : *never*	0.0636

Table 4: Estimated coefficients $\hat{u}^B$ concerning the common variables $\boldsymbol{X}$ and the specific variables $\boldsymbol{Z}$, estimated on B.

variable name(s)	category	$\hat{u}^B$
∅	(*intercept*)	2.5522
SEX	*female*	0.2933
SPORT	*often*	0.4878
EMPLOYED	*unemployed*	-0.6993
AGE	45 – 59 *years*	0.2209
AGE	≥ 60 *years*	-0.8103
DOCTOR	*seldom or never*	0.3646
SEX : SPORT	*female* : *often*	-0.5702
EMPLOYED : AGE	*unemployed* : 45 – 59 *years*	-0.4393
EMPLOYED : AGE	*unemployed* : ≥ 60 *years*	2.3137
AGE : DOCTOR	45 – 59 *years* : *seldom or never*	-0.5433
AGE : DOCTOR	≥ 60 *years* : *seldom or never*	-1.4249

C.4 Results with two additional edges in the graph

We have also analyzed the statistical matching results after adding the following two (substantively plausible) further edges in the Markov network: (AGE, SMOKE), and (DOCTOR, EMPLOYED). Figure 5 shows the resulting graph, and Figure 6 the bivariate corrected contingency coefficients computed on basis of this network structure. The results indicate that the structure learning algorithm has a considerable impact on the statistical matching results. This effect should be examined in detail in future studies.

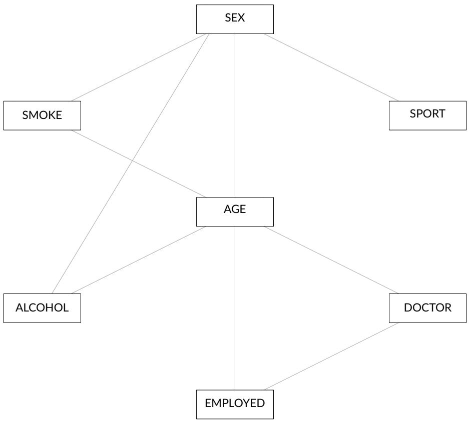
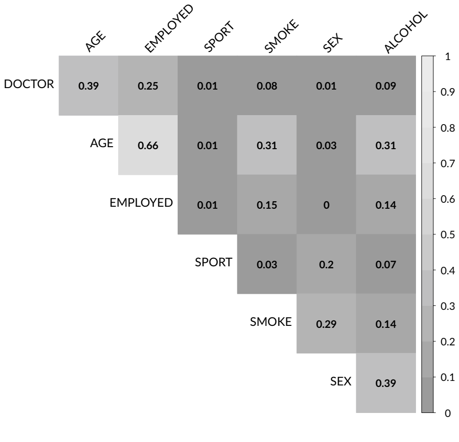


Figure 5: Markov network with two additional edges in between the specific variables and the common variables.



	AGE	EMPLOYED	SPORT	SMOKE	SEX	ALCOHOL
DOCTOR	0.39	0.25	0.01	0.08	0.01	0.09
AGE		0.66	0.01	0.31	0.03	0.31
EMPLOYED			0.01	0.15	0	0.14
SPORT				0.03	0.2	0.07
SMOKE					0.29	0.14
SEX						0.39

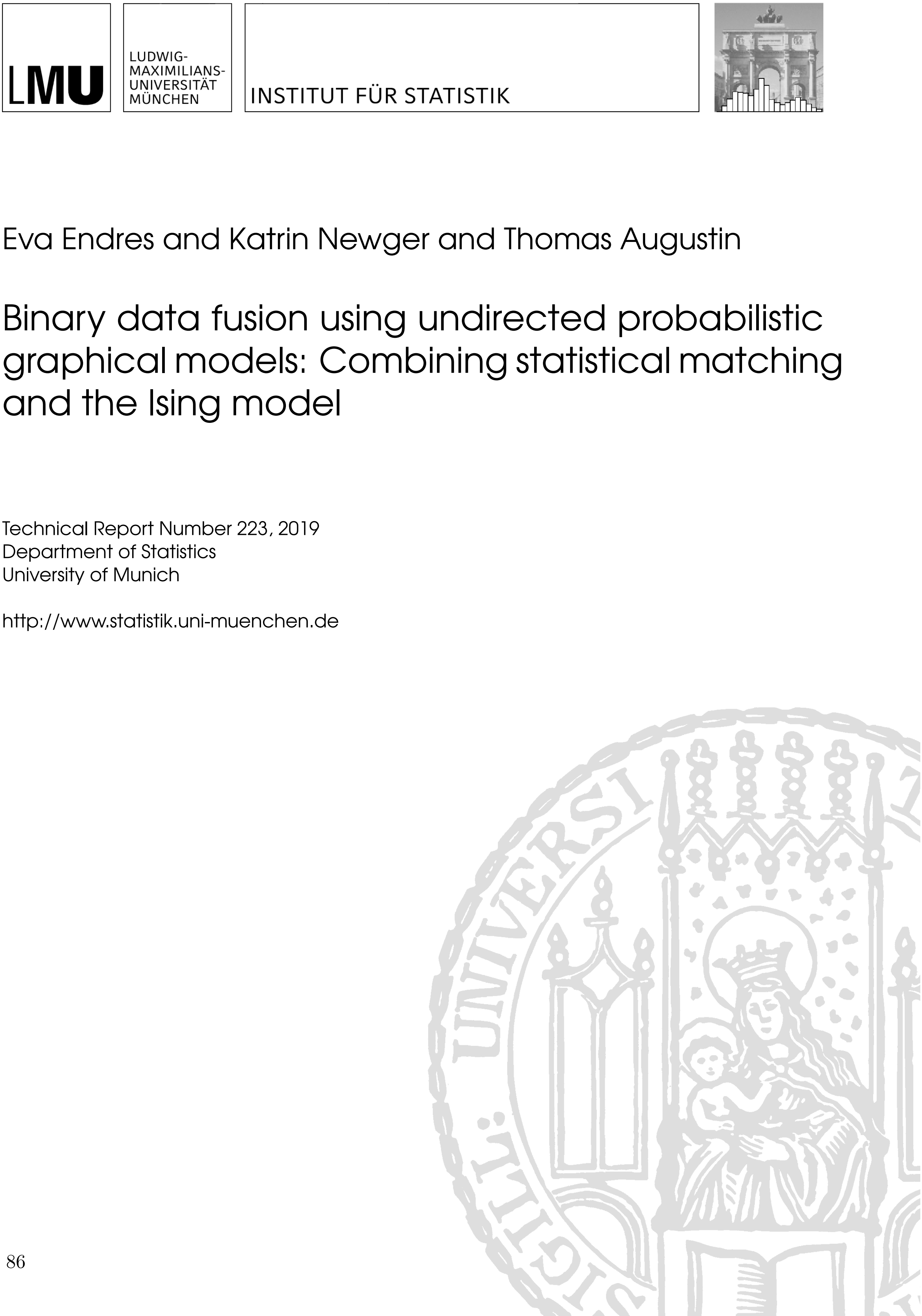
Figure 6: Markov network with two additional edges in between the specific variables and the common variables.

Contribution 3: pp. 86–105

Endres, E., Newger, K. and Augustin, T. (2019). Binary data fusion using undirected probabilistic graphical models: Combining statistical matching and the Ising model, Technical Report 223, Department of Statistics, LMU Munich.

The original publication is available at

`https://epub.ub.uni-muenchen.de/61732/`



LMU

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

INSTITUT FÜR STATISTIK

Eva Endres and Katrin Newger and Thomas Augustin

Binary data fusion using undirected probabilistic graphical models: Combining statistical matching and the Ising model

Technical Report Number 223, 2019
Department of Statistics
University of Munich

http://www.statistik.uni-muenchen.de

Binary data fusion using undirected probabilistic graphical models: Combining statistical matching and the Ising model

Eva Endres[*] Katrin Newger Thomas Augustin[†]

Department of Statistics, LMU München

23rd April 2019

Abstract

Graphical models can prove quite powerful for statistical matching, making secondary data analysis feasible also in situations where joint information about variables that were not collected together is sought. Without any constraints regarding the direction of influence of variables, we develop a method that uses the graphical Ising model to merge two or more data files containing binary data only. To this end, we rely on the conditional independence assumption commonly made in statistical matching to learn a joint Markov network graph structure over all variables from the given data. Based on this joint graph, the probability distribution is estimated by an adapted version of the Ising model. The quality of our new data fusion method is assessed on basis of a simulation study, sampling data from random Ising models. We investigate which parameters influence the quality of data integration, and how violations of the conditional independence assumption affect the results.

Keywords: statistical matching; data fusion; Markov network; Ising model; conditional independence

1 Introduction

With the ever growing flow of data, methods like statistical matching increase in relevance. Despite the mass of data, we may still face the problem that we need joint information about variables that have not been jointly observed. Statistical matching is a powerful tool and a relevant method of today's data analysis which tackles this problem (e.g. D'Orazio et al.,

[*] eva.endres@stat.uni-muenchen.de

[†] augustin@stat.uni-muenchen.de

2006a). The goal of statistical matching is to aggregate at least two independent data sets, A and B, containing only partly overlapping sets of variables, to achieve *joint information* about separately observed variables.

Several methods are available for this aim that differ in assumption, the presence of auxiliary inforamtion, and the type of results (e.g. D'Orazio et al., 2006a). Our work concentrates on the commonly used assumption of conditional independence of the so-called *specific variables*, given the *common variables.* This assumption leads to a factorization of the joint probability distribution in a form such that the problem of matching two disjoint data sets becomes solvable.

A framework that uses the decomposition of data by decoding independencies in the data is that of probabilistic graphical models (e.g. Koller and Friedman, 2009). For the aim of statistical matching it offers a great opportunity to handle the matching problem itself, but also to get an intuitive access to the structure of the data.

In this paper we will make a case for using Markov networks – an undirected variation of probabilistic graphical models – to perform statistical matching. The proceedings of this paper will be as follows. We will start with a recap of statistical matching in Section 2, followed in Section 3 by a brief summary of later needed aspects of undirected probabilistic graphical models. After recalling some general aspects in Subsection 3.1, we focus in Subsection 3.2 on our case of binary data and cover the Ising model. After that we will reformulate probabilistic graphical models for the aim of statistical matching in Section 4. To provide a general frame of how two independent binary files can be matched with Ising models, we adjust the Ising model to fit the data situation of statistical matching. To test our newly developed method, in Section 5 we simulate data from random Ising models. Knowing the true data, we split the original simulated data set into two disjoint i.i.d. data files A and B to perform statistical matching with probabilistic graphical models. Finally, we summarize and discuss our findings in Section 6.

2 Statistical matching

Data fusion, which is also known as *statistical matching* or *data integration*, means the integration of (at least) two data files A and B. File A is a data matrix containing n_{A} binary observations $(y_{a1}, \ldots, y_{aq}, x_{a1}, \ldots, x_{ap})$, where the index a is an element of the index set $\mathcal{I}_{\mathsf{A}}$ and refers to the a-th observation. Analogously, B contains n_{B} binary observations $(x_{b1}, \ldots, x_{bp}, z_{b1}, \ldots, z_{br})$, indexed by $b \in \mathcal{I}_{\mathsf{B}}$. The index sets $\mathcal{I}_{\mathsf{A}}$ and $\mathcal{I}_{\mathsf{B}}$ are disjoint. Altogether, we consider three sets of random variables: the set of *common variables* $\mathbf{X} = \{X_1, \ldots, X_p\}$, and the sets of *specific variables* $\mathbf{Y} = \{Y_1, \ldots, Y_q\}$ and $\mathbf{Z} = \{Z_1, \ldots, Z_r\}$. The sets of possible realizations are the Cartesian

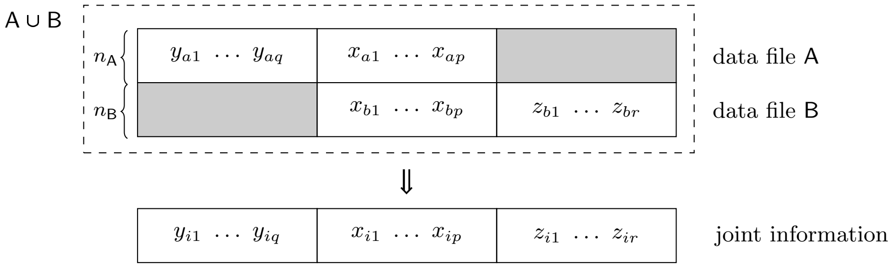


Figure 1: Graphical illustration of the data setting for statistical matching (cf. D'Orazio et al., 2006a, p. 5 (modified)).

products of the sets of possible realizations of the single elements of **X**, **Y**, and **Z**, respectively, and denoted by $\mathcal{X} = \bigtimes_{j=1}^{p} \mathcal{X}_j$, $\mathcal{Y} = \bigtimes_{k=1}^{q} \mathcal{X}_k$, and $\mathcal{Z} = \bigtimes_{\ell=1}^{r} \mathcal{Z}_\ell$.

Achieving the aim of statistical matching, namely the estimation of joint information of the specific variables, is considerably cumbered by a crucial identification problem. The missingness of any joint information on **Y** and **Z** makes the joint distribution unidentified. Even if we had an infinite number of observations, the relationship between the specific variables in **Y** and **Z** could not be estimated from the data without further assumptions or additional information.

Figure 1 shows the data situation graphically and indicates that statistical matching can also be interpreted as a missing data problem. However, the missing mechanism in the context of statistical matching can justifiably assumed to be ignorable (e.g. D'Orazio et al., 2006a, p. 6). Throughout the paper, we solely consider binary data and assume that the observations in A and B are independently and identically distributed, following a joint probability distribution

$$\pi(\mathbf{x}, \mathbf{y}, \mathbf{z}) := \mathbb{P}(X_1 = x_1, \dots, X_p = x_p, Y_1 = y_1, \dots, Y_q = y_q, Z_1 = z_1, \dots, Z_r = z_r).$$

This means that the union A ∪ B of the two files A and B can be viewed as a single data file where the observations $\mathbf{z}_a = (z_{a1}, \dots, z_{ar})$ and $\mathbf{y}_b = (y_{b1}, \dots, y_{bq})$ are missing in a block-wise pattern.

As previously mentioned, the aim of statistical matching is the collection of joint information on either **Y** and **Z**, or **X**, **Y**, and **Z**. According to D'Orazio et al. (2006a, p. 2), the term *joint information* refers to

1. the joint probability mass distribution or any of its characteristics (*macro approach*), or
2. a complete but synthetic data file with observations of **X**, **Y**, and **Z** (*micro approach*).

For instance, D'Orazio et al. (2006a) consider three different groups of approaches how these aims can be reached:

1. The oldest and probably most commonly used approach is based on the assumption of conditional independence of the specific variables $\mathbf{Y}$ and $\mathbf{Z}$ given the common variables $\mathbf{X}$. However, the validity of this assumption cannot be tested because of the missing joint information of the specific variables.

2. The second group of approaches is based on auxiliary information on the relationship between the specific variables $\mathbf{Y}$ and $\mathbf{Z}$. For instance, an additional data file might be available containing joint observations of the specific variables. Using a parametric approach, we could also have information about the parameters concerning the relation between $\mathbf{Y}$ and $\mathbf{Z}$.

3. The last group of approaches can be summarized under the umbrella term *partial identification.* In the absence of auxiliary information on the specific variables, these approaches do not force potentially unjustified assumptions to achieve a point-identified model for all variables of interest. In particular, this means that these approaches aim at finding all models which are compatible with the available data and rely on tenable assumptions only. These approaches yield a set of complete, synthetic data files for the micro approach or sets of plausible parameter estimates for the macro approach.

See, for instance, Di Zio and Vantaggi (2017), D'Orazio et al. (2006b), or Endres et al. (2018) for methods regarding the last type of statistical matching approaches. An approach belonging to the second type, which uses auxiliary information, is, for example, considered in Singh et al. (1993). For an overview of approaches that rely on the assumption of conditional independence, see, for instance, D'Orazio et al. (2006a, Chap. 2). Furthermore, Landes and Williamson (2016), and Endres and Augustin (2016) show how statistical matching can be incorporated into the context of Bayesian networks, under the assumption of conditional independence. A statistical matching method based on Markov networks for arbitrary categorical data is introduced in Endres and Augustin (2019).

With this paper, we will introduce a new statistical matching procedure for binary data. To tackle the identification problem, we work with the first type of approaches listed above, hence assuming conditional independence of the specific variables given the common variables. More precisely, we will embed the statistical matching task into the framework of the undirected probabilistic graphical Ising model, and derive an expression for the joint distribution of all specific and common variables that allows estimating it from the available data. Using the Ising model, the estimation of the joint distribution is markedly simplified compared to the more general approach in Endres and Augustin (2019). The factorization of the joint distribution cannot uniquely be determined from the graph structure. The Ising model is a pairwise Markov network that only considers connections between neighbouring variables. This simplifies the estimation of the joint distribution and the graph can intuitively be interpreted by potential users. In order to provide a basis for our way to proceed, in the next section we will first recapitulate a general definition for undirected graphical models, and then connect graphical models to statistical matching.

3 Undirected probabilistic graphical models

3.1 General aspects

In general, there are two kinds of probabilistic graphical models. Directed acyclic graphical models, which are also known under the term Bayesian networks, and undirected models, which are known as Markov networks or Markov random fields. Both types of models are suitable for dealing with categorical variables. For information on Bayesian networks, see, for instance, Koller and Friedman (2009). In the sequel we will focus on undirected probabilistic graphical models, which are discussed in more detail below.

