Statistical matching by Bayesian Networks
L'uso delle reti Bayesiane nello Statistical Matching

Daniela Marella and Paola Vicard and Vincenzina Vitale

Abstract The goal of statistical matching is the estimation of the joint distribution of variables not jointly observed in a sample survey but separately available from independent sample surveys. The lack of joint information on the variables of interest leads to uncertainty about the data generating model. In this paper we propose the use of Bayesian networks to deal with the statistical matching problem since they admit a recursive factorization of a joint distribution useful for evaluating the statistical matching uncertainty in the multivariate context.
Abstract Lo scopo dello statistical matching è stimare una distribuzione congiunta di variabili osservate separatamente in due campioni indipendenti. La mancanza di osservazioni congiunte sulle variabili di interesse causa incertezza sul modello che ha generato i dati: l'informazione campionaria non è in grado di discriminare tra un insieme di modelli plausibili. In questo lavoro il problema dello statistical matching è analizzato utilizzando le reti Bayesiane che consentono non solo di descrivere la struttura di dipendenza in distribuzioni multivariate ma ammettono una fattorizzazione della distribuzione congiunta utile ai fini della valutazione dell'incertezza.

Daniela Marella
Dipartimento di Scienze della Formazione, via del Castro Pretorio 20, 00185 Roma, e-mail: daniela.marella@uniroma3.it

Paola Vicard
Dipartimento di Economia, Via Silvio D'Amico 77, 00145 Roma, e-mail: paola.vicard@uniroma3.it

Vincenzina Vitale
Dipartimento di Economia, Via Silvio D'Amico 77, 00145 Roma, e-mail: vincenzina.vitale@uniroma3.it

1 Introduction

Statistical matching aims at combining information obtained from different non-overlapping sample surveys, referred to the same target population. The main target is in constructing a complete synthetic data set where all the variables of interest are jointly observed, see [3].

Formally, let (X,Y,Z) be a random variable (rv) with joint discrete distribution P. Without loss of generality, let $X = (X_1,\ldots,X_H)$, $Y = (Y_1,\ldots,Y_K)$ and $Z = (Z_1,\ldots,Z_T)$ be vectors of rvs of dimension H, K, T, respectively. Furthermore, let A and B be two independent samples of n_A and n_B independent and identically distributed records from (X,Y,Z). Assume that (X,Y) are observed in sample A while (X,Z) are observed in sample B. The main goal of statistical matching consists in estimating the joint distribution of (X,Y,Z) from the samples A and B. Such a distribution is not identifiable due to the lack of joint information on Z and Y given X.

In order to overcome this problem, the following approaches have been considered. The first approach uses techniques based on the conditional independence assumption between Y and Z given X (henceforth CIA assumption) see, e.g., [9]. The second approach uses techniques based on external auxiliary information regarding the statistical relationship between Y and Z , e.g. an additional file C where (X,Y,Z) are jointly observed is available, as in [12].

However, it is possible that neither case is appropriate, then the third group of techniques addresses the so called *identification problem*. The lack of joint information on the variables of interest is the cause of uncertainty about the model of (X,Y,Z). In other terms, the sample information provided by A and B is actually unable to discriminate among a set of plausible models for (X,Y,Z). For instance, in a parametric setting and for $K = T = 1$ the estimation problem cannot be "pointwise", only ranges of values containing all the pointwise estimates obtainable by each model compatible with the available sample information can be detected. Such intervals are uncertainty intervals. Uncertainty in statistical matching is analyzed in [8],[11], [4],[1] and [2].

In this paper we propose the use of Bayesian networks (BNs) to deal with statistical matching in the identification problem framework for multivariate categorical data. The first attempt in such direction is in [5] where the CIA is assumed.

The use of BNs is motivated by the following advantages: (i) BNs are widely used to describe dependencies among variables in multivariate distributions; (ii) BNs admit convenient recursive factorizations of their joint probability useful for the uncertainty evaluation in a multivariate context.

The paper is organized as follows. In section 2 the concept of uncertainty in statistical matching when BNs are used is discussed.

