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 $n_{\mathsf{A}}$ 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 $n_{\mathsf{B}}$ 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

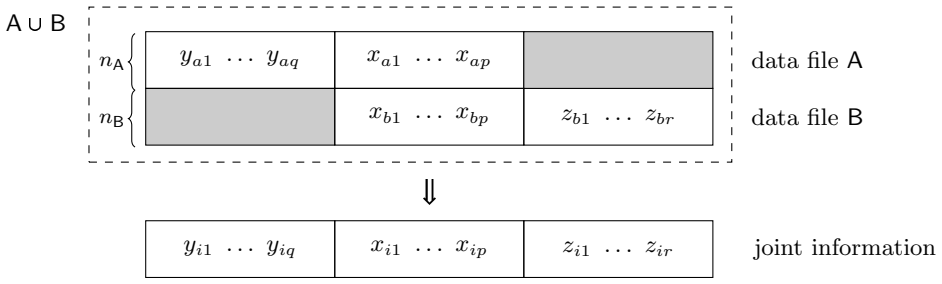


Figure 1: Graphical illustration of the data setting for statistical matching (cf. D'Orazio et al., 2006a, p. 5 (modified)).

products of the sets of possible realizations of the single elements of $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$, 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 $\mathbf{Y}$ and $\mathbf{Z}$ makes the joint distribution unidentified. Even if we had an infinite number of observations, the relationship between the specific variables in $\mathbf{Y}$ and $\mathbf{Z}$ 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

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

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 $\mathbf{z}_a = (z_{a1}, \ldots, z_{ar})$ and $\mathbf{y}_b = (y_{b1}, \ldots, y_{bq})$ are missing in a block-wise pattern.

As previously mentioned, the aim of statistical matching is the collection of joint information on either $\mathbf{Y}$ and $\mathbf{Z}$, or $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$. 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 $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ (*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 $\mathbf{C}_1, \ldots, \mathbf{C}_m$. 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 $\tau_j$ 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.

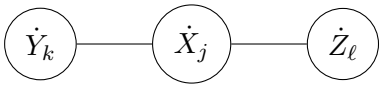


Figure 2: Basic form of the Ising model for statistical matching, reflecting the conditional independence assumption $\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} \mid \mathbf{X}$.

The estimation of the graph structure takes place in two steps, starting with the separate estimation of the graph structures on $\mathsf{A}$ and on $\mathsf{B}$. 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 $Z_\ell$, 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 $\mathsf{A}$, from $\mathsf{B}$, or from $\mathsf{A} \cup \mathsf{B}$. The joint probability distribution arises as

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

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

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

# 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 $\mathsf{A}$ and $\mathsf{B}$ with a known joint distribution, we simulate a complete file containing $n_{\mathsf{A}} + n_{\mathsf{B}}$ observations, and randomly allocate the observations into $\mathsf{A}$ or $\mathsf{B}$. Subsequently, the observations of the specific variables $\mathbf{Z}$ are removed from $\mathsf{A}$, and the observations of $\mathbf{Y}$ are removed from $\mathsf{B}$. 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.

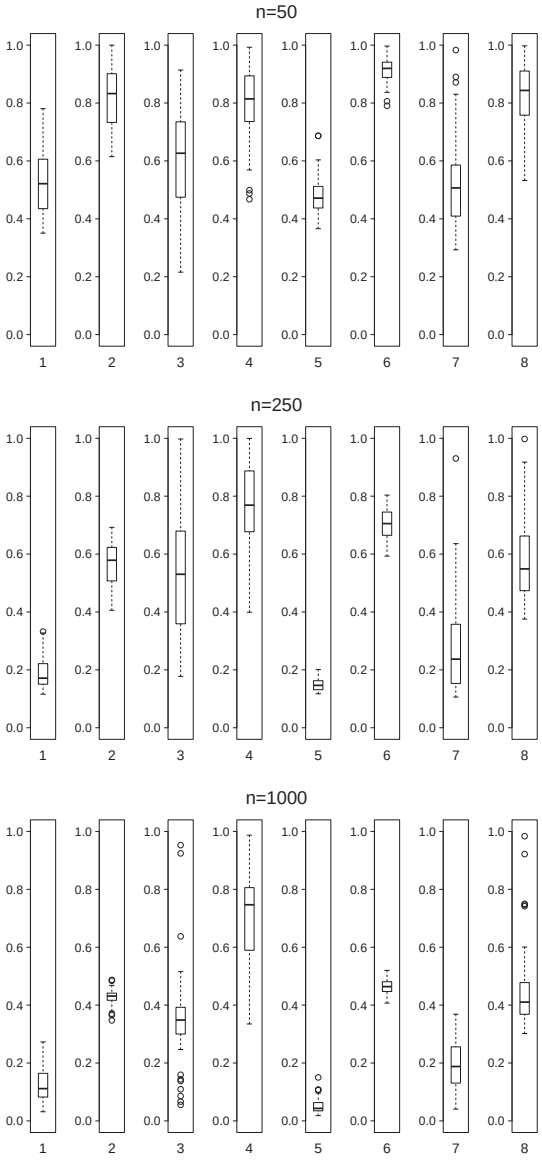


Figure 3: Jensen-Shannon divergences for the simulation setups, where the conditional independence assumption applies. The different rows indicate different sample sizes.

