LSTA #1010005, VOL 0, ISS 0

How far from identifiability? A systematic overview of the statistical matching problem in a non parametric framework

Pier Luigi Conti, Daniela Marella, and Mauro Scanu

QUERY SHEET

This page lists questions we have about your paper. The numbers displayed at left can be found in the text of the paper for reference. In addition, please review your paper as a whole for correctness.

Q1. Au: Please provide missing MSC for this article.
Q2. Au: Please provide missing publisher name and publisher location for the reference "Gazzelloni et al., 2007."
Q3. Au: Please provide missing publisher location for the reference "Manski, 1995."
Q4. Au: Please provide missing volume number and page number for the reference "Wolfson et al., 1989."

TABLE OF CONTENTS LISTING

The table of contents for the journal will list your paper exactly as it appears below:

How far from identifiability? A systematic overview of the statistical matching problem in a non parametric framework
Pier Luigi Conti, Daniela Marella, and Mauro Scanu

COMMUNICATIONS IN STATISTICS—THEORY AND METHODS
2016, VOL. 0, NO. 0, 1–28
http://dx.doi.org/10.1080/03610926.2015.1010005

Taylor & Francis
Taylor & Francis Group

How far from identifiability? A systematic overview of the statistical matching problem in a non parametric framework

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

[a]Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Roma, Italy; [b]Dipartimento di Scienze della Formazione, Università "Roma Tre", Roma, Italy; [c]ISTAT, Italian National Statistical Institute, Roma, Italy

ABSTRACT

Statistical matching consists in estimating the joint characteristics of two variables observed in two distinct and independent sample surveys, respectively. In a parametric setup, ranges of estimates for non identifiable parameters are the only estimable items, unless restrictive assumptions on the probabilistic relationship between the non jointly observed variables are imposed. These ranges correspond to the uncertainty due to the absence of joint observations on the pair of variables of interest. The aim of this paper is to analyze the uncertainty in statistical matching in a non parametric setting. A measure of uncertainty is introduced, and its properties studied: this measure studies the "intrinsic" association between the pair of variables, which is constant and equal to 1/6 whatever the form of the marginal distribution functions of the two variables when knowledge on the pair of variables is the only one available in the two samples. This measure becomes useful in the context of the reduction of uncertainty due to further knowledge than data themselves, as in the case of structural zeros. In this case the proposed measure detects how the introduction of further knowledge shrinks the intrinsic uncertainty from 1/6 to smaller values, zero being the case of no uncertainty. Sampling properties of the uncertainty measure and of the bounds of the uncertainty intervals are also proved.

ARTICLE HISTORY
Received 11 April 2014
Accepted 15 January 2015

KEYWORDS
Constrained matching; Fréchet classes; statistical matching; uncertainty; unidentifiability.

MATHEMATICS SUBJECT CLASSIFICATION

Q1

1. Introduction and setting

The data deluge we are experiencing in these last years allows researchers to give an answer to questions never addressed in the past by exploiting information already available in existing data sources without setting up new surveys, hence reducing costs and improving timeliness. Anyway, the usual statistical inference tools are not always suitable, and methodological advances are sometimes necessary. This is the case of the statistical matching problem: instead of constructing a new, complete survey containing a couple of variables of interest, statistical matching tackles the case of data integration of two already existing samples, drawn from the same population, and composed by non overlapping sets of units (i.e. the same unit is not observed in both the surveys). Each sample owns information on just one of the variables of interest, so that these variables are not jointly observed, as described in Table 1.

From a methodological point of view, a first example of this approach has been considered in Anderson (1957), where the case of a three-variate (X, Y, Z) normal distribution with

CONTACT Daniela Marella daniela.marella@uniroma3.it Dipartimento di Scienze della Formazione, Università "Roma Tre", Roma, Italy.

Table 1. General statistical matching problem, A and B are two independent sample surveys, the objective is to learn something on the relationship (or on the distribution) of the Y and Z r.v.s, even if they are never jointly observed.

Sample	Y_1	...	Y_Q	X_1	...	X_P	Z_1	...	Z_R
	y^A_{11}	...	y^A_{1Q}	x^A_{11}	...	x^A_{1P}			
	...			...					
A	y^A_{a1}	...	y^A_{aQ}	x^A_{a1}	...	x^A_{aP}			
	...			...					
	$y^A_{n_A 1}$	...	$y^A_{n_A Q}$	$x^A_{n_A 1}$	...	$x^A_{n_A P}$			
				x^B_{11}	...	x^B_{1P}	z^B_{11}	...	z^B_{1R}
				...				...	
B				x^B_{b1}	...	x^B_{bP}	z^B_{b1}	...	z^B_{bR}
				...				...	
				$x^B_{n_B 1}$	...	$x^B_{n_B P}$	$z^B_{n_B 1}$	...	$z^B_{n_B R}$

observations as in Table 1 is considered. In that paper, the maximum likelihood estimators (MLEs) of the unidentifiable parameters are studied. In practice, Table 1 setup has been met in many other situations, shortly listed in the sequel.

- Okner (1972), probably the first statistical matching application, aimed at reconstructing a unique file with both income subject to tax (say Z) from a sample B of individual tax returns and a total money income concept (say Y, which includes non taxable transfer payments but excludes taxable realized capital gains) from the sample A observed by the Census Bureau Current Population Survey. The study of economic variables observed in distinct sources is a ubiquitous goal, because of the difficulties in observing different economic phenomena by means of a unique survey. One of the latest examples is 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)".
- Advertisers and media planners study customers' behavior by fusing information describing people's characteristics, product and brand usage, and media exposure. Such information enables media planners and marketing researchers to pursue such objectives as increasing sales by formulating the right campaign and selecting the most appropriate media for it. An example is the fusion of the Broadcasters Audience Research Board (BARB, say A in the framework of Table 1) and the Target Group Index (TGI, say B) in the United Kingdom, see Adamek (1994).
- Official statistics is a context where the necessity to cut costs and the increasing informative needs push the National Statistical Institutes toward the massive use of data integration methods. Examples are in Gazzelloni et al. (2007), aiming at studying jointly variables observed in the Italian Labour Force Survey and in the Time Use Survey, in D'Orazio et al. (2006b), chapter 7, where there is the description of the reconstruction of the Social Accounting Matrix useful for the National Accounts objectives, and in Torelli et al. (2008) describing the use of statistical matching techniques for jointly estimating contingency tables of pairs of variables observed in a structural and in an economical survey on farms, respectively. All these examples are compliant with Table 1 structure.
- Microsimulation is an increasingly important tool in economics, consisting of an accounting model which processes each individual and family in a country, calculates taxes and transfers using legislated or proposed programs and algorithms, and reports

on the results. Microsimulation models usually give a high degree of control over the inputs and outputs to the model and can allow the user to modify existing tax/transfer programs or test proposals for entirely new programs. These models need very complete databases, containing information on household and personal characteristics ranging from income and personal properties to health, housing, education conditions. In order to obtain such databases sometimes researchers consider statistical matching of two or more sample surveys as in Table 1 (see Wolfson et al. , 1989).

A problematic feature of a problem as depicted in Table 1 is that the model (X, Y, Z) is not generally identifiable given the sample observations in $A \bigcup B$. This problem was first noted in Sims (1972), whose criticism on Okner (1972) focused on the fact that his computations were based on a specific model assumption for (X, Y, Z): the conditional independence of Y and Z given X (CIA henceforth). The CIA assumption produces an identifiable model for $A \bigcup B$, but this model (and hence the CIA itself) cannot be tested on the basis of available data sets. Kadane (1978) further showed the lack of identification of the model for the case of a three-variate normal distribution, stating that the unidentifiable parameter in this case is only the correlation coefficient $\rho_{yz|x}$ of Y and Z given X, and that this correlation is free to move from -1 up to 1. Anyway, the pairwise correlation ρ_{yz} between Y and Z is not free to move in the same interval, because of the presence of the common variable X. This aspect is investigated in a number of papers: Rubin (1986) applies the multiple imputation methodology to explore the set of possible values of ρ_{yz}, an idea further developed in Raessler (2002) in a proper Bayesian framework, and studied in Reiter (2012) as far as the validity of the standard multiple imputation variance estimator for assessing sampling variability in data fusion given specification of the imputation models is concerned; Moriarity and Scheuren (2001) use consistent estimates of the estimable parameters for establishing intervals for ρ_{yz} (although the use of coherent estimates, as the ones obtained through maximum likelihood, could allow the possibility to avoid unpleasant situations as negative definite variance and covariance matrices).

The previous framework was extended at first in D'Orazio et al. (2006a, b). The interval of values taken by non identifiable parameters was termed *uncertainty*. Uncertainty corresponds to the set of distributions plausible according to a criteria (e.g. the maximum likelihood approach) given the available data $A \bigcup B$. This kind of uncertainty, due to the lack of joint observations on Y and Z, can be reduced when knowledge related to the association of Y and Z is also included. D'Orazio et al. (2006a, b) use knowledge usually considered for editing purposes for microdata, while proving it useful also for estimation purposes in statistical matching, as the structural zeros. The authors also consider other kinds of additional knowledge, as the one given in terms of possible orderings between probabilities of the (Y, Z) or $(Y, Z|X)$ joint distribution. With both these kinds of constraint, uncertainty shrinks and the CIA is not included among the plausible solutions any more. D'Orazio et al. (2006a) further study the case of categorical r.v.s, a relevant case in many applied settings.

Final extensions have been defined in Conti et al. (2012, 2013). Conti et al. (2013) extend the results in D'Orazio et al. (2006a) for the case of categorical ordered r.v.s. In this case, uncertainty can be stated in terms of the whole probability distributions of Y and Z, as the joint cumulative distribution function of (Y, Z), leading to bounds of cell probabilities of the (Y, Z) contingency table that are shorter than the ones that can be obtained as in D'Orazio et al. (2006a). Uncertainty restriction by means of the availability of structural zeros is also studied. Conti et al. (2012) formally define uncertainty in different inferential settings, i.e. in parametric and non parametric frameworks. The non parametric framework is especially important in order to overcome the usual assumption of normality of the three r.v.s, an aspect

useful in applications. For instance, economic r.v.s as household income and consumption can be modeled in a non parametric setup without assuming normality or categorizing the distribution.

The aim of this paper is to analyze further the uncertainty in statistical matching in a non parametric setting. The paper is organized as follows. In Section 2 the concept of uncertainty in statistical matching is discussed. In Section 3, starting from results in Conti et al. (2012), the model uncertainty in a non parametric setting is investigated, and the role of Fréchet classes is stressed. In Section 4 the non parametric estimation of Fréchet bounds is analyzed. A first important original result consists in defining confidence regions for the Fréchet class. This is the first time that uncertainty due to lack of joint information on Y and Z and sampling uncertainty are studied together in a statistical matching problem outside the Bayesian framework (Reiter, 2012). Another important advance is the definition of an overall intrinsic measure of uncertainty for non identifiable models (Section 5). This intrinsic measure of uncertainty does not depend on the support and form of the marginal distributions of Y and Z, but only on the lack of joint observations. For this reason, this measure is always equal to 1/6 when only $A \bigcup B$ is the available source of information. Whenever additional information in terms of a structural zero or other restrictions are introduced, such an intrinsic measure of uncertainty decreases, zero being the case of absence of uncertainty (i.e. unidentifiable parameters become identifiable given $(A \bigcup B)$ and the imposed logical restrictions). Section 6 focuses on the introduction of structural zeros in this non parametric setting, i.e. restrictions on the support of the joint distribution of (Y, Z) given X. For each constraint an estimator is proposed and its asymptotic behavior is studied.