Markov networks aim at the graphical representation of the dependence structure among a set of categorical[1] random variables. They are composed of a graph $\mathcal{H} = (\dot{\mathbf{X}}, \mathbf{E})$ and a probability distribution $\mathbb{P}$ containing only positive components. In this notation, $\dot{\mathbf{X}}$ refers to a set of nodes which represent the random variables of the set $\mathbf{X}$, and $\mathbf{E} \subseteq \dot{\mathbf{X}} \times \dot{\mathbf{X}}$ refers to the set of undirected edges in the graph. If two random variables $X_j,\ X_{j'} \in \mathbf{X}$, $j \neq j'$, are dependent, there is an edge between them, and $(\dot{X}_j, \dot{X}_{j'})$ is an element of $\mathbf{E}$. Iff there is only an indirect path from $\dot{X}_j$ to $\dot{X}_{j'}$, the random variables X_j and $X_{j'}$ are conditionally independent, given the variables that are traversed by the path. The nodes $\dot{X}_j$ and $\dot{X}_{j'}$ are then said to be separated. If the graph is an I-map for the joint distribution of the variables, which means that all (conditional) independencies that can be read-off the graph are present in the distribution, the graph structure can be used to find a suitable factorization of the joint distribution.

In general, a Gibbs distribution is suitable to reflect the factorization of $\mathbb{P}$ according to the corresponding graph structure. It represents the distribution as a product of so-called factors f, one for each maximal clique $\mathbf{C}_1, \dots, \mathbf{C}_m$. A clique is defined as a subset of $\dot{\mathbf{X}}$, where all pairwise edges between the nodes in the clique are in $\mathbf{E}$. The joint distribution of $\mathbf{X}$ is

[1] In general, Markov networks can handle continuous data as well. However, we restrict ourselves to categorical data in this paper.

given as

$$\pi(\mathbf{x}) := \frac{1}{N} \prod_{o=1}^{m} f(\mathbf{C}_o), \text{ with } N = \sum_{\mathbf{x} \in \mathcal{X}} \left\{ \prod_{o=1}^{m} f(\mathbf{C}_o) \right\}, \tag{1}$$

which is a normalized product over m factors f, where f is a function from the set of possible realizations corresponding to the nodes forming a certain clique to the positive real numbers. This means that, although not explicitly visible in Equation (1), the factors are indeed dependent on the realizations $\mathbf{x}$. The normalizing constant[2] N is needed to ensure that $\pi(\mathbf{x})$ is a probability mass function.

In the following, we will focus on pairwise Markov networks. Within these models, factors are either over single nodes (node potentials $f(x_j); j = 1, \ldots, p$) or over pairwise edges (edge potentials $f(x_j, x_{j'})$; $j, j' \in \{1, \ldots, p\}$; $j \neq j'$). This results in two being the highest order of interaction terms. Moreover, our research is based on a special type of pairwise Markov networks, which is limited to binary random variables. This class of models can be described by the so-called Ising model[3]. With this constraint, the unhandy normalizing constant N will be much easier to tackle, as we will show later on.

3.2 The Ising model for binary data

The Ising model, originally developed by Ernst Ising (1925), comes from statistical physics and was used to describe ferromagnetism under the assumption of solely pairwise interacting neighbouring atoms. The basis is a magnetic field that is arranged in a grid. The magnetic field consists of elements which can take values in $\{0; 1\}$. They represent whether an atom's spin[4] is positive or negative. The spin of an atom is influenced by two factors: each atom has a ground level that affects the direction of the atom charge, and additionally each atom is influenced by the charge of its direct neighbouring atoms (Kindermann and Snell, 1980, pp. 1ff.). In summary, a ferromagnetic field consisting of p atoms referred to as $x_1, \ldots, x_p$ can have $|\mathcal{X}| = 2^p$ different states. The field remains in the state that costs the least energy. The herein used term energy refers to the physical quantity. In physical theory it is common to assume that an object prefers the state that

[2] The normalizing constant is in some literature also called partition function (e.g. Koller and Friedman, 2009, Chap. 4).

[3] A generalization of the Ising model with arbitrary numbers of categories is covered by the Potts model (e.g. Koller and Friedman, 2009, p. 127).

[4] The original Ising model is based on an effect coding where the elements are either -1 or 1. The coding with realizations in the set $\{0; 1\}$ can be attributed to the Boltzmann distribution. However, it can be shown that the energy functions of the two representations are equivalent. We use the dummy coding throughout this paper.

costs the least energy; this is exactly what the Ising model expresses (McCoy and Wu, 1973, pp. 2ff.).

This Ising model can easily be used to describe a probabilistic model of p binary random variables with 2^p possible realizations. As Kindermann and Snell (1980, p. 2) write, it means to put a probability measure on the set of possible realizations $\mathcal{X}$. In the following, we will briefly recall how the joint probability mass distribution of the variables $\mathbf{x} = (x_1, \ldots, x_p)$ can be derived.

By rewriting the factors of Equation (1) with the aid of energy functions e, we derive

$$\pi(\mathbf{x}) = \frac{1}{N} \cdot \exp\Big\{ - \sum_{o=1}^{m} e(\mathbf{C}_o) \Big\}, \tag{2}$$

with $f(\mathbf{C}) := \exp\{-e(\mathbf{C})\}$. Since we are considering the special case of pairwise Markov networks with binary variables, this leads to the following node potentials and edge potentials:

$$f(x_j) = \exp\{-e(x_j)\} \quad \text{and} \quad f(x_j, x_{j'}) = \exp\{-e(x_j, x_{j'})\}. \tag{3}$$

Hence, the joint distribution is

$$\pi(\mathbf{x}) = \frac{1}{N} \cdot \exp\Big\{ - \sum_{\dot{X}_j \in \dot{\mathbf{X}}} e(x_j) - \sum_{(\dot{X}_j, \dot{X}_{j'}) \in \mathbf{E}} e(x_j, x_{j'}) \Big\}. \tag{4}$$

The overall energy of this distribution can be expressed by a Hamiltonian function (e.g. van Borkulo et al., 2014, supplementary information) of the form

$$H(\mathbf{x}) = - \sum_{\dot{X}_j \in \dot{\mathbf{X}}} \tau_j \; x_j - \sum_{(\dot{X}_j, \dot{X}_{j'}) \in \mathbf{E}} \beta_{j,j'} \; x_j \; x_{j'}, \tag{5}$$

where τ_j is the weight for the j-th node in the graph, and $\beta_{j,j'}$ is the weight of the edge between $\dot{X}_j$ and $\dot{X}_{j'}$. This yields the following form for the joint distribution of the Ising model:

$$\pi(\mathbf{x}) = \frac{1}{N} \cdot \exp\Big\{ - H(\mathbf{x}) \Big\} = \frac{1}{N} \cdot \exp\Big\{ \sum_{\dot{X}_j \in \dot{\mathbf{X}}} \tau_j \; x_j + \sum_{(\dot{X}_j, \dot{X}_{j'}) \in \mathbf{E}} \beta_{j,j'} \; x_j \; x_{j'} \Big\}, \tag{6}$$

$$\text{with} \quad N = \sum_{\mathbf{x}} \exp\Big\{ \sum_{\dot{X}_j \in \dot{\mathbf{X}}} \tau_j \; x_j + \sum_{(\dot{X}_j, \dot{X}_{j'}) \in \mathbf{E}} \beta_{j,j'} \; x_j \; x_{j'} \Big\}. \tag{7}$$

4 Using the Ising model to integrate data

As previously mentioned, probabilistic graphical models consist of a graph structure and a probability distribution. Given the graph structure, we can find a factorization of the probability distribution. The single components of this factorization can be subsequently estimated from the data. Thus, if the true graph structure is unknown, the first issue we have to tackle is the estimation of the graph structure of the joint Markov network of $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ on $\mathsf{A} \cup \mathsf{B}$. Thus, the assumption of conditional independence will be crucial.

4.1 Estimating a joint network structure for X, Y and Z

When it comes to the estimation of the joint Markov network for $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$, we have to consider the special data situation. The problem we are still confronted with is the missing joint information on the specific variables. To address this problem, the assumption of conditional independence of the specific variables given the common variables comes into play. Thinking of a Markov network that represents this assumption, it must hold that there is no direct path between any $\dot{Y}_k \in \dot{\mathbf{Y}}$ and $\dot{Z}_\ell \in \dot{\mathbf{Z}}$. Note that paths from nodes in $\dot{\mathbf{Y}}$ to nodes in $\dot{\mathbf{Z}}$ over at least one $\dot{X}_j \in \dot{\mathbf{X}}$ are allowed after all, and even wanted. The simplest conceivable situation is sketched in Figure 2. In these cases, at least one $\dot{X}_j \in \dot{\mathbf{X}}$ separates the specific variables, which is the graphical counterpart for conditional independence. When it comes to the graph structure, we have to ensure that every path from $\dot{\mathbf{Y}}$ to $\dot{\mathbf{Z}}$ leads over at least one $\dot{X}_j \in \dot{\mathbf{X}}$, or vice versa.

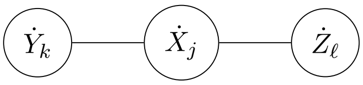


Figure 2: Basic form of the Ising model for statistical matching, reflecting the conditional independence assumption $\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} \mid \mathbf{X}$.

The estimation of the graph structure takes place in two steps, starting with the separate estimation of the graph structures on A and on B. With this procedure, we will receive two graphs, $\hat{\mathcal{H}}^{\mathsf{A}}_{\dot{X},\dot{Y}} = (\{\dot{\mathbf{X}}, \dot{\mathbf{Y}}\}, \hat{\mathbf{E}}_{\{\dot{\mathbf{X}},\dot{\mathbf{Y}}\}})$ and $\hat{\mathcal{H}}^{\mathsf{B}}_{\dot{X},\dot{Z}} = (\{\dot{\mathbf{X}}, \dot{\mathbf{Z}}\}, \hat{\mathbf{E}}_{\{\dot{\mathbf{X}},\dot{\mathbf{Z}}\}})$, containing the dependence structures among $\mathbf{X}$ and $\mathbf{Y}$, or $\mathbf{X}$ and $\mathbf{Z}$. However, it cannot be guaranteed that the estimated structure of the common variable is identical for both graphs. This is because the information for $\mathbf{X}$ in the sample is not necessarily the same due to random variations, even though it is assumed to be from the same population. In the event that the structures are different, we propose a procedure described by Endres and Augustin (2016, p. 5) for obtaining the joint network $\hat{\mathcal{H}}^{\mathsf{A}\cup\mathsf{B}}_{\dot{\mathbf{X}}\dot{\mathbf{Y}}\dot{\mathbf{Z}}} = (\{\dot{\mathbf{X}}, \dot{\mathbf{Y}}, \dot{\mathbf{Z}}\}, \mathbf{E}_{\{\dot{\mathbf{X}},\dot{\mathbf{Y}},\dot{\mathbf{Z}}\}})$. The set of nodes $\{\dot{\mathbf{X}}, \dot{\mathbf{Y}}, \dot{\mathbf{Z}}\}$ simply equals the union of the single node sets, i.e. $\dot{\mathbf{X}} \cup \dot{\mathbf{Y}} \cup \dot{\mathbf{Z}}$, while the set of edges $\mathbf{E}_{\{\dot{\mathbf{X}},\dot{\mathbf{Y}},\dot{\mathbf{Z}}\}}$ is the union of all edges found in the two separate graphs. Using the union $\mathbf{E}_{\{\dot{\mathbf{X}},\dot{\mathbf{Y}}\}} \cup \mathbf{E}_{\{\dot{\mathbf{X}},\dot{\mathbf{Z}}\}}$ ensures that the subsequent factorization of the probability distribution contains all the dependencies found in the data. That is, if a dependence was found in one file but not in the other file, the edges will still appear in the joint network. Thus, the risk of random independencies yielding a faulty factorization for the probability distribution decreases.

4.2 Parameter estimation in the statistical matching context

As any nodes $\dot{Y}_k \in \dot{\mathbf{Y}}$ and $\dot{Z}_\ell \in \dot{\mathbf{Z}}$ are separated by at least one $\dot{X}_j \in \dot{\mathbf{X}}$, the interaction terms between Y_k and Z_ℓ, i.e. the edge potentials, are always zero. In summary, the overall energy of the Ising model within the statistical matching framework is, including the assumption of conditional independence, given by the following equation:

$$\begin{aligned} H(\mathbf{x},\mathbf{y},\mathbf{z}) = &- \sum_{\dot{X}_j \in \dot{\mathbf{X}}} \tau_j \; x_j - \sum_{\dot{Y}_k \in \dot{\mathbf{Y}}} \upsilon_k \; y_k \\ &- \sum_{\dot{Z}_\ell \in \dot{\mathbf{Z}}} \phi_\ell \; z_\ell - \sum_{(\dot{X}_j,\dot{X}_{j'}) \in \mathbf{E}_{\{\dot{\mathbf{X}},\dot{\mathbf{Y}},\dot{\mathbf{Z}}\}}} \beta_{j,j'} \; x_j \; x_{j'} \\ - \sum_{(\dot{Y}_k,\dot{Y}_{k'}) \in \mathbf{E}_{\{\dot{\mathbf{X}},\dot{\mathbf{Y}},\dot{\mathbf{Z}}\}}} \gamma_{k,k'} \; y_k \; y_{k'} &- \sum_{(\dot{Z}_\ell,\dot{Z}_{\ell'}) \in \mathbf{E}_{\{\dot{\mathbf{X}},\dot{\mathbf{Y}},\dot{\mathbf{Z}}\}}} \delta_{\ell,\ell'} \; z_\ell \; z_{\ell'} \\ - \sum_{(\dot{X}_j,\dot{Y}_k) \in \mathbf{E}_{\{\dot{\mathbf{X}},\dot{\mathbf{Y}},\dot{\mathbf{Z}}\}}} \epsilon_{j,k} \; x_j \; y_k &- \sum_{(\dot{X}_j,\dot{Z}_\ell) \in \mathbf{E}_{\{\dot{\mathbf{X}},\dot{\mathbf{Y}},\dot{\mathbf{Z}}\}}} \zeta_{j,\ell} \; x_j \; z_\ell. \end{aligned} \tag{8}$$

This Hamiltonian function contains a main effect (node potential) for each node in the corresponding graph, and one interaction effect (edge potential) for every pair of neighbouring nodes. Due to the assumption of the conditional independence of the specific variables given the common variables, it contains no term that depends on any $Y_k \in \mathbf{Y}$ and $Z_\ell \in \mathbf{Z}$ at the same time. This fact yields the solution for the initial statistical matching problem. Every term of the energy function can be estimated from a subset of the available data, namely either from A, from B, or from $\mathsf{A} \cup \mathsf{B}$. The joint probability distribution arises as

$$\pi(\mathbf{x},\mathbf{y},\mathbf{z}) = \frac{1}{N} \cdot \exp\Big\{ - H(\mathbf{x},\mathbf{y},\mathbf{z})\Big\}, \tag{9}$$

$$\begin{aligned} \text{where } N = \sum_{\mathbf{x},\mathbf{y},\mathbf{z}} \exp\Big\{ & \sum_{\dot{X}_j \in \dot{\mathbf{X}}} \tau_j \; x_j + \sum_{\dot{Y}_k \in \dot{\mathbf{Y}}} \upsilon_k \; y_k \\ & + \sum_{\dot{Z}_\ell \in \dot{\mathbf{Z}}} \phi_\ell \; z_\ell + \sum_{(\dot{X}_j, \dot{X}_{j'}) \in \mathbf{E}_{\{\dot{\mathbf{X}},\dot{\mathbf{Y}},\dot{\mathbf{Z}}\}}} \beta_{j,j'} \; x_j \; x_{j'} \\ & + \sum_{(\dot{Y}_k, \dot{Y}_{k'}) \in \mathbf{E}_{\{\dot{\mathbf{X}},\dot{\mathbf{Y}},\dot{\mathbf{Z}}\}}} \gamma_{k,k'} \; y_k \; y_{k'} + \sum_{(\dot{Z}_\ell, \dot{Z}_{\ell'}) \in \mathbf{E}_{\{\dot{\mathbf{X}},\dot{\mathbf{Y}},\dot{\mathbf{Z}}\}}} \delta_{\ell,\ell'} \; z_\ell \; z_{\ell'} \\ & + \sum_{(\dot{X}_j, \dot{Y}_k) \in \mathbf{E}_{\{\dot{\mathbf{X}},\dot{\mathbf{Y}},\dot{\mathbf{Z}}\}}} \epsilon_{j,k} \; x_j \; y_k + \sum_{(\dot{X}_j, \dot{Z}_\ell) \in \mathbf{E}_{\{\dot{\mathbf{X}},\dot{\mathbf{Y}},\dot{\mathbf{Z}}\}}} \zeta_{j,\ell} \; x_j \; z_\ell \Big\} \end{aligned} \tag{10}$$

denotes the corresponding partition function. More specifically, using this notation, the parameters τ_j and $\beta_{j,j'}$ are estimated from $\mathsf{A} \cup \mathsf{B}$, the parameters υ_k, $\gamma_{k,k'}$, and $\epsilon_{j,k}$ are estimated from A, and the parameters ϕ_ℓ, $\delta_{\ell,\ell'}$, and $\zeta_{j,\ell}$ are estimated from B.

5 Simulation study

To investigate statistical matching of binary data with a graphical Ising model, we have performed a simulation study whose basis is the log-linear model in Equation (9). Altogether, we varied the following simulation parameters:

1. the number of nodes in the graph:
   (a) a total of seven variables (three common variables, two specific variables in each file),
   (b) a total of twelve variables (four common variables, four specific variables in each file);
2. the (in)dependence structure:
   (a) the assumption of conditional independence applies (all interaction terms between the specific variables are zero),
   (b) the assumption of conditional independence is violated for some variables (an interaction term between two specific variables is zero with probability 0.2),
   (c) the assumption of conditional independence is violated for all variables (all interaction terms between the specific variables are not equal to zero);
3. the number of observations n with $n_{\mathsf{A}} = n_{\mathsf{B}} = n/2$ ($n = 50$, $n = 250$, n=1000);
4. the sizes of the interaction coefficients are sampled from a uniform distribution:

   (a) $U(0.5; 2)$,
   (b) $U(2; 5)$;

5. the adjacency of two nodes in the graph is determined randomly either with probability 0.7 or with probability 0.3.

In summary, this leads to 72 simulation designs, each of which has been repeated 50 times.

