

Uncertainty in statistical matching under logical constraints: a nonparametric approach

Pier Luigi Conti, Daniela Marella and Mauro Scanu

Abstract Statistical matching consists in estimating the joint characteristics of two or more variables observed in two distinct sample surveys. The absence of joint information on the pair of variables of interest leads to uncertainty on the data generating model. The aim of this paper is to analyze the uncertainty in statistical matching in a nonparametric setting. More specifically, an overall measure of uncertainty for non identifiable models is introduced and the effect on model uncertainty due to the introduction of logical constraints is evaluated.

Key words: Combining data from different sources, Frechét classes, non identifiability

1 Introduction

Let (X, Z, Y) 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, Z, Y) . Assume that (X, Z, Y) are not completely observed in the samples. More specifically, we suppose that (X, Y) are observed in A and (X, Z) are observed in B . The aim of statistical matching is the reconstruction of a complete data set where each record includes not only the common variables X , but also the non-jointly observed variables Y and Z . Generally speaking, two solutions have been considered in the literature. In the

Pier Luigi Conti
Dipartimento di Statistica, Probabilità e Statistiche Applicate, Università di Roma “La Sapienza”,
e-mail: pierluigi.conti@uniroma1.it

Daniela Marella
Dipartimento di Scienze dell’Educazione, Università “Roma Tre”, Via del Castro Pretorio 20,
00185, Rome, Italy e-mail: dmarella@uniroma3.it

Mauro Scanu
Istituto Nazionale di Statistica, e-mail: scanu@istat.it

former, techniques based on the conditional independent assumption between Y and Z given X (CIA) are considered. Appropriateness of CIA is criticized in several papers. We quote, among the others [12, 10]. In the latter, techniques using external auxiliary information on the statistical relationship between Y and Z are considered, e.g. an additional file C where (X, Z, Y) are jointly observed is available, as in [13].

When either the CIA is a misspecified assumption or external auxiliary information is not available, the statistical model for (X, Z, Y) is not identifiable due to the absence of joint information on Z and Y given X [2]. In other words, the sample information in A and B is unable to discriminate among a set of plausible models for (X, Z, Y) . This problem leads to the third group of techniques which does not directly aim at the reconstruction of a complete data set, but addresses the so-called *identification problem* [6].

In a parametric setting the main consequence of the lack of identifiability is that some parameters of the model for (X, Z, Y) cannot be estimated on the basis of the available sample information. In practice, in a parametric setting the estimation problem cannot be “pointwise”. In fact, only ranges of values containing all the pointwise estimates obtainable by each model compatible with the available sample information can be detected; see [4, 11, 7, 9, 2]. Such intervals are *uncertainty intervals*.

The aim of this paper is to analyze the uncertainty in statistical matching in a nonparametric setting. The paper is organized as follows. In Section 2 the model uncertainty in a nonparametric setting is investigated, and the role of Fréchet classes is stressed.

In Section 3 an overall measure of uncertainty for non identifiable models is introduced. In Section 4 the effect on model uncertainty due to the introduction of logical constraints is evaluated.

2 Uncertainty in a nonparametric setting : the role of Fréchet bounds and their estimation

Uncertainty is defined as the set of probability distributions of (X, Y, Z) compatible with the sample information provided by A and B . In a parametric setting, such a set of probability distributions can be identified by a set of parameters. In a nonparametric setting, the assessment of uncertainty is more difficult since the data generating model is not identified by a finite number of parameters. Let the r.v.’s Z and Y be absolutely continuous and without loss of generality assume that the matching variable X is discrete. Also in this case, conditionally on X we have a set of plausible statistical models, the natural way to describe such a set of distributions consists in using the notion of Fréchet class.