2. The kind of uncertainty affecting statistical matching

In a parametric setting the main consequence of the lack of identifiability is that some parameters of the model for (X, Y, Z) 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. Such intervals are *uncertainty intervals*. Uncertainty in a statistical matching problem is a special case of estimation problems for general partially identifiable models (e.g. not assuming a specific missing data mechanism) as in Manski (1995), Horowitz (2000), Chernozhukov et al. (2007), and references therein for general inferential problems on a partially observed sample. Also in those cases, estimation is not pointwise, but consists of ranges. Another context characterized by parameter uncertainty is the case of categorical data (e.g. k-way contingency tables) where 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.

Statistical matching is also related to the so-called "ecological inference" problem; see King (1997), Cross and Manski (2002). An important difference is that in ecological inference marginal distribution functions come from population counts, and are not sample-based.

The first papers tackling the problem of assessing how uncertain some parameters are in statistical matching problems are Kadane (1978), Rubin (1986), Moriarity and Scheuren (2001), Raessler (2002). Assuming that (X, Y, Z) is a three-variate normal r.v. with parameters

$$\boldsymbol{\theta} = \left[\begin{pmatrix} \mu_x \\ \mu_y \\ \mu_z \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \sigma_{xy} & \sigma_{xz} \\ \sigma_{xy} & \sigma_y^2 & \sigma_{yz} \\ \sigma_{xz} & \sigma_{yz} & \sigma_z^2 \end{pmatrix} \right]$$

the only non estimable parameter is σ_{YZ}, whose range of plausible values given what has been observed (i.e. given the estimates of the other estimable parameters) can be determined by imposing that the determinant of the covariance matrix is not smaller than 0. As a result, the correlation coefficient between Y and Z can assume values only in the interval with extremes:

$$\hat{\rho}_{xy}\hat{\rho}_{xz} \pm \sqrt{\left(1 - \hat{\rho}_{xy}^2\right)\left(1 - \hat{\rho}_{xz}^2\right)}$$

Cell frequencies θ_{yz} of the (Y, Z) contingency table, given estimates $\hat{\theta}_{y|x}$ and $\hat{\theta}_{z|x}$ from A and B, respectively, and on $\hat{\theta}_x$ from $A \bigcup B$, for any x, y, z, can be obtained by means of the Fréchet bounds, D'Orazio et al. (2006a):

$$\max\left\{0; \hat{\theta}_{y|x} + \hat{\theta}_{z|x} - 1\right\} \leq \theta_{yz|x} \leq \min\left\{\hat{\theta}_{y|x}; \hat{\theta}_{z|x}\right\} \tag{1}$$

Raessler and Kiesl (2009), focusing again on normal distributions, combine the restrictions relative to the non negativity of the variance matrix and the Fréchet bounds.

Evaluation of the uncertainty in a statistical matching problem is also used for validation purposes. Raessler (2002) evaluates for the normal multivariate models the length of the uncertainty intervals for unidentifiable parameters in order to define a measure of the reliability of estimates under CIA. When uncertainty intervals are "short", the parameter estimates obtained under the different models compatible with the available sample information slightly differ from the ones estimated under the CIA. Let θ_k, $k = 1, \ldots, K$, be the unidentifiable parameters in a parametric model on (X, Y, Z). Raessler (2002) defines an overall measure of uncertainty as

$$\Delta = \frac{1}{K} \sum \left(\widehat{\theta}_k^{(U)} - \widehat{\theta}_k^{(L)}\right) \tag{2}$$

where $\widehat{\theta}_k^{(U)}$ and $\widehat{\theta}_k^{(L)}$ are the estimated lower and upper extrema of the uncertainty intervals of θ_k, $k = 1, \ldots, K$, in (1). 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. D'Orazio et al. (2006a) in the case of categorical data and in D'Orazio et al. (2006b) for general parametric models show that the uncertainty measure (2) is generally too wide. Furthermore, 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 "central" (or better, middle) point in the uncertainty interval for each parameter. Furthermore, unlike the Bayesian approach, in a likelihood-based approach it becomes important to include all the information that can reduce the uncertainty space. The best kind of information is given by structural zeros between Y and Z.

3. Uncertainty in a non parametric setting

Uncertainty in a non parametric setting is still described by a class of models, or better, by a *class of distributions*, for $(X,\ Y,\ Z)$. If compared to the parametric case (either multinormal or multinomial), there are two main sources of trouble. First of all, since the class of distributions for $(X,\ Y,\ Z)$ are not identified by a finite number of parameters, we need a technical tool to *describe* them. In the second place, in order to *measure*, to *quantify* uncertainty, we need to "summarize" the class of all possible distributions for $(X,\ Y,\ Z)$ with a single number. An even more important problem is the quantification of the uncertainty in the presence of auxiliary information on the model. This kind of information is frequently used, for instance, in imputation and editing (see Luzi et al., 2007).

The problem of describing uncertainty in a non parametric setting is shortly outlined in Ridder and Moffitt (2007), Kiesl and Raessler (2008), Raessler and Kiesl (2009). The basic ideas on how to describe and measure uncertainty under auxiliary information are presented in Conti et al. (2013), although they have studied in detail discrete distributions, the continuous case being outlined in Conti et al. (2012). Unfortunately, some cases of considerable practical interest (for instance, $Y =$ income and $Z =$ consumption expenses) are not covered by the methodologies developed in the above-mentioned papers.

In the present paper, we aim at studying how to describe and measure uncertainty (again, under auxiliary information) in a full generality, without additional restrictions on Y and/or Z. The starting point consists in observing that the natural way to describe classes of distributions consists in using the notion of Fréchet class. As a consequence, a measure of uncertainty is nothing more than a suitable functional that quantifies "how large" is such a class.

Consider a three-dimensional r.v. (X, Y, Z). Its joint distribution function (d.f.), denoted by $F(x, y, z)$, can be written as

$$dF(x, y, z) = dQ(x)dH(y,\ z\,|x) \tag{3}$$

where $Q(x)$ is the marginal d.f. of X and $H(y,\ z\,|x)$ is the d.f. of $(Y,\ Z)$ given X. In what follows, the r.v.s Z and Y will be assumed continuous, and the matching variable X is discrete.

Then, conditionally on X we have a set of plausible statistical models, namely the Fréchet class of all distribution functions $H(y, z|x)$ compatible with the univariate d.f.s $G(z|x), F(y|x)$. For every (y, z) the pair of inequalities

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

holds, where the bounds

$$L^x(F(y|x), G(z|x)) = \max(G(z|x) + F(y|x) - 1,\ 0)$$
$$U^x(F(y|x), G(z|x)) = \min(G(z|x),\ F(y|x))$$

are themselves joint d.f.s with margins $G(z|x)$ and $F(y|x)$. The set of d.f.s

$$\mathcal{H}^x = \{H(y, z|x) : L^x(F(y|x), G(z|x)) \leq H(y, z|x) \leq U^x(F(y|x), G(z|x))\} \tag{4}$$

is the Fréchet class of marginal d.f.s $G(z|x), F(y|x)$.

In the present case all the d.f.s belonging to the Fréchet class (4) 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 certain conclusions on the model.

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

$$\mathcal{H} = \{H(y,z):\ E_x[L^x(F(y|x), G(z|x))] \le H(y,z) \le E_x[U^x(F(y|x), G(z|x))]\} \tag{5}$$

As remarked in Ridder and Moffitt (2007), the Jensen inequality implies that the Fréchet class (5) is narrower than the "naive" Fréchet class

$$\{H(y,z):\ \max(F(y)+G(z)-1,\ 0) \le H(y,z) \le \min(F(y),\ G(z))\} \tag{6}$$

that does not use the common information X available on A and B, respectively.

We stress that the lower bound of the Fréchet class (4) corresponds to the maximal negative association between Y and Z, given X; this comes true if and only if (iff) Y is a strictly decreasing function of Z (and vice versa), given X. Similarly, the upper bound of the Fréchet class (4) corresponds to the maximal positive association between Y and Z, given X; this comes true iff Y is a strictly increasing function of Z (and vice versa), given X. As a consequence, in the absence of any further information about relationships between Y and Z, the Fréchet class (4) is "maximally wide".

4. Non parametric estimation of Fréchet bounds

If 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 (4) is given by

$$[\max(\widehat{G}_{n_B}(z|x) + \widehat{F}_{n_A}(y|x) - 1, 0),\ \min(\widehat{G}_{n_B}(z|x), \widehat{F}_{n_A}(y|x))] \tag{7}$$

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. More specifically, consider the indicator function of a set D

$$I_{(x\in D)} = \begin{cases} 1 & \text{if } x \in D \\ 0 & \text{otherwise} \end{cases}$$

and let $n_{A,x}$, $n_{B,x}$ be defined as

$$n_{A,x} = \sum_{i=1}^{n_A} I_{(X_i=x)}, \quad n_{B,x} = \sum_{i=1}^{n_B} I_{(X_i=x)}$$

The conditional d.f.s of Y, Z given X can be estimated by

$$\widehat{F}_{n_A}(y|x) = \frac{1}{n_{A,x}} \sum_{i=1}^{n_A} I_{(Y_i\le y,\, X_i=x)}, \quad \widehat{G}_{n_B}(z|x) = \frac{1}{n_{B,x}} \sum_{i=1}^{n_B} I_{(Z_i\le z,\, X_i=x)} \tag{8}$$

As a consequence, the unconditional Fréchet bounds (5) can be estimated by

$$\left[\sum_x \widehat{p}(x)\ \max(\widehat{G}_{n_B}(z|x) + \widehat{F}_{n_A}(y|x) - 1, 0),\ \sum_x \widehat{p}(x)\ \min(\widehat{G}_{n_B}(z|x), \widehat{F}_{n_A}(y|x))\right] \tag{9}$$

where

$$\widehat{p}(x) = \left(\frac{n_{A,x} + n_{B,x}}{n_A + n_B}\right) \tag{10}$$

is an estimate of $P(X = x)$.

According to the empirical likelihood approach as discussed by Owen (1991), the e.d.f.s (8) are non parametric maximum likelihood estimators (NPMLEs) of F and G, respectively.

Then, for the invariance property of MLE in the non parametric setting the estimators (7) and (9) represent the NPMLEs of Fréchet classes (4) and (5), respectively.

In order to display the sampling variability associated with the Fréchet class estimator (7), it is necessary to construct a confidence region, i.e. a pair of functions with a prespecified probability of containing the true Fréchet class. A confidence band can be constructed from the confidence bands for $F(y|x)$ and $G(z|x)$, respectively, from the Kolmogorov–Smirnov statistic.

Conditionally on X, large sample $(1-\alpha)$ confidence bands for $F(y|x)$ and $G(z|x)$ are given by

$$\mathcal{F}_{n_{A,x}} = \left(\widehat{F}_{n_A}(y|x) - \frac{k_\alpha}{\sqrt{n_{A,x}}},\ \widehat{F}_{n_A}(y|x) + \frac{k_\alpha}{\sqrt{n_{A,x}}};\quad y \in \mathbb{R} \right)$$

$$\mathcal{G}_{n_{B,x}} = \left(\widehat{G}_{n_B}(z|x) - \frac{k_\alpha}{\sqrt{n_{B,x}}},\ \widehat{G}_{n_B}(z|x) + \frac{k_\alpha}{\sqrt{n_{B,x}}};\quad z \in \mathbb{R} \right)$$

respectively, where k_α is the $1-\alpha$-quantile of the Kolmogorov–Smirnov distribution (i.e. the distribution of the supremum of the modulus of a Brownian bridge).

