Computational Statistics and Data Analysis 68 (2013) 311–325

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Uncertainty analysis for statistical matching of ordered categorical variables

Pier Luigi Conti[a], Daniela Marella[b,*], Mauro Scanu[c]

[a] Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Italy

[b] Dipartimento di Scienze della Formazione, Università "Roma Tre", Italy

[c] ISTAT, Italian National Statistical Institute, Roma, Italy

ARTICLE INFO

Article history:
Received 26 January 2012
Received in revised form 3 July 2013
Accepted 3 July 2013
Available online 15 July 2013

Keywords:
Statistical matching
Contingency tables
Structural zeros
Nonidentifiability
Uncertainty

ABSTRACT

The aim is to analyze the uncertainty in statistical matching for ordered categorical variables. Uncertainty in statistical matching consists in estimating a joint distribution by observing only samples from its marginals. Unless very restrictive conditions are met, observed data do not identify the joint distribution to be estimated, and this is the reason of uncertainty. The notion of uncertainty is first formally introduced, and a measure of uncertainty is then proposed. Moreover, the reduction of uncertainty in the statistical model due to the introduction of logical constraints is investigated and evaluated via simulation.

1. Introduction

In current practice, information needed for statistical analysis is frequently available in different data sources, each containing a subset of the variables of interest. This is the case of the *ecological inference* problem (cfr. King, 1997), where the main interest is estimating the joint probabilities of a contingency table, when the marginals are known from population counts. A typical example consists of contingency tables where rows and columns correspond to votes to political parties (known from election results) and race (known from population registers), respectively. One is then interested in the voters' behavior for different races. When rows' and columns' proportions are (separately) estimated from two different samples, then the problem becomes a genuine *statistical matching* problem. Another example of statistical matching problem is studied in Tonkin and Webber (2012), where the authors "statistically match expenditures for the Household Budget Survey (HBS) with income and material deprivation contained within EU Statistics on Income and Living Conditions (EU-SILC)".

Formally speaking, the statistical matching problem can be described as follows. Let (X, Y, Z) be a three-dimensional random variable (r.v.), and let A and B be two independent samples of n_A and n_B i.i.d. records from (X, Y, Z), respectively. Assume that the marginal (bivariate) (X, Y) is observed in A, and that the marginal (bivariate) (X, Z) is independently observed in B. The main goal of statistical matching, at a macro level, consists in estimating the joint distribution of (X, Y, Z). Such a distribution is not identifiable due to the absence of joint information on Z and Y given X, see D'Orazio et al. (2006b) and references therein, and Aluja-Banet et al. (2007) and Saporta (2002) for alternative approaches based on general multivariate analyses, as well as Conti et al. (2008) for an evaluation of statistical matching based on the matching noise.

[*] Corresponding author.
E-mail address: daniela.marella@uniroma3.it (D. Marella).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.07.004

Generally speaking, two approaches have been considered to ensure the identifiability of the joint distribution of (X, Y, Z):

- techniques based on the conditional independence assumption between Y and Z given X (CIA assumption), see, *e.g.*, Okner (1972) or on other kinds of identifiable models as independence of Y and Z given a latent variable (Kamakura and Wedel, 1997);
- techniques based on the 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 Singh et al. (1993).

Unfortunately, since CIA assumption is rarely met in practice (Rodgers, 1984; Sims, 1972), and external auxiliary information is hardly ever available, the lack of joint information on the variables of interest is the cause of *uncertainty* on the model of (X, Y, Z). In other terms, the sample information provided by A and B does not allow one to identify the joint distribution of (X, Y, Z), but only a *class of possible distributions* of (X, Y, Z) (*identification problem*). Such distributions are compatible with the available information, namely they may have generated the observed data. Note that, even if the marginal distributions of (X, Y) and (X, Z) were perfectly known, it will not be possible to draw certain conclusion on the model of (X, Y, Z).