The simulation and all analyses are conducted in R (R Core Team, 2018), using the packages *IsingSampler* (Epskamp, 2015) for data simulations, and *IsingFit* (van Borkulo et al., 2016) for structure and parameter learning. The basis of the learning algorithms in *IsingFit* is the so-called *eLasso*. It integrates the extended Bayesian information criterion into the estimation of (conditional) logistic regression models to find relevant edges in the graph structure (van Borkulo et al., 2014). Former simulation studies showed that the eLasso performs very well and that errors are mainly due to 'the suppression of very weak edges to zero' (van Borkulo et al., 2014). Details on the eLasso method can, for instance, be found in van Borkulo et al. (2014) and van Borkulo (2018).

To generate data files A and B with a known joint distribution, we simulate a complete file containing $n_{\mathsf{A}} + n_{\mathsf{B}}$ observations, and randomly allocate the observations into A or B. Subsequently, the observations of the specific variables $\mathbf{Z}$ are removed from A, and the observations of $\mathbf{Y}$ are removed from B. The resulting files fit the context of statistical matching and they can be integrated to assess the performance of our proposed method.

5.1 Simulation results

To assess the quality of the statistical matching results obtained by our proposed method, we analyze the Jensen-Shannon divergence (e.g. Lin, 1991) between the distribution in the complete simulation file and the distribution in the synthetic file achieved by statistical matching. The investigation of the divergence between these two distributions corresponds to the second quality criterion for statistical matching developed by Rässler (2002). Overall, Rässler (2002) determines the quality of a statistical matching procedure by investigating whether the individual values, the joint distribution, the correlation structure, and the marginal distributions have been preserved. As already stated by D'Orazio et al. (2006a, p. 10), the preservation of the individual values is not crucial for statistical analysis since the relevant information lies within the joint distribution, and the third and fourth quality levels are per se not sufficient to assess the statistical matching quality. Thus, the second level, which ensures that all statistical information of the joint distribution from the complete sample is preserved in the joint distribution of the synthetic file, is our means of choice.

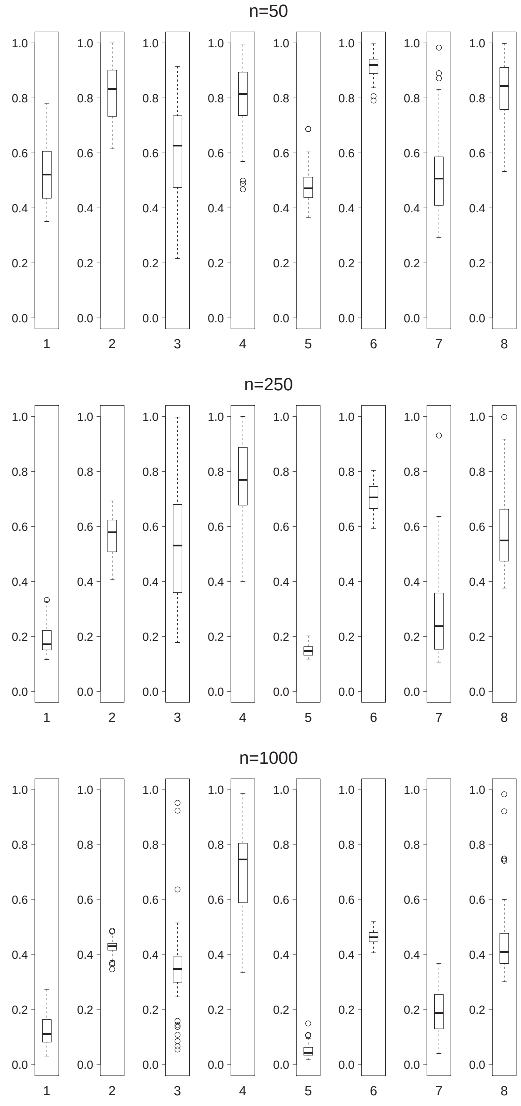


Figure 3: Jensen-Shannon divergences for the simulation setups, where the conditional independence assumption applies. The different rows indicate different sample sizes.

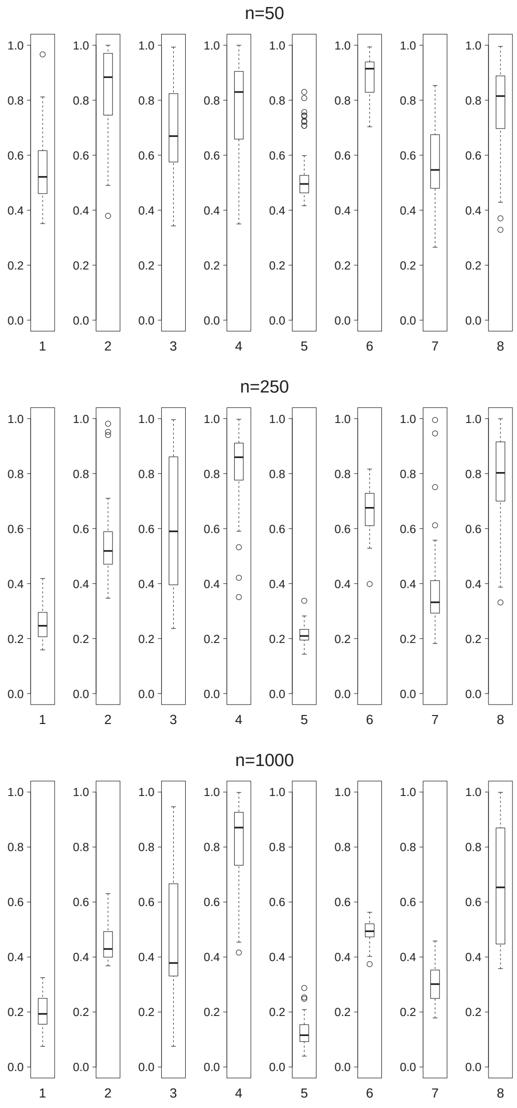


Figure 4: Jensen-Shannon divergences for the simulation setups, where the conditional independence assumption is violated for some variables. The different rows indicate different sample sizes.

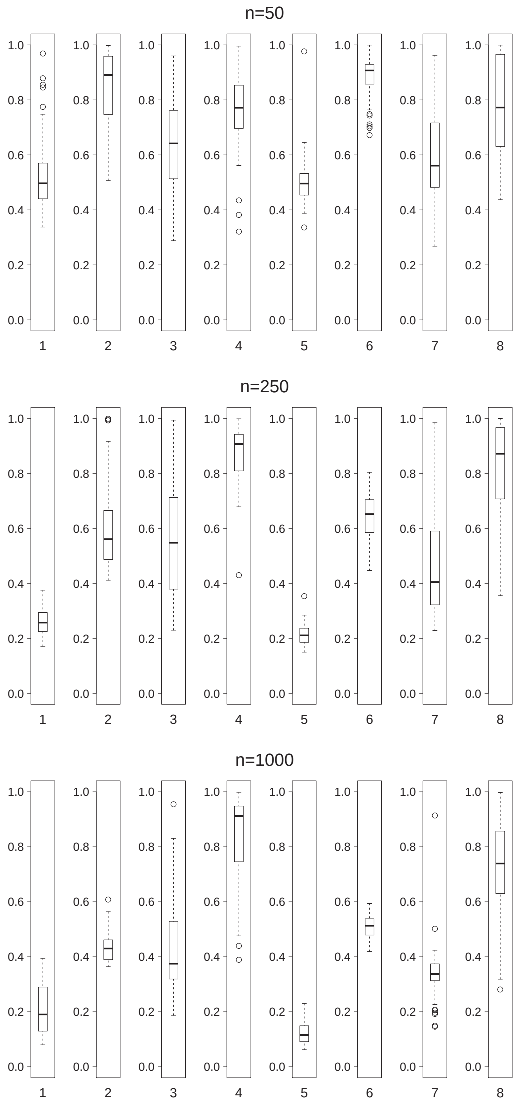


Figure 5: Jensen-Shannon divergences for the simulation setups, where the conditional independence assumption is violated for all variables. The different rows indicate different sample sizes.

The results of the Jensen-Shannon divergences are shown in Figures 3–5, separately for the different settings of the conditional independence assumption. Each figure contains three rows, each of which shows results for different numbers of observations. In every row, eight boxplots are displayed, which can be interpreted according to the parameter combinations listed in Table 1.

boxplot	number of nodes	interaction coefficients	adjacency probability
1	7	$U(0.5;2)$	0.7
2	12	$U(0.5;2)$	0.7
3	7	$U(2;5)$	0.7
4	12	$U(2;5)$	0.7
5	7	$U(0.5;2)$	0.3
6	12	$U(0.5;2)$	0.3
7	7	$U(2;5)$	0.3
8	12	$U(2;5)$	0.3

Table 1: Parameter combinations needed to interpret the boxplots in Figures 3–5.

All simulation scenarios support the statement that the higher the number of observations, the closer the distribution obtained by statistical matching is to the complete sample distribution. This effect can easily be explained: proportionally, we lose less statistical information when removing $\mathbf{Z}$ from A and $\mathbf{Y}$ from B if the overall number of observations is higher. Furthermore, we can observe that – as expected – the conditional independence has an influence on the quality of statistical matching. If the assumption holds, the Jensen-Shannon divergence between the complete sample distribution and the synthetic statistical matching distribution is in all cases smaller than in scenarios where the assumption is violated. Moreover, a slight violation of the assumption yields indeed better results than scenarios where the assumption is violated for all specific variables. This effect is most visible in boxplots 7–8, where the interaction coefficients are large and the adjacency probability is small. Interestingly, also all scenarios show that the number of nodes in the graph has a strong influence on the results. An overall number of seven nodes performs much better than a number of twelve nodes regarding the Jensen-Shannon divergence. This effect can indirectly also be attributed to the number of observations that is available for the estimation of node and edge potentials. Having more nodes and edges means that proportionally fewer observations are at hand that can be used for the estimation. Interaction coefficients drawn from the uniform distribution $U(0.5;2)$ lead to better results than the higher values drawn from $U(2;5)$, especially in cases where the total number of observations is 250 or 1000. Further research should consider whether this is due to the fact that methods using the conditional independence assumption also establish conditional independence in the matched, synthetic distribution (e.g. Rässler, 2002, p. 4). The generation of conditional independence may possibly result in the underestimation of large interaction coefficients. In most of the scenarios, a small adjacency probability seems to reduce the Jensen-Shannon divergence.

Since the simulation results were analysed by comparing the synthetic, matched distribution with the distribution estimated from the simulated complete sample, we particularly investigate the influence of the identification uncertainty on the Jensen-Shannon divergence. We can see that the smaller the sample sizes and the larger the number of nodes, the larger the divergences. This can be explained simply by the fact that the missing data has a stronger effect on smaller sample sizes, since markedly less data is available for estimation. Dependencies that are present in the complete sample are lost due to the block-wise lack of observations in the incomplete sample. Furthermore, high interaction effects get moderated if a lot of data is missing.

Summing up, the best results are obtained with a small number of nodes, combined with small interaction coefficients sampled from $U(0.5; 2)$. This parameter combination, moreover, affects the Jensen-Shannon divergence between the complete sample distribution and the synthetic statistical matching distribution in a very positive way. Even in situation where the conditional independence assumption is violated, this parameter combination yields divergences that are comparatively small and in the best case smaller than 0.1. This is a relevant finding since we face the problem that there is no way to test the assumption of conditional independence before matching the data. With this in mind, we were able to show that especially in a setting with less nodes (seven), interaction coefficients within $U(0.5; 2)$, an adjacency probability of 0.3, and a large sample size, the results obtained by statistical matching are still very good.

6 Summary, limitations, and outlook

The goal of this paper is to investigate the application and performance of a special type of Markov networks, namely the Ising model, as a method to perform statistical matching. Users are facing one main issue when matching data sets: the absence of any joint information about the specific variables. One popular option to solve this identification problem is to assume conditional independence of the specific variable blocks given the common variables. On basis of this assumption, we connect statistical matching of binary data with the probabilistic graphical Ising model, which uses the conditional independence assumption to derive a joint probability distribution of a set of binary variables. Beside the performance of the Ising model for the aim of statistical matching, the intuitive interpretation of Markov networks speaks for itself. Conclusions about the joint probability distribution can very easily be drawn. The user is not confronted with a set of parameters that is hard to understand, but rather with an intuitive graph. This undirected graph reveals the estimated dependence structure of the variables at first sight.

After a short recap of the theory of statistical matching and undirected probabilistic models, we presented the Ising model, which is the state-of-the-art model when fitting a Markov network for binary data. It has two main computational advantages compared to the more general Markov models: the computationally intensive normalization constant, which guarantees the characteristics of a density function, simplifies greatly with the help of the Ising model, and the model equation contains interaction effects of a maximum order of two. Our adapted version of the Ising model ensures that the block-wise missing data will not lead to any intractable problem. To achieve this goal no additional assumptions are made; only the conditional independence assumption is used. Although, critics may argue that this assumption is unjustified, we know that the stronger the relationship between the common and specific variables, and thus the higher the predictive power of the common variables for **Y** and **Z**, the is higher the chance of obtaining a good result for data fusion. To see how the graphical Ising model performs as a tool for statistical matching, we conducted a broad simulation study. On the basis of the adjusted log-linear model in Equation (9) we simulated data, which shows that the Ising model handles the task of matching two data sets very well. As we showed, the central assumption of conditional independence is relevant for the performance of the matching process. The best results are obtained in situations where the assumption holds. However, a main result of the simulation study is that the violation of the conditional independence assumption has less impact on the performance of statistical matching than expected. Even in settings that violated the conditional independence assumption for all variables, we found combinations of parameters that still gave good results.

As it could be expected, we are also facing limitations. The assumption of conditional independence is a strong one. When having serious doubts, that the assumption is fulfilled, the validity of results should be doubted as well. In this case another way of performing statistical matching is to prefer. Although, we investigated the influence of this assumption on the results, the simulation study cannot cover all possible parameters which might affect the statistical matching results. Moreover, the comparison of our proposed method to other statistical matching methods should be conducted in further simulation studies. A further natural progression of this work is to assess whether and how the Potts model, which is a generalization of the Ising model, can be used for statistical matching task. In connection with this, one could also investigate how this procedure theoretically and practically differs from the approach in Endres and Augustin (2019).

Right now, statistical matching is mostly used for official statistics. But with improving methods and better interpretation, statistical matching will become more relevant for applicants from other fields. Especially in areas like marketing research, where statistical matching has already been used in the past (see D'Orazio et al., 2006a, p. 174, for an overview of applications in this area), it is still of relevance to bring together data from surveys on an individual level. With statistical matching, the survey data can be used to get new results. Furthermore statistical matching can be a chance for biostatistics or medicine. In those areas it is often hard to collect meaningful data, especially where humans are involved. The secondary analysis of data can be a chance to reduce the number of respondents or variables. Taking this thought one step further, also personalized medicine can benefit from statistical matching. By putting together data sets with individual data, for example, forecasts on the success of certain treatments can be made.

References

M. Di Zio and B. Vantaggi. Partial identification in statistical matching with misclassification. *International Journal of Approximate Reasoning*, 82:227–241, 2017. doi: 10.1016/j.ijar.2016.12.015.

M. D'Orazio, M. Di Zio, and M. Scanu. *Statistical Matching: Theory and Practice*. Wiley, Chichester, United Kingdom, 2006a. doi: 10.1002/0470023554.

M. D'Orazio, M. Di Zio, and M. Scanu. Statistical matching for categorical data: Displaying uncertainty and using logical constraints. *Journal of Official Statistics*, 22(1):137–157, 2006b.

E. Endres and T. Augustin. Statistical matching of discrete data by Bayesian networks. In A. Antonucci, G. Corani, and C. P. de Campos, editors, *Proceedings of the Eighth International Conference on Probabilistic Graphical Models*, volume 52 of *Proceedings of Machine Learning Research*, pages 159–170, Lugano, Switzerland, 06–09 Sep 2016. PMLR. URL `http://proceedings.mlr.press/v52/endres16.html`.

E. Endres and T. Augustin. Utilizing log-linear Markov networks to integrate categorical data files. Technical Report 222, Department of Statistics, LMU Munich, 2019. URL `https://epub.ub.uni-muenchen.de/61678/`.

E. Endres, P. Fink, and T. Augustin. Imprecise imputation: A nonparametric micro approach reflecting the natural uncertainty of statistical matching with categorical data. Technical Report 214, Department of Statistics, LMU Munich, 2018. URL `https://epub.ub.uni-muenchen.de/42423/`.

S. Epskamp. *IsingSampler: Sampling Methods and Distribution Functions for the Ising Model*, 2015. URL `https://CRAN.R-project.org/package=IsingSampler`. R package version 0.2.

E. Ising. Beitrag zur Theorie des Ferromagnetismus. *Zeitschrift für Physik*, 31:253–258, 1925.

R. Kindermann and J. L. Snell. *Markov Random Fields and Their Applications.* Amercian Mathematical Society, Providence, 1980.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques.* MIT Press, Cambridge, MA, 2009.

J. Landes and J. Williamson. Objective Bayesian nets from consistent datasets. In A. Giffin and K. H. Knuth, editors, *AIP Conference Proceedings*, volume 1757, pages 020007–1 – 020007–8, Potsdam, NY, USA, 2016. doi: 10.1063/1.4959048.

J. Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. doi: 10.1109/18.61115.

B. M. McCoy and T. T. Wu. *The Two-Dimensional Ising Model.* Harvard University Press, Cambridge, MA, 1973.

R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2018. URL `https://www.R-project.org`.

S. Rässler. *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches.* Springer, New York, NY, 2002. doi: 10.1007/978-1-4613-0053-3.

A. C. Singh, H. J. Mantel, M. D. Kinack, and G. Rowe. Statistical matching: Use of auxiliary information as an alternative to the conditional independence assumption. *Survey Methodology*, 19(1):59–79, 1993.

C. van Borkulo. *Symptom network models in depression research: From methodological exploration to clinical application.* PhD thesis, University of Groningen, 2018.

C. van Borkulo, D. Borsboom, S. Epskamp, T. Blanken, L. Boschloo, R. Schoevers, and L. Waldorp. A new method for constructing networks from binary data. *Scientific Reports*, 4:1–10, 2014. doi: 10.1038/srep05918.