From now on, we denote by $H(z, y|x)$ the d.f. of (Z, Y) given $X = x$, and by $G(z|x) = H(z, +\infty|x)$, $F(y|x) = H(+\infty, y|x)$ its marginal d.f.s (again, conditionally on X). Furthermore, let $Q(x) = P(X \leq x)$ be the marginal d.f. of X . Conditionally on X , the Fréchet class of all distribution functions $H(z, y|x)$ compatible with $G(z|x)$

and $F(y|x)$, is given by:

$$L^x(F(y|x), G(z|x)) \leq H(z, y|x) \leq U^x(F(y|x), G(z|x)) \quad (1)$$

where the bounds

$$\begin{aligned} L^x(F(y|x), G(z|x)) &= \max(G(z|x) + F(y|x) - 1, 0) \\ U^x(F(y|x), G(z|x)) &= \min(G(z|x), F(y|x)) \end{aligned}$$

are themselves joint d.f.s with marginal d.f.s $G(z|x)$ and $F(y|x)$.

All the d.f.s $H(z, y|x)$ belonging to the Fréchet class (1) are compatible with the available information, namely they may have generated the observed data. Note that even if $F(y|x)$, $G(z|x)$ were perfectly known, it will not be possible to draw well-definite conclusions on the model.

Taking the expectation w.r.t. the distribution of X , we obtain the unconditional Fréchet class

$$E_x[L^x(F(y|x), G(z|x))] \leq H(z, y) \leq E_x[U^x(F(y|x), G(z|x))] \quad (2)$$

Since X is a categorical variable, and if each category is observed in B as well as in A , the natural estimator of the Fréchet class (1) is given by

$$[\max(\widehat{G}_{n_B}(z|x) + \widehat{F}_{n_A}(y|x) - 1, 0), \min(\widehat{G}_{n_B}(z|x), \widehat{F}_{n_A}(y|x))] \quad (3)$$

where $\widehat{G}_{n_B}(z|x)$ and $\widehat{F}_{n_A}(y|x)$ are the empirical distribution functions (e.d.f.s) of $G(z|x)$ and $F(y|x)$, respectively. As a consequence, the unconditional Fréchet bounds (2) can be estimated by

$$\left[\sum_x \widehat{p}(x) \max(\widehat{G}_{n_B}(z|x) + \widehat{F}_{n_A}(y|x) - 1, 0), \sum_x \widehat{p}(x) \min(\widehat{G}_{n_B}(z|x), \widehat{F}_{n_A}(y|x)) \right] \quad (4)$$

where

$$\widehat{p}(x) = \left(\frac{n_{A,x} + n_{B,x}}{n_A + n_B} \right) \quad (5)$$

is an estimate of $P(X = x)$. According to the empirical likelihood approach as discussed by [8], the e.d.f.s $\widehat{G}_{n_B}(z|x)$ and $\widehat{F}_{n_A}(y|x)$ are nonparametric maximum likelihood estimates (NPMLE) of F and G respectively. As a consequence, for the invariance property of MLE in the nonparametric setting the estimators (3), (4) represent the NPMLEs of Fréchet classes (1), (2), respectively.

3 Measures of uncertainty for non identifiable models

As described in Section 2, the lack of joint information on the variables of interest is the cause of *uncertainty* on the model for (X, Y, Z) . The problem is that sample information provided by A and B is actually unable to discriminate among a set of plausible models for (X, Y, Z) . The statistical model is not identifiable on the basis of sample data. In this setting, the main task consists in constructing a measure that can reasonably quantify the uncertainty about the data generating model. Formally speaking, a measure of uncertainty quantifies how “large” is the class of models compatible with the available sample information.

An even more important goal is the quantification of the uncertainty when auxiliary information regarding the statistical model for (X, Y, Z) is available. This point is investigated in Section 4.