Define now

$$\underline{\widehat{H}}(y,z|x) = \max\left\{ \widehat{G}_{n_B}(z|x) - \frac{k_\alpha}{\sqrt{n_{B,x}}} + \widehat{F}_{n_A}(y|x) - \frac{k_\alpha}{\sqrt{n_{A,x}}} - 1,\ 0 \right\}$$

$$\widehat{\overline{H}}(y,z|x) = \min\left\{ \widehat{G}_{n_B}(z|x) + \frac{k_\alpha}{\sqrt{n_{B,x}}},\ \widehat{F}_{n_A}(y|x) + \frac{k_\alpha}{\sqrt{n_{A,x}}} \right\}$$

A confidence region for the Fréchet class (4) is then

$$\mathcal{H}_n^x = \left\{ H(y,z|x) :\ \underline{\widehat{H}}(y,z|x) \le H(y,z|x) \le \widehat{\overline{H}}(y,z|x) \right\} \tag{11}$$

A lower bound for the confidence level of the region (11) can be easily evaluated. In fact, it is possible to write

$$\begin{aligned} P(\mathcal{H}^x \subset \mathcal{H}_n^x) &= P(\{\underline{\widehat{H}}(y,z|x) \le L^x(F(y|x), G(z|x))\}, \{\widehat{\overline{H}}(y,z|x) \ge U^x(F(y|x), G(z|x))\}) \\ &\ge P((F(y|x) \subset \mathcal{F}_{n_{A,x}}))P((G(z|x) \subset \mathcal{G}_{n_{B,x}})) \\ &= (1-\alpha)^2 \end{aligned}$$

Analogously, it is possible to define a confidence region for the "naive" Fréchet class (6) from the $(1-\alpha)$ confidence bands for the margins $F(y)$ and $G(z)$, respectively.

5. Measures of uncertainty for non identifiable models

In view of the Fréchet bounds (4), the interval

$$\left[L^x(F(y|x), G(z|x)),\ U^x(F(y|x), G(z|x)) \right] \tag{12}$$

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 (12), i.e. $U^x(F(y|x), G(z|x)) - L^x(F(y|x), G(z|x))$. We have a different measure of uncertainty for every triple x, y, z. Of course, if the model is identifiable, then the interval (12) reduces to a single point, with length zero.

The problem is now to summarize all pointwise measures of uncertainty into a unique, overall measure. If $T(x,\ y,\ z)$ is a weight function on $\mathbb{R}^3$, as an overall measure of uncertainty we may take the average length:

$$\int_{\mathbb{R}^3} \left[U^x(F(y|x), G(z|x)) - L^x(F(y|x), G(z|x))\right] dT(x,\ y,\ z) \tag{13}$$

As far as the weight function T is concerned, a "natural" choice consists in taking

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

With such a choice, the overall uncertainty measure (13) becomes

$$\begin{aligned}\Delta(F, G) &= \int_{\mathbb{R}} \left\{ \int_{\mathbb{R}^2} \left[U^x(F(y|x), G(z|x)) - L^x(F(y|x), G(z|x))\right] d[F(y|x)\, G(z|x)] \right\} dQ(x) \\ &= \int_{\mathbb{R}} \Delta^x(F, G)\, dQ(x) \\ &= E_x\left[\Delta^x(F, G)\right] \end{aligned} \tag{14}$$

where

$$\Delta^x(F, G) = \int_{\mathbb{R}^2} [U^x(F(y|x), G(z|x)) - L^x(F(y|x), G(z|x))]\, d[F(y|x)\, G(z|x)] \tag{15}$$

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

Relationships (14) and (15) show that the unconditional uncertainty measure can be expressed as a weighted mean of conditional uncertainty measures.

As it is often the case when dealing with multivariate distribution functions, the use of copulas simplifies matters, as shown, for instance, in Nelsen (1999). Let H be a d.f. with margins F and G, respectively. Then there exists a copula C^x such that for all (z, y)

$$H(y, z|x) = C^x(F(y|x), G(z|x)) \tag{16}$$

From Sklar's theorem, if F and G are continuous then $C^x(\cdot, \cdot)$ is unique, and it is equal to $H\left(F^{-1}(y|x),\ G^{-1}(z|x)\right)$. From an intuitive point of view, the copula (16) represents the "intrinsic" association between Y and Z, disregarding their marginal d.f.s. In other words, knowledge of (i) the copula (16) and of (ii) the marginal d.f.s $F(y|x)$, $G(z|x)$ allows one to reconstruct the joint d.f. $H(y,\ z|x)$. In matching problems the marginal d.f.s $F(y|x)$ and $G(z|x)$ can be estimated on the basis of sample data; the "actual uncertainty" only involves the association expressed by the copula (16). The copula version of the Fréchet bounds (4) is given by

$$\left\{C^x(u, v) :\ W^x(u, v) \le C^x(u, v) \le M^x(u, v),\ \forall\, u,\ v \in I\right\} \tag{17}$$

Conditionally on x, both $U = F(Y|x)$ and $V = G(Z|x)$ do have uniform distribution in $(0, 1)$, regardless of the shapes of F and G. Furthermore, $W^x(u, v) = \max(0, u + v - 1)$ and $M^x(u, v) = \min(u, v)$ are themselves copulas; they represent perfect dependence, either negative or positive. As a consequence, the uncertainty measure (15) becomes the volume between the surfaces $M^x(u, v)$ and $W^x(u, v)$:

$$\Delta^x(F, G) = \int\int_{I^2} [M^x(u, v) - W^x(u, v)]\, dudv \tag{18}$$

Then, the larger $\Delta^x(F, G)$ the more uncertain the data generating statistical model. The measure $\Delta^x(F, G)$ does not depend on F and G, as shown in Proposition 1, whose proof is straightforward.

Proposition 1. $\Delta^x(F, G) = 1/6$ *for any* F *and* G*, and for every* x.

The value $\Delta^x(F, G) = 1/6$ represents the uncertainty achieved when no external auxiliary information beyond knowledge of margins $F(y|x)$ and $G(z|x)$ is available. As an easy consequence of Proposition 1, also the unconditional uncertainty measure computed as in (14) takes the value 1/6. Proposition 1 tells us that when the only available information are sample data, then the uncertainty measure is equal to 1/6, independently of the marginal d.f.s $F(y|x)$ and $G(z|x)$. This is in accordance with intuition, because on the one hand uncertainty *only depends on the maximal and minimal values of the copula* $C^x(u, v)$*, and not on the marginals* $F(y|x)$ *and* $G(z|x)$, and on the other hand, *sample data do not provide any information on* $C^x(u, v)$. When no extra-sample auxiliary information on $C^x(u, v)$ is available, the minimal and maximal values of $C^x(u, v)$ are $\max(0,\ u + v - 1)$ and $\min(u,\ v)$, respectively, *independently of the sample data*. In a sense, this is the case of *maximal* uncertainty. Proposition 1 simply establishes that the maximal uncertainty is 1/6.

This situation seems to contradict what happens when $(X,\ Y,\ Z)$ are jointly normally distributed, as in Raessler (2002). In that 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_{yx} or ρ_{zx} go continuously to 1 (or -1).

6. The use of constraints for the reduction of model uncertainty

The main goal of the present section is to evaluate the effect on the model uncertainty due to the availability of auxiliary, extra-sample information in form of *logical constraints*, i.e. constraints on the support of the d.f. $H(y, z|x)$. Intuitively speaking, they "forbid" some specific parts of the support of $H(y, z|x)$. More precisely, a logical constraint implies that $H(y, z|x)$ is forced to give zero probability to some specific parts of $\mathbb{R}^2$. In this way they imply, as expected, a smaller degree of uncertainty since some models for (X, Y, Z) must be excluded from the set of plausible distribution functions. In other terms, some models for (X, Y, Z) become illogical because they give positive probability to "forbidden" parts of $\mathbb{R}^2$, and must be excluded from the set of plausible distribution functions. This means that the Fréchet bounds can be improved by logical constraints. As a consequence, the statistical model for the data becomes less uncertain. Clearly, the reduction of the model uncertainty depends on how informative the imposed constraints are.

The use of logical constraints goes back to D'Orazio et al. (2006a), where the important case of structural zeros in contingency tables is dealt with. The idea was then extended to general discrete distributions with ordered categories in Conti et al. (2013). The extension to continuous variates requires considerable changes in *defining* logical constraints, as well as the use of completely different techniques to study their effect. This is done in the sequel of the paper.

6.1. Constraint $a_x \le f(Y, Z) \le b_x$ given X

The kind of constraints we consider is $a_x \le f(Y, Z) \le b_x$ given $X = x$, where $f(Y, Z)$ is a monotone function of Y (Z) for each Z (Y).

In other terms, the constraint $a_x \le f(Y, Z) \le b_x$ tells us that the support of $H(y,\ z\,|x)$ is a subset (either proper or improper) of $\{(y,\ z) :\ a_x \le f(y,\ z) \le b_x\}$.

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

$$\begin{aligned} H(y, z|x) &= P(Z \le z, Y \le y\,|x) \\ &= P(Z \le z, Y \le y,\ f(Y, Z) \le b_x,\ f(Y, Z) \ge a_x|x) \\ &= P(Z \le z, Z \le \gamma_y(a_x), Y \le y, Y \le \delta_z(b_x)|x) \\ &= P(Z \le (z \wedge \gamma_y(a_x)), Y \le (y \wedge \delta_z(b_x))|x) \\ &= H(z \wedge \gamma_y(a_x), y \wedge \delta_z(b_x)|x) \end{aligned} \tag{19}$$

Hence, the Fréchet bounds (4) now become

$$\begin{aligned} K^x_+(y,\ z) &= U^x(G(z \wedge \gamma_y(a_x)|x), F(y \wedge \delta_z(b_x)|x)) \\ &= \min(G(z \wedge \gamma_y(a_x)|x), F(y \wedge \delta_z(b_x)|x)) \\ &= \min(G(z|x), G(\gamma_y(a_x)|x), F(y|x), F(\delta_z(b_x)\,|x)) \end{aligned} \tag{20}$$

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

and the whole class becomes smaller than (4).

To be more explicit, the introduction of a constraint of the form $a_x \le f(Y, Z) \le b_x$ modifies the support of the joint d.f. $H(y,\ z\,|x)$, which now becomes a subset of $\{(y,\ z) \in \Re^2 :\ a_x \le f(y,\ z) \le b_x\}$. As a consequence, all the d.f.s in the Fréchet class (4) that do not satisfy this condition are now impossible, since they are not compatible with the constraint itself. Hence, the Fréchet class (4) becomes the set of all bivariate d.f.s having marginal d.f.s $F(y\,|x)$, $G(z\,|x)$ and such that

$$\{H(y,\ z\,|x) : K^x_-(y,\ z) \le H(y,\ z\,|x) \le K^x_+(y,\ z)\} \tag{22}$$

The conditional measure of uncertainty for the "constrained Fréchet class" (22) is now given by

$$\Delta^x_c(F,\ G) = \int_{\boldsymbol{R}^2} \left(K^x_+(y,\ z) - K^x_-(y,\ z)\right)\ d[F(y\,|x)\, G(z\,|x)] \tag{23}$$

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

$$\Delta_c(F,\ G) = \sum_x p(x)\ \Delta^x_c(F,\ G) \tag{24}$$

As it clearly appears from (20) and (21), the measure of uncertainty $\Delta^x_c(F,\ G)$ depends on the marginal d.f.s $F(y\,|x)$ and $G(z\,|x)$. The same holds for $\Delta_c(F,\ G)$.

