Journal of Official Statistics, Vol. 28, No. 1, 2012, pp. 69–88

Uncertainty Analysis in Statistical Matching

Pier Luigi Conti[1], *Daniela Marella*[2], *and Mauro Scanu*[3]

Among the goals of statistical matching, a very important one is the estimation of the joint distribution of variables not jointly observed in a sample survey but separately available from independent sample surveys. The absence of joint information on the variables of interest leads to uncertainty about the data generating model. The present article reviews the concept of uncertainty in statistical matching and how to measure it by providing a unified framework for the parametric and nonparametric setting. Furthermore, the reduction of uncertainty due to the introduction of logical constraints is investigated and a simulation experiment is performed.

Key words: Data fusion; synthetical matching.

1. Introduction

A key issue in decision processes is the availability of information. Rich and detailed information will help a decision maker in making good decisions. Information may be gathered through a survey but the cost of implementing a new survey and the availability of information in archives or in other surveys may lead researchers to combine available information from different data sources. Statistical matching (sometimes called data fusion) aims at combining information available in distinct sample surveys referred to the same target population. Formally, let Y and Z be two random variables (rv). Statistical matching is defined as the estimation of the (Y, Z) joint distribution (e.g., its cumulative distribution function – cdf – $F(y, z)$) or of some of its parameters when:

- Y and Z are not jointly observed in a survey, but
- Y is observed in a sample A, of size n_A,
- Z is observed in a sample B, of size n_B,
- A and B are independent, and the sets of observed units in the two samples do not overlap (it is not possible to use record linkage),
- A and B both observe a set of additional variables X.

A detailed list of statistical matching applications is to be found in D'Orazio et al. (2006b) and Ridder and Moffitt (2007).

Generally speaking, two approaches have been considered. At first, e.g., Okner (1972), techniques based on the conditional independence assumption between Y and Z given X (CIA assumption) were considered. Appropriateness of CIA is discussed in several papers.

[1] University of Rome, La Sapienza, p. Aldo Moro 5, IT 00185 Rome, Italy. Email: pierluigi.conti@uniroma1.it
[2] University of Rome, Roma Tre, via Milazzo 11B, IT 00185 Rome, Italy. Email: dmarella@uniroma3.it
[3] ISTAT, via Cesare Balbo 16, 00184 Rome, Italy. Email: scanu@istat.it

We cite, among others, Sims (1972) and Rodgers (1984). The second group of techniques uses external auxiliary information on the statistical relationships 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)). Most of these approaches aim at making an inference as to the distribution of Y and Z by reconstructing, through imputation procedures, a synthetic file containing X, Y, Z data.

These approaches are actually theoretically justified when the joint probability distribution of the variables of interest in the population coincides with the probability distribution of the same variables in the synthetic data file, or at least when these two distributions are "very close". The discrepancy between the joint distribution of the variables of interest (a) in the population, and (b) in the synthetic data file, is usually referred to as matching noise (cf. Paass 1986). Attempts at evaluating the "closeness" of the empirical distribution of imputed data to the empirical distribution of "real" data have been performed in the literature, see D'Orazio et al. (2006b). In a nonparametric setting an important role is played by hot deck methods, as well as k-Nearest Neighbour (kNN, for short) methods. Their properties are studied in Marella et al. (2008) and in Conti et al. (2009a), where both theoretical and simulation results are obtained.

As a matter of fact, the CIA is usually a misspecified assumption, while external auxiliary information is hardly ever available. The lack of joint information on the variables of interest is the cause of *uncertainty* about the model of (X, Y, Z), since the sample information provided by A and B is actually unable to discriminate among a set of plausible models for (X, Y, Z). In other terms, the adopted statistical model is not identifiable on the basis of sample data. Hence, a third group of techniques that does not directly aim at reconstructing a complete data set is introduced. This group of techniques addresses the so-called identification problem. The main consequence of the lack of identifiability is that some parameters of the model cannot be estimated on the basis of the available sample information. For instance, in a parametric setting, instead of point estimates, one can only reasonably construct sets of "possible estimates", compatible with what can be actually estimated. These sets (usually intervals) formally provide a representation of uncertainty about the model parameters.

In this setting, the main task 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.

This article aims at reviewing uncertainty in statistical matching providing a unified framework for the parametric and nonparametric approach. More specifically, in Section 2 model uncertainty is defined and uncertainty measures are introduced. Furthermore, several examples are illustrated for both parametric and nonparametric settings. Section 3 shows the effect on model uncertainty due to the introduction of logical constraints (frequently used in imputation and editing). Finally, in Section 4 a simulation experiment is performed.

2. What is Uncertainty in Statistical Matching

Let (X, Y, Z) be a trivariate random variable with joint (cumulative) distribution function (df) $M(x, y, z)$. Denote further by $H(y, z|x)$ the joint df of (Y, Z) conditionally on X, $F(y|x)$

and $G(z|x)$ the marginal d.f.s of Y and Z conditionally on X, respectively, and $Q(x)$ the marginal df of X.

Let $\mathcal{M} = \{M(x, y, z)\}, \mathcal{H}^x = \{H(y, z|x)\}, \mathcal{F}^x = \{F(y|x)\}, \mathcal{G}^x = \{G(z|x)\}, \mathcal{Q} = \{Q(x)\}$ be the sets of df's in which M, H, F, G and Q lie. Knowledge of a specific distribution in $\mathcal{M}$ is equivalent to knowledge of the corresponding distributions in $\mathcal{H}^x$ and $\mathcal{Q}$.

We further assume that the observation mechanism allows one to identify the classes $\mathcal{F}^x$, $\mathcal{G}^x$ and $\mathcal{Q}$, but not the class $\mathcal{H}^x$, unless special assumptions are made. The most important is the conditional independence assumption, under which we have $\mathcal{H}^x = \{F(y|x)G(z|x); F \in \mathcal{F}^x, G \in \mathcal{G}^x\}$, i.e., $\mathcal{H}^x = \mathcal{F}^x \times \mathcal{G}^x$.

We distinguish two main cases:

1. Parametric case: the classes $\mathcal{H}^x$ and $\mathcal{Q}$ are indexed by real or vector parameters. As a consequence we may write:

$$\mathcal{Q} = \{Q_\xi(x), \xi \in \Xi\}$$

$$\mathcal{F}^x = \{F_\phi(y|x), \phi \in \Phi^x\}$$

$$\mathcal{G}^x = \{G_\psi(z|x), \psi \in \Psi^x\}$$

$$\mathcal{H}^x = \{H_\gamma(y, z|x), \gamma \in \Gamma^x\}$$

2. Nonparametric case: the above classes $\mathcal{H}^x$ and $\mathcal{Q}$ cannot be indexed by real or vector parameters.