In view of the Fréchet bounds (1), the interval

$$[L^x(F(y|x), G(z|x)), U^x(F(y|x), G(z|x))] \quad (6)$$

summarizes the pointwise uncertainty (w.r.t. x, y, z) about the statistical model under consideration. As a pointwise measure of uncertainty, it is then intuitive to take the length of the interval (6), *i.e.*

$$I^x(F(y|x), G(z|x)) = U^x(F(y|x), G(z|x)) - L^x(F(y|x), G(z|x)).$$

We have a different measure of uncertainty for every triple x, y, z . A natural way to summarize all pointwise measures of uncertainty into an overall measure is to take the average length. Formally, if $T(x, y, z)$ is a weight function on $\mathbb{R}^3$ (*i.e.* a finite measure with total mass one), as an overall measure of uncertainty we may take the average length given by

$$\int_{\mathbb{R}^3} I^x(F(y|x), G(z|x)) dT(x, y, z). \quad (7)$$

In particular, by taking

$$dT(x, y, z) = dQ(x) d[F(y|x)G(z|x)],$$

the overall uncertainty measure (7) becomes

$$\begin{aligned} \Delta(F, G) &= \int_{\mathbb{R}} \left\{ \int_{\mathbb{R}^2} I^x(F(y|x), G(z|x)) d[F(y|x)G(z|x)] \right\} dQ(x) \\ &= \int_{\mathbb{R}} \Delta^x(F, G) dQ(x) \\ &= E_x[\Delta^x(F, G)] \end{aligned} \quad (8)$$

where

$$\Delta^x(F, G) = \int_{\mathbb{R}^2} [U^x(F(y|x), G(z|x)) - L^x(F(y|x), G(z|x))] d[F(y|x) dG(z|x)] \quad (9)$$

is the uncertainty measure about the considered statistical model, *conditionally to* $X = x$. Relationships (8), (9) show that the unconditional uncertainty measure can be expressed as a weighted mean of conditional uncertainty measures. Then, the larger Δ^x the more uncertain the data generating statistical model. When the uncertainty measure (9) is expressed in terms of copulas, it is equal to $1/6$ for any F and G , and for any x . The value $\Delta^x = 1/6$ represents the maximum uncertainty achieved when no external auxiliary information beyond knowledge of $F(y|x)$ and $G(z|x)$ is available. As a consequence, also the unconditional uncertainty measure (8) is equal to $1/6$.

4 The use of constraints for the reduction of model uncertainty

Suppose that auxiliary information in the form of logical constraints regarding the statistical model for (X, Z, Y) is available. As expected, such constraints imply a smaller degree of uncertainty since some models for (X, Z, Y) become illogical and must be excluded from the set of plausible distribution functions compatible with the partial distribution of (X, Y) and (X, Z) . This implies that the Fréchet bounds can be improved when additional information about the joint distribution function of (Z, Y) given X is known. As a consequence, the statistical model for the data becomes less uncertain.

There are several kinds of constraints. For instance, one area where they are frequently used is statistical data editing and imputation [5]. In the sequel we consider the logical constraints including restrictions on the support of the joint distribution of (Z, Y) given X .

4.1 Uncertainty under the constraint $a_x \leq f(Y, Z) \leq b_x$ given X

A class of constraints frequently occurring in practice is $a_x \leq f(Y, Z) \leq b_x$ given $X = x$, where $f(Y, Z)$ is a monotone function of Y (Z) for each Z (Y). Let $\gamma_y(\cdot)$ and $\delta_z(\cdot)$ be the inverse functions of $f(Y, Z)$ for fixed y and z , respectively. Without loss of generality, suppose that $f(y, z)$ is an increasing function of y for fixed z and a decreasing function of z for fixed y . Then, we have

$$\begin{aligned} H(z, y|x) &= P(Z \leq z, Y \leq y) = P(Z \leq z, Y \leq y, f(Y, Z) \leq b_x, f(Y, Z) \geq a_x|x) \\ &= P(Z \leq z, Z \leq \gamma_y(a_x), Y \leq y, Y \leq \delta_z(b_x)|x) \\ &= P(Z \leq (z \wedge \gamma_y(a_x)), Y \leq (y \wedge \delta_z(b_x))|x) \\ &= H(z \wedge \gamma_y(a_x), y \wedge \delta_z(b_x)|x) \end{aligned} \quad (10)$$