Of course, the whole approach can be also developed in terms of copula. Let $u = F(y\,|x)$ and $v = G(z\,|x)$, so that $z = G^{-1}(v)$ and $y = F^{-1}(u)$. The copula version of the Fréchet bounds (20) and (21) is

$$V^x(u, v) \le C^x(u, v) \le N^x(u, v) \tag{25}$$

where

$$\begin{aligned} N^x(u, v) &= \min(v, G(\gamma_{F^{-1}(u)}(a_x)), u, F(\delta_{G^{-1}(v)}(b_x))) \\ V^x(u, v) &= \max(0, v \wedge G(\gamma_{F^{-1}(u)}(a)_x) + u \wedge F(\delta_{G^{-1}(v)}(b_x)) - 1) \end{aligned} \tag{26}$$

The conditional uncertainty measure $\Delta_c^x(F, G)$ is then

$$\Delta_c^x(F, G) = \int\int_{I^2} [N^x(u, v) - V^x(u, v)]\,du\,dv \tag{27}$$

and the corresponding unconditional uncertainty measure is

$$\Delta_c(F, G) = E_x\left(\Delta_c^x(F, G)\right) = \sum_x p(x)\Delta_c^x(F, G) \tag{28}$$

6.2. Estimation of the uncertainty measure under constraints

The uncertainty measures (23) and (24) can be easily estimated on the basis of the available data. Using the notation introduced in Section 3,

$$\widehat{K}_+^x(y,\ z) = \min\left\{\widehat{F}_{n_A}(y\,|x),\ \widehat{F}_{n_A}(\delta_z(b_x)\,|x),\ \widehat{G}_{n_B}(z\,|x),\ \widehat{G}_{n_B}(\gamma_y(a_x)\,|x)\right\} \tag{29}$$

$$\begin{aligned} \widehat{K}_-^x(y,\ z) = \max\big\{0,\ &\min(\widehat{F}_{n_A}(y\,|x),\ \widehat{F}_{n_A}(\delta_z(b_x)\,|x)) \\ &+ \min(\widehat{G}_{n_B}(z\,|x),\ \widehat{G}_{n_B}(\gamma_y(a_x)\,|x)) - 1\big\} \end{aligned} \tag{30}$$

As "natural" estimators of (23) and (24), we consider

$$\widehat{\Delta}_c^x = \int_{\boldsymbol{R}^2} \left(\widehat{K}_+^x(y,\ z) - \widehat{K}_-^x(y,\ z)\right)\, d[\widehat{F}_{n_A}(y\,|x)\,\widehat{G}_{n_B}(z\,|x)] \tag{31}$$

$$\widehat{\Delta}_c = \sum_x \widehat{p}(x)\,\widehat{\Delta}_c^x \tag{32}$$

In the next proposition, we show the consistency and asymptotic normality of the estimators (31) and (32).

Proposition 2. *Assume that $\gamma_y(a_x)$ and $\delta_z(b_x)$ are continuous functions of y and z, respectively, and that $n_A/(n_A + n_B) \to \alpha$ as n_A, n_B go to infinity, with $0 < \alpha < 1$. Then*

$$\widehat{\Delta}_c^x \xrightarrow{a.s.} \Delta_c^x(F,\ G) \ \ as\ n_A,\ n_B \to \infty \tag{33}$$

$$\widehat{\Delta}_c \xrightarrow{a.s.} \Delta_c(F,\ G) \ \ as\ n_A,\ n_B \to \infty \tag{34}$$

Proof. See Appendix A. □

The estimators (31) and (32) possess an asymptotic normal distribution, under suitable conditions. In order to write its asymptotic variance, which possesses rather a complicate form, we need to introduce some further symbols. Define the sets

$$T_1^x = \left\{(y,\ z) : K_+^x(y,\ z) = G(z\,|x)\right\},\ \ T_2^x = \left\{(y,\ z) : K_+^x(y,\ z) = G(\gamma_y(a_x)\,|x)\right\}$$

$$T_3^x = \left\{(y,\ z) : K_+^x(y,\ z) = F(y\,|x)\right\},\ T_4^x = \left\{(y,\ z) : K_+^x(y,\ z) = F(\delta_z(b_x)\,|x)\right\}$$
$$S_0^x = \left\{(y,\ z) : K_-^x(y,\ z) = 0\right\},\ S_1^x = \left\{(y,\ z) : K_-^x(y,\ z) = F(y\,|x) + G(z\,|x) - 1\right\}$$
$$S_2^x = \left\{(y,\ z) : K_-^x(y,\ z) = F(\delta_z(b_x)\,|x) + G(z\,|x) - 1\right\}$$
$$S_3^x = \left\{(y,\ z) : K_-^x(y,\ z) = F(y\,|x) + G(\gamma_y(a_x)\,|x) - 1\right\}$$
$$S_4^x = \left\{(y,\ z) : K_-^x(y,\ z) = F(\delta_z(b_x)\,|x) + G(\gamma_y(a_x)\,|x) - 1\right\}$$

and the functions

$$\tau_1^x(y,\ z) = \left\{I_{((y,z)\in T_3^x)} - I_{((y,z)\in S_1^x)} - I_{((y,z)\in S_3^x)}\right\} \tag{35}$$

$$\tau_2^x(y,\ z) = \left\{I_{((y,z)\in T_4^x)} - I_{((y,z)\in S_2^x)} - I_{((y,z)\in S_4^x)}\right\} \tag{36}$$

$$\tau_3^x(y,\ z) = \left\{I_{((y,z)\in T_1^x)} - I_{((y,z)\in S_1^x)} - I_{((y,z)\in S_2^x)}\right\} \tag{37}$$

$$\tau_4^x(y,\ z) = \left\{I_{((y,z)\in T_2^x)} - I_{((y,z)\in S_3^x)} - I_{((y,z)\in S_4^x)}\right\} \tag{38}$$

$$\beta^x(J;\ a,\ b) = \min(J(a\,|x),\ J(b\,|x)) - J(a\,|x)\,J(b\,|x);\quad J = F,\ G \tag{39}$$

$$A^x(y,\ z) = K_-^x(y,\ z) - K_+^x(y,\ z) \tag{40}$$

In the next proposition, the asymptotic normality of $\widehat{\Delta}_c^x$ is stated.

365 **Proposition 3.** *Assume that $\gamma_y(a_x)$ and $\delta_z(b_x)$ are continuous functions of y and z, respectively, and that $n_A/(n_A + n_B) \to \alpha$ as n_A and n_B go to infinity, with $0 < \alpha < 1$. Then*

$$\sqrt{\frac{n_{A,x}\,n_{B,x}}{n_{A,x} + n_{B,x}}}(\widehat{\Delta}_c^x - \Delta_c^x(F,\ G)) \xrightarrow{w} N(0,\ V(F, G;\ x))\ \ as\ n_A,\ n_B \to \infty \tag{41}$$

where

$$V(F, G;\ x) = (1-\alpha)V_1(F, G;\ x) + \alpha V_2(F, G;\ x) \tag{42}$$

with

$$\begin{aligned}
V_1(F, G;\ x) &= \int_{\boldsymbol{R}^4} G(z_1\,|x)\,G(z_2\,|x)\,\beta^x(F;\ y_1,\ y_2)\,dA^x(y_1, z_1)\,dA^x(y_2, z_2)\\
&+ \int_{\boldsymbol{R}^4} \tau_1^x(y_1,\ z_1)\,\tau_1^x(y_2,\ z_2)\beta^x(F;\ y_1,\ y_2)\,d[F(y_1\,|x)G(z_1\,|x)]\,d[F(y_2\,|x)G(z_2\,|x)]\\
&+ \int_{\boldsymbol{R}^4} \tau_2^x(y_1,\ z_1)\,\tau_2^x(y_2,\ z_2)\beta^x(F;\ \delta_{z_1}(b_x),\ \delta_{z_2}(b_x))\,d[F(y_1\,|x)G(z_1\,|x)]\\
&\times d[F(y_2\,|x)G(z_2\,|x)]\\
&+ 2\int_{\boldsymbol{R}^4} G(z_1\,|x)\,\tau_1^x(y_2,\ z_2)\beta^x(F;\ y_1,\ y_2)\,d[A^x(y_1,\ z_1)\,d[F(y_2\,|x)G(z_2\,|x)]\\
&+ 2\int_{\boldsymbol{R}^4} G(z_1\,|x)\,\tau_2^x(y_2,\ z_2)\beta^x(F;\ y_1,\ \delta_{z_2}(b_x))\,dA^x(y_1,\ z_1)\,d[F(y_2\,|x)G(z_2\,|x)]\\
&+ 2\int_{\boldsymbol{R}^4} \tau_1^x(y_1,\ z_1)\,\tau_2^x(y_2,\ z_2)\beta^x(F;\ y_1,\ \delta_{z_2}(b_x))\,d[F(y_1\,|x)G(z_1\,|x)]d[F(y_2\,|x)G(z_2\,|x)]
\end{aligned} \tag{43}$$

and

$$
\begin{aligned}
V_2(F, G;\ x) = &\int_{\boldsymbol{R}^4} F(y_1\,|x)\, F(y_2\,|x)\, \beta^x(G;\ z_1,\ z_2)\, dA^x(y_1, z_1)\, dA^x(y_2, z_2) \\
&+ \int_{\boldsymbol{R}^4} \tau_3^x(y_1,\ z_1)\, \tau_3^x(y_2,\ z_2) \beta^x(G;\ z_1,\ z_2)\, d[F(y_1\,|x) G(z_1\,|x)]\, d[F(y_2\,|x) G(z_2\,|x)] \\
&+ \int_{\boldsymbol{R}^4} \tau_4^x(y_1,\ z_1)\, \tau_4^x(y_2,\ z_2) \beta^x(G;\ \gamma_{y_1}(a_x),\ \gamma_{y_2}(a_x))\, d[F(y_1\,|x) G(z_1\,|x)] \\
&\times d[F(y_2\,|x) G(z_2\,|x)] \\
&+2 \int_{\boldsymbol{R}^4} F(y_1\,|x)\, \tau_3^x(y_2,\ z_2) \beta^x(G;\ z_1,\ z_2)\, d[A^x(y_1,\ z_1)\, d[F(y_2\,|x) G(z_2\,|x)] \\
&+2 \int_{\boldsymbol{R}^4} F(y_1\,|x)\, \tau_4^x(y_2,\ z_2) \beta^x(G;\ z_1,\ \gamma_{y_2}(a_x))\, dA^x(y_1,\ z_1)\, d[F(y_2\,|x) G(z_2\,|x)] \\
&+2 \int_{\boldsymbol{R}^4} \tau_3(y_1,\ z_1)\, \tau_4^x(y_2,\ z_2) \beta^x(G;\ z_1,\ \gamma_{y_2}(a_x)) d[F(y_1\,|x) G(z_1\,|x)]\, d[F(y_2\,|x) G(z_2\,|x)]
\end{aligned}
\tag{44}
$$

Proof. See Appendix A. □

As far as the unconditional measure of uncertainty (24) is concerned, the asymptotic normality of (32) can be obtained. Specifically, let $\boldsymbol{p}(x)$ be the vector of elements $p(x)$s, as x ranges in the support of X, and let $\widehat{\boldsymbol{p}}(x)$ be the corresponding vector of estimates $\widehat{p}(x)$s. Let further $\boldsymbol{\Sigma}$ be the squared matrix of elements

$$
\sigma(x,\ t) = \begin{cases} p(x)\,(1 - p(x)) & \text{if } t = x \\ -p(x)\, p(t) & \text{if } t \neq x \end{cases}
$$

as x, t range in the support of X. Finally, denote by $\boldsymbol{\Delta}_c^x(F,\ G)$ the vector of conditional uncertainty measures $\Delta_c^x(F,\ G)$, again as x ranges in the support of X. The following result holds.