Given the (known) marginal distributions $F(y|x)$ and $G(z|x)$, uncertainty is defined as the set of probability distributions of the random vector $(Y, Z|X)$ compatible with $F(y|x)$ and $G(z|x)$. Formally

$$H^-(y, z|x) \leq H(y, z|x) \leq H^+(y, z|x) \tag{1}$$

where

$$H^+(y, z|x) = \sup_{H \in \mathcal{H}^x} \{H(y, z|x) : H(y, +\infty|x) = F(y|x), H(+\infty, z|x) = G(z|x)\};$$

$$H^-(y, z|x) = \inf_{H \in \mathcal{H}^x} \{H(y, z|x) : H(y, +\infty|x) = F(y|x), H(+\infty, z|x) = G(z|x)\}.$$

In the parametric case, we may write

$$H^+(y, z|x) = H^+(y, z|x; \psi, \phi)$$

$$= \sup_{\gamma \in \Gamma^x} \{H(y, z|x; \gamma); H(y, +\infty|x; \gamma) = F(y|x; \phi), H(+\infty, z|x; \gamma) = G(z|x; \psi)\}$$

$$H^-(y, z|x) = H^-(y, z|x; \psi, \phi)$$

$$= \inf_{\gamma \in \Gamma^x} \{H(y, z|x; \gamma); H(y, +\infty|x; \gamma) = F(y|x; \phi), H(+\infty, z|x; \gamma) = G(z|x; \psi)\}$$

The df's belonging to the class $\mathcal{H}^x$ are compatible with the available information, namely they may have generated the observed data.

The interval (1) summarizes the pointwise uncertainty about the statistical model for every triple (x, y, z). It is intuitive to take the length of such an interval as a pointwise

measure of uncertainty. The wider the interval of extremes $[H^-(\cdot,\cdot|x), H^+(\cdot,\cdot|x)]$ the more uncertain the statistical model generating the data w.r.t. (x, y, z). Of course, if the model is identifiable, then the interval reduces to a single point, with length zero, and there is no uncertainty at all.

In this case, it is possible to compute the following measures of uncertainty. The first one is based on computing for each point (x, y, z) the distance between the extrema $H^-(y, z|x), H^+(y, z|x)$.
Formally

$$\Delta^{yz|x} = H^+(F(y|x), G(z|x)) - H^-(F(y|x), G(z|x)) \tag{2}$$

In order to summarize the differences in (2) into an overall measure of uncertainty we may take the average length

$$\Delta = \int_{\mathbb{R}^3} \Delta^{yz|x} dT(x, y, z)$$

where $T(x, y, z)$ is a weight function on $\mathbb{R}^3$, i.e., a measure having total mass 1. A "natural" choice consists in taking

$$dT(x, y, z) = dF(y|x)dG(z|x)dQ(x)$$

This distribution is "natural" because: i) it is the simplest choice given the available df's $F(y|x)$, $G(z|x)$, $Q(x)$ and makes the integral in Δ easily computable in many cases; ii) among all the possible associations between Y and Z, we consider a neutral position, i.e., we do not give preference to any specific positive or negative association. Hence, a conditional measure of uncertainty is

$$\Delta^x = \int_{\mathbb{R}^2} \Delta^{yz|x} dF(y|x)dG(z|x) \tag{3}$$

As a matter of fact, integrating (3) with respect to X, we obtain again the overall measure of uncertainty Δ

$$\Delta = \int_{\mathbb{R}} \Delta^x dQ(x) \tag{4}$$

Relationships (3), (4) show that the unconditional uncertainty measure (4) can be expressed as a weighted mean of conditional uncertainty measures (3). Then, the larger Δ^x the more uncertain the data generating statistical model.

If interest is only in the joint distribution of the (not jointly observed) variables (Y, Z), it is possible also to consider the unconditional Fréchet class

$$(E_x[H^-(F(y|x), G(z|x))], E_x[H^+(F(y|x), G(z|x))]) \tag{5}$$

A unique number can be obtained by using an appropriate weight function $T(y, z)$.

The uncertainty measure depends on the marginal df's F and G, that can be estimated by the available sample information. The asymptotic properties of the estimators of F and G (both in the parametric and nonparametric cases) determine the asymptotic properties of Δ. The limit distribution of Δ allows to construct confidence intervals and hypothesis tests for the uncertainty measure.

Example 1. Nonparametric Case – Assume that $\mathcal{H}^x$ cannot be indexed by real or vector parameters, so that $\mathcal{H}^x$ is the set of all bivariate df's having marginal df's $F(y|x)$ and $G(z|x)$. Denote further by $L(u,v) = \max(0, u+v-1)$, and by $U(u,v) = \min(u,v)$. Then the Fréchet inequalities

$$L(F(y|x), G(z|x)) \leq H(y,z|x) \leq U(F(y|x), G(z|x)) \tag{6}$$

hold true, $L(F(y|x), G(z|x))$ and $U(F(y|x), G(z|x))$ being bivariate df's corresponding to maximal negative and maximal positive association between Y and Z (given X). Furthermore

$$U(F(+\infty|x), G(z|x)) = L(F(+\infty|x), G(z|x)) = G(z|x)$$

and

$$U(F(y|x), G(+\infty|x)) = L(F(y|x), G(+\infty|x)) = F(y|x)$$

namely $L(F(y|x), G(z|x))$ and $U(F(y|x), G(z|x))$ both possess marginal df's $F(y|x)$ and $G(z|x)$.

As a consequence of Fréchet inequalities (6), we have

$$H^-(y,z|x) = L(F(y|x), G(z|x))$$

and

$$H^+(y,z|x) = U(F(y|x), G(z|x))$$

With such a choice, the overall uncertainty measure (4) becomes

$$\begin{aligned}\Delta &= \int_{\mathbb{R}} \left\{ \int_{\mathbb{R}^2} [U(F(y|x), G(z|x)) - L(F(y|x), G(z|x))] d[F(y|x) dG(z|x)] \right\} dQ(x) \\ &= \int_{\mathbb{R}} \Delta^x dQ(x) \quad = E_x[\Delta^x]\end{aligned} \tag{7}$$

where

$$\Delta^x = \int_{\mathbb{R}^2} [U(F(y|x), G(z|x)) - L(F(y|x), G(z|x))] dF(y|x) dG(z|x) \tag{8}$$

is the uncertainty measure about the considered statistical model, conditionally on $X = x$.

Further results and simplifications can be obtained when $F(y|x)$ and $G(z|x)$ are continuous df's, so that $R_x = F(Y|x)$ and $S_x = G(Z|x)$ both have uniform distribution in [0,1] for every x. From the Sklar theorem (Nelsen 1999), the joint df H may be uniquely represented as

$$H(y,z|x) = C_x(F(y|x), G(z|x)) \tag{9}$$

where $C_x(u,v)$ is the copula, i.e., the joint d.f. of R_x and S_x. In this case, the conditional

uncertainty measure simplifies to

$$\begin{aligned}\Delta^x &= \int_{[0,1]^2} [U(r,s) - L(r,s)]d[r\ s] \\ &= \int_0^1\int_0^1 \{\min(r,s) - \max(0, r+s-1)\}dr\ ds\end{aligned} \tag{10}$$

In this setting, it is straightforward to show that the uncertainty measure (10) 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 the knowledge of the marginals $F(y|x)$ and $G(z|x)$ is available. As a consequence, also the unconditional uncertainty measure (7) takes the value 1/6.

Example 2. Contingency Tables – Assume that, given a discrete rv X with I categories, Y and Z are discrete rv's with J and K categories, respectively, not necessarily ordered. In this case, the whole theory developed so far can still be applied by simply replacing the d.f.s with probability functions. With no loss of generality, the symbols $i = 1, \ldots, I$, $j = 1, \ldots, J$, and $k = 1, \ldots, K$, denote the categories taken by X, Y and Z, respectively.