C. van Borkulo, S. Epskamp, and with contributions from Alexander Robitzsch. *IsingFit: Fitting Ising Models Using the ELasso Method*, 2016. URL `https://CRAN.R-project.org/package=IsingFit`. R package version 0.3.1.

Contribution 4: pp. 108–134

Endres, E., Fink, P. and Augustin, T. (2018). Imprecise imputation: A nonparametric micro approach reflecting the natural uncertainty of statistical matching with categorical data.

Accepted for publication in the *Journal of Official Statistics.*

Imprecise Imputation: A Nonparametric Micro Approach Reflecting the Natural Uncertainty of Statistical Matching with Categorical Data

Eva Endres[*] Paul Fink[†] Thomas Augustin[‡]

Department of Statistics, Ludwig-Maximilians-Universität München

17th April 2019

Statistical matching is the term for the integration of two or more data files which share a partially overlapping set of variables. Its aim is to obtain joint information on variables collected in different surveys based on different observation units. This naturally leads to an identification problem since there is no observation which contains information on all variables of interest.

We develop the first statistical matching micro approach reflecting the natural uncertainty of statistical matching arising from the identification problem in the context of categorical data. A complete synthetic file is obtained by imprecise imputation, replacing missing entries by *sets* of suitable values. Altogether, we discuss three imprecise imputation strategies and propose ideas for potential refinements.

Additionally, we show how the results of imprecise imputation can be embedded into the theory of finite random sets, providing tight lower and upper bounds for probability statements. The results based on a newly developed simulation design – which is customised to the specific requirements for assessing the quality of a statistical matching procedure for categorical data – corroborate that the narrowness of these bounds is practically relevant and that these bounds almost always cover the true parameters.

Keywords: data fusion; data integration; finite random sets; hot deck imputation; (partial) identification;

Acknowlegement: We thank the associate editor for valuable remarks. We also want to express our gratitude to all participants of the research seminar of the *Foundations of Statistics Group* at LMU Munich for valuable discussions on this work. The first author also thanks the LMUMentoring program for financial support.

[*] eva.endres@stat.uni-muenchen.de

[†] paul.fink@stat.uni-muenchen.de

[‡] augustin@stat.uni-muenchen.de

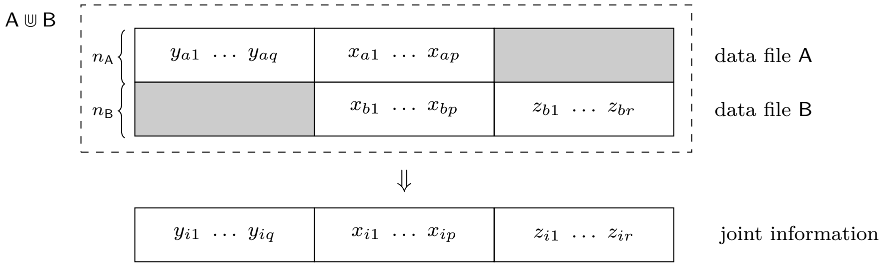


Figure 1: Schematic representation of the statistical matching problem (see D'Orazio et al., 2006b, p. 5 (modified)).

1. Introduction

Nowadays, a tremendous amount of data is readily accessible, as generated by researchers, companies, and governments. Thus, instead of collecting new data to answer research questions, it is a more convenient alternative to use already-available data sources. However, there is often no single data source that includes all information of interest. Statistical matching (also called data integration or data fusion) furnishes a method with which researchers can integrate data collected in different surveys. For example, it was applied by Serafino and Tonkin (2017) to statistically match the data of the *EU Statistics on Income and Living Conditions* and the *Household Budget Survey*.

Assume that we are interested in three blocks of variables, $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$, while there are two data files, A and B, available. Data file A contains n_{A} observations of $(\boldsymbol{X}, \boldsymbol{Y})$, and data file B contains n_{B} observations of $(\boldsymbol{X}, \boldsymbol{Z})$. The observations in B come from the same population but are disjoint from the observations in A. The aim of statistical matching, namely the gain of joint information about variables not jointly observed, is twofold (e.g. D'Orazio et al., 2006b, p. 2):

(i) the estimation of the joint distribution of $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ or any of its characteristics (*macro approach*), and/or

(ii) the creation of a synthetic data file with complete observations on $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ (*micro approach*).

As the schematic representation in Figure 1 suggests, statistical matching can be interpreted as a missing data problem. The observations of the *specific variables* $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are missing in a special block-wise pattern in $\mathsf{A} \uplus \mathsf{B}$, which denotes the union of the two available data files. Following, for example, D'Orazio et al. (2006b, p. 6), the missingness is induced by the given allocation to a certain data file, and thus the missing data mechanism in the framework of statistical matching can convincingly be assumed to be missing completely at random. However, this absence of joint information on all variables results in a severe identification problem: the parameters which concern the relationship between $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are not directly estimable from $\mathsf{A} \uplus \mathsf{B}$. Throughout the paper, we use the term *parameter* to refer to a component of the (joint) probability distribution.

For instance, D'Orazio et al. (2006b) show various ways to remedy the issue of non-identifiability. On the basis of their underlying concepts, these methods can be allocated into three basic groups:
Approaches which

(i) assume the conditional independence of the specific variables given the *common variables* $\boldsymbol{X}$, in order to achieve a factorisation of the joint distribution whose components are estimable on $\mathsf{A} \uplus \mathsf{B}$,

(ii) require auxiliary information in terms of a third file or other external information about parameters concerning the relationship of $\boldsymbol{Y}$ and $\boldsymbol{Z}$,

(iii) refrain from aiming at precise point estimates and account for the uncertainty of the statistical matching problem by estimating a set of plausible parameters, resulting in lower and upper bounds for the parameters concerning the relationship between $\boldsymbol{Y}$ and $\boldsymbol{Z}$. These estimates can be interpreted as set-valued point estimates, not to be confused with confidence regions.

In practice, it is not testable whether the conditional independence assumption holds, and in most applications it might be contested. Manski's *Law of Decreasing Credibility* (Manski, 2007, p. 3), which states that the maintenance of unjustified assumptions reduces the credibility of analyses, makes a very strong argument against the first group of approaches. Auxiliary information, which is the basis of the second group of approaches, is often not available for a certain statistical matching task. Hence, the application of statistical matching taking the underlying uncertainty credibly into account is the means of choice in these situations.

In the context of statistical matching, typically the term *uncertainty* refers to the previously mentioned identification problem. It points to the fact that even if we have complete information on the marginal distributions of $(\boldsymbol{X}, \boldsymbol{Y})$ and $(\boldsymbol{X}, \boldsymbol{Z})$, the joint distribution of $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ cannot uniquely be determined (e.g. D'Orazio et al., 2006a). Thus, lower and upper bounds on the parameters (i.e. probability components) are the best which can be obtained without relying on strong untestable assumptions or external information.

The elaboration of the concept of uncertainty and how to measure it formed the central focus of the papers by Conti et al. (2012) and Conti et al. (2017). Much of the current literature on uncertainty regarding the statistical matching task pays attention to the continuous case, especially to normally distributed variables (e.g. D'Orazio et al., 2006b; Rässler, 2002; Ahfock et al., 2016). However, there is also a relatively small body of literature that is concerned with categorical data. For instance, D'Orazio et al. (2006a), Vantaggi (2008), and Di Zio and Vantaggi (2017) deal with statistical matching of categorical data considering different circumstances.

As emphasised by Conti et al. (2012, p. 70), the "third group of techniques" reflecting the natural uncertainty of statistical matching, does not [usually] "directly aim at reconstructing a complete data set". In the present paper, we introduce imprecise (single) imputation as the first micro approach for categorical data which directly accounts for the natural uncertainty of statistical matching. It is based on the imputation of *sets* of plausible values, which leads to a complete synthetic data file with partially set-valued observations. Furthermore, embedding imprecise imputation into the framework of *finite random sets* will allow us to derive lower and upper bounds for the parameters of the joint distribution. As we will highlight, imprecise imputation can be interpreted as a generalisation of multiple hot deck imputation (e.g. Little and Rubin, 2002) and fractional hot deck imputation (e.g. Kim and Fuller, 2004). The bounds, which we obtain by our imprecise imputation procedure, envelop the results from multiple hot deck imputation and fractional hot deck imputation.

The paper is structured as follows. Section 2 recalls the background of our work by giving a brief overview of the basic setting of statistical matching, its interpretation as a missing data problem, and hot deck imputation in this context. Section 3 describes the idea of imprecise imputation and introduces three imputation procedures. Subsequently, in Section 4, we embed imprecise imputation into the theory of finite disjunctive random sets and show how it can be utilised to estimate lower and upper bounds for the parameters of interest from our imputed data file. After providing the setting and results of a simulation study in Section 5, we conclude with a summary and outlook in Section 6. The appendix contains a more detailed description and justification of the design of the simulation study and graphics on the results of the simulation study.

2. Statistical matching

2.1. The basic setting and its missing data interpretation

Let us assume that we have two data files, A and B, indexed by $\mathcal{I}_{\mathsf{A}}$ and $\mathcal{I}_{\mathsf{B}}$, respectively, with n_{A} and n_{B} disjoint observation units. Without loss of generality, we assume that the index sets are disjoint: $\mathcal{I}_{\mathsf{A}} = \{1, \ldots, n_{\mathsf{A}}\}$ and $\mathcal{I}_{\mathsf{B}} = \{n_{\mathsf{A}} + 1, \ldots, n_{\mathsf{A}} + n_{\mathsf{B}}\}$. Furthermore, let $\boldsymbol{X} = (X_1, \ldots, X_p)$ be the vector of common variables, and $\boldsymbol{Y} = (Y_1, \ldots, Y_q)$ and $\boldsymbol{Z} = (Z_1, \ldots, Z_r)$ be the vectors of specific variables. Denote the domains of the possible values of X_ℓ, $\ell = 1, \ldots, p$, by $\mathcal{X}_\ell$, their corresponding Cartesian product by $\mathcal{X}$, and proceed analogously for the specific variables, defining $\mathcal{Y}_1, \ldots, \mathcal{Y}_q$, $\mathcal{Z}_1, \ldots, \mathcal{Z}_r$, as well as $\mathcal{Y}$ and $\mathcal{Z}$.

As displayed in Figure 1, data file A exclusively contains information on $(\boldsymbol{X},\ \boldsymbol{Y})$ as observations $(\boldsymbol{x}_a, \boldsymbol{y}_a)_{a \in \mathcal{I}_{\mathsf{A}}}$, while data file B comprises information on $(\boldsymbol{X},\ \boldsymbol{Z})$ only, as observations $(\boldsymbol{x}_b, \boldsymbol{z}_b)_{b \in \mathcal{I}_{\mathsf{B}}}$. Consequently, there is no observation that contains simultaneous information on $\boldsymbol{Y}$ and $\boldsymbol{Z}$. In the following, the available information will be consolidated in the incomplete sample $\mathsf{A} \uplus \mathsf{B}$, representing the union of files A and B (see Figure 1) with $n := n_{\mathsf{A}} + n_{\mathsf{B}}$ observations, indexed by $\mathcal{I} = \mathcal{I}_{\mathsf{A}} \cup \mathcal{I}_{\mathsf{B}}$.

Furthermore, we assume that all observations are independently and identically distributed, each following the joint probability distribution $P(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{Z} = \boldsymbol{z})$, where the realisations for a certain observation $i \in \mathcal{I}$ are depicted as $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$, $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iq})$, and $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ir})$. By collecting all probability components of the underlying distribution, we derive the parameter vector consisting of the probability entries of the multidimensional probability table of $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$.

As previously mentioned, statistical matching may be regarded as a missing data problem. Hence, a natural strategy to solve the statistical matching task is imputation, i.e. the substitution of the missing entries with suitable real or artificial values to derive a complete (but partially synthetic) data file. To prepare our method, we focus in the following section on *hot deck imputation*, where the missing entries of an observation (*recipient*) are replaced by records from a similar observation (*donor*) of the same sample. Hot deck imputation ensures that only so-called *live* values, i.e. actually observed and no artificial values, are substituted, and that the marginal and conditional distributions are preserved well for large samples (e.g. Conti et al., 2008). Hot deck imputation methods are frequently used in practice, comparatively easy to apply, and non-parametric (e.g. Andridge and Little, 2010); for a general missing data case, see, for example, Little and Rubin (2002, p. 66).

2.2. Hot deck imputation for statistical matching

In the context of statistical matching, hot deck imputation belongs to the group of non-parametric micro approaches. In the following, we will recall and formalise an example for four variables (X_1, X_2, Y_1, Z_1) from D'Orazio et al. (2006b, Chapter 2.4) and also explain our notation. The data samples A and B are assigned to the roles of *recipient file* and *donor file*. Since it is a symmetric problem, D'Orazio et al. (2006b) only describe the case where A is the recipient file and B the donor file. The reverse case works analogously. The choice of whether only A, only B, or $\mathsf{A} \uplus \mathsf{B}$ should be imputed depends on many factors. In this paper, we impute $\mathsf{A} \uplus \mathsf{B}$ without loss of generality. See, for instance, D'Orazio et al. (2006b, pp. 35–36) for a discussion on this issue.

Random hot deck imputation means that for each missing entry in the recipient file, a donor record from the donor file is randomly chosen by simple random sampling and its corresponding values are used to replace the missing entries in the recipient file. Every missing entry of the specific variable Z_1 in the recipient file A, i.e. z_{a1}, $a \in \mathcal{I}_\mathsf{A}$, is replaced by the synthetic value $\tilde{z}_{a1} := z_{b1}$, $b \in \mathcal{I}_\mathsf{B}$, where b is the randomly chosen observation unit from the index set $\mathcal{I}_\mathsf{B}$ of data file B and, hence, $\tilde{z}_{a1} \in \{z_{b1} : b \in \mathcal{I}_\mathsf{B}\}$. The a-th observation of complete, synthetic data file A is composed of $(x_{a1}, x_{a2}, y_{a1}, \tilde{z}_{a1})$, where the tilde marks the imputed and thus synthetic value.

However, simple random sampling gives all observation units in the donor file the same probability of being selected. Thus, it implicitly induces the independence of both the common and specific variables.

A more promising procedure is the assignment of donor and recipient records within groups of similar (homogeneous) records that are created by exploiting the information of the common variables. The realisations of selected categorical common variables are used to generate groups of similar records in both the recipient file and the donor file. Little and Rubin (2002) call these groups *adjustment cells.* Following D'Orazio et al. (2006b), we will call them *donation classes.* The choice of the common variables that are actually used to perform statistical matching (the so-called *matching variables*) is of high impact on the resulting matching quality. It is desirable that the common variables are highly correlated with, or good predictors for the specific variables (Rässler, 2002, p. 10). See, for instance, D'Orazio et al. (2017) on how to choose the *matching variables.*

Consider again data file A as the recipient. The first step of hot deck imputation within homogenous groups is the assignment of all observations in $\mathsf{A} \uplus \mathsf{B}$ to donation classes. For this purpose, we partition the index set $\mathcal{I}$ into $D \leq |\mathcal{X}|$ index sets $\mathcal{I}^d$, $d = 1, \ldots, D$, such that for any d, all observation units in $\mathcal{I}^d$ have the same realisations of $\boldsymbol{X}$. Moreover, define $\mathcal{I}^d_\mathsf{A} := \mathcal{I}^d \cap \mathcal{I}_\mathsf{A}$ and $\mathcal{I}^d_\mathsf{B} := \mathcal{I}^d \cap \mathcal{I}_\mathsf{B}$. Every missing entry for the specific variable Z_1 of an observation unit from A in the d-th donation class, i.e. z_{a1}, $a \in \mathcal{I}^d_\mathsf{A}$, is replaced by $\tilde{z}_{a1} := z_{b1}$, $b \in \mathcal{I}^d_\mathsf{B}$, which is the corresponding value of a randomly chosen observation from the donation class $\mathcal{I}^d_\mathsf{B}$, and hence $\tilde{z}_{a1} \in \{z_{b1} : b \in \mathcal{I}^d_\mathsf{B}\}$ for all $a \in \mathcal{I}^d_\mathsf{A}$.

Using donation classes, the imputation of $\boldsymbol{Z}$ is conditional on $\boldsymbol{X}$, thus reproducing the empirical conditional distribution of $\boldsymbol{Z}$ given $\boldsymbol{X}$ in A. Since there are no joint observations of all variables, additionally conditioning on $\boldsymbol{Y}$ is not possible. Thus, a conditional independence – between the imputed values of $\boldsymbol{Z}$ and the values of $\boldsymbol{Y}$, given $\boldsymbol{X}$ – is implicitly (empirically) established in the synthetic parts of the resulting complete file (see Rässler, 2002, pp. 200–204).

Every complete synthetic data file consisting of observations $(\boldsymbol{x}_a, \boldsymbol{y}_a, \tilde{\boldsymbol{z}}_a)_{a \in \mathcal{I}_\mathsf{A}}$ and $(\boldsymbol{x}_b, \tilde{\boldsymbol{y}}_b, \boldsymbol{z}_b)_{b \in \mathcal{I}_\mathsf{B}}$ straightforwardly delivers estimates of the underlying joint distribution by evaluating the observed relative frequencies. Written in a form preparing for the generalisation developed in Section 4.3, we obtain for an event $\mathcal{E} = \mathcal{E}_\mathcal{X} \times \mathcal{E}_\mathcal{Y} \times \mathcal{E}_\mathcal{Z}$

with $\mathcal{E}_{\mathcal{X}} \subseteq \mathcal{X}$, $\mathcal{E}_{\mathcal{Y}} \subseteq \mathcal{Y}$ and $\mathcal{E}_{\mathcal{Z}} \subseteq \mathcal{Z}$,

$$\begin{aligned}\widehat{P}(\mathcal{E}) := \widehat{P}\big((\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}) \in \mathcal{E}\big) &= \frac{1}{n}\Big|\big\{a \in \mathcal{I}_{\mathsf{A}} : (\boldsymbol{x}_a, \boldsymbol{y}_a, \tilde{\boldsymbol{z}}_a) \in \mathcal{E}\big\} \cup \big\{b \in \mathcal{I}_{\mathsf{B}} : (\boldsymbol{x}_b, \tilde{\boldsymbol{y}}_b, \boldsymbol{z}_b) \in \mathcal{E}\big\}\Big| \\ &= \frac{1}{n}\Big|\big\{a \in \mathcal{I}_{\mathsf{A}} : \boldsymbol{x}_a \in \mathcal{E}_{\mathcal{X}}, \boldsymbol{y}_a \in \mathcal{E}_{\mathcal{Y}}, \tilde{\boldsymbol{z}}_a \in \mathcal{E}_{\mathcal{Z}}\big\}\Big| \\ &\quad + \frac{1}{n}\Big|\big\{b \in \mathcal{I}_{\mathsf{B}} : \boldsymbol{x}_b \in \mathcal{E}_{\mathcal{X}}, \tilde{\boldsymbol{y}}_b \in \mathcal{E}_{\mathcal{Y}}, \boldsymbol{z}_b \in \mathcal{E}_{\mathcal{Z}}\big\}\Big| \,. \end{aligned} \tag{1}$$

Any event which is not directly representable as a Cartesian product can be decomposed into the union of disjoint events of the previous form.