Hence, the Fréchet bounds (1) are given by

$$\begin{aligned} K_+^x(y, z) &= U^x(G(z \wedge \gamma_y(a_x)|x), F(y \wedge \delta_z(b_x)|x)) \\ &= \min(G(z \wedge \gamma_y(a_x)|x), F(y \wedge \delta_z(b_x)|x)) \\ &= \min(G(z|x), G(\gamma_y(a_x)|x), F(y|x), F(\delta_z(b_x)|x)) \end{aligned} \quad (11)$$

$$\begin{aligned} K_-^x(y, z) &= L^x(G(z \wedge \gamma_y(a_x)|x), F(y \wedge \delta_z(b_x)|x)) \\ &= \max(0, G(z \wedge \gamma_y(a_x)|x) + F(y \wedge \delta_z(b_x)|x) - 1) \\ &= \max(0, G(z|x) \wedge G(\gamma_y(a_x)|x) + F(y|x) \wedge F(\delta_z(b_x)|x) - 1) \end{aligned} \quad (12)$$

The conditional measure of uncertainty is then given by

$$\Delta_c^x(F, G) = \int_{\mathbb{R}^2} (K_+^x(y, z) - K_-^x(y, z)) d[F(y|x)G(z|x)] \quad (13)$$

where c represents the constraint $a_x \leq f(Y, Z) \leq b_x$. The corresponding unconditional measure of uncertainty is then given by

$$\Delta_c(F, G) = \sum_x p(x) \Delta_c^x(F, G). \quad (14)$$

As it clearly appears from (11), (12), the measure of uncertainty $\Delta_c^x(F, G)$ depends on the marginal d.f.s $F(y|x)$, $G(z|x)$. The same holds for $\Delta_c(F, G)$.

The uncertainty measures (13), (14) can be easily estimated on the basis of the available data. Using the notation introduced in Section 2, estimators of (11), (12) are given by

$$\widehat{K}_+^x(y, z) = \min \left\{ \widehat{F}_{n_A}(y|x), \widehat{F}_{n_A}(\delta_z(b_x)|x), \widehat{G}_{n_B}(z|x), \widehat{G}_{n_B}(\gamma_y(a_x)|x) \right\} \quad (15)$$

$$\begin{aligned} \widehat{K}_-^x(y, z) &= \max \left\{ 0, \min(\widehat{F}_{n_A}(y|x), \widehat{F}_{n_A}(\delta_z(b_x)|x)) \right. \\ &\quad \left. + \min(\widehat{G}_{n_B}(z|x), \widehat{G}_{n_B}(\gamma_y(a_x)|x)) - 1 \right\}. \end{aligned} \quad (16)$$

respectively. Hence, empirical estimators of (13), (14) can be defined as follows

$$\widehat{\Delta}_c^x = \int_{\mathbb{R}^2} (\widehat{K}_+^x(y, z) - \widehat{K}_-^x(y, z)) d[\widehat{F}_{n_A}(y|x) \widehat{G}_{n_B}(z|x)] \quad (17)$$

$$\widehat{\Delta}_c = \sum_x \widehat{p}(x) \widehat{\Delta}_c^x. \quad (18)$$

In [1], the consistency and asymptotic normality of the estimators (17), (18) is proved.

Example 1. Assume that there exists constants a_x, b_x such that $a_x \leq Y/Z \leq b_x$. For instance, in business surveys, X could be the type of activity, Y the total sales and Z

the number of employees. Let $f(Y, Z) = Y/Z$, $\gamma_y(a_x) = y/a_x$, $\delta_z(b_x) = b_x z$ and from the results (11), (12) we obtain the following Fréchet bounds (1)

$$\begin{aligned} K_+^x(y, z) &= \min \left(G \left(z \wedge \frac{y}{a_x} \middle| x \right), F(y \wedge b_x z | x) \right) \\ &= \min \left(G(z|x) \wedge G \left(\frac{y}{a_x} \middle| x \right), F(y|x) \wedge F(b_x z | x) \right) \\ &= \min \left(G(z|x), G \left(\frac{y}{a_x} \middle| x \right), F(y|x), F(b_x z | x) \right) \end{aligned} \quad (19)$$

$$\begin{aligned} K_-^x(y, z) &= \max \left(0, G \left(z \wedge \frac{y}{a_x} \middle| x \right) + F(y \wedge b_x z | x) - 1 \right) \\ &= \max \left(0, G(z|x) \wedge G \left(\frac{y}{a_x} \middle| x \right) + F(y|x) \wedge F(b_x z | x) - 1 \right) \end{aligned} \quad (20)$$

From (19), (20), it is possible to compute both conditional and unconditional uncertainty measures.