Proposition 4. *Under the conditions of Proposition 3, we have*

$$
\sqrt{\frac{n_A\, n_B}{n_A + n_B}} (\widehat{\Delta}_c - \Delta_c(F,\ G)) \xrightarrow{w} N(0,\ T(F, G)) \quad as\ n_A,\ n_B \to \infty \tag{45}
$$

where

$$
T(F, G) = \alpha(1 - \alpha)\, \boldsymbol{\Delta}_c^x(F,\ G)^{'}\, \boldsymbol{\Sigma}\, \boldsymbol{\Delta}_c^x(F,\ G) + \sum_x p(x) V(F, G;\ x) \tag{46}
$$

and $V(F, G;\ x)$ is given by (42).

Variances (42) and (46) are awkward, and need to be estimated in order to make operational Propositions 3 and 4. The simplest way to estimate (42) and (46) consists in resorting to bootstrap. In practice, in the present case bootstrap works as follows.

1. Generate from $\widehat{F}_{n_A}(y\,|x)$ a sample of size n_A.
2. Generate from $\widehat{G}_{n_B}(z\,|x)$ a sample of size n_B.
3. Use samples generated in steps 1 and 2 to compute the "bootstrap version" $\widetilde{\Delta}_c^x$ of $\widehat{\Delta}_c^x$.

Steps 1–3 are repeated M times, so that the M bootstrap values $\widetilde{\Delta}^x_{c,m}$, $m = 1, \ldots, M$, are obtained. Let $\overline{\Delta}^x_{c,M}$ be their average, and S^{2x}_M be their variance:

$$\overline{\Delta}^x_{c,M} = \frac{1}{M}\sum_{m=1}^{M}\widetilde{\Delta}^x_{c,m}, \quad S^{2x}_M = \frac{1}{M-1}\sum_{m=1}^{M}(\widetilde{\Delta}^x_{c,m} - \overline{\Delta}^x_c)^2$$

As an estimate of $V(F, G;\ x)$, we may take

$$\widehat{V}^x_M = \left(\frac{n_{A,x}\, n_{B,x}}{n_{A,x} + n_{B,x}}\right)^{-1} S^{2x}_M \tag{47}$$

From (47) it is also easy to construct an estimate of the variance (46). It is actually enough to replace $V(F, G;\ x)$ by $\widehat{V}^x_M$, and $p(x)$, $\Delta^x_c(F,\ G)$ by their sample estimates $\widehat{p}(x)$, $\widehat{\Delta}^x_c$.

The above results are useful to construct point and interval estimates of the uncertainty measures Δ^x_c and Δ_c. They are also useful to test the hypothesis that the Fréchet class with lower and upper bounds (20) and (21), respectively, is "narrow enough", when the constraint $a_x \le f(Y,\ Z) \le b_x$ is considered. To clarify this point, let us concentrate on the conditional uncertainty measure Δ^x_c (similar considerations hold for the unconditional uncertainty measure Δ_c). Let ϵ be a "small" real number (for instance, $\epsilon = 0.01$, or less), and consider the hypothesis problem

$$\begin{cases} H_0: \ \Delta^x_c(F, G) \le \epsilon \\ H_1: \ \Delta^x_c(F, G) > \epsilon \end{cases} \tag{48}$$

Intuitively speaking, the null hypothesis postulates that the constraint $a_x \le f(Y,\ Z) \le b_x$ makes the Fréchet class narrow, and hence "close" to identifiability. In this case, the whole Fréchet class can be replaced by one of its bivariate d.f.s, which can be adopted as an approximation of the "true" joint d.f. of Y, Z (given X). The narrower the Fréchet class (namely, the smaller Δ^x_c), the smaller the bias introduced by this approximation. As a consequence of Proposition 3 and (47), a test for the hypothesis problem (48), with an asymptotic significance level γ, consists in accepting H_0 whenever

$$\widehat{\Delta}^x_c \le \epsilon + z_\gamma \left(\frac{n_{A,x}\, n_{B,x}}{n_{A,x} + n_{B,x}}\right)^{-1/2} \sqrt{S^{2x}_M}$$

where z_γ is the γth quantile of the standard normal distribution.

6.3. Examples

Example 1. A kind of constraint frequently occurring in practice is $Y \ge Z$, given X. For instance, this is the case of Okner (1972) where Y plays the role of total income and Z plays the role of income subject to taxation. The constraint $Y \ge Z$ means that, for each given x, the support of $(Y,\ Z)$ is (a subset of) the half-plane below the straight line $z = y$, *i.e.* it is (a subset, either proper or improper of) $\{(y,\ z) \in \boldsymbol{R}^2 :\ y \ge z\}$.

The constraint $Y \ge Z$ is equivalent to assume that, for each real y, the region $\{(u,\ v) :\ u \le y,\ y < v \le z\}$ does have null probability, conditionally on X. On the other hand, these relationships turn out to be equivalent to

$$H(y,\ z\,|x) = H(y,\ \min(y,\ z)\,|x) \tag{49}$$

for each y and z.

Relationship (49) modifies the Fréchet bounds in (4). The new Fréchet bounds can be obtained from the theory developed in Section 6.1. Set $f(Y, Z) = Y/Z$ with $a_x = 1$ and $b_x \to \infty$, from the results (20) and (21) we obtain the following new Fréchet bounds (4):

$$\begin{aligned} K_+^x(y,\ z) &= \min(G(z|x) \wedge G(y|x), F(y|x)) \\ &= \min(F(y|x), G(z|x), G(y|x)) \end{aligned} \tag{50}$$

$$K_-^x(y,\ z) = \max(0, G(z|x) \wedge G(y|x) + F(y|x) - 1) \tag{51}$$

According to (23) the conditional uncertainty measure is then

$$\begin{aligned} \Delta_{Y \geq Z}^x(F, G) = \int_{\boldsymbol{R}^2} &\big(\min(F(y|x), G(z|x), G(y|x)) \\ &- \max(0, G(z|x) \wedge G(y|x) + F(y|x) - 1)\big)\ d[F(y\,|x)\, G(z\,|x)] \end{aligned} \tag{52}$$

The value of (52), of course, depends on the marginal d.f.s $F(y\,|x)$ and $G(z\,|x)$. Finally, the unconditional uncertainty measure is obtained from (52) by averaging w.r.t. the distribution of X.

Example 2. Assume that there exist constants a_x and 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. For the sake of simplicity, assume further that both Y and Z are positive r.v.s. The constraint introduced above means that, for each given x, the support of $(Y,\ Z)$ is in between the two straight lines $z = y/b_x$ and $z = y/a_z$, *i.e.* is (a subset either proper or improper of) the cone $\{(y,\ z):\ y \leq 0,\ y/b_x \leq z \leq y/a_x\}$.

The constraint $a_x \leq Y/Z \leq b_x$ is equivalent to say that all regions of the form $\{(u,\ v):\ u \leq y,\ y/a_x \leq v < z\}$ and $\{(u,\ v):\ b_x z < u \leq y,\ v \leq z\}$ do have null probability, for each y, z.

Again, the relationships illustrated above modify the Fréchet bounds in (4). The new Fréchet bounds can be obtained from the theory developed in Section 6.1. Let $f(Y, Z) = Y/Z$, $\gamma_y(a_x) = y/a_x$, $\delta_z(b_x) = b_x z$, and from the results (20) and (21) we obtain the following Fréchet bounds (4):

$$\begin{aligned} K_+^x(y,\ z) &= \min\left(G\left(z \wedge \frac{y}{a_x}|x\right), F(y \wedge b_x z|x)\right) \\ &= \min\left(G(z|x) \wedge G\left(\frac{y}{a_x}|x\right), F(y|x) \wedge F(b_x z|x)\right) \\ &= \min\left(G(z|x), G\left(\frac{y}{a_x}|x\right), F(y|x), F(b_x z|x)\right) \end{aligned} \tag{53}$$

$$\begin{aligned} K_-^x(y,\ z) &= \max\left(0, G\left(z \wedge \frac{y}{a_x}|x\right) + F(y \wedge b_x z|x) - 1\right) \\ &= \max\left(0, G(z|x) \wedge G\left(\frac{y}{a_x}|x\right) + F(y|x) \wedge F(b_x z|x) - 1\right) \end{aligned} \tag{54}$$

From (53) and (54), it is possible to compute both conditional and unconditional uncertainty measures that now depend on the marginal d.f.s $F(y\,|x)$ and $G(z\,|x)$.

7. Simulation study

In this section we perform a simulation experiment in order to evaluate the effects on model uncertainty due to introduction of logical constraints. 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 $\mu_{xy} = (0, 20)$ and covariance matrix given by

$$\mathbf{\Sigma}_{xy} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$

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 $\mu_{xz} = (0, 22)$ and covariance matrix given by

$$\mathbf{\Sigma}_{xz} = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$$

3. The continuous variable X has been discretized by partitioning the range of observed values $x = x_A \cup x_B$ into $k = 4$ intervals according to the hth percentiles of data, for $h = 25 - 75(25)$. As a consequence, the discretized variable will assume the values x, for $x = 1, 2, 3, 4$.
4. Conditionally on x (for $x = 1, 2, 3, 4$) the pointwise measure of uncertainty in (x, y, z) is estimated by

$$\hat{\Delta}^{yz|x} = min(\widehat{G}_{n_B}(z|x),\ \widehat{F}_{n_A}(y|x)) - max(\widehat{G}_{n_B}(z|x) + \widehat{F}_{n_A}(y|x) - 1,\ 0) \tag{55}$$

where $\widehat{F}_{n_A}(y|x)$ and $\widehat{G}_{n_B}(z|x)$ 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 $x = j$ in samples A and B, and denote by $\boldsymbol{y} = (y_{1,x}, \ldots, y_{n_{A,x}})$ and $\boldsymbol{z} = (z_{1,x}, \ldots, z_{n_{B,x}})$ the corresponding sample values of Y and Z, respectively. Then for $x = j$ we obtain $n_{A,x}n_{B,x}$ pointwise uncertainty measures (55).
5. Conditionally on x (for $x = 1, 2, 3, 4$), the uncertainty measure when no external auxiliary information is available is obtained by averaging the $n_{A,x}n_{B,x}$ pointwise uncertainty measures (55). Formally

$$\widehat{\Delta}^{x} = \frac{1}{n_{A,x}n_{B,x}} \sum_{y \in \boldsymbol{y}} \sum_{z \in \boldsymbol{z}} \hat{\Delta}^{yz|x} \tag{56}$$

6. The overall unconditional uncertainty measure is a weighted mean of conditional uncertainty measure (56)

$$\widehat{\Delta} = \sum_{x} \Delta^{x=j} \widehat{p}(x) \tag{57}$$

where $\widehat{p}(x) = \frac{n_{A,x}+n_{B,x}}{n_A+n_B}$.
7. Suppose that the constraint regarding the statistical model for (X, Y, Z) is $Y \geq Z$. Conditionally on x and under the constraint $Y \geq Z$, the pointwise uncertainty measure in (x, y, z) is estimated by

$$\hat{\Delta}_c^{yz|x} = \widehat{K}_+^x(y,\ z) - \widehat{K}_-^x(y,\ z)$$

where the subscript c represents the constraint and $\widehat{K}^x_+(y,\ z)$ and $\widehat{\mathrm{K}}^{\mathrm{x}}_-(\mathrm{y},\ \mathrm{z})$ are given by (50) and (51), respectively.

