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.

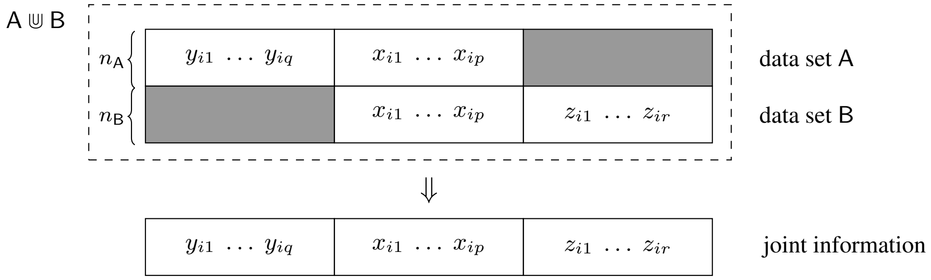

Figure 1: Schematic representation of the statistical matching problem (adapted from D'Orazio et al., 2006a, p. 5).

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 $\mathsf{A} \uplus \mathsf{B}$, 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

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

$x_j \in \mathcal{X}_j$, $y_k \in \mathcal{Y}_k$, $z_\ell \in \mathcal{Z}_\ell$, $j = 1, \ldots, p$, $k = 1, \ldots, q$, $\ell = 1, \ldots, r$, where $\mathcal{X} = \mathcal{X}_1 \times \ldots \mathcal{X}_p$, $\mathcal{Y} = \mathcal{Y}_1 \times \ldots \times \mathcal{Y}_q$, and $\mathcal{Z} = \mathcal{Z}_1 \times \ldots \times \mathcal{Z}_r$ denote the domains of $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$, and $(\mathbf{x},\mathbf{y},\mathbf{z}) := (x_1, \ldots, x_p, y_1, \ldots, y_q, z_1, \ldots, z_r)$. 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 $\mathbf{p}^{\mathsf{A} \uplus \mathsf{B}} := (\mathbf{p}_{\mathbf{Y}|\mathbf{X}}, \mathbf{p}_{\mathbf{Z}|\mathbf{X}}, \mathbf{p}_{\mathbf{X}})'$ 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 $\mathsf{A} \uplus \mathsf{B}$ becomes

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

by selecting the appropriate component of $\mathbf{p}^{\mathsf{A} \uplus \mathsf{B}}$ 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 $\mathbf{W} = (W_1, \ldots, W_s)'$ is composed of a global probability distribution $P(\mathbf{W} = \mathbf{w}) = P(W_1 = w_1, \ldots, W_s = w_s)$ and a directed acyclic graph (DAG) $\mathcal{G}_{\mathbf{W}}$, where each random variable W_m, $m = 1, \ldots, s$ 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] $\hat{\mathfrak{E}}^{\mathsf{A}\Cup\mathsf{B}}_{\mathbf{X}} \cup \hat{\mathfrak{E}}^{\mathsf{A}}_{\mathbf{X},\mathbf{Y}} \cup \hat{\mathfrak{E}}^{\mathsf{B}}_{\mathbf{X},\mathbf{Z}}$.

Procedure 2 (individual graph structures $\mathcal{G}^{\mathsf{A}}_{\mathbf{X}}$ and $\mathcal{G}^{\mathsf{B}}_{\mathbf{X}}$ for the common variables in A and B): For the second procedure for estimating and combining the graph structures, we separately estimate two DAGs $\hat{\mathcal{G}}^{\mathsf{A}}_{\mathbf{X},\mathbf{Y}}$ and $\hat{\mathcal{G}}^{\mathsf{B}}_{\mathbf{X},\mathbf{Z}}$ with sets of edges $\hat{\mathfrak{E}}^{\mathsf{A}}_{\mathbf{X},\mathbf{Y}}$ and $\hat{\mathfrak{E}}^{\mathsf{B}}_{\mathbf{X},\mathbf{Z}}$ 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. $\hat{\mathfrak{E}}^{\mathsf{A}}_{\mathbf{X}} \neq \hat{\mathfrak{E}}^{\mathsf{B}}_{\mathbf{X}}$. To obtain one joint graph structure for all common variables, we suggest to union $\hat{\mathfrak{E}}^{\mathsf{A}\Cup\mathsf{B}} = \hat{\mathfrak{E}}^{\mathsf{A}}_{\mathbf{X},\mathbf{Y}} \cup \hat{\mathfrak{E}}^{\mathsf{B}}_{\mathbf{X},\mathbf{Z}}$ or intersect $\hat{\mathfrak{E}}^{\mathsf{A}\Cup\mathsf{B}} = (\hat{\mathfrak{E}}^{\mathsf{A}}_{\mathbf{X}} \cap \hat{\mathfrak{E}}^{\mathsf{B}}_{\mathbf{X}}) \cup \hat{\mathfrak{E}}^{\mathsf{A}}_{\mathbf{Y},\mathbf{Y}-\mathbf{X}} \cup \hat{\mathfrak{E}}^{\mathsf{B}}_{\mathbf{Z},\mathbf{Z}-\mathbf{X}}$ the sets of edges within the common variables of $\hat{\mathcal{G}}^{\mathsf{A}}$ and $\hat{\mathcal{G}}^{\mathsf{B}}$, where, for example, $\hat{\mathfrak{E}}^{\mathsf{A}}_{\mathbf{Y},\mathbf{Y}-\mathbf{X}}$ 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 $\hat{\mathfrak{E}}^{\mathsf{A}}_{\mathbf{X}}$ and $\hat{\mathfrak{E}}^{\mathsf{B}}_{\mathbf{X}}$ correspond to two directed acyclic graphs, for the intersection of both holds that $\hat{\mathfrak{E}}^{\mathsf{A}\Cup\mathsf{B}}_{\mathbf{X}} \subseteq \hat{\mathfrak{E}}^{\mathsf{A}}_{\mathbf{X}}$ and $\hat{\mathfrak{E}}^{\mathsf{A}\Cup\mathsf{B}}_{\mathbf{X}} \subseteq \hat{\mathfrak{E}}^{\mathsf{B}}_{\mathbf{X}}$ and therefore, $\hat{\mathfrak{E}}^{\mathsf{A}\Cup\mathsf{B}}_{\mathbf{X}}$ 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

