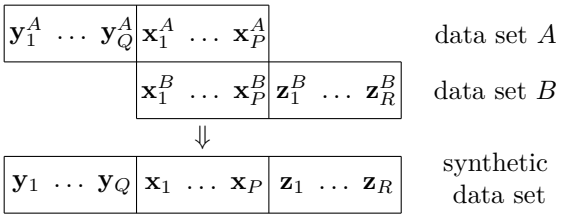
Probabilistic Graphical Models for Statistical Matching

Eva Endres
Department of Statistics, LMU Munich
eva.endres@stat.uni-muenchen.de

Thomas Augustin
Department of Statistics, LMU Munich
augustin@stat.uni-muenchen.de

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.

The initial situation of statistical matching [2, e.g.] 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. The observation units in the different data sets 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*).



The most popular statistical matching strategies are premised on the restrictive assumption of conditional independence, 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 and, thus, estimable 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 $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$ without joint information on $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ [2, cf.].

Here, it is proposed to perform statistical matching by graphical network models. This might be a promising alternative to existing statistical matching approaches, since it provides a natural form of representing conditional independence. In addition, the use of auxiliary information for solving the statistical matching problem remains possible.

In a first step, one Bayesian network [3, e.g.] has to be created on each of the data sets to be matched. Random variables are represented by nodes and the dependencies between them are displayed by arcs.

Afterwards, the individual networks can be linked to a single one by means of graph union or graph intersection, respectively.

The second step will be the application of credal networks [1, e.g.] in this setting. Thereby, the uncertainty of the statistical matching process can be taken into consideration by sets of compatible contingency tables. Moreover, the strict conditional independence assumption can be weakened by using independence concepts for sets of conditional probabilities.

Keywords. Statistical matching, Bayesian networks, credal networks, independence.

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, 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.