In the context of missing data, it is a well-known problem that single imputations are not able to reflect the uncertainty which arises from the missingness of joint information on $\boldsymbol{Y}$ and $\boldsymbol{Z}$. Therefore, it is commonly recommended to apply *multiple imputation* techniques (e.g. Little and Rubin, 2002, chap. 5.4), where the replacement of missing entries is performed several times. The obtained complete data files are then analysed by common methods for complete data and the results are subsequently pooled to achieve point estimates. Such multiple imputation techniques have been further developed by Rässler (2002, chap. 4) for application in statistical matching with the intention to estimate lower and upper bounds for the parameters of interest in the spirit of Manski (1995). However, Rässler (2002) only considers normally distributed data and, as stated in Ahfock et al. (2016, p. 82), by applying multiple imputation "there is no guarantee that the range of imputed datasets fully captures the uncertainty over the partially identified parameters".

3. Imprecise imputation

3.1. Basic idea and terminology

Based on these considerations, we will now develop the concept of imprecise imputation, where we suggest imputing a *set* of plausible values for a missing entry. This leads to precise observations $(\boldsymbol{x}_a,\ \boldsymbol{y}_a)_{a\in\mathcal{I}_{\mathsf{A}}}$ in A and $(\boldsymbol{x}_b,\ \boldsymbol{z}_b)_{b\in\mathcal{I}_{\mathsf{B}}}$ in B, and to *imprecise*, i.e. set-valued, synthetic observations $(\tilde{\mathfrak{z}}_a)_{a\in\mathcal{I}_{\mathsf{A}}}$ in A and $(\tilde{\mathfrak{y}}_b)_{b\in\mathcal{I}_{\mathsf{B}}}$ in B. Please note that our aim is *not* to identify a single element of these imprecise observations for the purpose of precise single imputation but rather to regard the whole set as the final piece of indivisible information. In Section 4.3 we show how the set-valued imprecise observations can be directly used to obtain estimates for the probability components of the joint distribution.

The following subsections detail and illustrate imprecise imputation. Three different ways of determining the sets of plausible values to be imputed are introduced, each taking into account the variations in how strong and trustworthy the underlying relationship between the common and specific variables is. Without loss of generality, again let A be the recipient and B the donor file, and let the donor classes be defined as in Section 2.2.

- **D** *Domain imputation* replaces every missing entry $z_{a\ell}$, $a \in \mathcal{I}_{\mathsf{A}}$, of a variable Z_ℓ, $\ell = 1, \ldots, r$, with its domain, i.e.

$$\tilde{\mathfrak{z}}_{a\ell} := \mathcal{Z}_\ell\,, \qquad \forall a \in \mathcal{I}_{\mathsf{A}}\,,\ \ell = 1, \ldots, r\,. \tag{2}$$

- **VW** *Variable-wise imputation* on the basis of donation classes replaces every missing entry $z_{a\ell}$, $a \in \mathcal{I}_{\mathsf{A}}^d$, of a variable Z_ℓ, $\ell = 1, \dots, r$, with the set of live values of Z_ℓ within the corresponding class $\mathcal{I}_{\mathsf{B}}^d$. Thus,

$$\tilde{\mathfrak{z}}_{a\ell} := \left\{ z_{b\ell} : b \in \mathcal{I}_{\mathsf{B}}^d \right\}, \qquad \forall a \in \mathcal{I}_{\mathsf{A}}^d\,,\ d = 1, \dots, D,\ \ell = 1, \dots, r\,. \tag{3}$$

- **CW** *Case-wise imputation*, i.e. the simultaneous imputation of all missing entries of an observation a in $\mathcal{I}_{\mathsf{A}}^d$, where every tuple $\boldsymbol{z}_a = (z_{a1}, \dots, z_{ar})$, $a \in \mathcal{I}_{\mathsf{A}}^d$ is replaced with the set of live tuples in the corresponding class $\mathcal{I}_{\mathsf{B}}^d$. Consequently,

$$\tilde{\boldsymbol{\mathfrak{z}}}_a := \left\{ (z_{b1}, \dots, z_{br}) : b \in \mathcal{I}_{\mathsf{B}}^d \right\}, \qquad \forall a \in \mathcal{I}_{\mathsf{A}}^d\,,\ d = 1, \dots, D\,. \tag{4}$$

3.2. Illustration and discussion of the different types of imprecise imputation

3.2.1. Domain imputation

The most conservative way to determine the set of plausible values which are candidate values for the substitution of a missing entry is to use the whole domain of the corresponding variable. Concretely, this means that every missing entry $z_{a\ell}$, $a \in \mathcal{I}_{\mathsf{A}}$, $\ell = 1, \dots, r$ is substituted by the set of all possible realisations of Z_ℓ, i.e. its domain $\mathcal{Z}_\ell$. Hence, $\tilde{\mathfrak{z}}_{a\ell} := \mathcal{Z}_\ell$, $\forall a \in \mathcal{I}_{\mathsf{A}}$ becomes a set-valued entry in data file A, where all elements of the set are treated as equally plausible, but without a further reduction in the complexity by some (arbitrary) weighting or aggregation of the elements. The imputed sets for one variable are equal for all observations. This procedure is briefly illustrated in the following running toy example.

Minimal Example 1 *Consider two data files,* A *and* B*, which consist of* $n_{\mathsf{A}} = 2$ *observations of* (Y_1, Y_2, X_1, X_2) *and* $n_{\mathsf{B}} = 3$ *observations of* (X_1, X_2, Z_1, Z_2)*, respectively. The corresponding domains of the variables are* $\mathcal{X}_1 = \mathcal{X}_2 = \mathcal{Y}_1 = \mathcal{Z}_1 = \{0, 1\}$ *and* $\mathcal{Y}_2 = \mathcal{Z}_2 = \{0, 1, 2\}$*. Domain imputation results in the following completed data file.*

Y_1	Y_2	X_1	X_2	Z_1	Z_2
1	**2**	**1**	**0**	$\{0;1\}$	$\{0;1;2\}$
0	**2**	**0**	**0**	$\{0;1\}$	$\{0;1;2\}$
$\{0;1\}$	$\{0;1;2\}$	**1**	**0**	**0**	**0**
$\{0;1\}$	$\{0;1;2\}$	**1**	**0**	**1**	**1**
$\{0;1\}$	$\{0;1;2\}$	**0**	**0**	**1**	**2**

Numbers in bold represent the original data. The files A *and* B *are visually divided by the dashed line. The numbers in curly brackets depict the sets of possible realisations of the corresponding variables, i.e. the domains, which are here the replacements for the previously missing entries.*

This imputation procedure resembles the approach of Ramoni and Sebastiani (2001), who use an incomplete sample to estimate bounds for the parameters of conditional probability distributions in the context of Bayesian networks.

Applying domain imputation, it is guaranteed that the true (but missing) value is always an element of the imputed set. As previously mentioned, domain imputation is very conservative, and thus it can also be applied if the common variables are not good predictors for the specific variables. However, it neglects any available dependence structure between the common and specific variables in the available data. In the following, we will introduce two other methods to determine the set of values for imputation which both take these dependencies into account, albeit to a different extent.

3.2.2. Variable-wise imputation

If $q \geq 2$ or $r \geq 2$, with due regard to the association between the common and specific variables, imputation can be performed on two different levels, either by treating each of the specific variables separately or by treating the specific variables within each of the two blocks simultaneously (see, e.g. Joenssen, 2015, chap. 3, for precise imputation). In this section, we describe imprecise imputation on the separate level, while the simultaneous level will be addressed in the next section.

The imputation of live values only within donation classes ensures that associations between the common and specific variables are incorporated. As a consequence, the preservation of the dependence structure is improved and the estimated bounds for the parameters of interests become more narrow.

Without loss of generality, again let A be the recipient file and B the donor file. All observations $i \in \mathcal{I}_{\mathsf{A}} \cup \mathcal{I}_{\mathsf{B}}$ are allocated into donation classes depending on their realisations of the matching variables selected from the common variables $\boldsymbol{X}$, following the notation as introduced in Section 2.2. For every observation $a \in \mathcal{I}_{\mathsf{A}}^{d}$, the missing entry $z_{a\ell}$ of the variable Z_ℓ, $\ell = 1, \dots, r$ is substituted by the set of all live values of this variable from the same donation class in the donor file B, resulting in Equation (3).

Minimal Example 2 *Consider the same data situation as in Example 1. Now we will illustrate the application of the just-described variable-wise imputation. The different backgrounds display the different donation classes based on the combinations of the realisations of X_1 and X_2. Both common variables are used as matching variables in this example.*

Y_1	Y_2	X_1	X_2	Z_1	Z_2
1	**2**	**1**	**0**	$\{0;1\}$	$\{0;1\}$
0	**2**	**0**	**0**	$\{1\}$	$\{2\}$
$\{1\}$	$\{2\}$	**1**	**0**	**0**	**0**
$\{1\}$	$\{2\}$	**1**	**0**	**1**	**1**
$\{0\}$	$\{2\}$	**0**	**0**	**1**	**2**

This procedure preserves the dependencies between the common and the specific variables; however, the successive imputation of single variables breaks the dependence structure among the specific variables. Little and Rubin (see 2002, p. 72), for instance, have already stated that imputation should be multivariate to preserve the dependencies between the variables. If one attaches high value to this requirement, the imputation should be performed simultaneously for all variables in the data file as described in the following section. Nevertheless, variable-wise imputation is a good compromise between the very conservative domain imputation and the more data-driven case-wise imputation procedure detailed in the following section.

3.2.3. Case-wise imputation

For case-wise imputation, we interpret the missing entries of one observation $a \in \mathcal{I}_{\mathsf{A}}^d$ out of the d-th donation class in the recipient file as tuple of the form $(z_{a1}, \dots, z_{ar})$. This tuple of missing entries is replaced by the set of tuples $\tilde{\mathfrak{z}}_a$, which have been observed in the donor file B and the same donation class d, as in Equation (4). This strategy ensures that also the dependencies among the specific variables $\boldsymbol{Z}$ remain unchanged. The following example illustrates this imputation procedure.

Minimal Example 3 *Consider again the situation of Example 1 as a starting point. Interpret the empty cells z_{a1} and z_{a2} as tuples (z_{a1}, z_{a2}), $a = 1, 2$, and analogously y_{b1} and y_{b2} as tuples (y_{b1}, y_{b2}), $b = 3, 4, 5$. The result of case-wise imputation in this example is displayed in the following.*

(Y_1, Y_2)	X_1	X_2	(Z_1, Z_2)
$(\mathbf{1}, \mathbf{2})$	$\mathbf{1}$	$\mathbf{0}$	$\{(0,0); (1,1)\}$
$(\mathbf{0}, \mathbf{2})$	$\mathbf{0}$	$\mathbf{0}$	$\{(1,2)\}$
$\{(1,2)\}$	$\mathbf{1}$	$\mathbf{0}$	$(\mathbf{0}, \mathbf{0})$
$\{(1,2)\}$	$\mathbf{1}$	$\mathbf{0}$	$(\mathbf{1}, \mathbf{1})$
$\{(0,2)\}$	$\mathbf{0}$	$\mathbf{0}$	$(\mathbf{1}, \mathbf{2})$

3.2.4. General remarks

A potential issue arises if at least one donation class in the donor file is empty. If so, variable-wise and case-wise imputation cannot directly be applied and we then recommend to impute the domains $\mathcal{Z}_1, \dots, \mathcal{Z}_r$ or the Cartesian product of the domains $\mathcal{Z}$.

The partially set-valued data files produced by imprecise imputation can be interpreted as a set of underlying precise data files. On closer inspection, the sets produced by the three imputation procedures are nested: the largest set of underlying precise data files is obtained by domain imputation, while case-wise imputation yields the smallest set. Equation (15) shows this relationship formally.

Fractional hot deck imputation (see, e.g. Kim and Fuller, 2004), which is also an imputation approach that is based on set-valued imputations, produces precise results that are contained in the sets obtained by imprecise imputation. It uses a weighting scheme, which is transferred onto the set of values to impute. This strategy reduces complexity by circumventing the direct handling of the imputed set-valued observation by creating a single completed data file with accordingly down-weighted precise pseudo-observations. This kind of precise data allows the direct use of common statistical models and methods. The variability, introduced by having multiple values to be imputed, is, in the situation of fractional hot deck imputation, accounted for in the variance estimation of the precise estimator. However, variance estimation in the context of fractional hot deck imputation may be argued to be more complex yet more reliable in comparison to multiple imputation (see, e.g. Yang and Kim, 2016).

During the imprecise imputation process, variable-wise and, in particular, domain imputation may create combinations of variable realisations which are contextually unjustified. For instance, D'Orazio et al. (2006b), distinguishes between two types of so-called *logical constraints* to exclude impossible or unlikely combinations in the synthetic categorical data:

(i) *existence of some quantities* on the basis of the individual observation unit, and

(ii) *inequality constraints* on the level of the estimated probability distributions.

Especially the first case can easily be incorporated into the imputation step. Single, implausible values or tuples of values containing the unjustified combinations can easily be removed from the synthetic file. As an extension to both types of constraints, the set of values to be imputed can be restricted further removing not only contextually impossible values but also combinations of values that showed to be very rare within the data file or the population, motivated by the approach of Cattaneo (2013), developed in a decision-theoretic context. This means that the set of (variable-wise or case-wise) live values is restricted to the set of all values whose relative frequencies exceed a certain threshold δ, which may be dependent on the donation class. Increasing δ would gradually eclipse our conservative perspective, resulting, in the extreme case, in a precise single-valued imputation.

We propose to build upon the set-valued data directly, without reducing their complexity via a weighting scheme. In contrast to widely adopted imputation procedures yielding single-valued data, we are now in the situation of statistical analysis of partially set-valued data. To frame imprecise imputation formally, it will be embedded into the concept of finite disjunctive random sets, which allows the estimation of tight lower and upper bounds for the parameters.

In order to allow for a concise description in the following sections, we will take the observation-wise perspective on the imputed sets (i.e. the notation in terms of tuples), which corresponds to the perspective taken by the case-wise imputation. The imputation results of the other procedures can be transferred by taking the Cartesian product, e.g. $\tilde{\boldsymbol{\mathfrak{z}}}_a = \tilde{\mathfrak{z}}_{a1} \times \ldots \times \tilde{\mathfrak{z}}_{ar}$.

4. Imprecision imputation and finite disjunctive random sets

Imprecise imputation provides us with partially set-valued data. To prepare a well-founded statistical analysis, we have to formalise imprecise imputation probabilistically. For this purpose, the direct formalisation of $\boldsymbol{X}, \boldsymbol{Y}$, and $\boldsymbol{Z}$ as collections of random variables and corresponding realisations is no longer sufficient. Starting from an applied point of view, two types of generalisations, which will indeed prove compatible among each other, could be imagined. Firstly, we could abstractly look for a concept of set-valued variables with corresponding set-valued realisations. Secondly, we could assume that every set represents outcomes of various random variables, one of which is the true underlying, yet not precisely observable, random variable. (Throughout this paper we use the term *random variable* to refer to a mapping to the real numbers as well as to some non-numerical finite space. In context of the latter, the term *random element* is sometimes used for the sake of distinction (e.g. Nguyen, 2006).)

In this section it will be shown how set-valued observations, and thus the resulting data files of the three imprecise imputation procedures in particular, are covered by the concept of *disjunctive random sets*, also known as *ill-perceived random variables* (Couso et al., 2014; Nguyen, 2006). This embedding allows for the assessment of probability statements and the construction of corresponding estimates from the partially set-valued synthetic file derived from imprecise imputation. The interpretation of the set-valued quantities as disjunctive random sets corresponds to the view of Dempster (1967), on which the so-called Dempster-Shafer theory of belief functions (Shafer, 1976) is built, which has become very popular in artificial intelligence (see, e.g. Denœux, 2016).

4.1. Random set formulation of imprecise imputation

The true random variables $\boldsymbol{X}, \boldsymbol{Y}$, and $\boldsymbol{Z}$ map from the underlying population space, denoted by Ω in the sequel, into the domains $\mathcal{X}, \mathcal{Y}$ and $\mathcal{Z}$, yielding realisations $\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{z}_i$ with $i \in \mathcal{I}$, respectively. Now, neither $\boldsymbol{y}_b$ nor $\boldsymbol{z}_a$ are available, but are replaced by synthetic observations $\tilde{\mathfrak{y}}_b$ and $\tilde{\mathfrak{z}}_a$, respectively, according to either Equation (2), (3), or (4), depending on the chosen imprecise imputation procedure. To formalise this situation, we follow the common practice in statistical matching, treating $\mathcal{I}_{\mathsf{A}}$ and $\mathcal{I}_{\mathsf{B}}$ as fixed. This allows us to globally replace $\boldsymbol{Y}$ and $\boldsymbol{Z}$ by the set-valued variables $\mathfrak{Y}$ and $\mathfrak{Z}$ (with realisations $\mathfrak{y}_i$ and $\mathfrak{z}_i$, $i \in \mathcal{I}$). The imputed values are already sets, so they fit in nicely, but in order to deal with the already observed realisations, we regard them now as singletons containing only the observed value, e.g. $\mathfrak{z}_{b\ell} = \{z_{b\ell}\}$, $\forall b \in \mathcal{I}_{\mathsf{B}}, \ell = 1, \ldots, r$. The variables $\mathfrak{Y}$ and $\mathfrak{Z}$ map into the corresponding power sets $2^{\mathcal{Y}}$ and $2^{\mathcal{Z}}$, whereby mapping into the empty set is excluded.