Let $\gamma_{jk|i}$ be the probability that $(Y = j, Z = k|X = i)$, with marginals $\phi_{j|i}$ representing the probability of the event $(Y = j|X = i)$ and $\psi_{k|i}$ representing the probability of the event $(Z = k|X = i)$. Since

$$L(\phi_{j|i}, \psi_{k|i}) \leq \gamma_{j,k|i} \leq U(\phi_{j|i}, \psi_{k|i}) \tag{11}$$

the conditional uncertainty measure turns out to be equal to

$$\Delta^{x=i} = \sum_{j=1}^{J}\sum_{k=1}^{K} \{U(\phi_{j|i}, \psi_{k|i}) - L(\phi_{j|i}, \psi_{k|i})\}\phi_{j|i}\psi_{k|i}$$

and the overall uncertainty measure is

$$\Delta = \sum_{i=1}^{I} \Delta^{x=i}\xi_i$$

where ξ_i represents the probability of the event $(X = i)$. Sharper results are obtained when the categories taken by (X, Y, Z) are ordered. For the sake of simplicity, we use the customary order for natural numbers. In this case, the cumulative df's are

$$H_{j,k|i} = \sum_{y=1}^{j}\sum_{z=1}^{k} \gamma_{yz|i}, \quad j = 1, \ldots, J,\ k = 1, \ldots, K, i = 1, \ldots, I$$

$$F_{j|i} = \sum_{y=1}^{j} \phi_{y|i}, \quad j = 1, \ldots, J,\ i = 1, \ldots, I$$

$$G_{k|i} = \sum_{z=1}^{k} \psi_{z|i}, \quad k = 1, \ldots, K,\ i = 1, \ldots, I$$

Using the same arguments as in Example 1, the inequalities

$$L(F_{j|i}, G_{k|i}) \leq H_{j,k|i} \leq U(F_{j|i}, G_{k|i}) \tag{12}$$

hold. Note that inequalities (12) imply that

$$\gamma^{-}_{jk|i} \leq \gamma_{jk|i} \leq \gamma^{+}_{jk|i} \tag{13}$$

where

$$\gamma^{-}_{jk|i} = L(F_{j|i}, G_{k|i}) - L(F_{j-1|i}, G_{k|i}) - L(F_{j|i}, G_{k-1|i}) + L(F_{j-1|i}, G_{k-1|i})$$

$$\gamma^{+}_{jk|i} = U(F_{j|i}, G_{k|i}) - U(F_{j-1|i}, G_{k|i}) - U(F_{j|i}, G_{k-1|i}) + U(F_{j-1|i}, G_{k-1|i})$$

Then, it is not difficult to realize that

$$\gamma^{-}_{jk|i} \geq L(\phi_{j|i}, \psi_{k|i})$$

$$\gamma^{+}_{jk|i} \leq U(\phi_{j|i}, \psi_{k|i})$$

so that inequalities (13) are sharper than (11). At any rate, the conditional uncertainty measure is

$$\Delta^{x=i} = \sum_{j=1}^{J} \sum_{k=1}^{K} \{U(F_{j|i}, G_{k|i}) - L(F_{j|i}, G_{k|i})\} \phi_{j|i} \psi_{k|i} \tag{14}$$

and the unconditional uncertainty measure is

$$\Delta = \sum_{i=1}^{I} \Delta^{x=i} \xi_i \tag{15}$$

In contrast with what was found in Example 1, Δ is not equal to 1/6, since the uncertainty measure depends on the marginal probabilities of $Y|X$ and $Z|X$.

Example 3. Multinormal Distribution – Let X, Y, Z, be jointly multinormally distributed, with mean vector and covariance matrix equal to

$$\mu = \begin{bmatrix} \mu_x \\ \mu_y \\ \mu_z \end{bmatrix}, \Sigma = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} & \sigma_{xz} \\ \sigma_{xy} & \sigma_y^2 & \sigma_{yz} \\ \sigma_{xz} & \sigma_{yz} & \sigma_z^2 \end{bmatrix} \tag{16}$$

respectively. All bivariate marginals are still multinormal, with mean vectors and covariance matrices easily obtained from (16). Furthermore, conditionally on X, (Y, Z) do have joint bivariate normal distribution, with mean vector and covariance matrix given by

$$\mu_{yz|x} = \begin{bmatrix} \mu_y + \beta_{y/x}(x - \mu_x) \\ \mu_z + \beta_{z/x}(x - \mu_x) \end{bmatrix},$$

$$\Sigma_{yz|x} = \begin{bmatrix} \sigma_y^2\left(1 - \rho_{xy}^2\right) & \sigma_y \sigma_z(\rho_{yz} - \rho_{xy}\rho_{xz}) \\ \sigma_y \sigma_z(\rho_{yz} - \rho_{xy}\rho_{xz}) & \sigma_z^2\left(1 - \rho_{xz}^2\right) \end{bmatrix} \tag{17}$$

where

$$\beta_{y|x} = \frac{\sigma_{xy}}{\sigma_x^2}, \beta_{z|x} = \frac{\sigma_{xz}}{\sigma_x^2}$$

are the regression coefficients of Y, Z w.r.t. X, respectively, and

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}, \rho_{xz} = \frac{\sigma_{xz}}{\sigma_x \sigma_z}, \rho_{yz} = \frac{\sigma_{yz}}{\sigma_y \sigma_z}$$

are the correlation coefficients between (X, Y), (X, Z), (Y, Z), respectively.

The only unidentified parameter is ρ_{yz}, the correlation coefficient between Y and Z. From (17) it is immediate to see that, given ρ_{xy} and ρ_{xz}, ρ_{yz} ranges in the interval

$$\left[\rho_{xy}\rho_{xz} - \sqrt{\left(1-\rho_{xy}^2\right)\left(1-\rho_{xz}^2\right)}, \rho_{xy}\rho_{xz} + \sqrt{\left(1-\rho_{xy}^2\right)\left(1-\rho_{xz}^2\right)}\right] \tag{18}$$

From the Slepian (1962) inequality, the joint d.f. of Y, Z given X is an increasing function of the correlation coefficient between Y and Z given X:

$$\rho_{yz|x} = \frac{\rho_{yz} - \rho_{xy}\rho_{xz}}{\sqrt{\left(1-\rho_{xy}^2\right)\left(1-\rho_{xz}^2\right)}}$$

i.e., it turns out to be a monotone function of ρ_{yz}. In other words, the conditional d.f.s $H(y, z|x)$ are totally ordered on the basis of ρ_{yz}. The case $\rho_{yz} = \rho_{xy}\rho_{xz} - \sqrt{(1-\rho_{xy}^2)(1-\rho_{xz}^2)}$ corresponds to $\rho_{yz|x} = -1$, so that conditionally on X, Y is a linear decreasing function of Z (and *vice versa*). As a consequence, $H^-(y, z|x) = \max(0, F(y|x) + G(z|x) - 1)$, as in Example 1. Similarly, the case $\rho_{yz} = \rho_{xy}\rho_{xz} + \sqrt{(1-\rho_{xy}^2)(1-\rho_{xz}^2)}$ corresponds to $\rho_{yz|x} = 1$, i.e., to $H^+(y, z|x) = \min(F(y|x), G(z|x))$, again as in Example 1. As a consequence, the computation of uncertainty measures parallels the nonparametric case.

