LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Probabilistic Graphical Models for Statistical Matching

Eva Endres and Thomas Augustin
Department of Statistics, University of Munich, Germany

1 Motivation

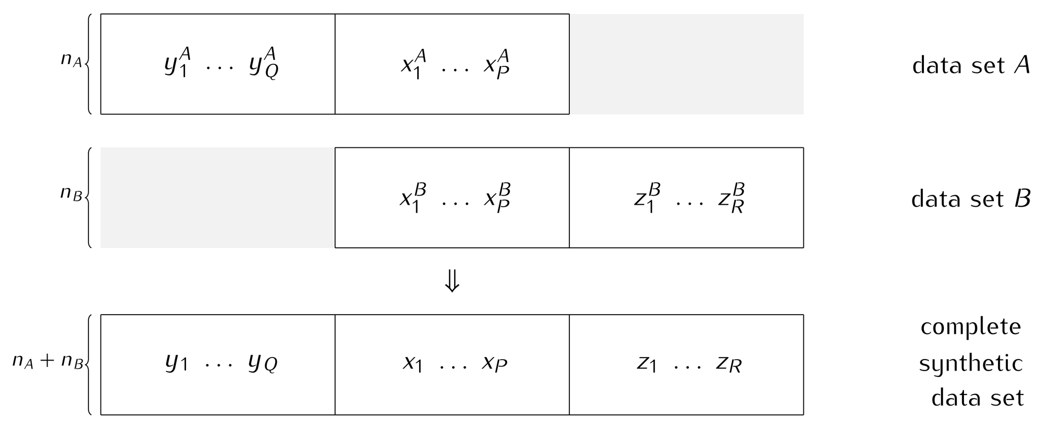
In the information age a massive amount of data is available. It can be of great benefit to use this existing data for secondary analysis instead of collecting new data, which might be time-consuming and expensive. But what can be done if the required variables are not all accessible in one single data set? The solution is given by statistical matching: With the aid of statistical matching, information from different surveys can be combined.

2 Statistical Matching

Statistical matching (or data fusion) aims at the achievement of joint information on variables that are on the one hand not jointly observed and on the other hand based on a disjoint set of observation units [e.g. 2, p. 2].

Initial situation of partially overlapping data sets

The initial situation of statistical matching [e.g. 2] are two (or more) data sets, e.g. A and B with n_A or n_B observations, respectively, that contain information on a set of common variables $\mathbf{X}$, and specific variables $\mathbf{Y}$ and $\mathbf{Z}$ which are not jointly observed. Furthermore, the observation units in A and B are not the same.
The objective is, on the one hand, to estimate the joint probability distribution of all common and specific variables (macro approach) or, on the other hand, to generate one synthetic data set, that contains information on all variables of interest (micro approach).



Conditional independence assumption to achieve an identifiable joint distribution

It is common practice to use statistical matching strategies that are premised on the restrictive assumption of conditional independence (CIA), i.e. the independence of $\mathbf{Y}$ and $\mathbf{Z}$ given $\mathbf{X}$. This technical assumption makes the joint distribution of $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ identifiable for $A \cup B \in \mathbb{R}^{(n_A+n_B)\times(P+Q+R)}$, where $A \cup B$ is an incomplete i.i.d. sample from $f(\mathbf{x}, \mathbf{y}, \mathbf{z}) \overset{CIA}{=} f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})\, f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x})\, f_{\mathbf{X}}(\mathbf{x})$ without joint information on $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ [e.g. 2, p. 13].