If we collect the random variables of interest in a variable $\boldsymbol{\Gamma}$ and define $\mathcal{W} := \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$, then

$$\boldsymbol{\Gamma} := (\boldsymbol{X}, \mathfrak{Y}, \mathfrak{Z}) \colon \; \Omega \;\longrightarrow\; 2^{\mathcal{W}} \setminus \{\emptyset\} \tag{5}$$

is a finite non-empty random set (see Definition 3.1 in Nguyen, 2006, p. 35), satisfying the required measurability condition by equipping $2^{\mathcal{W}} \backslash \{\emptyset\}$ with its power set. Since in our setting the imputed (synthetic) set-valued entries of the specific variables are understood as the collection of possible underlying true values, this random set has to be interpreted in the so-called disjunctive way (see, e.g. Couso et al., 2014; Couso and Dubois, 2014).

In general, any disjunctive random set $\boldsymbol{\Gamma}$ induces an upper inverse $\boldsymbol{\Gamma}^*$ and a lower inverse $\boldsymbol{\Gamma}_*$. When considering an event of interest $\mathcal{E} \subseteq \mathcal{W}$, which is now a singleton in the considered space $2^{\mathcal{W}}$, the upper inverse contains all the elements of the population whose image overlaps with $\mathcal{E}$, while the lower inverse contains only those elements of the population whose (non-empty) image is entirely contained within $\mathcal{E}$:

$$\boldsymbol{\Gamma}^*(\mathcal{E}) := \Big\{ \omega \in \Omega : \boldsymbol{\Gamma}(\omega) \cap \mathcal{E} \neq \emptyset \Big\} \tag{6}$$

and

$$\boldsymbol{\Gamma}_*(\mathcal{E}) := \Big\{ \omega \in \Omega : \boldsymbol{\Gamma}(\omega) \subseteq \mathcal{E} \Big\} \; . \tag{7}$$

In a heuristic formulation, the upper inverse considers all aspects that do not entirely contradict $\mathcal{E}$, while the lower inverse collects all aspects that necessarily imply $\mathcal{E}$. By using the probability measure $\mathbb{P}$ defined on the original probability space involving Ω, the upper and lower probabilities are then defined in terms of the upper and lower inverse, respectively:

$$P^*(\mathcal{E}) = \mathbb{P}\big(\boldsymbol{\Gamma}^*(\mathcal{E})\big) \quad \text{and} \quad P_*(\mathcal{E}) = \mathbb{P}\big(\boldsymbol{\Gamma}_*(\mathcal{E})\big) \quad \forall \mathcal{E} \subseteq \mathcal{W} \; . \tag{8}$$

In order to improve readability we have not marked the image probability measure induced by the random set $\boldsymbol{\Gamma}$, i.e. $P_{\boldsymbol{\Gamma}} = P$, and we proceed analogously with the corresponding set functions P^* and P_*. If we refer to a different image measure, the according inducing random quantity will be set as subscript to P. If we turn to the view of an underlying, ill-perceived random variable $\boldsymbol{W}_0 : \Omega \longrightarrow \mathcal{W}$, only knowing that the unobserved true value $\boldsymbol{W}_0(\omega)$ lies (with probability one) within the observed set $\boldsymbol{\Gamma}(\omega)$, it can be shown (see, e.g. Couso et al., 2014) that for every event $\mathcal{E} \subseteq \mathcal{W}$ the upper and lower probabilities induced by the random set enclose the probability of $\boldsymbol{W}_0$:

$$P_*(\mathcal{E}) \leq P_{\boldsymbol{W}_0}(\mathcal{E}) \leq P^*(\mathcal{E}) \quad \forall \mathcal{E} \subseteq \mathcal{W} \, .$$

This leads to another way to interpret a random set, namely as producing a family of compatible, precise probability measures $\mathcal{P}(\boldsymbol{\Gamma})$, which is a subset of the set $\mathcal{P}$ of all probability measures on $(2^{\mathcal{W}}, 2^{2^{\mathcal{W}}})$. Nguyen (1978) showed that if $\mathcal{W}$ is finite, the probability distribution induced by $\boldsymbol{\Gamma}$ corresponds to the so-called basic probability assignment in Dempster-Shafer theory and thus makes the belief function mathematically equivalent to P_*. Consequently, the technical results from that area may be used as well.

In the present special case of finite $\mathcal{W}$, the set $\mathcal{P}(\boldsymbol{\Gamma})$ coincides with the credal set $\mathcal{M}(P^*)$, i.e. those precise probability measures that respect the upper and lower bounds defined by P^* and P_* event-wise (see Miranda et al., 2010), which also embeds the situation considered here into the framework of imprecise probabilities (e.g. Walley, 1991; Augustin et al., 2014).

In particular, P_* and P^* are lower and upper probabilities that are envelopes of all probability measures P in $\mathcal{M}(P^*)$:

$$P_*(\mathcal{E}) = \inf_{P \in \mathcal{M}(P^*)} P(\mathcal{E}) \quad \text{and} \quad P^*(\mathcal{E}) = \sup_{P \in \mathcal{M}(P^*)} P(\mathcal{E}) \, .$$

Indeed, P^*, P_* and $\mathcal{M}(P^*)$ are three mathematically equivalent formulations that can be transferred into each other. Therefore, from an applied point of view, each of them can be seen as the core result of a probabilistic description of imprecise imputation. For any possibly true probability distribution $P_{\boldsymbol{W}_0}$, our embedding into random sets provides us with a set $\mathcal{M}(P^*)$ of distributions induced by $P_{\boldsymbol{W}_0}$ such that $\mathcal{M}(P^*)$ contains $P_{\boldsymbol{W}_0}$. By construction, this is the smallest set that is deducible from the concrete imputation procedure without adding further assumptions or knowledge. Dually, $P^*(\mathcal{E})$ and $P_*(\mathcal{E})$ are the narrowest bounds, deducible on the probabilities of an event $\mathcal{E}$.

4.2. Conditioning disjunctive random sets

The representation via the set $\mathcal{M}(P^*)$ of compatible probability distributions including the embedding into the framework of imprecise probabilities guides the further probabilistic analysis of the partially set-valued data file achieved by imprecise imputation. For instance, if the elements of $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ get eventually associated with real-valued outcomes, then a generalised expectation is logically defined via the infimum and supremum of all compatible traditional expectations based on image measures of elements of $\mathcal{M}(P^*)$.

A similar procedure suggests itself for conditioning, namely an element-wise application of conditioning for all $P \in \mathcal{M}(P^*)$, provided $P(\mathcal{C}) > 0$ for an conditioning event $\mathcal{C}$ (see, e.g. Dubois and Prade (1992) or Fagin and Halpern (1991) for a discussion and a comparison to an alternative). It can be shown (e.g. de Campos et al. (1990), Couso et al. (2014), and Fagin and Halpern (1991)) that this leads to the following closed-form results for the upper conditional probability

$$P^*(\mathcal{S}|\mathcal{C}) = \sup_{P\in\mathcal{M}(P^*)} P(\mathcal{S}|\mathcal{C}) = \frac{P^*(\mathcal{S}\cap\mathcal{C})}{P^*(\mathcal{S}\cap\mathcal{C}) + P_*(\bar{\mathcal{S}}\cap\mathcal{C})} \tag{9}$$

and the lower conditional probability

$$P_*(\mathcal{S}|\mathcal{C}) = \inf_{P\in\mathcal{M}(P^*)} P(\mathcal{S}|\mathcal{C}) = \frac{P_*(\mathcal{S}\cap\mathcal{C})}{P_*(\mathcal{S}\cap\mathcal{C}) + P^*(\bar{\mathcal{S}}\cap\mathcal{C})}\ , \tag{10}$$

where $\bar{\mathcal{S}}$ denotes the complement of $\mathcal{S}$.

4.3. Parameter estimation by means of disjunctive random sets based on imprecise imputation

So far, this approach has been described in a probabilistic setting, where every entity involved is known (besides the true hidden/ill-perceived random variable). In the following, the statistical perspective will be taken in which the probabilities corresponding to the random set need to be estimated from a finite sample. Consequently, we take our synthetic data file derived from imprecise imputation as consisting of $n = n_{\mathsf{A}} + n_{\mathsf{B}}$ realisations $\boldsymbol{\gamma}_i$, $i\in\mathcal{I}$, of the corresponding generic random set $\boldsymbol{\Gamma}$ from Equation (5). Referring to Equation (8), with Equations (6) and (7), we obtain, in generalisation of Equation (1), for our event $\mathcal{E} = \mathcal{E}_{\mathcal{X}}\times\mathcal{E}_{\mathcal{Y}}\times\mathcal{E}_{\mathcal{Z}}$:

$$\begin{aligned}
\widehat{P^*}(\mathcal{E}) &= \frac{1}{n}\Big|\big\{i\in\mathcal{I}:\boldsymbol{\gamma}_i\cap\mathcal{E}\neq\emptyset\big\}\Big| \\
&= \frac{1}{n}\bigg(\Big|\big\{a\in\mathcal{I}_{\mathsf{A}}:(\boldsymbol{x}_a,\boldsymbol{y}_a,\tilde{\boldsymbol{\mathfrak{z}}}_a)\cap\mathcal{E}\neq\emptyset\big\}\Big| + \Big|\big\{b\in\mathcal{I}_{\mathsf{B}}:(\boldsymbol{x}_b,\tilde{\boldsymbol{\mathfrak{y}}}_b,\boldsymbol{z}_b)\cap\mathcal{E}\neq\emptyset\big\}\Big|\bigg) \\
&= \frac{1}{n}\Big|\big\{a\in\mathcal{I}_{\mathsf{A}}:\boldsymbol{x}_a\in\mathcal{E}_{\mathcal{X}},\boldsymbol{y}_a\in\mathcal{E}_{\mathcal{Y}},\tilde{\boldsymbol{\mathfrak{z}}}_a\cap\mathcal{E}_{\mathcal{Z}}\neq\emptyset\big\}\Big| \\
&\quad+\frac{1}{n}\Big|\big\{b\in\mathcal{I}_{\mathsf{B}}:\boldsymbol{x}_b\in\mathcal{E}_{\mathcal{X}},\tilde{\boldsymbol{\mathfrak{y}}}_b\cap\mathcal{E}_{\mathcal{Y}}\neq\emptyset,\boldsymbol{z}_b\in\mathcal{E}_{\mathcal{Z}}\big\}\Big|
\end{aligned} \tag{11}$$

and

$$\begin{aligned}
\widehat{P_*}(\mathcal{E}) &= \frac{1}{n}\Big|\big\{i\in\mathcal{I}:\boldsymbol{\gamma}_i\subseteq\mathcal{E},\boldsymbol{\gamma}_i\neq\emptyset\big\}\Big| \\
&= \frac{1}{n}\bigg(\Big|\big\{a\in\mathcal{I}_{\mathsf{A}}:(\boldsymbol{x}_a,\boldsymbol{y}_a,\tilde{\boldsymbol{\mathfrak{z}}}_a)\subseteq\mathcal{E}\big\}\Big| + \Big|\big\{b\in\mathcal{I}_{\mathsf{B}}:(\boldsymbol{x}_b,\tilde{\boldsymbol{\mathfrak{y}}}_b,\boldsymbol{z}_b)\subseteq\mathcal{E}\big\}\Big|\bigg) \\
&= \frac{1}{n}\Big|\big\{a\in\mathcal{I}_{\mathsf{A}}:\boldsymbol{x}_a\in\mathcal{E}_{\mathcal{X}},\boldsymbol{y}_a\in\mathcal{E}_{\mathcal{Y}},\tilde{\boldsymbol{\mathfrak{z}}}_a\subseteq\mathcal{E}_{\mathcal{Z}}\big\}\Big| \\
&\quad+\frac{1}{n}\Big|\big\{b\in\mathcal{I}_{\mathsf{B}}:\boldsymbol{x}_b\in\mathcal{E}_{\mathcal{X}},\tilde{\boldsymbol{\mathfrak{y}}}_b\subseteq\mathcal{E}_{\mathcal{Y}},\boldsymbol{z}_b\in\mathcal{E}_{\mathcal{Z}}\big\}\Big|.
\end{aligned} \tag{12}$$

From $\widehat{P^*}(\mathcal{E})$ and $\widehat{P_*}(\mathcal{E})$ also an estimate of the induced underlying set of probability measures can be derived:

$$\widehat{\mathcal{M}}(P^*) = \big\{P\in\mathcal{P}:\widehat{P_*}(\mathcal{E})\leq P(\mathcal{E})\leq\widehat{P^*}(\mathcal{E})\,,\ \forall\mathcal{E}\subseteq\mathcal{W}\big\}\,. \tag{13}$$

For comparing the estimates resulting from the different types of imputation procedures, it is essential to recall that the different set-valued data files are by construction nested with respect to all compatible underlying precise data files. The set resulting from domain imputation is a (non-strict) superset of the set obtained from variable-wise imprecise imputation, which contains the set produced by case-wise imprecise imputation. Therefore, with the abbreviations introduced in Section 3.1, it holds that

$$\widehat{\mathcal{M}}\Big(P^{*^{CW}}\Big) \subseteq \widehat{\mathcal{M}}\Big(P^{*^{VW}}\Big) \subseteq \widehat{\mathcal{M}}\Big(P^{*^{D}}\Big) \tag{14}$$

and, for every event $\mathcal{E} \subseteq \mathcal{W}$,

$$\widehat{P_*}^{D}(\mathcal{E}) \leq \widehat{P_*}^{VW}(\mathcal{E}) \leq \widehat{P_*}^{CW}(\mathcal{E}) \leq \widehat{P^*}^{CW}(\mathcal{E}) \leq \widehat{P^*}^{VW}(\mathcal{E}) \leq \widehat{P^*}^{D}(\mathcal{E})\,. \tag{15}$$

This allows us to compare the results obtained by the different imputation approaches to the result under conditional independence, which yields a single precise probability distribution. It can be argued that the probability distribution under conditional independence is contained in any of the estimated sets. Furthermore, as can be seen from the relations between the different sets of probabilities in Equation (14), the set induced by case-wise imputation can be regarded as containing probability distributions neighbouring the one under conditional independence. The other sets can be interpreted to deviate even more from conditional independence, where domain imputation has the largest deviation. Domain imputation demonstrably neglects any conditional dependence structure in the construction of its bounds. Therefore, the bounds are maximal, but not vacuous, and thus constraining the parameter space.

Additional to logical constraints on the imputation level (see Section 3.2.4), constraints on the level of the estimated probability distribution can be regarded as a refinement of the estimated set $\widehat{\mathcal{M}}(P^*)$ of probabilities derived from our imprecise imputation (see Equation (13)). Since by construction $\widehat{\mathcal{M}}(P^*)$ is representable as a convex polyhedron in $\mathbb{R}^{|\mathcal{W}|-1}$, especially linear constraints can be incorporated very conveniently.

Minimal Example 4 *For demonstrative purpose let us estimate the bounds of conditional probabilities* $P(Y_1 = 1|Z_1 = 1)$ *for the case-wise imputed data of our toy example from Example 3. For the upper conditional probability we need to estimate* $P^*(Y_1 = 1, Z_1 = 1)$ *and* $P_*(Y_1 \neq 1, Z_1 = 1)$ *in accordance to Equation* (9)*. We estimate the upper joint probability with Equation* (11) *by counting how many observations have or could have realisation with* $y_1 = 1$ *and* $z_1 = 1$*. This holds for observations 1 and 4:* $\widehat{P^*}(Y_1 = 1, Z_1 = 1) = \frac{1}{5} \cdot 2 = 0.4$*. The lower joint probability is obtained by Equation* (12) *by counting how many observations only have realisations with* $Y_1 \neq 1$ *and* $Z_1 = 1$*. This holds for observations 2 and 5, and hence* $\widehat{P_*}(Y_1 \neq 1, Z_1 = 1) = \frac{1}{5} \cdot 2 = 0.4$ *and thus the upper conditional probability is* $\widehat{P^*}(Y_1 = 1|Z_1 = 1) = \frac{0.4}{0.4+0.4} = 0.5$*. Similarly, the lower and upper joint probabilities are estimated, occurring in Equation* (10)*:* $\widehat{P_*}(Y_1 = 1, Z_1 = 1) = 0.2$ *and* $\widehat{P^*}(Y_1 \neq 1, Z_1 = 1) = 0.4$*, resulting in the lower conditional probability* $\widehat{P_*}(Y_1 = 1|Z_1 = 1) = \frac{0.2}{0.4+0.2} = \frac{1}{3}$*. Thus,* $\hat{P}(Y_1 = 1|Z_1 = 1)$ *is within the interval* $[\frac{1}{3}; \frac{1}{2}]$*.*

5. Simulation study of imprecise imputation

To investigate the quality of imprecise imputation, we have performed a simulation study. It would have been possible to also match real data, but in a real-data application the true underlying distribution is unknown and assessing the statistical matching quality is possible only by checking whether the marginal distributions are preserved. Since this is clearly not sufficient as a sole quality criterion, we have simulated data. With the aid of a simulation study we have also been able to cover various data scenarios which make the results of our investigation of the quality criteria more credible. Moreover, the noise arising from the sampling procedure in the context of real-data applications is neutralised.