The assumption of multinormal distribution of (X, Y, Z) is largely debated in the statistical matching literature, since the seminal paper by Kadane (1978). Useful discussions and advances are to be found in Moriarity and Scheuren (2001), Rässler (2002), Kiesl and Rässler (2008).

At first glance, results of Examples 1, 3 seem to be contradictory. In the multinormal case, uncertainty affects the correlation coefficient ρ_{yz}. Uncertainty bounds for the correlation coefficients are determined by the condition that the correlation matrix of (X, Y, Z) must be positive definite. In this case, the uncertainty space restricts continuously to a single value when ρ_{xy} or ρ_{xz} go continuously to 1 (or -1). The width of the Fréchet class behaves differently. The lower and upper bounds of the Fréchet class only depend on the marginal df's of Y and Z (given X). The copula transformation makes such marginals uniform in the interval (0, 1), and hence the whole Fréchet class is always mapped on the same subset of the square $(0, 1)^2$. This is why the uncertainty measure remains constantly equal to 1/6. A discontinuity happens if either Y or Z is fully determined by X. In this case, the square $(0, 1)^2$ collapses to the segment (0, 1), and the uncertainty measure becomes equal to zero.

Example 4. Skew-normal Distribution – Suppose that the joint distribution of (Y, Z), conditionally on X, does possess a bivariate skew-normal distribution with parameters $(\delta, \Omega, \alpha, \tau)$, (see Capitanio et al. 2003), where

$$\Omega = \begin{pmatrix} \omega_{yy} & \omega_{yz} \\ \omega_{yz} & \omega_{zz} \end{pmatrix}$$

is a 2×2 matrix. In particular, ω_{yz} is the association parameter between Y and Z, given X, and ranges in the interval $(-1, 1)$. Of course, this means that also the distribution of (X, Y, Z) does possess (extended) skew-normal distribution as shown in Capitanio et al. (2003). As ω_{yz} tends to $+1$, $(Y, Z)|X$ tends to be perfectly positively correlated, so that $H^+(y, z|x) = \min(F(y|x), G(z|x))$, again as in Example 1.

However, as ω_{yz} tends to -1, Y and Z (given X) do not achieve the situation of perfect negative correlation. Hence, $H^-(y, z|x) \geq \max(0, F(y|x) + G(z|x))$.

Example 5. Farlie-Gumbel-Morgenstern Distribution – Suppose that X is distributed as a uniform in $[0, 1]$. Conditionally on X, let Y and Z be marginally distributed as uniform distributions in $[0, x]$. Furthermore, the joint d.f. of (Y, Z) given X is:

$$H(y, z|x) = \frac{y}{x}\frac{z}{x}\left[1 + \alpha\left(1 - \frac{y}{x}\right)\left(1 - \frac{z}{x}\right)\right] \tag{19}$$

where $-1 \leq \alpha \leq 1$ (otherwise (19) is not a distribution function), $0 \leq y \leq x$, $0 \leq z \leq x$. The maximal value $H^+(y, z|x)$ is obtained when $\alpha = 1$, and is equal to

$$H^+(y, z|x) = \frac{y}{x}\frac{z}{x}\left[1 + \left(1 - \frac{y}{x}\right)\left(1 - \frac{z}{x}\right)\right]$$

which is strictly smaller than $\min(F(y|x), G(z|x))$. On the other hand, the minimal value $H^-(y, z|x)$ is obtained when $\alpha = -1$, and is equal to

$$H^-(y, z|x) = \frac{y}{x}\frac{z}{x}\left[1 - \left(1 - \frac{y}{x}\right)\left(1 - \frac{z}{x}\right)\right]$$

which is strictly greater than $\max(0, F(y|x) + G(z|x))$.

3. Reducing Uncertainty

When auxiliary information in the form of logical constraints regarding the statistical model for (Y, Z) or $(Y, Z|X)$ is available, some models for (X, Y, Z) become illogical and must be excluded from the set of plausible distribution functions. As a consequence, the statistical model for the data becomes less uncertain. This kind of information is frequently used, for instance, in imputation and editing, see Luzi et al. (2007). The introduction of such constraints may complicate the estimation process, because there could be no probability distributions satisfying both the logical constraints and lying in the set of all estimated bivariate distributions.

There are essentially two types of constraints:

1. constraints on the values of the parameters;
2. constraints on the support of (Y, Z) or $(Y, Z|X)$.

The first kind of constraints is essentially used in the parametric case. They have been largely studied in D'Orazio et al. (2006a) when (X, Y, Z) are categorical. The constraints are in terms of inequalities between the cell probabilities of the contingency table (Y, Z) or $(Y, Z)|X$. In the normal case, these constraints can be used when information on the correlation coefficient ρ_{yz} or equivalently $\rho_{yz|x}$ is available, see Rässler (2002).

The second kind of constraints has been mainly used in the parametric-multinomial case (D'Orazio et al. 2006a). These constraints have been defined in terms of structural zeros on the joint (Y, Z) distribution. In Vantaggi (2008) a different estimation method to deal with the case of constraints in form of structural zeros in a coherent probability framework is proposed.

The use of structural zeros for reducing uncertainty has been confined to the case of categorical rv's. In fact, for continuous rv's the introduction of structural zeros corresponds to the restriction of the (Y, Z) or $(Y, Z|X)$ support to a subset of the Cartesian product of the Y and Z supports, i.e., (Y, Z) is fully concentrated on a set of points $(y, z) \in \mathbb{R}^2$. The relationship between the original marginal parameters and the joint (Y, Z) distribution is heavily dependent on the nature of this constraint and does not allow easy generalizations. For instance, even if both $(Y|X)$ and $(Z|X)$ follow a normal distribution, the introduction of a structural zero will prevent $(Y, Z|X)$ from being bivariate normal. As a matter of fact, structural zeros can be easily used in a nonparametric approach, and the estimation of uncertainty can be performed using Equations (3) and (4).

For the sake of simplicity, assume that Y and Z are continuous r.v.'s, that X is a discrete r.v., and that the support constraints have the following shape: $a_x \leq Y/Z \leq b_x$ given X. An example of this kind of constraint happens in household surveys, when X is a household socio-economic character, Y the household consumption and Z the household income. Conditionally on X, the inequality $a_x \leq Y/Z \leq b_x$ holds. Using the notation $a \wedge b = \min(a, b)$, it is not difficult to see that the pair of inequalities

$$L^x\left(G\left(z \wedge \frac{y}{a_x}|x\right), F(y \wedge b_x z|x)\right) \leq H(z, y|x) \leq U^x\left(G\left(z \wedge \frac{y}{a_x}|x\right), F(y \wedge b_x z|x)\right)$$

holds, where

$$\begin{aligned} U^x\left(G\left(z \wedge \frac{y}{a_x}|x\right), F(y \wedge b_x z|x)\right) &= min\left(G\left(z \wedge \frac{y}{a_x}|x\right), F(y \wedge b_x z|x)\right) \\ &= min\left(G(z|x) \wedge G\left(\frac{y}{a_x}|x\right), F(y|x)g \wedge F(b_x z|x)\right) \\ &= min\left(G(z|x), G\left(\frac{y}{a_x}|x\right), F(y|x), F(b_x z|x)\right) \end{aligned} \tag{20}$$

$$\begin{aligned} L^x\left(G\left(z \wedge \frac{y}{a_x}|x\right), F(y \wedge b_x z|x)\right) &= max\left(0, G\left(z \wedge \frac{y}{a_x}|x\right) + F(y \wedge b_x z|x) - 1\right) \\ &= max\left(0, G(z|x) \wedge G\left(\frac{y}{a_x}|x\right) + F(y|x) \wedge F(b_x z|x) - 1\right) \end{aligned} \tag{21}$$