Maximum likelihood estimation under the CIA

Given the CIA, the observed likelihood function of $A \cup B$ in the parametric framework is given by

$$L(\theta|A\cup B) = \prod_{a=1}^{n_A} f_{\mathbf{XY}}(\mathbf{x}_a, \mathbf{y}_a; \theta_{\mathbf{XY}}) \prod_{b=1}^{n_B} f_{\mathbf{XZ}}(\mathbf{x}_b, \mathbf{z}_b; \theta_{\mathbf{XZ}})$$
$$= \prod_{a=1}^{n_A} f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}_a|\mathbf{x}_a; \theta_{\mathbf{Y}|\mathbf{X}}) \prod_{b=1}^{n_B} f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}_b|\mathbf{x}_b; \theta_{\mathbf{Z}|\mathbf{X}}) \prod_{a=1}^{n_A} f_{\mathbf{X}}(\mathbf{x}_a; \theta_{\mathbf{X}}) \prod_{b=1}^{n_B} f_{\mathbf{X}}(\mathbf{x}_b; \theta_{\mathbf{X}}),$$

where f can either be a density function or a probability distribution.
Although $A \cup B$ is an incomplete data set, the maximum likelihood estimators $\hat{\theta}_{\mathbf{X}}$, $\hat{\theta}_{\mathbf{Y}|\mathbf{X}}$, and $\hat{\theta}_{\mathbf{Z}|\mathbf{X}}$ can directly be estimated from it [e.g. 2, p. 14], where $\theta_{\mathbf{Y}|\mathbf{X}}$ and $\theta_{\mathbf{Z}|\mathbf{X}}$ denote the parameters of the conditional distributions.

3 Probabilistic Graphical Models

Probabilistic graphical models aim at the compact representation of complex distributions over a possibly high-dimensional space by exploiting the (conditional) independences among the concerned random variables [e.g. 3, p. 3].

Bayesian networks

A Bayesian network over a set of random variables $\mathbf{X} = \{X_1, \dots, X_p\}$ is composed of a global probability distribution and a directed acyclic graph $\mathcal{G} = (\mathbf{N}, \mathbf{A})$, where

- each random variable $X_i \in \mathbf{X}$ is depicted by a node $n_i \in \mathbf{N}$, and
- the dependence relations among the random variables, i.e. the direct influence of one node on another [e.g. 3, pp. 51], are illustrated by the set of directed edges $\mathbf{A}$.

Global probability distribution

Taking into account the so-called Markov property, which states that every variable is conditionally independent of its non-descendants given its parents, and the chain rule, the joint probability distribution over $\mathbf{X}$ can be obtained by the product over the local conditional probability distributions as follows

$$P(\mathbf{x}) = P(x_1, \dots, x_p) = \prod_{i=1}^{p} P(x_i|pa(X_i)).$$

The term $P(x_i|pa(X_i))$ denotes the conditional probability of $X_i = x_i$, where $Pa(X_i)$ represents the parent nodes of X_i, and $pa(X_i)$ its realizations.

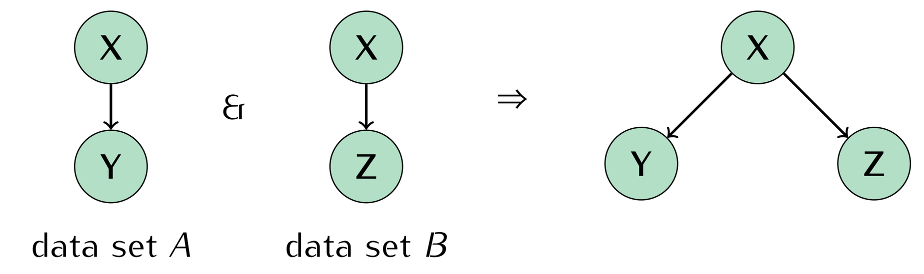
References

1 A. Antonucci, C. P. de Campos, and M. Zaffalon. Probabilistic graphical models. In T. Augustin, F. P. A. Coolen, G. de Cooman and M. C. M. Troffaes, editors, *Introduction to Imprecise Probabilities*, pages 207–229. Wiley, Chichester, UK, 2014.
2 M. D'Orazio, M. Di Zio, and M. Scanu. *Statistical Matching: Theory and Practice.* Wiley, 2006.
3 D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques.* MIT Press, 2009.