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

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 $\hat{\mathfrak{E}}^{\mathsf{A}\Cup\mathsf{B}}_{\mathbf{X}}$, $\hat{\mathfrak{E}}^{\mathsf{A}}_{\mathbf{X},\mathbf{Y}}$, and $\hat{\mathfrak{E}}^{\mathsf{B}}_{\mathbf{X},\mathbf{Z}}$ 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, $k = 1, \ldots, q$ or Z_ℓ, $\ell = 1, \ldots, r$, 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 ($\in \mathbb{R}^{400\times12}$) and B ($\in \mathbb{R}^{400\times11}$) which can then be

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

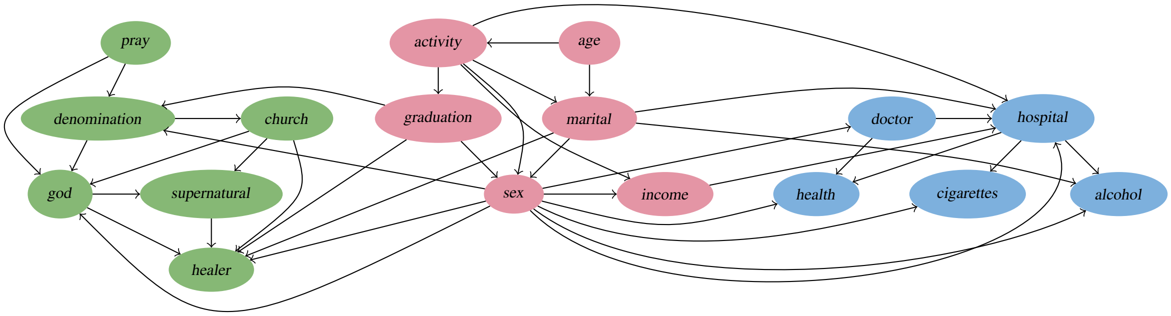

Figure 2: Graph structure of the matched Bayesian network $\hat{\mathcal{G}}^{\mathsf{A} \Cup \mathsf{B}}$ using Procedure 1 in Step 1 of the statistical matching approach.

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

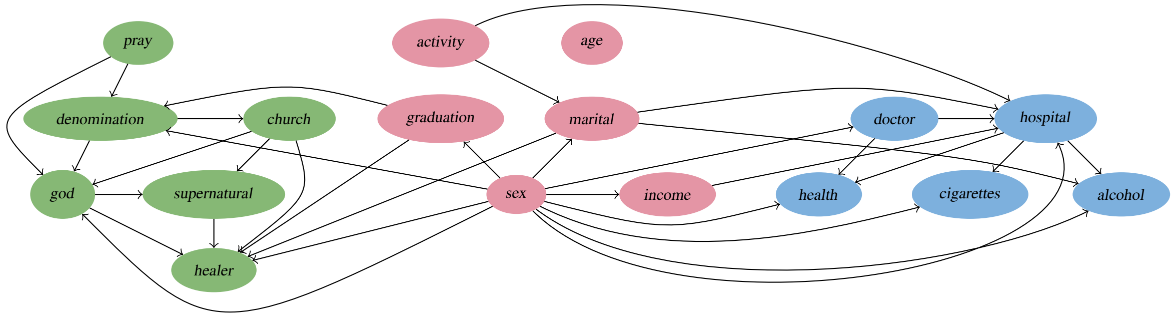

Figure 3: Graph structure of the matched Bayesian network $\hat{\mathcal{G}}^{\mathsf{A} \uplus \mathsf{B}}$ using edge intersection in Procedure 2 in Step 1 of the statistical matching approach.

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-

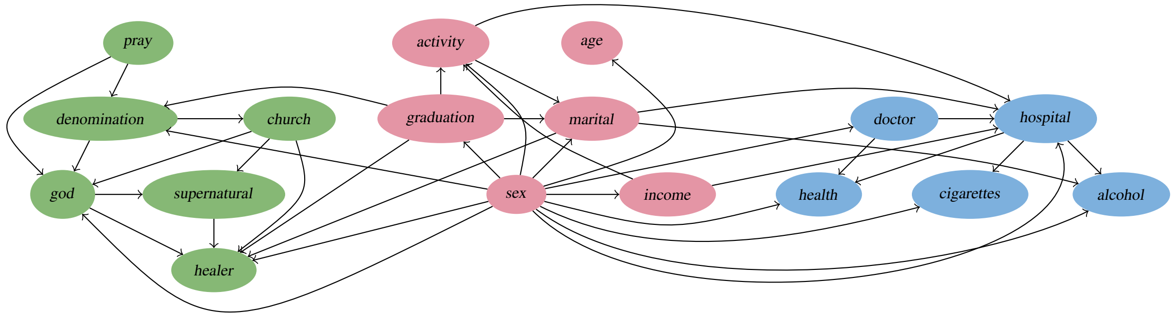

Figure 4: Graph structure of the matched Bayesian network $\hat{\mathcal{G}}^{\mathsf{A} \uplus \mathsf{B}}$ using edge union in Procedure 2 in Step 1 of the statistical matching approach.

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.