With obvious notation, according to (8) the conditional uncertainly measure is then

$$\Delta_c^x = \int_{\mathbb{R}^2} \left(U^x \left(G\left(z \wedge \frac{y}{a_x} | x \right), F(y \wedge b_x z | x) \right) - L^x \left(G\left(z \wedge \frac{y}{a_x} | x \right), F(y \wedge b_x z | x) \right) \right) dF(y|x) dG(z|x) \tag{22}$$

where c denotes the constraint $a_x \leq Y/Z \leq b_x$. The unconditional uncertainty measure is easily obtained from (22) by averaging w.r.t. the distribution of X.

As remarked above, in this case both conditional and unconditional uncertainty measures are strictly smaller than 1/6. In other words, the constraint $a_x \leq Y/Z \leq b_x$ reduces the uncertainty about the statistical model.

Clearly, the reduction in the model uncertainty depends on how informative the imposed constraints are. In some circumstances, the logical constraints can be so informative that the Fréchet class reduces to a unique distribution. This happens when Z can be exactly predicted by X (or alternatively by (X, Y)) due to a deterministic relationship between X and Z (or (X, Z, Y)). The effects on model uncertainty due to the introduction of logical constraints in a nonparametric setting are investigated in Conti et al. (2009b). Moreover, in Conti et al. (2009b) estimators of conditional and unconditional measures of uncertainty under logical constraints are proposed and their consistency and asymptotic normality are proved.

4. A Simulation Study

In this section we perform a simulation experiment in order to evaluate the effects on model uncertainty due to the introduction of logical constraints and in order to investigate the asymptotic behavior of the uncertainty measures proposed in Section 3. The simulation involves the following steps.

1. A sample A composed by n_A i.i.d. records has been generated according to a bivariate normal distribution (X, Y) with mean vector $\mu_{xy} = (0, 0)$ and covariance matrix given by

$$\Sigma_{xy} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$

2. A sample B composed by n_B i.i.d. records has been generated according to a bivariate normal distribution (X, Z) with mean vector $\mu_{xz} = (0, 0)$ and covariance matrix given by

$$\Sigma_{xz} = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$$

3. Since X is a continuous variable in order to compute the uncertainty measures we need to discretize it. The range of observed values $x = x_A \cup x_B$ has been divided into $I = 20$ intervals according to the hth percentiles of data, for $h = 5 - 95(5)$. Formally, the discretized variable will assume the value $x = i$, for $i = 1, \ldots, 20$.

4. Conditionally on $x = i$ (for $i = 1, \ldots, 20$) the pointwise measure of uncertainty in (x, y, z) is estimated by

$$\hat{\Delta}^{yz|x} = min(\hat{G}_{n_B}(z|x), \hat{F}_{n_A}(y|x)) - max(\hat{G}_{n_B}(z|x) + \hat{F}_{n_A}(y|x) - 1, 0) \tag{23}$$

where $\hat{F}_{n_A}(y|x)$ and $\hat{G}_{n_B}(z|x)$ are the empirical distribution functions of $F(y|x)$ and $G(z|x)$, respectively. Let $n_{A,x}$ and $n_{B,x}$ be the number of units such that $x = i$ in samples A and B, and denote by $\mathbf{y} = (y_{1,x}, \ldots, y_{n_A,x})$ and $\mathbf{z} = (z_{1,x}, \ldots, z_{n_B,x})$ the corresponding sample values of Y and Z, respectively. Then, for $x = i$ we obtain $n_{A,x}n_{B,x}$ pointwise uncertainty measures (23).

5. Conditionally on $x = i$ the uncertainty measure when no external auxiliary information is available is obtained by averaging the $n_{A,x}n_{B,x}$ pointwise uncertainty measures (23).
Formally

$$\hat{\Delta}^{x=i} = \frac{1}{n_{A,x}n_{B,x}} \sum_{y \in \mathbf{y}} \sum_{z \in \mathbf{z}} \hat{\Delta}^{yz|x} \tag{24}$$

6. The overall unconditional uncertainty measure is a weighted mean of the conditional uncertainty measure (24)

$$\hat{\Delta} = \sum_x \Delta^{x=i} \hat{p}(x) \tag{25}$$

where

$$\hat{p}(x) = \left(\frac{n_{A,x} + n_{B,x}}{n_A + n_B} \right) \tag{26}$$

7. Suppose that auxiliary information in the form of logical constraints regarding the statistical model for (X, Y, Z) is available. More specifically let us assume that there exist constants a_x and b_x such that $a_x \leq Y/Z \leq b_x$. In the simulation study we set $a_x = 0.1/x$ and $b_x = 1/x$. Then, conditionally on $x = i$ and under the constraint $0.1/x \leq Y/Z \leq 1/x$, the pointwise uncertainty measure in (x, y, z) is estimated by

$$\hat{\Delta}_c^{yz|x} = \hat{U}^x(y, z) - \hat{L}^x(y, z)$$

where the subscript c represents the constraint and

$$\hat{U}^x(y, z) = min\{\hat{F}_{n_A}(y|x), \ \hat{F}_{n_A}(b_x z|x), \ \hat{G}_{n_B}(z|x), \ \hat{G}_{n_B}(y/a_x|x)\} \tag{27}$$

$$\hat{L}^x(y, z) = max\{0, \ min(\hat{F}_{n_A}(y|x), \hat{F}_{n_A}(b_x z|x)) + min(\hat{G}_{n_B}(z|x), \hat{G}_{n_B}(y/a_x|x)) - 1\} \tag{28}$$

8. Conditionally on $x = i$ the estimator of conditional uncertainty measure under the constraint $0.1/x \leq Y/Z \leq 1/x$ is given by

$$\hat{\Delta}_c^{x=i} = \frac{1}{n_{A,x}n_{B,x}} \sum_{y \in \mathbf{y}} \sum_{z \in \mathbf{z}} \hat{\Delta}_c^{yz|x} \tag{29}$$

9. The overall unconditional uncertainty measure under the constraint $0.1/x \leq Y/Z \leq 1/x$ is obtained as a weighted mean of conditional uncertainty measures (29)

$$\hat{\Delta}_c = \sum_x \hat{\Delta}_c^{x=i} \hat{p}(x) \tag{30}$$

where $\hat{p}(x)$ is given by (26).

10. Steps 1 to 9 have been repeated 500 times and for different sample sizes $n_A = n_B = n = (1,000, 2,000, 5,000)$.

Given n, for each sample s (for $s = 1, \ldots, 500$) and for each category i (for $i = 1, \ldots, 20$) denote by $\hat{\Delta}^{x=i,s}$,$\hat{\Delta}^s$, $\hat{\Delta}_c^{x=i,s}$, $\hat{\Delta}_c^s$ the uncertainty measures (24), (25), (29) and (30), respectively.