4 Probabilistic Graphical Models for Statistical Matching

Here, it is proposed to perform statistical matching by graphical network models. This might be a promising alternative to existing statistical matching approaches, since probabilistic graphical models provide a natural form of representing conditional independence.
The basic idea is composed by the following two steps:

Step 1: Create one network on each of the data sets to be matched
Step 2: Link the networks to one single network and estimate the global probability distribution



5 Simulation

Simulation design

- Data set A with $n_A = 3500$ observations
- Data set B with $n_B = 1500$ observations

- (Conditional) probability distribution:

$x = 0$:

	$Z=0$	$Z=1$	
$Y=0$	0.8075	0.0425	0.85
$Y=1$	0.1425	0.0075	0.15
	0.95	0.05	1

$x = 1$:

	$Z=0$	$Z=1$	
$Y=0$	0.525	0.175	0.7
$Y=1$	0.225	0.075	0.3
	0.75	0.25	1

Macro approach

Estimation of the joint probability distribution

$$P_{X,Y,Z}(x,y,z) = P_X(x) P_{Y|X}(y|x) P_{Z|X}(z|x)$$

$P_X(x)$: can be estimated from A and B
$P_{Y|X}(y|x)$: can be estimated from A
$P_{Z|X}(z|x)$: can be estimated from B

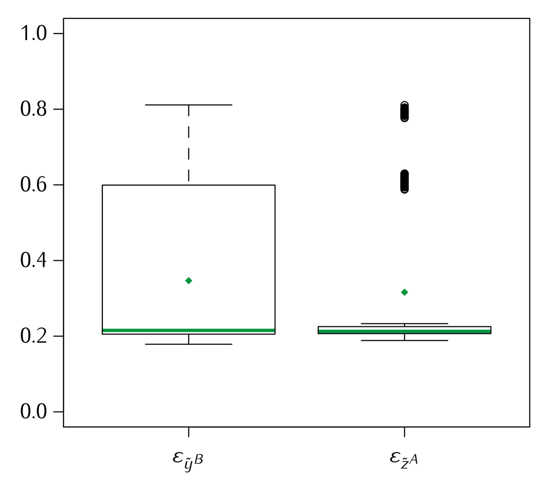
Micro approach

Substitution of the missing values of Z in A by draws from the posterior

$$P_{Z|X,Y}(Z|x,y) = \frac{P_{X,Y,Z}(x,y,Z)}{P_{X,Y}(x,y)} = \frac{P_{X,Y,Z}(x,y,Z)}{\sum_z P_{X,Y,Z}(x,y,z)}$$
$$= \frac{P_X(x)P_{Y|X}(y|x)P_{Z|X}(Z|x)}{P_X(x)P_{Y|X}(y|x)P_{Z|X}(0|x) + P_X(x)P_{Y|X}(y|x)P_{Z|X}(1|x)}$$

[Analogous procedure for the imputation of the missing values of Y in B.]

Prediction error



Further Research

Credal networks

The next step will be the application of credal networks [e.g. 1] to match partially overlapping data sets. This approach offers decisive advantages: On the one hand, the strict conditional independence assumption can be weakened by using independence concepts for conditional credal sets. On the other hand, the uncertainty of the statistical matching process can be taken into consideration by sets of compatible contingency tables.

Combination of differing network structures or parameter estimates

Furthermore, the combination of possibly differing network structures of $\mathbf{x}^A$ and $\mathbf{x}^B$ require further investigations. Feasible solutions are provided by graph union, graph intersection, or model averaging. Also the opportunity of varying parameter estimates on the two data sets A and B need to be taken into account.

Continuous and hybrid network models

Moreover, the extension to continuous and hybrid network models is planned, starting with Gaussian Bayesian networks and networks for exponential families.

eva.endres@stat.uni-muenchen.de www.statistik.lmu.de thomas.augustin@stat.uni-muenchen.de