We simulated a complete categorical data file $\mathsf{A} \uplus \mathsf{B}$ with i.i.d. observations and split it into two separate files, A and B, with $n_{\mathsf{A}} = n_{\mathsf{B}}$. Subsequently, the observations of $\boldsymbol{Z}$ and $\boldsymbol{Y}$ are deleted from A and B, respectively, and the two files are statistically matched by imprecise imputation. To assess the statistical matching quality, we analysed, on the one hand, whether the true parameters of the marginal distributions and the joint distributions are within their respective estimated bounds, and, on the other hand, the distance between the upper and lower bounds. This distance, which we will call *interval width* in the following, is an appropriate performance measure since the true parameters would always lie within the estimated bounds if we chose the unit interval as a trivial estimator of a probability component. Thus, the narrower the interval which covers the component of the true parameter, the better the procedure performs. In the following, we will detail the simulation design, parameters, and results. All simulations and analyses are conducted in R (R Core Team, 2018). The specific task presented in this paper is implemented in a published R-package *impimp* (Fink et al., 2019), which was also utilised in the simulation but is in the same way usable for real-data applications.

5.1. Simulation design

The starting point of our simulation analysis is two categorical data files, A and B. Both of them contain information on four common variables $\boldsymbol{X} = \{X_1, X_2, X_3, X_4\}$ and four specific variables $\boldsymbol{Y} = \{Y_1, Y_2, Y_3, Y_4\}$ or $\boldsymbol{Z} = \{Z_1, Z_2, Z_3, Z_4\}$, respectively, with domains $\mathcal{X}_1 = \mathcal{X}_2 = \mathcal{Y}_1 = \mathcal{Y}_2 = \mathcal{Z}_1 = \mathcal{Z}_2 = \{0, 1\}$ and $\mathcal{X}_3 = \mathcal{X}_4 = \mathcal{Y}_3 = \mathcal{Y}_4 = \mathcal{Z}_3 = \mathcal{Z}_4 = \{0, 1, 2\}$.

Altogether, we modify the following four simulation parameters:

1. The strength of the bivariate associations in terms of the corrected contingency coefficient C, also known as Sakoda's adjusted Pearson's C: $C \in [0, 0.2)$, $C \in [0.2, 0.6)$, or $C \in [0.6, 1)$;
2. The Jensen-Shannon divergence *JSD* (e.g. Lin, 1991) from the marginal distribution of the common variables to the discrete uniform distribution: $\mathit{JSD} > 0.15$ or $\mathit{JSD} \leq 0.015$;
3. The numbers of observations $n_{\mathsf{A}} = n_{\mathsf{B}} \in \{50, 100, 250\}$; and
4. the dependence structure among the variables (see Figure 2).

Altogether, we obtain 72 simulation scenarios. An explanation of the choice of the simulation parameters follows in the next section. An exhaustive justification and description of the simulation design can be found in Appendix A and Appendix B, respectively.

5.2. Simulation parameters

As already stated by Rässler (2002, p. 10), the common variables should be good predictors for the specific variables. This ensures that the donation classes are suitable to generate homogeneous groups of observations

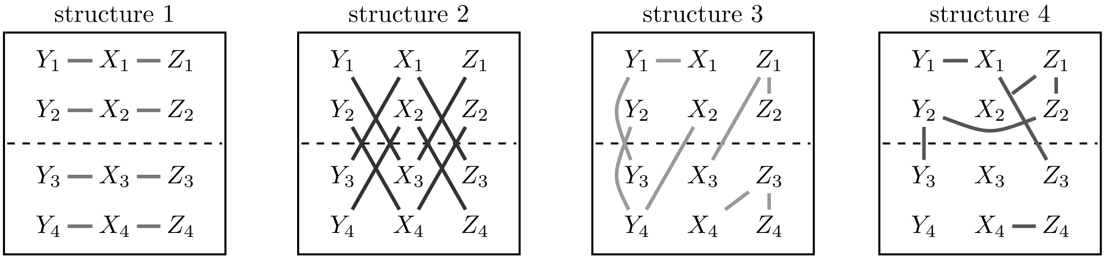


Figure 2: Four different dependence structures among the variables in the simulation study. A line between two variables displays dependence between them.

which lead to proper donor values for a missing entry. Taking this fact into account, we vary the dependence structure within a simulated data file in terms of its bivariate associations.

Figure 2 shows four different dependence structures which are covered by our simulation design. The upper six variables of each design represent the binary variables, and the six variables below the dashed line represent the variables with three categories. The connecting lines between the variables display the bivariate dependencies among these variables. For example, in the top line of structure 1, the variable X_1 is connected to variable Y_1 and also to variable Z_1. The strengths of these bivariate associations are controlled by the corrected contingency coefficient $C \in [0, 1]$. This association measure for categorical variables is based on the χ^2-coefficient for contingency tables but is corrected for the number of observations as well as for the number of categories.

At first sight, the number of observations plays a counterintuitive role in this simulation study. We expect that the distances between the lower and upper bounds for the parameters of interest increase in situations with a higher number of observation. This is due to the fact that a growth of the number of observations also causes an increase in the number of missing entries, which, in turn, leads to less precise estimations.

The Jensen-Shannon divergence from the marginal distributions of the common variables to the discrete uniform distribution is expected to have an indirect effect on the statistical matching quality. If one or more of these marginals are far away from the discrete uniform distribution, we obtain rare realisations of our matching variables which induce rare donation classes. This circumstance may likely lead to situations where certain rare donation classes of the recipient file do not exist in the donor file. In these cases, we impute, in accordance with the recommendation in Section 3.2.4, the domain for the missing entries that corresponds to a minimum of information which, in turn, leads to bounds which are (slightly) further apart.

5.3. Simulation results

As discussed, we use two measures of quality. Firstly, we investigate whether the true parameters of our simulation distributions lie within the corresponding lower and upper bounds estimated on the synthetic and partially set-valued data. Secondly, we report the mean interval widths which equal the mean distances between the upper and lower bounds. An interval width of 0 corresponds to a precise estimation.

Table 1 shows that the true values of the components of the marginal and the joint distributions almost always lie inside the estimated bounds.When considering the coverage of the marginal distributions (upper part of Table 1), the only visible difference is between the domain and donation based approaches with respect to the coverage of the true probability: While the intervals for domain imputation are always wide enough to cover the

Table 1: Relative number of probability table components for which the true parameter of the marginal distributions (top) / joint distributions (bottom) lies inside the estimated bounds, aggregated over all repetitions. The presented summary lists the result when pooling all simulation scenarios. The absence of decimal places for domain imputation highlights the numerically exact values.

imputation procedure	min.	1st quartile	median	3rd quartile	max.	mean
domain	1	1	1	1	1	1
variable-wise	0.9250	0.9613	0.9867	0.9967	1.0000	0.9792
case-wise	0.9250	0.9613	0.9867	0.9967	1.0000	0.9792

imputation procedure	min.	1st quartile	median	3rd quartile	max.	mean
domain	1	1	1	1	1	1
variable-wise	0.9975	0.9989	0.9994	0.9996	0.9998	0.9992
case-wise	0.9944	0.9985	0.9990	0.9993	0.9997	0.9987

true probability, for variable-wise and to the same extent for case-wise imputation the estimated intervals are sometimes too narrow. Regarding the joint distribution (lower part of Table 1), the intervals estimated on the domain-imputed data still always cover the true probability, but there is now also a slight difference between case-wise and variable-wise imputation, showing the hierarchy of the intervals as given in Equation (15). Nonetheless, the estimated intervals of the donation based imputation approaches still almost always cover the true probability. The difference between marginal and joint coverage is mostly due to the fact that by the simulation design the joint distribution had more components (46,656) than observations in the data file, which means that most of the underlying probability entries were zero. The marginal distributions, in contrast, consisted of only two to three entries, which made it harder to distinguish on the estimated level between the different imputation approaches. By and large, the results show a desirable output and also demonstrate the power of our method, which achieves high average coverage even across the diverse simulation scenarios.

The interval width was separately analysed for the components of the marginal distributions and joint distributions within the simulation. The aggregated results are displayed in the figures in Appendix C and summarised in the following.

The mean and maximal interval widths of the estimated intervals for the marginal distributions using domain imputation are always 0.5. This is the maximum interval width which can be achieved if we impute $\mathsf{A} \Cup \mathsf{B}$ under the constraint that $n_{\mathsf{A}} = n_{\mathsf{B}}$. Both variable-wise imputation and case-wise imputation yield intervals which are in most of the cases smaller than the intervals obtained by domain imputation. This also holds for the components of the joint distributions.

The interval widths of the marginals are conspicuously affected by the divergence of the marginal distributions to the discrete uniform distribution. If the marginals are close to the uniform distribution, the intervals are narrow. However, this effect decreases if there are few direct connections between the specific variables and the common variables. For the interval widths of the components of the joint distribution, we can observe a slightly contrary effect regarding the combination of marginals which are close to the uniform distribution and few direct connections between the specific variables and common variables. For the simulation designs with a higher divergence to the uniform distribution, the variation of the interval widths is considerably smaller. Moreover, in these cases, the median of the interval widths lies below the median of the design, with a smaller divergence to the uniform distribution. At first sight, this result appears somewhat counterintuitive, but can be explained as follows. Given a fixed value for the corrected contingency coefficient C, with marginal distributions of the common variables which are far away from the discrete uniform distribution, we obtain a probability table which has fewer combinatorial possibilities for each cell than with marginals close to the uniform distribution. This circumstance makes the estimation more precise in some cases, which in turn leads to smaller interval widths.

Furthermore, the results show that with a growing number of observations, the interval widths of the marginal distributions slightly increase. The interval widths also show higher variations in these cases. The interval widths for the components of the joint distribution show the same behaviour with respect to the number of observations.

The strengths of the bivariate associations in terms of the corrected contingency table also affects the widths of the intervals concerning the marginal distributions. In particular, the first dependence structure shows that the interval width decreases with a higher C. Nevertheless, the difference between low and high associations is, in few cases, (especially for marginals close to the uniform distribution) opposite, or only visible in the variations. Considering the interval widths for the components of the joint distribution, we can see that high associations improve the estimation.

The simulation results also show that, as expected, the dependence structure among the variables in a data file has an influence on the estimated lower and upper bounds of the parameters of the marginal distributions. The mean interval widths increase if the specific variables and the common variables have only few connections. The last dependence structure where there are only few connections between the common variables and the specific variables tends to lead to intervals with higher widths for the components of the joint distribution.

To sum up, all imputation procedures yield lower and upper bounds which almost always cover the components of the true parameter value. The number of cases where a component of the true parameter lies outside of the estimated interval is negligible. Additionally, the width of the intervals decreases the more the dependence structure among the variables in the data file are incorporated in the imputation procedure. This also holds for small associations and for structures where the specific variables only have few connections to the common variables.

6. Concluding remarks

We have presented the first micro approach for statistical matching of categorical data that reflects the natural uncertainty of statistical matching. Our approach relies on imprecise imputation, i.e. the idea to impute sets of plausible values. We suggested three types of imputation strategies: domain, variable-wise, and case-wise imprecise imputation. They can be distinguished by their ability to reproduce the available dependence structure between the common and the observed specific variables in the original files A and B into the synthetic file. They also differ in the amount of data constellations produced beyond those obtained by single or multiple imputation under the conditional independence assumption. Imprecise imputation can be seen as a set-valued generalisation of multiple (hot deck) imputation on the one hand, and fractional hot deck imputation on the other hand.

The most conservative approach, domain imputation, does not take any dependencies in the original data into account. Essentially, the dependencies present in the original files are diluted in the resulting complete synthetic file. This approach is suitable especially when there is little dependence between the common and specific variables. On the other hand, imprecise imputation based on donation classes is able to utilise the observed dependencies between the common and specific variables, and even, in the example of the case-wise variant, within the specific variables.

Embedding imprecise imputation into the framework of finite random sets allows us to derive set-valued estimates of the underlying true parameters. These estimates – possibly after their refinement by external information, see, e.g. Section 3.2.4 – reflect the uncertainty inherent to the identification problem of statistical matching. The estimation procedure utilises the set-valued information to full extent without artificially reducing the complexity of the imputed sets. Simulation results, based on a new simulation technique for dependent categorical data, corroborate that the true parameter values lie almost always inside the respective estimated bounds.

Imprecise imputation is an intuitive statistical matching micro approach which can easily be extended for more than two data files. In a strongly unbalanced statistical matching situation where, e.g. $n_{\mathsf{A}} \ll n_{\mathsf{B}}$, imprecise imputation can be applied straightforwardly to impute only the smaller file. If so, A takes the role of the recipient and the larger file, B, the role of the donor. In this special situation, the estimates for the specific variables $\boldsymbol{Y}$ are precise.

Moreover, the imprecise imputed data file with synthetic set-valued observations can be used as a starting point to derive one or multiple data files of the usual form. This would bring back the opportunity to use statistical procedures for the analysis of these now entirely single-valued data and to combine the results obtained from those data files by common multiple imputation techniques. However, one would then lose sight, to a considerable extent, of the conviction of this work, which is to produce a credible analysis by taking the full uncertainty into account.

Further studies need to be carried out to validate the performance of imprecise imputation. On the one hand, additional simulation parameters and dependence structures should be investigated in simulation studies. On the other hand, the performance of imprecise imputation should also be assessed by real-data applications. However, considerably more work will need to be done to find a definition of appropriate statistical matching quality criteria, since the true joint distribution is not available for comparisons. A further natural progression of this work is the comparison of imprecise imputation to existing statistical matching macro approaches which also address the identification problem. For this purpose, a comparison of the uncertainty measures introduced in Conti et al. (2012) or Conti et al. (2017) is desirable.

Finally, we should stress that imprecise imputation is not restricted to the block-wise missing pattern in the statistical matching framework: it is also applicable to general missing data problems. All three types of imprecise imputation promise considerable potential for a credible analysis of (non)randomly missing data far beyond statistical matching and are worthwhile to be elaborated upon and evaluated in detail.

References

Ahfock, D., S. Pyne, S.X. Lee, and G.J. McLachlan 2016. "Partial identification in the statistical matching problem." *Computational Statistics & Data Analysis* 104: 79–90. DOI: https://doi.org/10.1016/j.csda.2016.06.005.

Andridge, R.R. and R.J.A. Little 2010. "A review of hot deck imputation for survey non-response." *International Statistical Review* 78: 40–64. DOI: https://doi.org/10.1111/j.1751-5823.2010.00103.x.

Augustin, T., F.P.A. Coolen, G. de Cooman, and M.C.M. Troffaes (Eds.) 2014. *Introduction to Imprecise Probabilities*. Chichester: Wiley. DOI: https://doi.org/10.1002/9781118763117.

Barbiero, A. and P.A. Ferrari 2017. "An R package for the simulation of correlated discrete variables." *Communications in Statistics – Simulation and Computation* 46: 5123–5140. DOI: https://doi.org/10.1080/03610918.2016.1146758.

Cattaneo, M. 2013. "Likelihood decision functions." *Electronic Journal of Statistics* 7: 2924–2946. DOI: https://doi.org/10.1214/13-EJS869.

Conti, P.L., D. Marella, and M. Scanu 2008. "Evaluation of matching noise for imputation techniques based on nonparametric local linear regression estimators." *Computational Statistics & Data Analysis* 53: 354–365. DOI: https://doi.org/10.1016/j.csda.2008.07.041.

Conti, P.L., D. Marella, and M. Scanu 2012. "Uncertainty analysis in statistical matching." *Journal of Official Statistics* 28: 69–88.

Conti, P.L., D. Marella, and M. Scanu 2017. "How far from identifiability? A systematic overview of the statistical matching problem in a non parametric framework." *Communications in Statistics – Theory and Methods* 46: 967–94. DOI: https://doi.org/10.1080/03610926.2015.1010005.

Couso, I. and D. Dubois 2014. "Statistical reasoning with set-valued information: Ontic vs. epistemic views." *International Journal of Approximate Reasoning* 55: 1502–1518. DOI: https://doi.org/10.1016/j.ijar.2013.07.002.

Couso, I., D. Dubois, and L. Sánchez 2014. *Random Sets and Random Fuzzy Sets as Ill-Perceived Random Variables*. Cham: Springer. DOI: https://doi.org/10.1007/978-3-319-08611-8.

de Campos, L.M., M.T. Lamata, and S. Moral 1990. "The concept of conditional fuzzy measure." *International Journal of Intelligent Systems* 5: 237–246. DOI: https://doi.org/10.1002/int.4550050302.

Dempster, A.P. 1967. "Upper and lower probabilities induced by a multivalued mapping." *The Annals of Mathematical Statistics* 38: 325–339. DOI: https://doi.org/10.1214/aoms/1177698950.

Denœux, T. 2016. "40 years of Dempster-Shafer theory." *International Journal of Approximate Reasoning* 79: 1–6. DOI: https://doi.org/10.1016/j.ijar.2016.07.010.

Di Zio, M. and B. Vantaggi 2017. "Partial identification in statistical matching with misclassification." *International Journal of Approximate Reasoning* 82: 227–241. DOI: https://doi.org/10.1016/j.ijar.2016.12.015.

D'Orazio, M., M. Di Zio, and M. Scanu 2006a. "Statistical matching for categorical data: Displaying uncertainty and using logical constraints." *Journal of Official Statistics* 22: 137–157.

D'Orazio, M., M. Di Zio, and M. Scanu 2006b. *Statistical Matching: Theory and Practice.* Chichester: Wiley. DOI: https://doi.org/10.1002/0470023554.

D'Orazio, M., M. Di Zio, and M. Scanu 2017. "The use of uncertainty to choose matching variables in statistical matching." *International Journal of Approximate Reasoning* 90: 433–440. DOI: https://doi.org/10.1016/j.ijar.2017.08.015.

Dubois, D. and H. Prade 1992. "Evidence, knowledge, and belief functions." *International Journal of Approximate Reasoning* 6: 295–319. DOI: https://doi.org/10.1016/0888-613X(92)90027-W.

Fagin, R. and J.Y. Halpern 1991. "A new approach to updating beliefs." In *Uncertainty in Artificial Intelligence*, edited by P. Bonissone, M. Henrion, L. Kanal, and J. Lemmer, 347–374. New York: Elsevier.

Fink, P., E. Endres, and M. Schmoll 2019. *impimp: Imprecise Imputation for Statistical Matching.* URL: https://CRAN.R-project.org/package=impimp. Last access: April 2019.