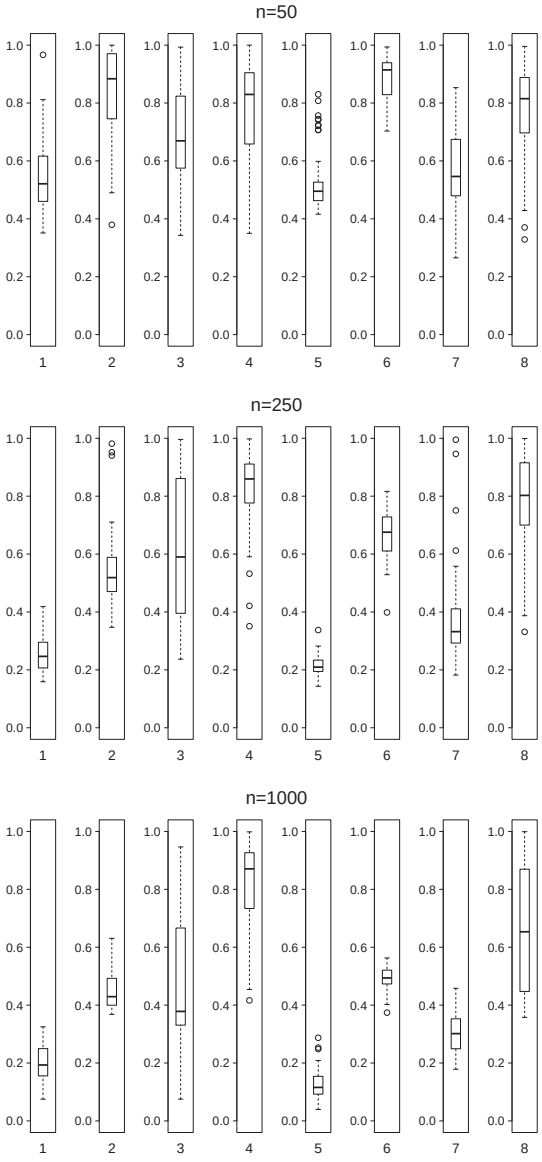


Figure 4: Jensen-Shannon divergences for the simulation setups, where the conditional independence assumption is violated for some variables. The different rows indicate different sample sizes.

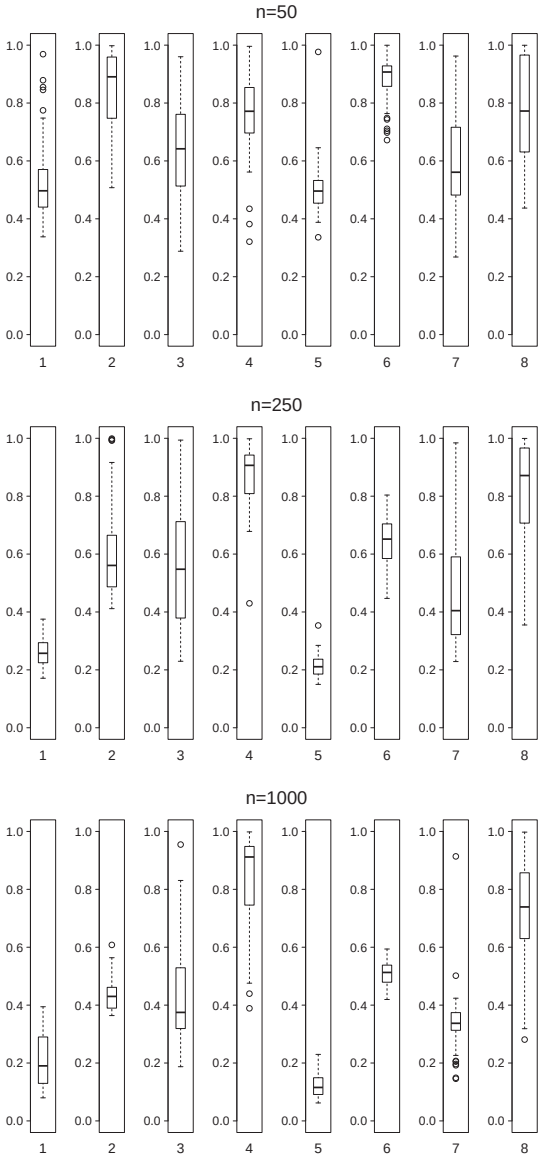


Figure 5: Jensen-Shannon divergences for the simulation setups, where the conditional independence assumption is violated for all variables. The different rows indicate different sample sizes.

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 $\mathsf{A}$ and $\mathbf{Y}$ from $\mathsf{B}$ 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 $\mathbf{Y}$ and $\mathbf{Z}$, 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.