8. Conditionally on x the estimator of conditional uncertainty measure under the constraint $Y \geq Z$ is given by

$$\widehat{\Delta}^x_c = \frac{1}{n_{A,x} n_{B,x}} \sum_{y \in \boldsymbol{y}} \sum_{z \in \boldsymbol{z}} \hat{\Delta}^{yz|x}_c \tag{58}$$

9. The overall unconditional uncertainty measure under the constraint $Y \geq Z$ is obtained as a weighted mean of conditional uncertainty measures (58)

$$\widehat{\Delta}_c = \sum_x \widehat{\Delta}^x_c \widehat{p}(x) \tag{59}$$

where $\widehat{p}(x) = \frac{n_{A,x}+n_{B,x}}{n_A+n_B}$.

10. Steps 1–9 have been repeated 500 times and for sample sizes $n_A = n_B = n = 1000$.

Given n, for each sample s (for $s = 1, \ldots, 500$) and for each category x (for $x = 1, 2, 3, 4$) denote by $\hat{\Delta}^{x,s}$,$\hat{\Delta}^s$, $\hat{\Delta}^{x,s}_c$, $\hat{\Delta}^s_c$ the uncertainty measures (56)–(59), respectively.

As for the conditional uncertainty measure estimates (56) their mean over simulation runs is

$$\overline{\Delta}^x = \frac{1}{500} \sum_{s=1}^{500} \hat{\Delta}^{x,s} \tag{60}$$

while

$$\overline{\Delta} = \frac{1}{500} \sum_{s=1}^{500} \hat{\Delta}^s \tag{61}$$

represents the corresponding overall uncertainty.

Analogously, if we refer to the conditional uncertainty measure estimates (31), we have that the mean over simulation runs is given by

$$\overline{\Delta}^x_c = \frac{1}{500} \sum_{s=1}^{500} \hat{\Delta}^{x,s}_c \tag{62}$$

while

$$\overline{\Delta}_c = \sum_{s=1}^{500} \overline{\Delta}^s_c \tag{63}$$

is the corresponding overall uncertainty under the constraint $Y \geq Z$. Finally, given Propositions 3 and 4 and resorting to bootstrap confidence intervals for the uncertainty measures $\Delta^x_c(F, G)$ and $\Delta_c(F, G)$ have been constructed. Furthermore, the hypothesis (48) both for the conditional and unconditional uncertainty measures $\Delta^x_c(F, G)$ and $\Delta_c(F, G)$ has been tested.

7.1. Simulation results

The uncertainty measure $\overline{\Delta}$ is equal to $1/6 = 0.166$. The introduction of the constraint $Y \geq Z$ reduces the model uncertainty for (X, Y, Z) to $\overline{\Delta}_c \cong 0.0154$.

In Table 2, conditionally on x, the uncertainty measures $\overline{\Delta}^x$ and $\overline{\Delta}^x_c$ given by (60) and (62), respectively, are reported for $x = 1, \ldots, 4$. Last column in Table 2 reports the percentage of

Table 2. Uncertainty measures as the category $x = j$ varies.

x	$\overline{\Delta}^x$	$\overline{\Delta}_c^x$	% *of support reduction*
1	0.1666	0.0339	92
2	0.1666	0.0130	96
3	0.1666	0.0083	97
4	0.1666	0.0063	98

sample observations that does not satisfy the constraint $Y \geq Z$. In other words, last column represents in percentage terms the effect of the constraint on the support reduction of the joint distribution of (Y, Z) given X.

Clearly, the larger the reduction of support induced by the constraint the larger is the effect of constraint on model uncertainty, that is more informative is the constraint. As Table 2 shows, the larger reduction regards the category $x = 4$ and the smaller regards the category $x = 1$, with a percentage equal to 98% and 92%, respectively.

In order to construct interval estimates for the uncertainty measure $\Delta_c^x(F, G)$ we resort to bootstrap. From sample $s = 1$, $M = 500$ bootstrap replications have been drawn and the M bootstrap estimates $\tilde{\Delta}_{c,m}^x$, $m = 1, \ldots, M$, their mean $\overline{\Delta}_{c,M}^x$, variance S_M^{2x}, and variance estimate $\hat{V}_M^x$ given by (47) have been computed.

Conditionally on x, the results are reported in Table 3 where inf^x and sup^x represent the lower and the upper endpoints of the $1 - \gamma = 0.95$ confidence interval, respectively.

The hypothesis $H_0 : \Delta_c^x(F, G) \leq \epsilon$ against $H_1 : \Delta_c^x(F, G) > \epsilon$ with an asymptotic significance level $\gamma = 0.05$ has been tested. Conditionally on $x = 1$, the hypothesis H_0 is accepted for $\epsilon = 0.05$. Conditionally on $x = 2$, the hypothesis H_0 is accepted for $\epsilon = 0.01$. Conditionally on $x = 3$ ($x = 4$), the hypothesis H_0 is accepted for $\epsilon = 0.005$.

Similar considerations hold for the unconditional uncertainty measure $\Delta_c(F, G)$. The estimate of the asymptotic variance $T(F, G)$ is 0.0016, where $\alpha = 0.5$. The estimate of $\Delta_c(F, G)$ over the bootstrap replications is 0.016. The confidence interval is (0.0134, 0.0205). Given $\gamma = 0.05$, the hypothesis $H_0 : \Delta_c(F, G) \leq \epsilon$ against $H_1 : \Delta_c(F, G) > \epsilon$ is accepted for $\epsilon = 0.02$.

Table 3. Confidence interval for $\Delta_c^x(F, G)$ as the category x varies.

x	$\overline{\Delta}_{c,M}^x$	$n_{A,x}$	$n_{B,x}$	$\hat{V}_M^x$	inf^x	sup^x
1	0.0413	244	256	0.0045	0.0296	0.0531
2	0.0138	251	249	0.0010	0.0082	0.0195
3	0.0078	245	239	0.0005	0.0039	0.0117
4	0.0050	250	250	0.0003	0.0022	0.0078

Table 4. Bounds of $H(y, z|X = 1)$ for some percentiles of the Y and Z marginal distributions.

	10		20		30		40	
$Y\|Z$	$L^x(y, z)$	$U^x(y, z)$	$L^x(y, z)$	$U^x(y, z)$	$L^x(y, z)$	$U^x(y, z)$	$L^x(y, z)$	$U^x(y, z)$
10	0	0.180	0	0.180	0	0.180	0.055	0.180
20	0	0.359	0	0.365	0.130	0.365	0.240	0.365
30	0	0.359	0.118	0.520	0.286	0.520	0.395	0.520
40	0.019	0.359	0.257	0.598	0.425	0.660	0.535	0.660
50	0.105	0.359	0.343	0.598	0.511	0.746	0.621	0.746
80	0.314	0.359	0.552	0.598	0.720	0.766	0.830	0.875

Table 5. Bounds of $H(y, z|X = 1)$ under the constraint $Y \geq Z$ for some percentiles of the Y and Z marginal distributions.

	10		20		30		40	
$Y\|Z$	$K^x_-(y,z)$	$K^x_+(y,z)$	$K^x_-(y,z)$	$K^x_+(y,z)$	$K^x_-(y,z)$	$K^x_+(y,z)$	$K^x_-(y,z)$	$K^x_+(y,z)$
10	0	0.004	0	0.004	0	0.004	0	0.004
20	0	0.016	0	0.016	0	0.016	0	0.016
30	0	0.027	0	0.027	0	0.027	0	0.027
40	0	0.059	0	0.059	0	0.059	0	0.059
50	0	0.121	0	0.121	0	0.121	0	0.121
80	0.314	0.359	0.400	0.445	0.400	0.445	0.400	0.445

Conditionally on $x = 1$, Tables 4 and 5 report the bounds $(L^x_-(y,z), U^x_+(y,z))$and $(K^x_-(y,z), K^x_+(y,z))$ corresponding to some percentiles of Y and Z marginal distributions, respectively.

Note that, if the point (y, z) does not satisfy the constraint $Y \geq Z$ (as for $y =$ 10th, 20th, 30th, 40th, 50th percentiles of Y marginal distribution and for $z =$ 10th, 20th, 30th, 40th percentiles of Z marginal distribution) then the conditional bounds $(K^x_-(y,z), K^x_+(y,z))$ are shorter than $(L^x(y,z), U^x(y,z))$. As a consequence, the pointwise uncertainty measure becomes smaller. On the other side, if the point (y, z) satisfies the constraint $Y \geq Z$, as happens for $y =$ 80th percentile of Y marginal distribution and for $z =$ 10th, 20th, 30th, 40th percentiles of Z marginal distribution, then the bounds $(K^x_-(y,z), K^x_+(y,z))$ are the same as $(L^x(y,z), U^x(y,z))$.

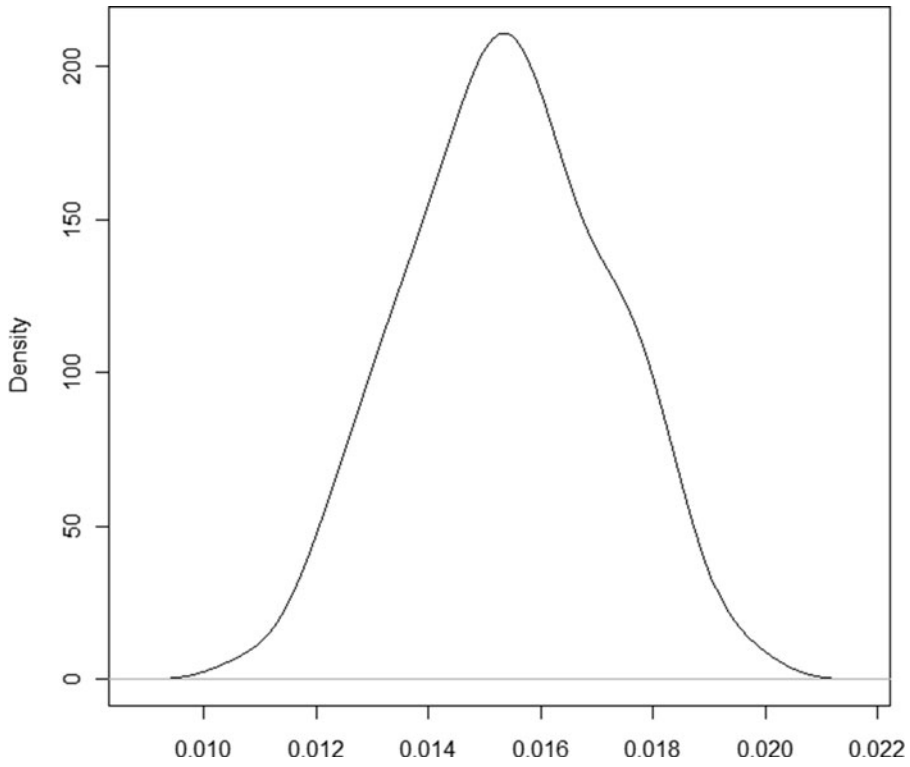

Figure 1. Density estimate of overall uncertainty measure under the constraint $Y \geq Z$.

Finally, in Figure 1 the Kernel density estimate of the overall conditional uncertainty measure $\hat{\Delta}_c$ (Proposition 4) under the constraint $Y \geq Z$ is shown. Such an estimate has been computed using the 500 values $\hat{\Delta}_c$ given by (32). Note that the uncertainty measure distribution tends to a normal distribution. The bandwidth selection rule is given by Sheather and Jones (1991).