2 Uncertainty in Statistical Matching using graphical models

BNs are multivariate statistical models satisfying sets of conditional independence statements contained in a directed acyclic graph (DAG), see [10]. The network consists of two components: the DAG where each node corresponds to a random variable, while edges represent direct dependencies; the set of all parameters in the network. For instance, with regard to the random vector $X = (X_1, \ldots, X_H)$, a BN encodes the joint probability distribution of X by specifying: (i) the set of conditional independence statements by means of a DAG and (ii) the set of conditional probability distributions associated to the nodes of the graph. The joint probability distribution can be factorized according to the DAG as follows

$$P(X_1, \ldots, X_H) = \prod_{h=1}^{H} P(X_h | \mathrm{pa}(X_h))$$

where $P(X_h|\mathrm{pa}(X_h))$ is the probability distribution associated to node X_h given its parents $\mathrm{pa}(X_h)$, $h = 1, \ldots, H$. Given two nodes $X_{h'}$ and X_h, linked by an arrow pointing from to $X_{h'}$ to X_h, $X_{h'}$ is said parent of X_h, and X_h is said child of $X_{h'}$. We say that two vertices X_h and $X_{h'}$ are adjacent if there is an edge connecting them. Let $fa(X_h) = X_h \bigcup pa(X_h)$ then the clan of X_h is defined as $clan(X_h) = fa(X_h \bigcup ch(X_h))$ where $ch(X_h)$ is the set of all children of X_h.

The non identifiability of a statistical model for (X,Y,Z) implies that both the DAG and its parameters can not be estimated from the available sample information. Two kinds of uncertainty can be distinguished: 1) uncertainty regarding the DAG, that is the dependence structure between the variables of interest; 2) uncertainty regarding the network parameters, given the DAG, *i.e.* the joint probability factorization.

2.1 Uncertainty in the dependence structure

Let P be the joint probability distribution of (X,Y,Z) associated to the DAG $G_{XYZ} = (V,E)$ consisting of a set of vertices V and a set E of directed edges between pairs of nodes. Let us denote by $G_{XY} = (V_{XY}, E_{XY})$ and $G_{XZ} = (V_{XZ}, E_{XZ})$ the DAGs estimated on sample A and B, respectively. As in [5] G_{XY} and G_{XZ} are estimated subject to the condition that the association structure of the common variables X is fixed. In particular, the DAG G_X is estimated on the overall sample $A \bigcup B$. Given G_X, we proceed to estimate the association structure between (X,Y) and (X,Z) on the basis of sample data in A and B, respectively.

As far as P is concerned, unless special assumptions are made, one can only say that it lies in the class of all joint probability distributions for (X,Y,Z) satisfying the estimate collapsibility over Y and Z, respectively. Formally, we say that the joint probability distribution P is *estimate collapsible* over Z_t if

$$\widehat{P}(X,Y,Z\backslash\{Z_t\}) = \widehat{P}_{G_{XYZ\backslash\{Z_t\}}}(X,Y,Z\backslash\{Z_t\}). \quad (1)$$

That is, the estimate $\widehat{P}(X,Y,Z\backslash\{Z_t\})$ of $P(X,Y,Z\backslash\{Z_t\})$ obtained by marginalizing the maximum likelihood estimate (MLE) of $\widehat{P}(X,Y,Z)$ under the original DAG model (G_{XYZ},P) coincides with the MLE under the DAG model $(G_{XYZ\backslash\{Z_t\}})$, see [7]. Estimate collapsibility over a set Z is defined similarly. In terms of graphs a concept equivalent to estimate collapsibility is the c-removability. A vertex Z_t is c-removable from G_{XYZ} if any two vertices in $clan(Z_t)$ are adjacent, except when both vertices belong to $pa(Z_t)$. Further, the set $Z = (Z_1,\ldots,Z_T)$ is sequentially c-removable if all vertices in Z can be ordered so that they can be c-removed according to that ordering. An analogous condition is required for the estimate collapsibility over Y. The class of plausible joint distributions for (X,Y,Z) can be described as follows

$$\mathscr{P}_{XYZ} = \{P : \widehat{P}(X,Y) = \widehat{P}_{G_{XY}}(X,Y), \widehat{P}(X,Z) = \widehat{P}_{G_{XZ}}(X,Z)\} \quad (2)$$

or equivalently by using the graph of the model structure, the class can also be defined as the class of plausible DAGs G_{XYZ} where the variables Z and Y are removable, respectively. Formally

$$\mathscr{G}_{XYZ} = \{G_{XYZ} : Z \quad is \quad removable, \quad Y \quad is \quad removable\} \quad (3)$$