Joenssen, D.W.H. 2015. *Hot-Deck-Verfahren zur Imputation fehlender Daten – Auswirkungen des Donor-Limits [Hot-Deck Procedures for the Imputation of Missing Data: Effects of the Donor Limit, translation by the authors].* Ph. D. thesis, Technische Universität Ilmenau. URL: https://www.db-thueringen.de/receive/dbt_mods_00026076. Last access: April 2019.

Kim, J.K. and W. Fuller 2004. "Fractional hot deck imputation." *Biometrika* 91: 559–578. DOI: https://doi.org/10.1093/biomet/91.3.559.

Lin, J. 1991. "Divergence measures based on the Shannon entropy." *IEEE Transactions on Information Theory* 37: 145–151. DOI: https://doi.org/10.1109/18.61115.

Little, R.J.A. and D.B. Rubin 2002. *Statistical Analysis with Missing Data* (2nd ed.). Hoboken: Wiley. DOI: https://doi.org/10.1002/9781119013563.

Manski, C.F. 1995. *Identification Problems in the Social Sciences.* Cambridge: Harvard University Press.

Manski, C.F. 2007. *Identification for Prediction and Decision.* Cambridge: Harvard University Press.

Miranda, E., I. Couso, and P. Gil 2010. "Approximations of upper and lower probabilities by measurable selections." *Information Sciences* 180: 1407–1417. DOI: https://doi.org/10.1016/j.ins.2009.12.005.

Nguyen, H.T. 1978. "On random sets and belief functions." *Journal of Mathematical Analysis and Applications* 65: 531–542. DOI: https://doi.org/10.1016/0022-247X(78)90161-0.

Nguyen, H.T. 2006. *An Introduction to Random Sets.* Boca Raton: Chapman & Hall/CRC.

R Core Team 2018. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. URL: https://www.R-project.org/. Last access: April 2019.

Ramoni, M. and P. Sebastiani 2001. "Robust learning with missing data." *Machine Learning* 45: 147–170. DOI: https://doi.org/10.1023/A:1010968702992.

Rässler, S. 2002. *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches.* New York: Springer.

Serafino, P. and R. Tonkin 2017. "Statistical matching of European Union statistics on income and living conditions (EU-SILC) and the household budget survey." In *Eurostat: Statistical working papers.* Luxembourg: Publications Office of the European Union. DOI: https://doi.org/10.2785/933460.

Shafer, G. 1976. *A Mathematical Theory of Evidence.* Princeton: Princeton University Press.

Vantaggi, B. 2008. "Statistical matching of multiple sources: A look through coherence." *International Journal of Approximate Reasoning* 49: 701–711. DOI: https://doi.org/10.1016/j.ijar.2008.07.005.

Walley, P. 1991. *Statistical Reasoning with Imprecise Probabilities.* London: Chapman and Hall.

Yang, S. and J.K. Kim 2016. "Fractional imputation in survey sampling: A comparative review." *Statistical Science* 31: 415–432. DOI: https://doi.org/10.1214/16-STS569.

Appendixes

A. Why we need a new simulation procedure

To generate simulated categorical data meeting all the desired properties, we propose a new procedure which we detail in the following section. But, first, we want to elucidate why conventional simulation approaches are not suitable for our requirements. The key aspects are listed as follows:

(i) One way to generate categorical data with predefined properties is to draw random observations from a multidimensional probability table, which, on the one hand, fulfils all of these properties and, which, on the other hand, represents the probability entries of the joint distribution of all variables. The main disadvantage of this procedure is that it can be very difficult to find a suitable joint distribution which fulfils all the desired properties. Furthermore, we would argue that it is necessary to consider several joint distributions in order to draw valid conclusions about the performance of imprecise imputation which in turn makes the problem of finding suitable distributions even harder.

(ii) Another option would be the simulation of categorical data based on a multidimensional (logit) regression model. However, a regression model cannot be used to control for the dependence structure and strength within the set of variables in the detail we wish to have.

(iii) The simulation of categorical data which imply a certain dependence structure can also be realised using a probabilistic graphical model such as a Bayesian network. The major problem with this way of proceeding is the resulting conditional independence among parts of our variables. If the – in real-world applications potentially unjustified – conditional independence assumption holds in our simulated data, statistical matching techniques directly utilising this assumption would unfairly outperform, making a fair comparison of procedures impossible.

(iv) A further feasible way to generate dependent categorical data is to employ a multivariate normal distribution with a predefined correlation matrix and discretise the data drawn from it. Nevertheless, the resulting simulated data have an ordinal scale instead of a nominal scale and we have no direct control on the strengths of the dependencies in terms of the corrected contingency coefficient. The same problems hold for simulation techniques which are based on a Gaussian copulas, such as the one suggested by Barbiero and Ferrari (2017).

To sum up, our goal is to use a simulation technique which takes all of our desired properties into account and avoid the problems described previously.

B. Simulation procedure

For this purpose, we invented a new simulation procedure which is directly based on two-way tables of relative frequencies and a suitable association measure. The bivariate associations within the simulated data can be expressed by this association measure on bivariate frequency tables of sizes 2×2, 2×3, and 3×3 reflecting the domains listed in Section 5. As also mentioned therein, we use the corrected contingency coefficient to express the strength of associations. Since – for a fixed and known number of observations – the absolute frequencies can be directly derived by the relative frequencies, and vice versa, this association measure is also suitable for tables of relative frequencies and leads to the same results.

In a first step, we generate a set S of relative frequency tables which represents the set of all possible frequency tables of above-mentioned sizes. S is created by taking all combinations of two discrete (marginal) probability distributions, whose event probabilities are strictly positive and on a one-percent grid. This strict positivity is needed because zero entries in the marginal distributions lead to zero entries in the table under independence. This entails that the χ^2 coefficient and all association measures based on it are not defined. S covers a large variety of marginal distributions and association measures ($|S| = 48\ 044\ 502$).

In a second step, we randomly draw one frequency table from $S^\star$ for each bivariate association depicted in Figure 2, where $S^\star \subseteq S$ denotes the set of probability tables which meets all predefined requirements for a specific simulation setting. Afterwards, we multiply the selected tables of relative frequencies with the desired number of observations and create a data file with complete observations $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$. To meet the challenges of a statistical matching framework, we split this data file into two parts which represent the files A and B, with $n_\mathsf{A} = n_\mathsf{B}$, and remove the observations $\boldsymbol{z}$ from A and $\boldsymbol{y}$ from B, respectively.

C. Simulation results

Figures 3 – 7 show the interval widths of the parameter estimates on the partially set-valued synthetic data, aggregated for 20 simulation runs. The graphics are grouped by the different dependence designs (see Figure 2) and the numbers of observations. The results are displayed separately for the parameters of the marginal distributions and the parameters of the joint distributions. The whiskers range from the minimum to the maximum to ensure better readability. Please note that while the interval widths for the components of the joint distribution are reported on a square root scale to spread the values and make the different results more visible, the values themselves are not transformed.

The figure showing the mean and maximal interval widths of the components of the marginal distributions of the specific variables for domain imputation is not shown here since the interval widths are 0.5 for all simulation scenarios. This is no coincidence but results deterministically from the numbers of observations n_A and n_B.

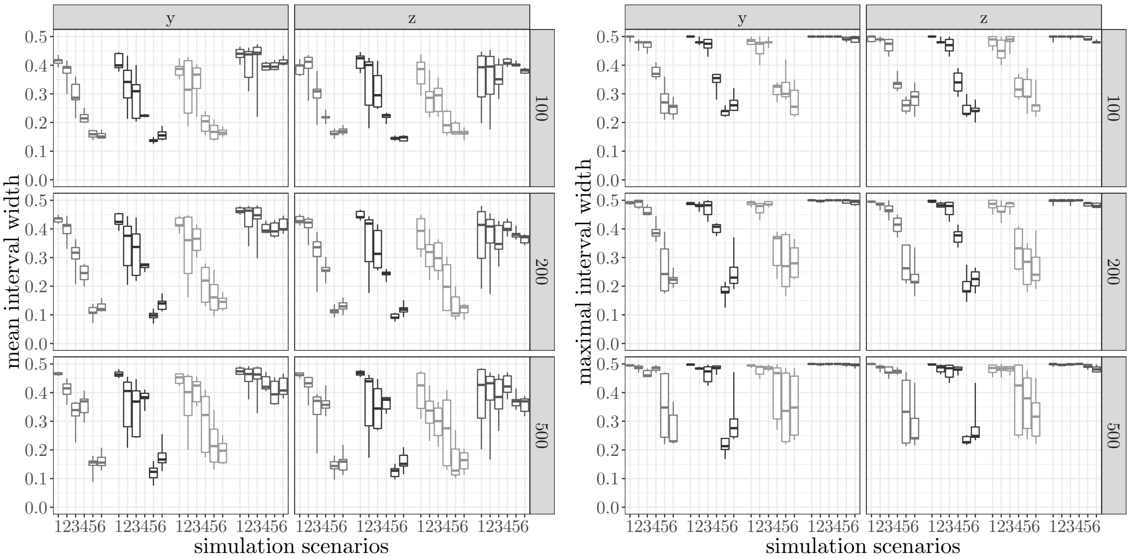


Figure 3: Mean and maximal interval widths of the components of the marginal distributions of the specific variables for variable-wise imputation. The two columns display the pooled results for the marginals of the specific variables $\boldsymbol{Y}$ and $\boldsymbol{Z}$, respectively.

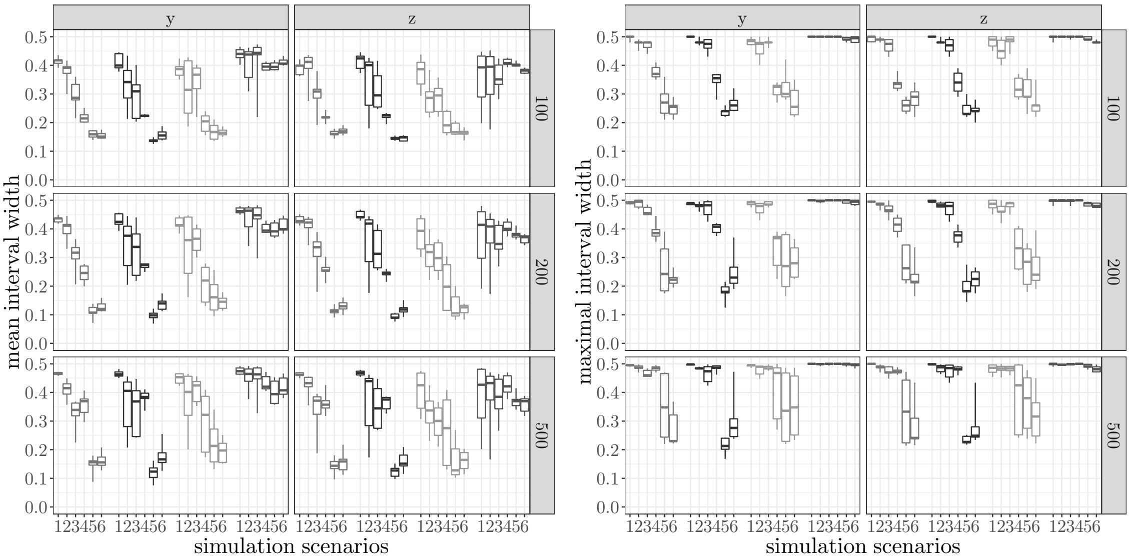


Figure 4: Mean and maximal interval widths of the components of the marginal distributions of the specific variables for case-wise imputation. The two columns display the pooled results for the marginals of the specific variables $\boldsymbol{Y}$ and $\boldsymbol{Z}$, respectively.

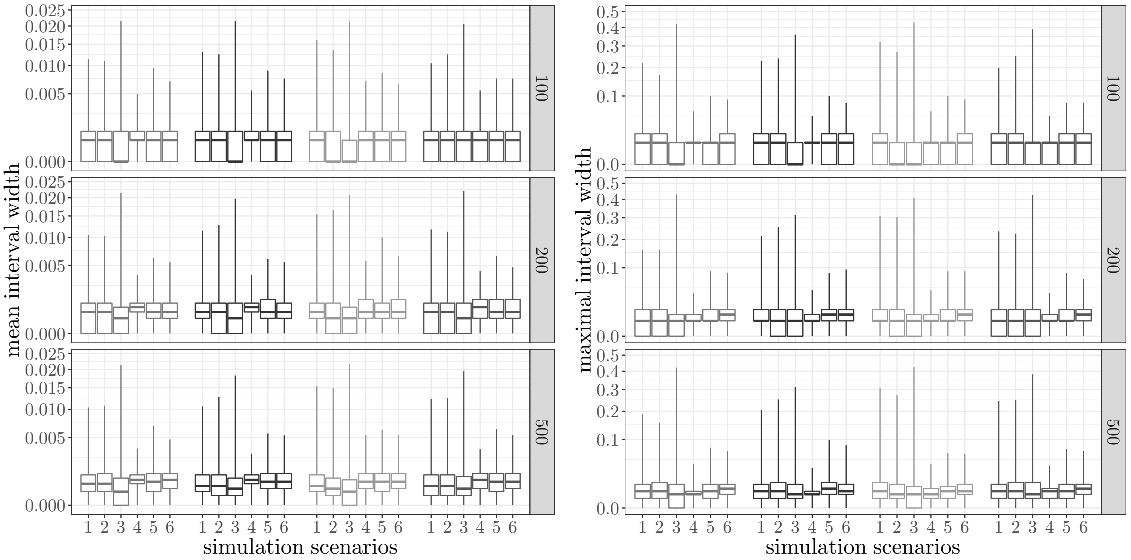


Figure 5: Mean and maximal interval widths (on the square-root scale) of the components of the joint distributions of $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$ for domain imputation.

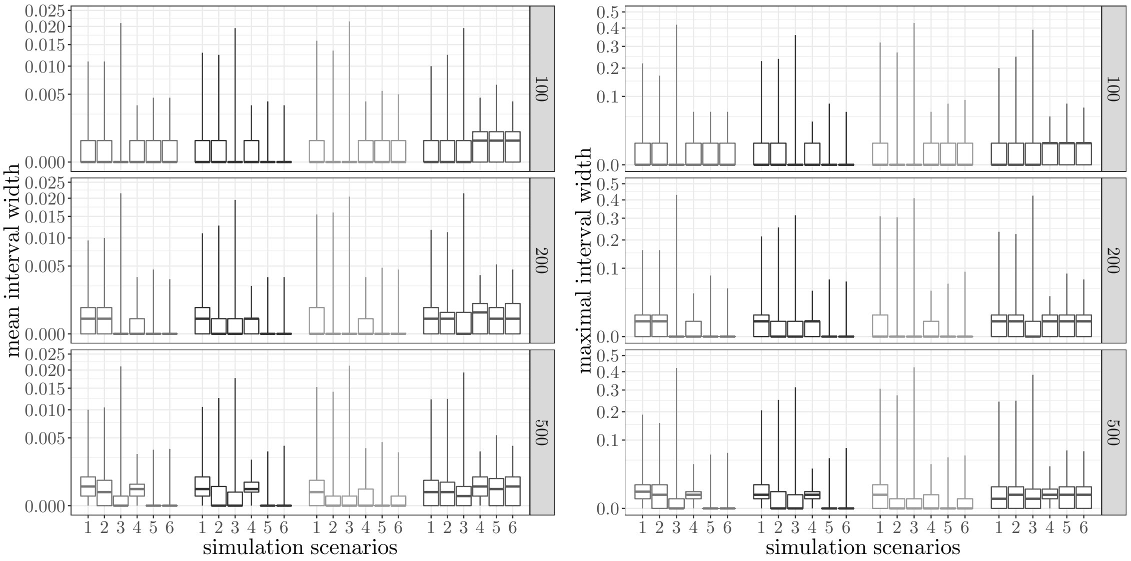


Figure 6: Mean and maximal interval widths (on the square-root scale) of the components of the joint distributions of $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$ for variable-wise imputation.

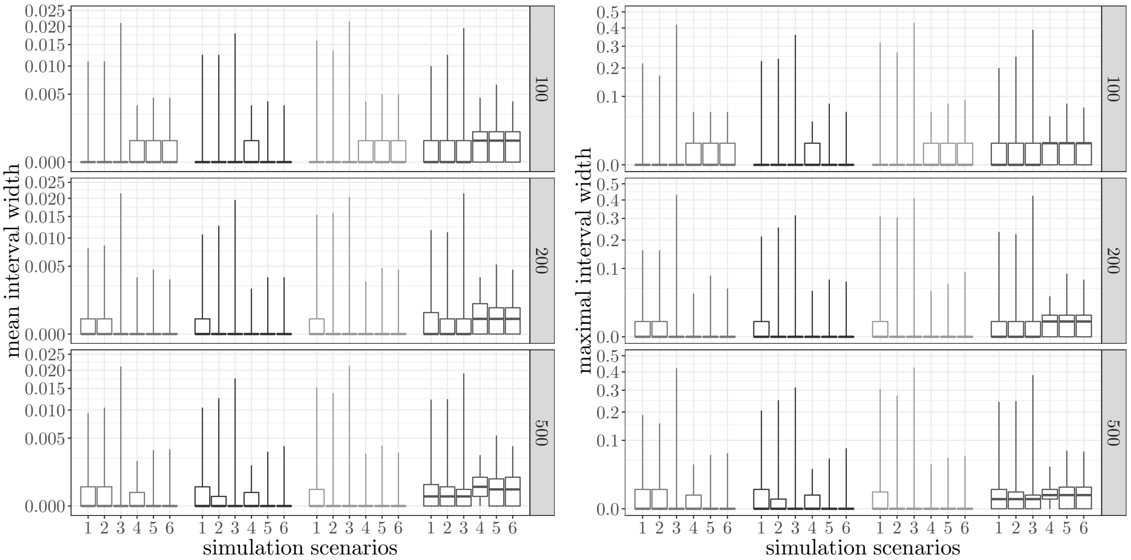


Figure 7: Mean and maximal interval widths (on the square-root scale) of the components of the joint distributions of $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$ for case-wise imputation.

Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12.07.11, § 8, Abs. 2 Pkt. 5.)

Hiermit erkläre ich, Eva-Marie Christina Endres, an Eidesstatt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

München, 02. Juli 2019

(Eva-Marie Christina Endres)