Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Partial identification in statistical matching with misclassification

Marco Di Zio [a], Barbara Vantaggi [b,*]

[a] *Istituto Nazionale di Statistica ISTAT, Rome, Italy*
[b] *Dip. S.B.A.I., "La Sapienza" Università di Roma, Italy*

ARTICLE INFO

Article history:
Received 11 July 2016
Received in revised form 20 December 2016
Accepted 21 December 2016
Available online xxxx

Keywords:
Data fusion
Synthetical matching
Consistency
Partial identifiability
Envelopes of probabilities
Inference

ABSTRACT

The main target of statistical matching is to make inference on variables observed in different sources by using information on common variables. The partial information implies generally that the model is not identifiable. The aim of this paper is to study the case when common variables are affected by a misclassification. The partially identifiable region, that is the class of probabilities extending the conditional probabilities obtained by the information in different sources, is determined. These regions are determined in the general case and under some specific restrictions on the misclassification mechanism. An application to real data is used to show in practice the results achieved in the paper.

1. Introduction

In the last years, integration of data coming from different sources is becoming an important way for gaining knowledge on phenomena under study exploiting as much as possible all the available information. This is particularly important in these times when the amount of information, gathered in different frameworks for different scopes, is increasing. Statistical inference based on integrated data resorts to different techniques according to the nature of the data sources to be integrated. A specific case is when data sources do not contain the same units. This generally happens when the data sources are sample surveys. Some applications exploit the informative power of the common variables, i.e., variables observed in both the data sources (we refer to two data sources without loss of generality), to make inference on the variables observed only in one of the two data sets, respectively. For instance, an application is concerned with the integration of surveys gathering information on household income and consumption expenditure, see [11]. This integration process is named differently according to the reference context of application, *statistical matching*, *data fusion*, or *data combination*, for more details see [7,27]. After the first applications, see e.g., [25] and [28], it was made clear that most of the methods used for inferring on the not jointly observed variables were based on their conditional independence given the common variables. Unfortunately, in this context, data at hand do not allow to test the conditional independence or the copula linking the variables observed only in one data set. Hence, the use of any copula, in particular the product one giving rise to the independence assumption, makes the statistical matching an hard and risky problem.

Recently, techniques to avoid the conditional independence assumption are introduced. Their aim is to estimate the set of values that make the unidentifiable parameters of interest consistent with the estimable parameters. These situations

* Corresponding author.
E-mail addresses: dizio@istat.it (M. Di Zio), barbara.vantaggi@sbai.uniroma1.it (B. Vantaggi).

http://dx.doi.org/10.1016/j.ijar.2016.12.015
0888-613X/

are common in econometrics, decision science, medicine, official statistics, see e.g. [12,13,18,20,27]. In official statistics this kind of approach is named *uncertainty analysis* [8], in other application domains it takes other names, for instance *partial identification* in econometrics and social sciences, see e.g., [22,24]. In a more theoretical framework it relates to the concept of *lower and upper probabilities*, see e.g. [5,10,33].

Statistical matching is based on the assumption that the common variables are observed in both the data sets. In surveys, variables can be affected by measurement errors, i.e., the observed value of the variable is possibly contaminated by an error. In case of categorical variables we refer to misclassification.

Approaches based on partial identification with misclassified data are available in literature, see e.g. [21,24]. Those papers aim at making inference on the misclassified variables.

In this paper we focus on statistical matching without imposing any fixed copula in favor of partial identification analysis. In particular, we deal with common variables affected by misclassification. When the common variable is misclassified according to the same mechanism in both sources the inferential problem on not jointly observed variables essentially falls in the usual statistical matching problem. In this case we describe partially identifiable region by means of lower and upper envelopes of the consistent probabilities. Then, we focus on the case where the misclassified variables are only those in one of the two data sets. We look for the partially identifiable region induced by the whole set of consistent probabilities: we provide a characterization of the lower and upper envelopes of this class. This problem is studied in the general context and under suitable restrictions on the misclassification mechanism. Although some of the probabilities characterizing the partially identifiable regions are generally estimated, we do not deal with the uncertainty due to sampling variability.

The motivating case comes from a study performed jointly by Istat (Italian National Institute of Statistics) and CEIS (Centre for Economic and International Studies) of University Tor Vergata. Istat gathers information on Health Conditions through the Multipurpose survey on Households, while CEIS manages a database named Health Search-SiSSI (HS) which contains patient-level data collected routinely by General Practitioners.