8. Conclusions

The first statistical matching procedures (e.g. Okner , 1972) were based on intrinsic non parametric methods of imputation, as those belonging to the hot-deck class. Anyway, contrary to what happens in the parametric Gaussian case (as in Kadane, 1978; Moriarity and Scheuren, 2001; Raessler, 2002) or in the categorical case (as in D'Orazio et al., 2006a), there has never been a discussion of how uncertain the statistical matching results are when a non parametric setup is considered. This paper addresses this issue, defining appropriate measures of uncertainty in a non parametric framework and properties of uncertainty width estimators useful for confidence intervals and tests. Furthermore, emphasis is given to the possibility to reduce uncertainty when logical constraints between the never jointly observed variables are introduced.

References

Adamek, J.C. (1994). Fusion: combining data from separate sources. *Market. Res.: Mag. Manage. Appl.* 6:48–50.

Anderson, T.W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *J. Am. Stat. Assoc.* 52:200–203.

Billingsley, P. (1968). *Convergence of Probability Measures*. New York: Wiley.

Chernozhukov, V., Hong, H., Tamer, E. (2007). Estimation and confidence regions for parameter sets in econometric models. *Econometrica* 75:1243–1284.

Conti, P.L., Marella, D., Scanu, M. (2012). Uncertainty analysis in statistical matching. *J. Off. Stat.* 28:69–88.

Conti, P.L., Marella, D., Scanu, M. (2013). Uncertainty analysis for statistical matching of ordered categorical variables. *Comput. Stat. Data Anal.* 68:311–325.

Cross, P.J., Manski, C.F. (2002). Regressions, short and long. *Econometrica* 70:357–368.

D'Orazio, M., Di Zio, M., Scanu, M. (2006). Statistical matching for categorical data: Displaying uncertainty and using logical constraints. *J. Off. Stat.* 22:137–157.

D'Orazio, M., Di Zio, M., Scanu, M. (2006). *Statistical Matching: Theory and Practice*. New York: Wiley.

Dobra, A., Fienberg, S.E. (2001). Bounds for cell entries in contingency tables induced by fixed marginal totals with applications to disclosure limitation. *Stat. J. United Nations ECE* 18:363–371.

Fan, Y., Park, S.S. (2009a). Sharp bounds on the distribution of treatment effects and their statistical inference. *Economet. Theory* doi: 10.1017/S0266466609990168.

Fan, Y., Park, S.S. (2009b). Partial identification of the distribution of treatment effects and its confidence sets. In: Li, Q., Racine, J.S., eds. *Advances in Econometrics: Nonparametric Econometric Methods*, Vol. 25 (pp. 3–70). Bingley, UK: Emerald Group Publishing Limited.

Gazzelloni, S., Romano, M.C., Corsetti, G., Di Zio, M., D'Orazio, M., Pintaldi, F., Scanu, M., Torelli, N. (2007). Time use and labour force: A proposal to integrate the data through statistical matching. In: Romano, M.C., ed. *Time Use in Daily Life*, collana Argomenti, 32 (pp. 297–320).

Gaenssler, P., Stute, W. (1979). Empirical processes: A survey of results for independent and identically distributed random variables. *Ann. Probab.* 7:193–243.

Hall, P., Wolff, R.C.L, Yao, M. (1999). Methods for estimating a conditional distribution function. *J. Am. Stat. Assoc.* 94:154–163.

Hildebrandt, T.H. (1963). *Introduction to the Theory of Integration*. New York: Academic Press.

Horowitz, J.L., Manski, C.F. (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. *J. Am. Stat. Assoc.* 95:77–84.

Kadane, J.B. (1978). Some statistical problems in merging data files. In Compendium of tax research, Department of Treasury, U.S. Government Printing Office, Washington, D.C., 159–179 (Reprinted in 2001). *J. Off. Stat.* 17:423–433.

Kiesl, H., Raessler, S. (2008). The validity of data fusion. CENEX-ISAD workshop, Vienna 20–30 May 2008. Available at: http://cenex-isad.istat.it.

King, G. (1997). *A Solution to the Ecological Inference Problem*. Princeton: Princeton University Press.

Luzi, O., Di Zio, M., Guarnera, U., Manzari, A., De Waal, T., Pannekoek, J., Hoogland, J., Tempelman, C., Hulliger, B., Kilchmann, D. (2007). Recommended practices for editing and imputation in cross-sectional business surveys, Istat, CBS, SFSO, Eurostat.

Q3 Manski, C.F. (1995). *Identification Problems in the Social Sciences*. Harvard University Press.

Moriarity, C., Scheuren, F. (2001). Statistical matching: A paradigm of assessing the uncertainty in the procedure. *J. Off. Stat.* 17:407–422.

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

Nelsen, R.B., Quesada Molina, J.J., Rodriguez Lallena, J.A., Úbeda Flores, M. (2001). Bounds on bivariate distribution functions with given margins and measures of association. *Commun. Stat. - Theory Methods* 30:1155–1162.

Okner, B.A. (1972). Constructing a new data base from existing microdata sets: The 1966 merge file. *Ann. Econ. Soc. Meas.* 1:325–342.

Owen, A.B. (1995). Nonparametric likelihood confidence bands for a distribution function. *J. Am. Stat. Assoc.* 90:516–521.

Owen, A.B. (2001). *Empirical Likelihood*. London: Chapman & Hall.

Qin, J., Lawless, J. (1994). Empirical likelihood and general estimating equations. *Ann. Stat.* 22:300–325.

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

Raessler, S., Kiesl, H. (2009). How useful are uncertainty bounds? Some recent theory with an application to Rubin's causal model. *Proceedings of the 57th Session of the International Statistical Institute*, Durban, South Africa, 16–22 August, 2009.

Reiter, J.P. (2012). Bayesian finite population imputation for data fusion. *Stat. Sin.* 22:795–811.

Ridder, G., Moffitt, R. (2007). The econometrics of data combination. In: Heckmann, J.J., Leamer, E.E., eds. *Handbook of Econometrics*, Vol. 6 (pp. 5469–5547). Elsevier.: Amsterdam

Rodgers, W.L. (1984). An evaluation of statistical matching. *J. Bus. Econ. Stat.* 2:91–102.

Rubin, D.B. (1974). Characterizing the estimation of parameters in incomplete-data problems. *J. Am. Stat. Assoc.* 69:467–474.

Rubin, D.B. (1986). Statistical matching with adjusted weights and multiple imputations. *J. Bus. Econ. Stat.* 4:87–94.

Sheather, S.J., Jones, M.C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. Royal Stat. Soc. Ser. B* 53:683–690.

Sims, C.A. (1972). Comments on: "Constructing a new data base from existing microdata sets: the 1966 merge file", by B.A. Okner. *Ann. Econ. Soc. Meas.* 1:343–345.

Singh, A.C., Mantel, H., Kinack, M., Rowe, G. (1993). Statistical matching: Use of auxiliary information as an alternative to the conditional independence assumption. *Surv. Methodol.* 19:59–79.

Tonkin, R., Webber, D. (2012). Statistical matching of EU-SILC and Household Budget Survey to compare poverty estimates using income, expenditures and material deprivation, EU-SILC International Conference, Vienna, 6–7 December 2012.

Torelli, N., Ballin, M., D'Orazio, M., Di Zio, M., Scanu, M., Corsetti, G. (2008). Statistical matching of two surveys with a nonrandomly selected common subset. Workshop of the CENEX-ISAD (Integration of Survey and Administrative Data) project, Vienna 29–30 May 2008. http://cenex-isad.istat.it.

Wolfson, M., Gribble, S., Bordt, M., Murphy, B., Rowe, G., Scheuren, F. (1989). The social policy simulation database and model: An example of survey and administrative data integration. *Surv. Curr. Bus.* May, 1989.

Q4

A. Appendix

Proof of Proposition 2. First of all, the inequality

$$|\widehat{\Delta}_c^x - \Delta_c^x(F,\ G)| \le A_1 + A_2 \tag{A.1}$$

holds, where

$$A_1 = \sup_{y,z} \left|\widehat{K}_+^x(y,\ z) - K_+^x(y,\ z)\right| + \sup_{y,z} \left|\widehat{K}_-^x(y,\ z) - K_-^x(y,\ z)\right|$$

$$A_2 = \left| \int_{\boldsymbol{R}^2} (K_+^x(y,\ z) - K_-^x(y,\ z))\, d[\widehat{F}_{n_A}(y\,|x)\, \widehat{G}_{n_B}(z\,|x)] - \Delta_c^x(F,\ G) \right|$$

From Glivenko–Cantelli theorem and the continuity assumption on $\gamma_y(a_x), \delta_z(b_x)$, it is not difficult to see that

$$\sup_y \left|\widehat{F}_{n_A}(y\,|x) - F(y\,|x)\right| \xrightarrow{a.s.} 0,\ \sup_z \left|\widehat{F}_{n_A}(\delta_z(b_x)\,|x) - F(\delta_z(b_x)\,|x)\right| \xrightarrow{a.s.} 0 \tag{A.2}$$

$$\sup_z \left|\widehat{G}_{n_B}(z\,|x) - G(z\,|x)\right| \xrightarrow{a.s.} 0,\ \sup_y \left|\widehat{G}_{n_B}(\gamma_y(a_x)\,|x) - G(\gamma_y(b_x)\,|x)\right| \xrightarrow{a.s.} 0 \tag{A.3}$$

as n_A and n_B go to infinity. Hence, taking into account the definition of K_+, K_-, $\widehat{K}_+$, $\widehat{K}_-$, we have

$$A_1 \xrightarrow{a.s.} 0 \ \ as\ n_A,\ n_B \to \infty$$

In the second place, from the strong law of large numbers it is immediate to see that

$$\int_{\boldsymbol{R}^2} (K_+^x(y,\ z) - K_-^x(y,\ z))\, d[\widehat{F}_{n_A}(y\,|x)\, \widehat{G}_{n_B}(z\,|x)] \xrightarrow{a.s.} \Delta_c^x(F,\ G) \ \ as\ n_A,\ n_B \to \infty$$

which also implies that A_2 tends a.s. to 0 as n_A and n_B increase. As a consequence, (33) follows. Result (34) follows from (33) and the strong law of large numbers applied to $\widehat{p}(x)$s. □

Proof of Proposition 3. The asymptotic normality (41) is deduced from some basic results on empirical processes. A survey is in Gaenssler and Stute (1979). As well known (see, for instance, Billingsley, 1968, pp. 144–145), as n_A and n_B go to infinity, the two sequences of (independent) empirical processes

$$\left(W_{1n_A}^x(y);\ y \in \boldsymbol{R}\right) = \left(\sqrt{n_{A,x}}(\widehat{F}_{n_A}(y\,|x) - F(y\,|x));\ y \in \boldsymbol{R}\right)$$

$$\left(W_{2n_B}^x(z);\ z \in \boldsymbol{R}\right) = \left(\sqrt{n_{B,x}}(\widehat{G}_{n_B}(z\,|x) - G(z\,|x));\ z \in \boldsymbol{R}\right)$$

converge weakly to independent Gaussian processes $(W_1^x(y);\ y \in \boldsymbol{R})$ and $(W_2^x(z);\ z \in \boldsymbol{R})$, with null mean functions and covariance kernels $\min(F(y_1\,|x),\ F(y_2\,|x)) - F(y_1\,|x)\, F(y_2\,|x)$ and $\min(G(z_1\,|x),\ G(z_2\,|x)) - G(z_1\,|x)\, G(z_2\,|x)$, respectively.