5 Remarks on possible extensions

The approach outlined so far can be extended to the multivariate case. Assume that $\mathbf{Y}$, $\mathbf{Z}$ are k -variate and m -variate r.v.s, respectively. Furthermore, conditionally on $X = x$, let $H(\mathbf{y}, \mathbf{z}|x)$ be the joint d.f. of $\mathbf{Y}$, $\mathbf{Z}$, and let $F(\mathbf{y}|x)$, $G(\mathbf{z}|x)$ be their marginals. Similarly to Section 2, the inequality

$$\max(0, F(\mathbf{y}|x) + G(\mathbf{z}|x) - 1) H(\mathbf{y}, \mathbf{z}|x) \leq \min(F(\mathbf{y}|x), G(\mathbf{z}|x))$$

still stands. In principle, with a price of a considerable complication in both notation and formal developments, the whole theory developed so far can be extended to this case. However, there is an important *caveat*. When k and/or m are “large”, the empirical estimators of Section 2 become largely inefficient. In such cases, it would be sensible to restrict the statistical model for the triplet $(X, \mathbf{Z}, \mathbf{Y})$, for instance by adopting a semiparametric approach. This problem will be the subject of a further study.

A second extension of interest is when the common r.v. X is not discrete, for instance when it possesses a density. In this case the empirical estimators of Section 2 cannot be used. A promising approach, to be developed, consists in estimating the conditional d.f.s $F(y|x)$, $G(z|x)$ by the methods proposed in [3].

References

1. Conti, P.L., Marella, D., Scanu, M.: How far from identifiability? A nonparametric approach to uncertainty in statistical matching under logical constraints. *Technical Report 22*, DSPSA, Università di Roma “La Sapienza” (2009).
2. D’Orazio, M., Di Zio, M., Scanu, M.: *Statistical Matching: Theory and Practice*. Wiley, New York (2006)
3. Hall, P., Wolffe, R.C.L., Yao, Q.: Methods for Estimating a Conditional Distribution Function. *Journal of the American Statistical Association* **94**, 154-163 (1999)
4. Kadane, J.B.: Some statistical problems in merging data files. In Department of Treasury, U.S. Government Printing Office: *Compendium of tax research*. Washington D.C., 159–179 (1978) (Reprinted in *Journal of Official Statistics* **17**, 423–433 (2001))
5. Luzi, O., Di Zio, M., Guarnera, U., Manzari, A., De Waal, T., Pannekoek, J., Hoogland, J., Tempelman, C., Hulliger, B., Kilchmann, D.: *Recommended practices for editing and imputation in cross-sectional business surveys*. Istat, CBS, SFSO, Eurostat (2007)
6. Manski, C.F.: *Identification problems in the social sciences*. Harvard University Press, Harvard (1995)
7. Moriarity, C., Scheuren, F.: Statistical Matching: A Paradigm of Assessing the Uncertainty in the Procedure. *Journal of Official Statistics* **17**, 407–422 (2001)
8. Owen, A.B.: *Empirical likelihood*. Chapman & Hall, London (2001)
9. Rässler, S.: *Statistical matching: A frequentist theory, practical applications and alternative bayesian approaches*. Springer Verlag, New York (2002)
10. Rodgers, W.L.: An evaluation of statistical matching. *Journal of Business and Economic Statistics* **2**, 91–102 (1984)
11. Rubin, D.B.: Statistical matching with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics* **4**, 87-94 (1986)
12. Sims, C.A.: Comments on: Constructing a new data base from existing microdata sets: the 1966 merge file, by Okner, B.A. *Annals of Economic and Social Measurements* **1**, 343–345 (1972)
13. 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).