The most favorable case, that for instance happens under CIA, occurs when the class (2) is composed by a single joint probability distribution defined as $P(X,Y,Z) = P(X)P(Y|X)P(Z|X)$. In an equivalent manner, this means that the class (3) collapses into a single graph given by $G^{CIA}_{XYZ} = G_{XY} \bigcup G_{XZ}$ where Y and Z are d-separated by the set X. Note that such a network always belongs to the class (3). Under the CIA, both the dependence structure and the BN parameters are estimable from the sample data.

Clearly, when the CIA does not hold, in order to choose a plausible DAG from the class $\mathscr{G}_{XYZ}$, it is important to have extra-sample information on the dependence structure. This is generally available or can be elicited by experts. As stressed in [6], *qualitative dependencies among variables can often be asserted with confidence, whereas numerical assessments are subject to a great deal of hesitacy*. For example, for $K = T = 1$ an expert may willingly state that the variable Y is related to variable Z, however he/she would not provide a numeric quantification of this relationships.

2.2 Uncertainty in the parameter estimation

Suppose that a DAG G^*_{XYZ} has been selected from the class $\mathscr{G}_{XYZ}$. Let P^* the joint probability distribution associated to G^*_{XYZ}. According to G^*_{XYZ} the distribution P^* can be factorized into local probability distributions some of which can be estimated from the available sample information while other not. In the case of categorical variables, uncertainty is dealt with in [4] where parameters uncertainty is estimated

according to the maximum likelihood principle. The parameter estimate maximizing the likelihood function is not unique and the set of maximum likelihood estimates is called likelihood ridge.

Assume that, X_h, Y_k and Z_t are discrete rvs with I, J and L categories, respectively and that their joint distribution is multinomial with vector parameter $\theta^* = \{\theta^*_{ijl}\}$, for $i = 1,\ldots,I$, $j = 1,\ldots,J$, and $l = 1,\ldots,L$. Suppose that from the factorization of P^*, the unique parameter that can not be estimated is the joint probability $P(X_h, Y_k, Z_t)$. Analogously to (2), as far as θ^* is concerned, one can only say that it lies in the following set:

$$\Theta^* = \{\theta^* : \sum_l \theta^*_{ijl} = \widehat{\theta}_{ij.}, \sum_j \theta^*_{i.l} = \widehat{\theta}_{i.l}, \theta^*_{ijl} \geq 0, \sum_{ijl} \theta^*_{ijl} = 1\} \tag{4}$$

where

$$\widehat{\theta}_{ij.} = \frac{n^A_{ij.}}{n^A_{i..}} \frac{n^A_{i..} + n^B_{i..}}{n_A + n_B}, \quad \widehat{\theta}_{i.l} = \frac{n^A_{i.l}}{n^A_{i..}} \frac{n^A_{i..} + n^B_{i..}}{n_A + n_B} \tag{5}$$

are the marginal distribution ML estimates of (X_h, Y_k) and (X_h, Z_t) from samples A and B, respectively. The maximum of the observed likelihood in θ^*_{ijl} is not unique, all the distributions in the likelihood ridge are equally informative, given the data. For details, see [4].

In order to exclude some parameter vectors in Θ^* it is important to introduce constraints characterizing the phenomenon under study. These constraints can be defined in terms of structural zero ($\theta^*_{ijl} = 0$ for some (i, j, l)) and inequality constraints between pairs of distribution parameters ($\theta^*_{ijl} < \theta^*_{i'j'l'}$ for some $(i, j, l), (i', j', l')$). Their introduction is useful for reducing the overall parameter uncertainty. Clearly, the amount of reduction depends on the informativeness of the imposed constraints. The problem of the likelihood function maximization when constraints are imposed may be solved through a modified EM algorithm, see [13].