As for the conditional uncertainty measure estimates (24), their average over simulation runs is

$$\bar{\Delta}^{x=i} = \frac{1}{500} \sum_{s=1}^{500} \hat{\Delta}^{x=i,s} \tag{31}$$

while

$$\bar{\Delta} = \frac{1}{500} \sum_{s=1}^{500} \hat{\Delta}^s \tag{32}$$

represents the corresponding overall uncertainty. The standard deviation of $\hat{\Delta}^{x=i,s}$, again over simulation runs, is equal to

$$SD(\hat{\Delta}^{x=i}) = \sqrt{\frac{1}{499} \sum_{s=1}^{500} (\hat{\Delta}^{x=i,s} - \bar{\Delta}^{x=i})^2} \tag{33}$$

and the corresponding mean squared error is given by

$$MSE(\hat{\Delta}^{x=i}) = [Sd(\hat{\Delta}^{x=i})]^2 + (\bar{\Delta}^{x=i} - \Delta^{x=i})^2 \tag{34}$$

where $\Delta^{x=i} = 1/6$.

Analogously, if we refer to the conditional uncertainty measure estimates (29), we have that the average over simulation runs is given by

$$\bar{\Delta}^{x=i} = \frac{1}{500} \sum_{s=1}^{500} \hat{\Delta}_c^{x=i,s} \tag{35}$$

while

$$\bar{\Delta}_c = \sum_{s=1}^{500} \bar{\Delta}_c^s \tag{36}$$

is the corresponding overall uncertainty under the constraint $0.1/x \leq Y/Z \leq 1/x$. The standard deviation over simulation runs is

$$SD\left(\hat{\Delta}_c^{x=i}\right) = \sqrt{\frac{1}{499}\sum_{s=1}^{500}\left(\hat{\Delta}_c^{x=i,s} - \bar{\Delta}_c^{x=i}\right)^2} \tag{37}$$

and finally the corresponding mean squared error is

$$MSE\left(\hat{\Delta}_c^{x=i}\right) = \left[Sd\left(\hat{\Delta}_c^{x=i}\right)\right]^2 + \left(\bar{\Delta}^{x=i} - \Delta_c^{x=i}\right)^2 \tag{38}$$

where $\Delta_c^{x=i}$ is the actual conditional uncertainty measure for the ith category under the constraint c. The values $\Delta_c^{x=i}$ have been numerically computed. More specifically $N = 30,000$ i.i.d. records have been generated from (X, Y) and (X, Z) according to the distributions specified in Steps 1 and 2. The actual overall uncertainty measure under the constraint is given by

$$\Delta_c = \frac{1}{20}\sum_{i=1}^{20}\Delta_c^{x=i} \cong 0.11225 \tag{39}$$

4.1. Simulation Results

In Table 1 the uncertainty measures $\bar{\Delta}$ and $\bar{\Delta}_c$ given by (32) and (36), respectively, are reported for different values of the sample size n. Note that $\bar{\Delta}$ is always smaller than 1/6, and that it tends to 1/6 as the sample size n increases. Furthermore, the constraint $0.1/x \leq Y/Z \leq 1/x$ reduces the model uncertainty for (X, Y, Z) from $\bar{\Delta} \cong 0.16$ to $\bar{\Delta}_c \cong 0.11$.

In Figure 1 the expectation of the estimated uncertainty (31) is reported. As the category $x = i$ varies, the expected estimated uncertainty is essentially the same. Furthermore, as the sample size n increases from 1,000 to 5,000, for each category $x = i$ the expectation of the estimated mean uncertainty tends to 1/6 and the precision of our estimator increases as shown in Figure 2, where the standard deviation (33) is reported. As shown in Figure 3, as n increases, the mean squared error (34) decreases for each category $x = i$. Such a property comes from the consistency of the empirical distribution functions $\hat{F}_{n_A}(y|x)$ and $\hat{G}_{n_B}(z|x)$ as estimators of $F(y|x)$ and $G(z|x)$, respectively.

The same analysis has been performed under the constraint $0.1/x \leq Y/Z \leq 1/x$ where $a_x = 0.1/x$ and $b_x = 1/x$. Conditional and unconditional uncertainty measures

Table 1. Uncertainty measures as the sample size n varies

n	$\bar{\Delta}$	$\bar{\Delta}_c$
1,000	0.16653	0.10946
2,000	0.16663	0.11068
5,000	0.16666	0.11163

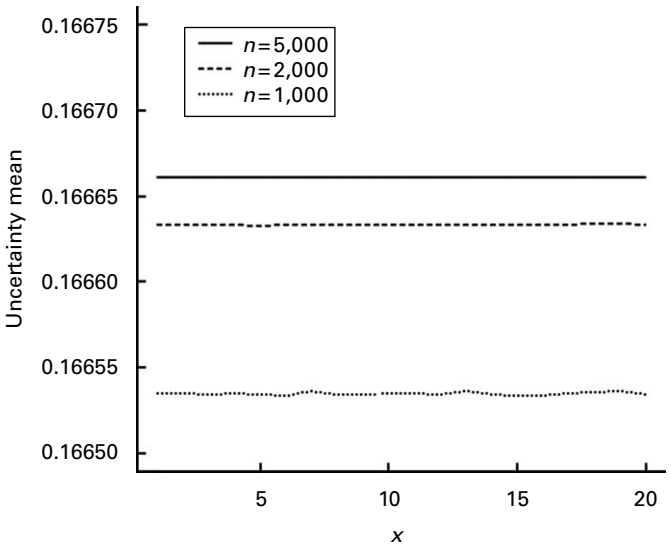

Fig. 1. Expected values of estimated uncertainty measures in the unconstrained case

(numerically computed) are reported in Figure 4. The horizontal line is the unconditional uncertainty measure, while the dashed line depicts the conditional uncertainty measure as $x = i$ varies. Conditional uncertainty takes small values, about $3 \times 10^{-2} - 7 \times 10^{-2}$, for $x = 1$–$5(1)$, while it is about 0.15 for the categories $x = 13 - 18(1)$. The behavior of the corresponding estimates is shown in Figures 5, 6 and 7. In particular, in Figure 5 the expectations of the estimated uncertainty for sample sizes $n = 1,000, 2,000, 5,000$, i.e., $\bar{\Delta}_c^{x=i}$ given by (35), are reported. The absolute bias of estimators is approximately constant

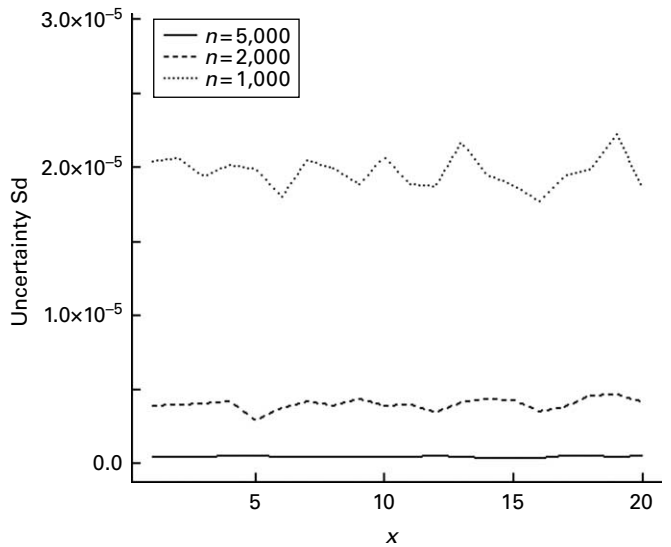

Fig. 2. Standard deviation of estimated uncertainty measures in the unconstrained case

