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

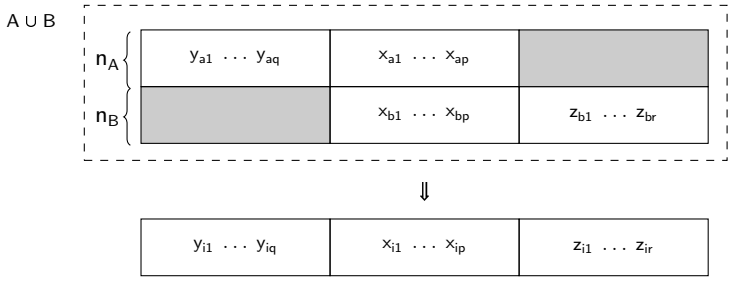


Figure 1: Schematic representation of the statistical matching problem (see D'Orazio et al., 2006a, p.5 (modified)).

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 $(x_1, \ldots, x_p, y_1, \ldots, y_q)$ of the variables in the sets $\boldsymbol{X} = \{X_1, \ldots, X_p\}$ and $\boldsymbol{Y} = \{Y_1, \ldots, Y_q\}$, and B a data file with n_B categorical observations $(x_1, \ldots, x_p, z_1, \ldots, z_r)$ of the variables in the sets $\boldsymbol{X}$ and $\boldsymbol{Z} = \{Z_1, \ldots, Z_r\}$. 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 $j = 1, \ldots, p$, $k = 1, \ldots, q$, and $\ell = 1, \ldots r$.

If we treat the files A and B as a single data source $A \Cup B$ 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 $A \Cup B$. 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 $A \Cup B$. 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 $A \cup\!\!\!\cup B$ 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 $\boldsymbol{X} = \{X_1, \ldots, X_p\}$ 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, $j \in \{1, \ldots, p\}$, while certain subsets of $\boldsymbol{X}$ are characterized by

an index set $\boldsymbol{j} \subseteq \{1, \ldots, p\}$ such that $\boldsymbol{X_j} := \{X_j : X_j \in \boldsymbol{X}, j \in \boldsymbol{j}\}$. The corresponding realizations $\boldsymbol{x} = (x_1, \ldots, x_p)$ are analogously indexed and $x_j \in \mathcal{X}_j = \{0, 1, \ldots, d_j - 1\}$ for every $j \in \{1, \ldots, p\}$. Referring to the d_j different categories of the j-th variable as $\{0, 1, \ldots, d_j - 1\}$ does not imply any ordering.

A Markov network over the set of categorical variables $\boldsymbol{X} = \{X_1, \ldots, X_p\}$ 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 $\times$ 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 $\pi(\boldsymbol{x}) = \mathbb{P}(X_1 = x_1, \ldots, X_p = x_p)$ 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 $A \uplus B$ 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 $\mathcal{H}^{A \uplus B}$, is supposed to be known, or at least that the information in A and B is sufficient to estimate an error-free version $\hat{\mathcal{H}}^{A \uplus B}$ of the true network structure. $\mathcal{H}^{A \uplus B}$ is composed of a set of undirected edges $\boldsymbol{E}^{A \uplus B}$ and a set of nodes $\mathring{\boldsymbol{V}}^{A \uplus B} \equiv \mathring{\boldsymbol{X}} \cup \mathring{\boldsymbol{Y}} \cup \mathring{\boldsymbol{Z}}$ 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 $\boldsymbol{E}^{A \uplus B}$, $k = 1, \ldots, q$, $\ell = 1, \ldots, r$. 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

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

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 $\mathcal{H}^{A \uplus B}$ (i.e. the pairwise edges are not in

$\boldsymbol{E}^{A \Cup B}$). 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 $A \Cup B$. This means that we now have an identifiable model for the incomplete sample $A \Cup B$, 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:

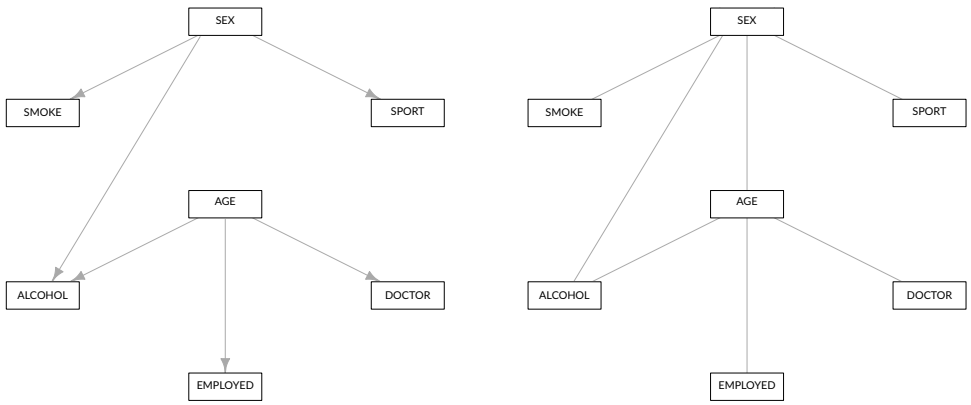
- file A has 688 observations of SEX, AGE, EMPLOYED, SMOKE, and ALCOHOL
- file B has 687 observations of SEX, AGE, EMPLOYED, DOCTOR, and SPORT.