Example Suppose that an expert can elicit the association structure between the variables of interest $(X_1, Y_1, Y_2, Z_1, Z_2)$. The BN is reported in Figure 1. Note that, $Y = (Y_1, Y_2)$ is sequentially c-removable according to the ordering (Y_1, Y_2), and $Z = (Z_1, Z_2)$ is sequentially c-removable according to the ordering (Z_1, Z_2).

The joint distribution P can be factorized according to the graph as follows

$$P(X,Y,Z) = P(Z_1)P(X_1|Z_1)P(Y_1|X_1)P(Z_2|X_1,Z_1)P(Y_2|X_1,Z_1,Z_2) \tag{6}$$

The parameter $P(Y_2|X_1, Z_1, Z_2)$ can not be estimated from the sample available information in A and B. Nevertheless, such a distribution can be estimated following the approach described in Section 2.2 using an iterative procedure starting by $P(Z_1, Y_2|X_1)$ and ending with $P(Y_2, Z_2|X_1, Z_1)$.

Clearly, the larger is the number of directed edges between the components of Y and Z, the larger is the number of uncertain parameters needed to be estimated in the factorization of the joint distribution P^*.

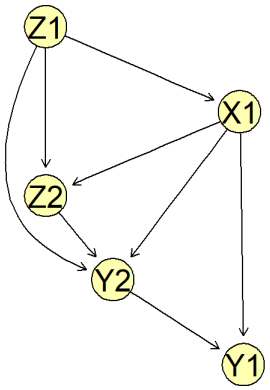

Fig. 1 BN for $(X_1, Y_1, Y_2, Z_1, Z_2)$

References

1. Conti, P.L., Marella, D., Scanu, M.: Uncertainty analysis in statistical matching.*Journal of Official Statistics*, 28, 69-88, (2012)
2. Conti, P.L., Marella, D., Scanu, M.: Statistical matching analysis for complex survey data with applications. Journal of the American Statistical Association. DOI:10.1080/01621459.2015.1112803, (2015)
3. D'Orazio, M., Di Zio, M., Scanu, M.: Statistical Matching: Theory and Practice. Chichester: Wiley, (2006)
4. D'Orazio, M., Di Zio, M., Scanu, M.: Statistical Matching for Categorical Data: Displaying Uncertainty and Using Logical Constraints. *Journal of Offcial Statistics*, 22, 137–157, (2006)
5. Endres, E., Augustin, T.: Statistical matching of Discrete Data by Bayesian Networks. JMLR: Workshop and Conference Proocedings, **52**, 159–170, (2016)
6. Geiger, D., Verma, T., Pearl, J.: Identifying Independence in Bayesian Networks. *Networks*, **20**, 507–534, (1990)
7. Kim, S.H., Kim, S.H.: A Note on Collapsibility in DAG Models of Contingency Tables. Scandinavian Journal of Statistics, **33**, 575-590, (2006)
8. Moriarity, C., Scheuren, F.: Statistical Matching: A Paradigm of Assessing the Uncertainty in the Procedure. Journal of Official Statistics, **17**, 407–422, (2001)
9. Okner, B.: Constructing a new data base from existing microdata sets: the 1966 merge file. Annals of Economic and Social Measurement, **1**, 325–342, (1972)
10. Pearl, J.: *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, (1998)
11. Rässler, S.: Statistical Matching: A Frequentist Theory, Practical Applications and Alternative Bayesian Approaches. Springer, New York, (2002)
12. Singh, A.C., Mantel, H., Kinack, M., Rowe, G.: Statistical Matching: Use of Auxiliary Information as an Alternative to the Conditional Independence Assumption. Survey Methodology, **19**, 59–79, (1993).
13. Winkler, W.E.: Improved Decision Rules in the Fellegi-Sunter Model of Record Linkage. Proceedings of the American Statistical Association, Section on Survey Research Methods, 274–279, (1993).