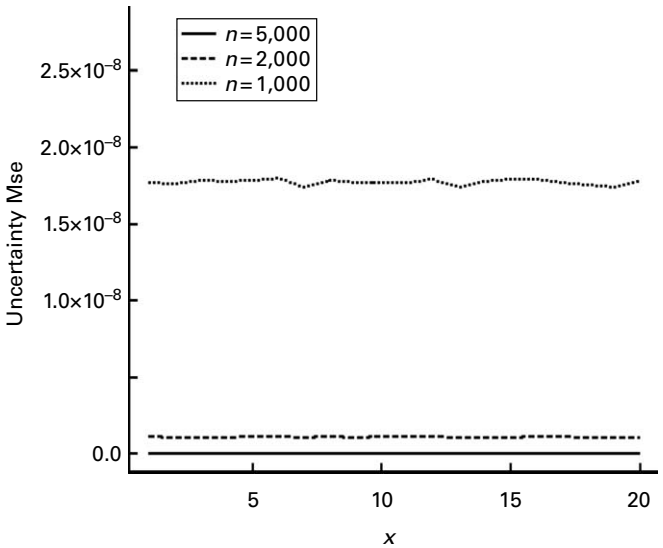

Fig. 3. Mean squared error of estimated uncertainty measures in the unconstrained case

and small, if compared to the corresponding true uncertainty measures. In Figures 6, 7 the standard deviation (37) and the mean squared error (38) of our uncertainty estimators are reported. The highest efficiency of estimation is obtained for the classes $x = 1 - 5(1)$. The smallest efficiency is obtained for the classes $x = 12 - 16(1)$. At any rate, we stress that the mean squared errors of conditional uncertainty estimators are all considerably small, ranging from 10^{-3} to 1.3×10^{-2}.

In Figure 8 the Kernel density estimate of the overall uncertainty measure under the constraint $0.1/x \leq Y/Z \leq 1/x$ is shown. Such an estimate has been computed using for

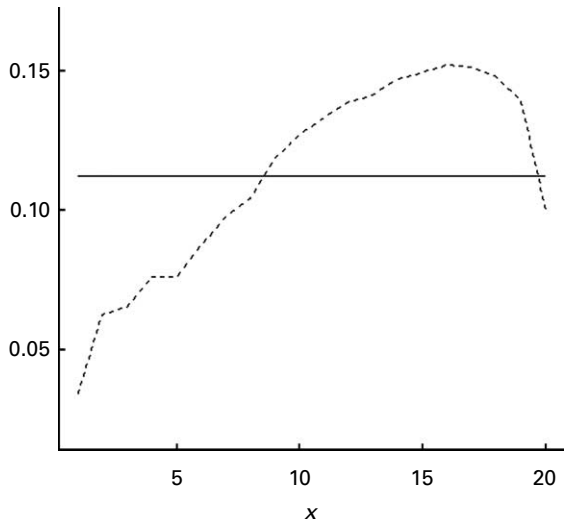

Fig. 4. Uncertainty measure Δ_c and Δ_c^i for each $x = i$ under the constraint $0.1/x \leq Y/Z \leq 1/x$

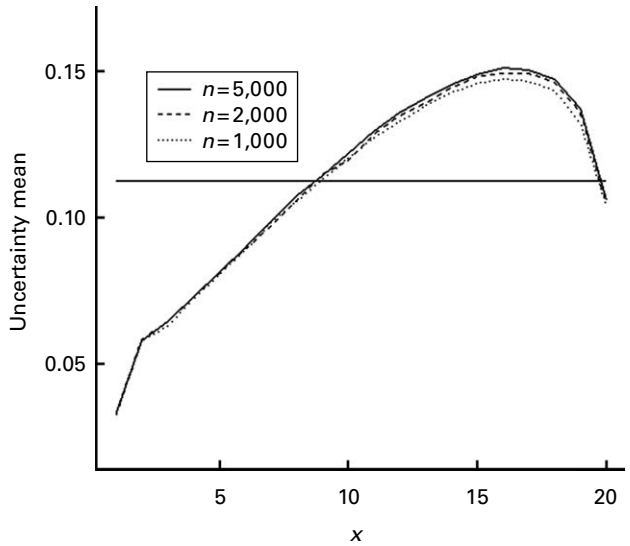

Fig. 5. Expectation of estimated uncertainty measures under the constraint $0.1/x \leq Y/Z \leq 1/x$

each sample size n the 500 value $\bar{\Delta}_c$ given by (30). Note that as the sample size n increases the uncertainty measure distribution tends to a normal distribution as proved in Conti et al. (2009b). The bandwidth selection rule is given by Sheather and Jones (1991). Similar considerations hold for the estimated conditional uncertainty measures.

5. Concluding Remarks

The approach described in the above sections seems weird in statistical inference: the result of an estimation process is not a single value, or a single distribution, but a set of

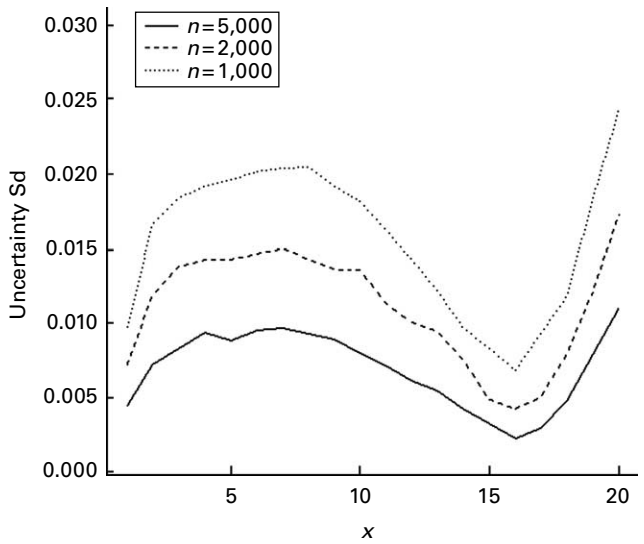

Fig. 6. Standard deviation of estimated uncertainty measures under the constraint $0.1/x \leq Y/Z \leq 1/x$