The Istat survey detects the amount of health services consumed, but not their costs. The latter information can be derived by the CEIS database. Those two Institutes started to perform a process of integration of these two data sources in order to combine information on prices/expenditures by socio-demographic characteristics.

The paper is structured as follows. Section 2 introduces the statistical matching problem with categorical variables. It focuses on the identification regions. The role of common variable is shown. In fact, it allows to go beyond the Fréchet–Hoeffding bounds.

Section 3 studies statistical matching when common variables are misclassified. In the general misclassification case the lower and upper sharp bounds coincide with Fréchet–Hoeffding bounds. Then, additional assumptions on the misclassification process are needed to reduce the space of consistent probabilities. Identification regions with respect to some misclassification models are outlined in Section 4. Section 5 illustrates estimation methods with an application to health conditions and health expenditures data in Italy. Conclusions are reported in Section 6.

2. Statistical matching

Let A and B be two data sources where a variable(s) X is observed in both the data sources, while a variable(s) Y is observed only in A and Z is observed only in B. A and B are independent, and the sets of observed units in the two samples do not overlap. The general aim of statistical matching is to infer on the probability distribution of (Y, Z) or $(Y, Z|X)$, (X, Y, Z). We restrict to the case where variables are categorical.

Notice that we have information to estimate the conditional probability distributions $P_{Y|X}$ from A and $P_{Z|X}$ from B and the marginal probability distribution P_X from both data sources. For instance, assuming the units in A and B are independent and identically distributed observations from the probability distributions P_{XY} and P_{XZ} respectively, and assuming that the categorical variables (Y, X) take categories in the set $\mathcal{X} \times \mathcal{Y}$ and (Z, X) in $\mathcal{X} \times \mathcal{Z}$, the maximum likelihood estimates of the probabilities $P_{Y|X}, P_{Z|X}, P_X$ can be easily obtained from the corresponding observed frequencies, e.g.,

$$\hat{p}(y|x) = \frac{n^A_{xy}}{n^A_x},\ \hat{p}(z|x) = \frac{n^B_{xz}}{n^B_x},\ \hat{p}(x) = \frac{n^A_x + n^B_x}{n^A + n^B},$$

where n^A_{xy} is the number of the units in A characterized by categories (x, y), n^B_{xz} the units in B assuming (x, z), n^A_x and n^B_x the absolute frequency of observations with category x, respectively, in A and B.

The aim of statistical matching is to compute all the probabilities P' on the algebra $\mathcal{B}(X, Y, Z)$ (generated by the random variables X, Y, Z) **consistent with** the conditional and marginal probability distributions $P_{Y|X}, P_{Z|X}, P_X$, that means all the probabilities having as marginal P_X on $\mathcal{X}$ and conditional distributions $P_{Y|X}$ on $\mathcal{Y} \times \mathcal{X}$ and $P_{Z|X}$ on $\mathcal{Z} \times \mathcal{X}$. For any such probability P' it follows trivially $P_{Z|X}(z|x) = \frac{\sum_{y\in\mathcal{Y}} P'(x,y,z)}{\sum_{y\in\mathcal{Y},z\in\mathcal{Z}} P'(x,y,z)}$, $P_{Y|X}(y|x) = \frac{\sum_{z\in\mathcal{Z}} P'(x,y,z)}{\sum_{y\in\mathcal{Y},z\in\mathcal{Z}} P'(x,y,z)}$ whenever $\sum_{y\in\mathcal{Y},z\in\mathcal{Z}} P'(x, y, z) = P_X(x) > 0$. When such probability P' exists the assessment $P_{Y|X}, P_{Z|X}, P_X$ is said to be **consistent**.

Consistency coincides with coherence [9] (see also [26,5]) and natural extension for precise assessments [33]. One of the main features of coherent assessments is their coherent extendibility, that leads in general to a class of coherent extensions. When the range $\mathcal{V}$ of the vector (X, Y, Z) coincides with the cartesian product $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ of the range of its components (i.e., $\mathcal{B}(X, Y, Z) = \mathcal{B}(X) \bigotimes \mathcal{B}(Y) \bigotimes \mathcal{B}(Z)$), there is at least one probability P' on $\mathcal{B}(X, Y, Z)$ consistent with the whole assessment

Please cite this article in press as: M. Di Zio, B. Vantaggi, Partial identification in statistical matching with misclassification, Int. J. Approx. Reason. (2017), http://dx.doi.org/10.1016/j.ijar.2016.12.015