The lack of identifiability of the distribution of (X, Y, Z) is due to the sampling mechanism that is actually unable to identify the conditional distribution of (Y, Z) given X. Hence, considering uncertainty about the conditional distribution of (Y, Z) given X is equivalent to consider uncertainty on the distribution of (X, Y, Z).

An important task, in this setting, consists in constructing a coherent measure that can reasonably quantify the uncertainty about the (estimated) model. From an operational point of view, a measure of uncertainty essentially quantifies how "large" is the class of models estimable on the basis of the available sample information. The smaller the measure of uncertainty, the smaller the class of estimated models. For a review on uncertainty in statistical matching providing a unified framework for the parametric and nonparametric approach, see Conti et al. (2012). A specific approach dealing with the case of dichotomous Y and Z is in Gilula et al. (2006), where also a Bayesian analysis of the uncertainty space is proposed.

In our setting, the main task consists in providing a precise definition of uncertainty on the (estimated) model, and in constructing a coherent measure that can reasonably quantify such an uncertainty.

The paper is organized as follows. In Section 2 the model uncertainty for ordered categorical variables is investigated. More specifically, model uncertainty is defined and uncertainty measures are introduced. In Section 3 the effect on model uncertainty due to the introduction of logical constraints is evaluated. In Section 4 estimators of uncertainty measures are proposed and their asymptotic behavior is studied. Finally, in Section 5 a simulation experiment is performed.

2. Uncertainty in statistical matching for ordered categorical variables

As stressed above, the statistical matching problem is characterized by uncertainty on the statistical model for the joint distribution of all variables of interest. Uncertainty in statistical matching can be viewed as a special case of estimation problems for general partially identifiable models, as in Manski (1995, 2003), and references therein. In those cases, estimation is not pointwise, but consists of ranges. Another example comes from the so called "disclosure problem" for confidentiality protection. In the case of categorical data (*e.g.* k-way contingency tables) upper and lower bounds on cell counts induced by a set of released margins play an important role in the disclosure limitation techniques; see Dobra and Fienberg (2001). In that context, for each suppressed cell we get an uncertainty interval called "feasibility interval". Such an interval should be sufficiently wide in order to ensure adequate confidentiality protection.

Uncertainty in statistical matching for parametric models, mainly in the multinormal case, is studied in Kadane (1978), Rubin (1986), Moriarity and Scheuren (2001) and Raessler (2002). The basic feature common to all those papers is that a multivariate distribution is not completely observed; only (some of) its marginals are observed. Sample observations cannot identify the statistical model generating data. As already said, this produces a kind of uncertainty about the model. Such an uncertainty is quantified by taking the range of an association parameter (e.g. the correlation coefficient in the normal bivariate case) between the non-jointly observed variables. In the case of categorical (non-ordinal) variables, uncertainty is dealt with in D'Orazio et al. (2006a).

Evaluation of the uncertainty in a statistical matching problem is also used for validation purposes. In particular, Raessler (2002) evaluates for multinormal models the length of the uncertainty intervals for unidentifiable parameters in order to define a measure of the reliability of estimates under CIA. "Small" uncertainty intervals imply that parameter estimates obtained under the different models compatible with the available sample information slightly differ from the ones estimated under the CIA.

The attention to the estimates under the CIA is justified by the fact that when (X, Y, Z) are multinormal, estimates under the CIA are the midpoint of the uncertainty interval of the inestimable parameters, usually the correlation coefficients between Y and Z. For other parametric models this property of the estimates under the CIA does not hold. Generalizations have been considered in D'Orazio et al. (2006a) in the case of categorical data, and in D'Orazio et al. (2006b) for general parametric models. They consider a maximum likelihood approach, and a related general measure of uncertainty given by the (hyper)volume of the likelihood ridge (in this case called "uncertainty space"). Formally, the parameter estimate which maximizes the likelihood function is not unique, the set of maximum likelihood estimates is called likelihood ridge. Statistical analysis of the likelihood ridge determines the middle point in the uncertainty interval for each parameter.