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.



a) b)

Figure 2: Joint DAG based on $A \Cup B$ on the left side. Joint undirected graph based on $A \Cup B$ 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 $A \Cup B$, 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 $A \Cup B$. On the right hand side, we see the moralized graph, i.e. the structure of the joint Markov network on $A \Cup B$. 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

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

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

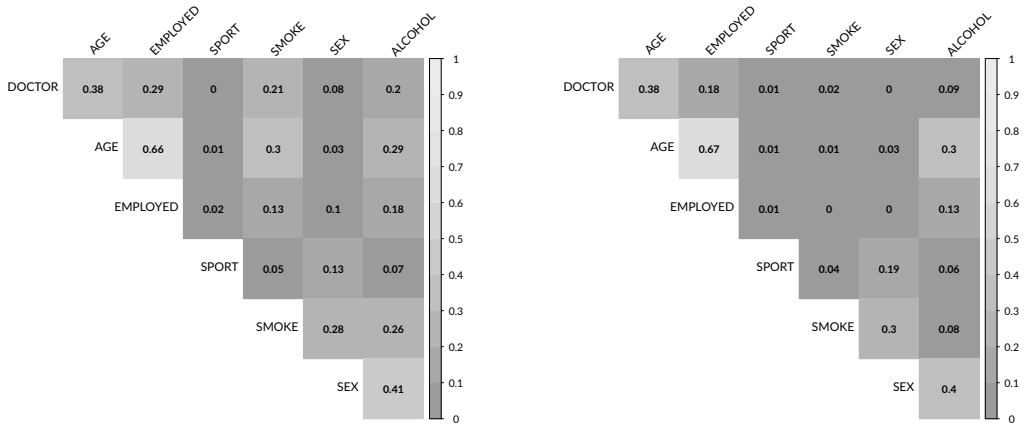
3. the preservation of the joint distribution.

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

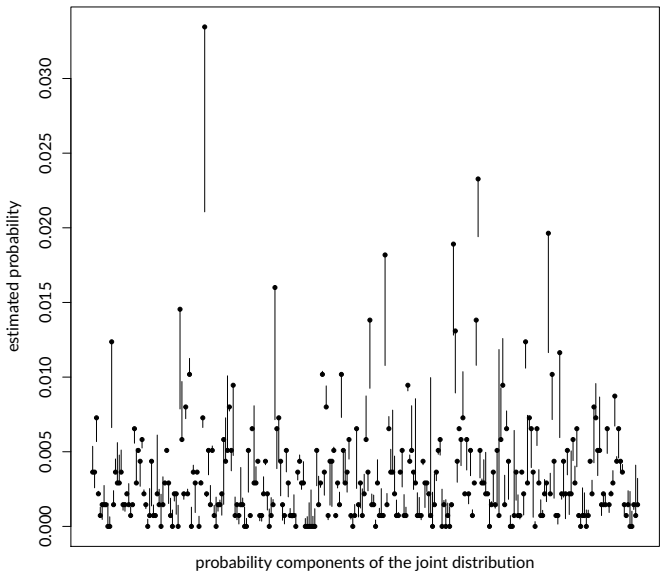


Figure 4: Absolute difference between the sample distribution and the matched distribution in the GGSS data example separately for all probability components of the joint distribution. The black points are the estimates for the components of $\hat{\pi}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$ 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

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

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

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

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

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

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

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

with

$$u_X(x) = \begin{cases} u_x & , x = 1 \\ 0 & , x = 0, \end{cases}$$

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

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

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 $2 \times 2 \times 3$-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

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

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

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

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 con-

ditional 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:

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

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 $u_{\emptyset}$

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

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

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

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

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

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

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

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

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

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 $\hat{u}^{A \cup\!\!\cup B}$ concerning the common variables $\boldsymbol{X}$, estimated on $A \cup\!\!\cup B$.

variable name(s)	category	$\hat{u}^{A \cup\!\!\cup B}$
∅	(*intercept*)	5.1896
EMPLOYED	*unemployed*	-0.6674
AGE	$45-59$ *years*	-0.0973
AGE	$\geq$ 60 *years*	-1.5395
SEX	*female*	-0.0800
EMPLOYED : AGE	*unemployed* : $45-59$ *years*	-0.6129
EMPLOYED : AGE	*unemployed* : $\geq$ 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	$\geq$ 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* : $\geq$ 60 *years*	-0.3784
SEX : ALCOHOL	*female* : *never*	1.1629
AGE : ALCOHOL	$45-59$ *years* : *never*	0.2623
AGE : ALCOHOL	$\geq$ 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	$\geq$ 60 *years* : *unemployed*	2.2951
SEX : AGE : ALCOHOL	*female* : $45-59$ *years* : *never*	0.2749
SEX : AGE : ALCOHOL	*female* : $\geq$ 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	$\hat{u}^B$
$\varnothing$	$(intercept)$	2.5522
SEX	$female$	0.2933
SPORT	$often$	0.4878
EMPLOYED	$unemployed$	-0.6993
AGE	$45-59$ $years$	0.2209
AGE	$\geq$ 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	$unemployed : \geq$ 60 $years$	2.3137
AGE : DOCTOR	$45-59$ $years : seldom\ or\ never$	-0.5433
AGE : DOCTOR	$\geq$ 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.

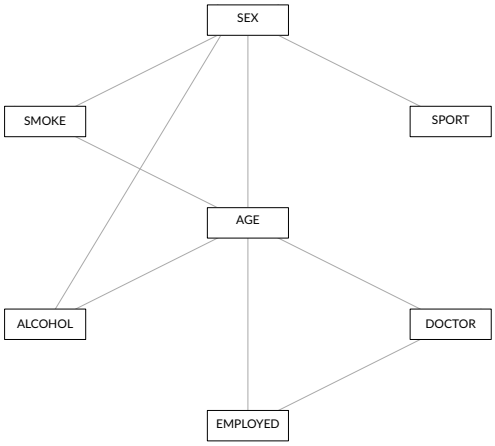


Figure 5: Markov network with two additional edges in between the specific variables and the common variables.

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