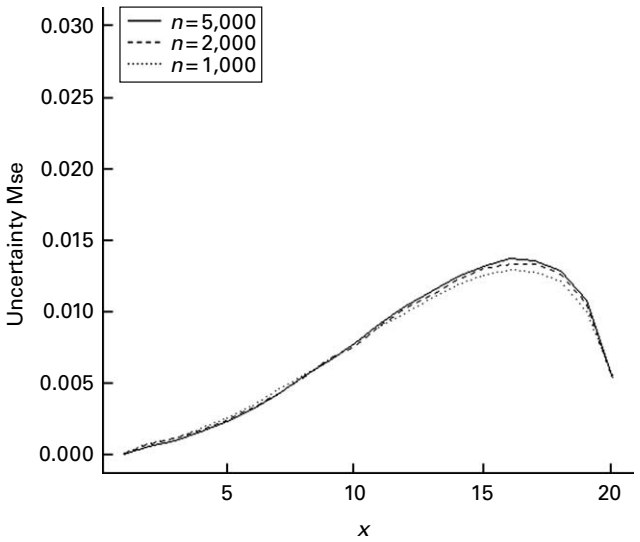

Fig. 7. Mean squared error of estimated uncertainty measures under the constraint $0.1/x \leq Y/Z \leq 1/x$

values or a set of distributions. Such sets are justified by the lack of joint information on the rv's of interest Y and Z, and can be mitigated (although not avoided) by strong statistical relationships of X with Y and/or Z, as well as by constraints justified by logical rules on the (Y, Z) joint distribution. If a set is not acceptable, it is then necessary to include additional assumptions, such as the CIA. However, the use of additional assumptions, sometimes neither theoretically justified nor testable on the available data, is not a benefit, as stated in the so-called "law of decreasing credibility" (cf. Manski 2003): *the credibility of inference decreases with the strength of the assumption maintained.*

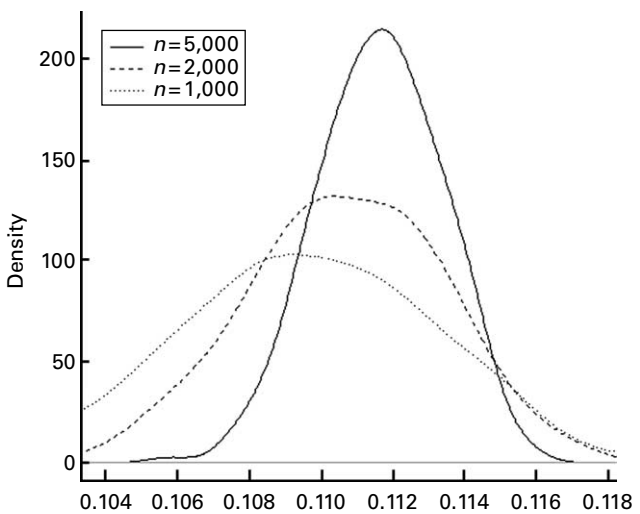

Fig. 8. Density estimate of overall uncertainty measure under the constraint $0.1/x \leq Y/Z \leq 1/x$

The present article essentially acknowledges the importance of the law of decreasing credibility. The measure of uncertainty defined in Section 2 aims at measuring the "credibility of inferences" that can be made on the basis of available data and prior information. Similar approaches have also been used in the areas of statistical disclosure control (cf. Fienberg and Slavkovic 2005), ecological inference (cf. King 1997) and analysis of partially observed data sets when neither MAR nor MCAR assumptions are made (cf. Imbens and Manski 2004).

6. References

Capitanio, A., Azzalini, A., and Stanghellini, E. (2003). Graphical Models for Skew-Normal Variates. Scandinavian Journal of Statistics, 30, 129–144.

Conti, P.L., Marella, D., and Scanu, M. (2009a). Evaluation of Matching Noise for Imputation Techniques based on Nonparametric Local Linear Regression Estimators. Computational Statistics & Data Analysis, 53, 354–365.

Conti, P.L., Marella, D., and Scanu, M. (2009b). How Far from Identifiability? A Nonparametric Approach to Uncertainty in Statistical Matching under Logical Constraints. Technical Report n.22, DSPSA, Università di Roma "La Sapienza".

D'Orazio, M., Di Zio, M., and Scanu, M. (2006a). Statistical Matching for Categorical Data: Displaying Uncertainty and Using Logical Constraints. Journal of Official Statistics, 22, 137–157.

D'Orazio, M., Di Zio, M., and Scanu, M. (2006b). Statistical Matching: Theory and Practice. Chichester: Wiley.

Fienberg, S.E. and Slavkovic, A.B. (2005). Preserving the Confidentiality of Categorical Data Bases When Releasing Information for Association Rules. Data Mining and Knowledge Discovery, 11, 155–180.

Imbens, G. and Manski, C. (2004). Confidence Intervals for Partially Identified Parameters. Econometrica, 72, 1845–1857.

Kadane, J.B. (1978). Some Statistical Problems in Merging Data Files. Compendium of Tax Research, Department of Treasury, U.S. Government Printing Office, Washington D.C., 159–179 (Reprinted in 2001, Journal of Official Statistics, 17, 423–433).

Kiesl, H. and Rässler, S. (2008). The Validity of Data Fusion. Insights on Data Integration Methodologies, ESSnet-ISAD workshop, Vienna, 29–30 May 2008, 59–67.

King, G. (1997). A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data. Princeton: Princeton University Press.

Luzi, O., Di Zio, M., Guarnera, U., Manzari, A., De Waal, T., Pannekoek, J., Hoogland, J., Tempelman, C., Hulliger, B., and Kilchmann, D. (2007). Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys, ISTAT, CBS, SFSO, Eurostat. Available at http://edimbus.istat.it.

Manski, C.F. (2003). Partial Identification of Probability Distributions. New York: Springer Verlag.

Marella, D., Scanu, M., and Conti, P.L. (2008). On the Matching Noise of Some Nonparametric Imputation Procedures. Statistics and Probability Letters, 78, 1593–1600.

Moriarity, C. and Scheuren, F. (2001). Statistical Matching: A Paradigm for Assessing the Uncertainty in the Procedure. Journal of Official Statistics, 17, 407–422.

Nelsen, R.B. (1999). An Introduction to Copulas. New York: Springer Verlag.

Okner, B.A. (1972). Constructing a New Microdata Base from Existing Microdata Sets: The 1966 Merge File. Annals of Economic and Social Measurement, 1, 325–362.

Paass, G. (1986). Statistical Match: Evaluation of Existing Procedures and Improvements by Using Additional Information. Microanalytic Simulation Models to Support Social and Financial Policy, G.H. Orcutt and H. Quinke (eds). Amsterdam: Elsevier.

Rässler, S. (2002). Statistical Matching: A Frequentist Theory, Practical Applications and Alternative Bayesian Approaches. Lecture Notes in Statistics. New York: Springer Verlag.

Ridder, G. and Moffitt, R. (2007). The Econometrics of Data Combination. Handbook of Econometrics, J.J. Heckmann and E.E. Leamer (eds). vol. 6A. Amsterdam: Elsevier.

Rodgers, W.L. (1984). An Evaluation of Statistical Matching. Journal of Business and Economic Statistics, 2, 91–102.

Sheather, S.J. and Jones, M.C. (1991). A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation. Journal of the Royal Statistical Society, Series B, 53, 683–690.

Sims, C.A. (1972). Comments and Rejoinder (On Okner (1972)). Annals of Economic and Social Measurement, 1, 343–345, 355–357.

Singh, A.C., Mantel, H., Kinack, M., and Rowe, G. (1993). Statistical Matching: Use of Auxiliary Information as an Alternative to the Conditional Independence Assumption. Survey Methodology, 19, 59–79.

Slepian, D. (1962). The One-Sided Barrier Problem for Gaussian Noise. Bell System Technical Journal, 41, 463–501.

Vantaggi, B. (2008). Statistical Matching of Multiple Sources: A Look Through Coherence. International Journal of Approximate Reasoning, 49, 701–711.

Received November 2009
Revised November 2011