Furthermore, from Skorokhod's representation theorem there exist versions $\widetilde{W}_{1n_A}^x$, $\widetilde{W}_{2n_B}^x$, $\widetilde{W}_1^x$, $\widetilde{W}_2^x$ of $W_{1n_A}^x$, $W_{2n_B}^x$, W_1^x, W_2^x , respectively, defined on an appropriate probability space $(\widetilde{\Omega},\ \widetilde{\mathcal{F}},\ \widetilde{P})$, such that

$$\widetilde{W}_{1n_A}^x \overset{d}{=} W_{1n_A}^x,\ \widetilde{W}_{2n_B}^x \overset{d}{=} W_{2n_B}^x\ \ \forall\, n_{A,x},\ n_{B,x} \ge 1;\ \ \widetilde{W}_1^x \overset{d}{=} W_1^x,\ \widetilde{W}_2^x \overset{d}{=} W_2^x$$

and

$$\sup_y \left|\widetilde{W}_{1n_A}^x(y) - \widetilde{W}_1^x(y)\right| \to 0,\ \sup_z \left|\widetilde{W}_{2n_A}^x(z) - \widetilde{W}_2^x(z)\right| \to 0 \ \ as\ n_A,\ n_B \to \infty,\ \ a.s. - \widetilde{P} \tag{A.4}$$

In order to prove (41), consider first the term

$$\sqrt{\frac{n_{A,x}n_{B,x}}{n_{A,x}+n_{B,x}}}\left(\int_{\boldsymbol{R}^2}\widehat{K}^x_+(y,\,z)\,d[\widehat{F}_{n_A}(y\,|x)\widehat{G}_{n_B}(z\,|x)]-\int_{\boldsymbol{R}^2}K^x_+(y,\,z)\,d[F(y\,|x)G(z\,|x)]\right)$$
$$=I_1+I_2 \tag{A.5}$$

where

$$I_1=\sqrt{\frac{n_{A,x}n_{B,x}}{n_{A,x}+n_{B,x}}}\int_{\boldsymbol{R}^2}\widehat{K}^x_+(y,\,z)\,d[\widehat{F}_{n_A}(y\,|x)\widehat{G}_{n_B}(z\,|x)-F(y\,|x)G(z\,|x)]$$

$$I_2=\sqrt{\frac{n_{A,x}n_{B,x}}{n_{A,x}+n_{B,x}}}\int_{\boldsymbol{R}^2}\left(\widehat{K}^x_+(y,\,z)-K^x_+(y,\,z)\right)\,d[F(y\,|x)G(z\,|x)]$$

In a similar way, it is possible to write

$$\sqrt{\frac{n_{A,x}n_{B,x}}{n_{A,x}+n_{B,x}}}\left(\int_{\boldsymbol{R}^2}\widehat{K}^x_-(y,\,z)\,d[\widehat{F}_{n_A}(y\,|x)\widehat{G}_{n_B}(z\,|x)]-\int_{\boldsymbol{R}^2}K^x_-(y,\,z)\,d[F(y\,|x)G(z\,|x)]\right)$$
$$=I_3+I_4 \tag{A.6}$$

where

$$I_3=\sqrt{\frac{n_{A,x}n_{B,x}}{n_{A,x}+n_{B,x}}}\int_{\boldsymbol{R}^2}\widehat{K}^x_-(y,\,z)\,d[\widehat{F}_{n_A}(y\,|x)\widehat{G}_{n_B}(z\,|x)-F(y\,|x)G(z\,|x)]$$

$$I_4=\sqrt{\frac{n_{A,x}n_{B,x}}{n_{A,x}+n_{B,x}}}\int_{\boldsymbol{R}^2}\left(\widehat{K}^x_-(y,\,z)-K^x_-(y,\,z)\right)\,d[F(y\,|x)G(z\,|x)]$$

The proof can be split into four steps.

Claim 1. As n_A and n_B tend to infinity, we have

$$I_1\xrightarrow{w}\sqrt{1-\alpha}\int_{\boldsymbol{R}}W^x_1(y)\,dF(y\,|x)+\sqrt{\alpha}\int_{\boldsymbol{R}}W^x_2(z)\,dG(z\,|x)$$
$$-\sqrt{1-\alpha}\int_{\boldsymbol{R}^2}G(z\,|x)W^x_1(y)\,dK^x_+(y,\,z)$$
$$-\sqrt{\alpha}\int_{\boldsymbol{R}^2}F(y\,|x)W^x_2(z)\,dK^x_+(y,\,z) \tag{A.7}$$

An integration by parts (see, for instance, Hildebrandt , 1963, p. 127) shows that

$$I_1=\sqrt{\frac{n_{A,x}n_{B,x}}{n_{A,x}+n_{B,x}}}\int_{\boldsymbol{R}}(\widehat{F}_{n_A}(y\,|x)-F(y\,|x))\,d\widehat{K}^x_+(y,\,+\infty)$$
$$+\sqrt{\frac{n_{A,x}n_{B,x}}{n_{A,x}+n_{B,x}}}\int_{\boldsymbol{R}}(\widehat{G}_{n_B}(z\,|x)-G(z\,|x))\,d\widehat{K}^x_+(+\infty,\,z)$$
$$-\sqrt{\frac{n_{A,x}n_{B,x}}{n_{A,x}+n_{B,x}}}\int_{\boldsymbol{R}^2}\left\{\widehat{G}_{n_B}(z\,|x)\,(\widehat{F}_{n_A}(y\,|x)-F(y\,|x))\right.$$
$$\left.+F(y\,|x)\,(\widehat{G}_{n_B}(z\,|x)-G(z\,|x))\right\}\,d\widehat{K}^x_+(y,\,z)$$
$$=\sqrt{\frac{n_{B,x}}{n_{A,x}+n_{B,x}}}\int_{\boldsymbol{R}}W^x_{1n_A}(y)\,d\left(F(y\,|x)+n^{-1/2}_{A,x}W^x_{1n_A}(y)\right)$$

$$
\begin{aligned}
&+\sqrt{\frac{n_{A,x}}{n_{A,x}+n_{B,x}}}\int_{\boldsymbol{R}} W^x_{2n_B}(y)\,d\left(G(z\,|x)+n_{B,x}^{-1/2}W^x_{2n_B}(z)\right)\\
&-\sqrt{\frac{n_{B,x}}{n_{A,x}+n_{B,x}}}\int_{\boldsymbol{R}^2}(G(z\,|x)+n_{B,x}^{-1/2}W^x_{2n_B}(z))W^x_{1n_A}(y)\,d\widehat{K}^x_+(y,\,z)\\
&-\sqrt{\frac{n_{A,x}}{n_{A,x}+n_{B,x}}}\int_{\boldsymbol{R}^2}F(y\,|x)W^x_{2n_B}(z)\,d\widehat{K}^x_+(y,\,z)\\
\overset{d}{=}&\sqrt{\frac{n_{B,x}}{n_{A,x}+n_{B,x}}}\int_{\boldsymbol{R}}\widetilde{W}^x_{1n_A}(y)\,d\left(F(y\,|x)+n_{A,x}^{-1/2}\widetilde{W}^x_{1n_A}(y)\right)\\
&+\sqrt{\frac{n_{A,x}}{n_{A,x}+n_{B,x}}}\int_{\boldsymbol{R}}\widetilde{W}^x_{2n_B}(z)\,d\left(G(z\,|x)+n_{B,x}^{-1/2}\widetilde{W}^x_{2n_B}(z)\right)\\
&-\sqrt{\frac{n_{B,x}}{n_{A,x}+n_{B,x}}}\int_{\boldsymbol{R}^2}\left(G(z\,|x)+n_{B,x}^{-1/2}\widetilde{W}^x_{2n_B}(z)\right)\widetilde{W}^x_{1n_A}(y)\,d\widetilde{K}^x_+(y,\,z)\\
&-\sqrt{\frac{n_{A,x}}{n_A+n_B}}\int_{\boldsymbol{R}^2}F(y\,|x)\widetilde{W}^x_{2n_B}(z)\,d\widetilde{K}^x_+(y,\,z)\\
=&\sqrt{\frac{n_{A,x}}{n_{A,x}+n_{B,x}}}\int_{\boldsymbol{R}}\widetilde{W}^x_1(y)\,dF(y\,|x)+\sqrt{\frac{n_{B,x}}{n_{A,x}+n_{B,x}}}\int_{\boldsymbol{R}}\widetilde{W}^x_2(z)\,dG(z\,|x)\\
&-\sqrt{\frac{n_{B,x}}{n_{A,x}+n_{B,x}}}\int_{\boldsymbol{R}^2}G(z\,|x)\widetilde{W}^x_1(y)\,d\widetilde{K}^x_+(y,\,z)\\
&-\sqrt{\frac{n_{A,x}}{n_{A,x}+n_{B,x}}}\int_{\boldsymbol{R}^2}F(y\,|x)\widetilde{W}^x_2(z)\,d\widetilde{K}^x_+(y,\,z)+Rem_1 \qquad \text{(A.8)}
\end{aligned}
$$

where $\widetilde{K}^x_+(y,\,z)$ is defined exactly as $\widehat{K}^x_+(y,\,z)$, except that $\widehat{F}_{n_A}(\cdot|x)$ and $\widehat{G}_{n_B}(\cdot|x)$ are replaced by their $\widetilde{P}$ versions $F(\cdot|x)-n_{A,x}^{-1/2}\widetilde{W}^x_{1n_A}(\cdot)$, $G(\cdot\,|x)+n_{B,x}^{-1/2}\widetilde{W}^x_{2n_B}(\cdot)$, respectively, and

$$
|Rem_1|\le 2\left\{\sup_y\left|\widetilde{W}^x_{1n_A}(y)-\widetilde{W}^x_1(y)\right|+\sup_z\left|\widetilde{W}^x_{2n_B}(z)-\widetilde{W}^x_2(z)\right|\right\} \qquad \text{(A.9)}
$$

Now, $n_{A,x}/(n_{A,x}+n_{B,x})$ and $n_{B,x}/(n_{A,x}+n_{B,x})$ converge a.s. to α and $1-\alpha$, respectively. Furthermore $\widetilde{K}^x_+(y,\,z)$ converges pointwise to $K^x_+(y,\,z)$ as n_A and n_B go to infinity, a.s.-$\widetilde{P}$. Since $\widetilde{W}^x_1$ and $\widetilde{W}^x_2$ possess a.s. continuous and bounded trajectories, from the Helly–Bray theorem it follows that

$$
\begin{aligned}
&\int_{\boldsymbol{R}^2}G(z\,|x)\widetilde{W}^x_1(y)\,d\widetilde{K}^x_+(y,\,z)\to\int_{\boldsymbol{R}^2}G(z\,|x)\widetilde{W}^x_1(y)\,dK^x_+(y,\,z)\\
&\int_{\boldsymbol{R}^2}F(y\,|x)\widetilde{W}^x_2(z)\,dK^x_+(y,\,z)\to\int_{\boldsymbol{R}^2}F(y\,|x)\widetilde{W}^x_2(z)\,dK^x_+(y,\,z)
\end{aligned}
$$

a.s.-$\widetilde{P}$, as n_A and n_B tend to infinity. Furthermore, from (A.9) it follows that Rem_1 tends to zero a.s.-$\widetilde{P}$, as n_A and n_B increase. Hence, in view of (A.8) convergence (A.7) is proved.

Claim 2. As n_A and n_B tend to infinity, we have

$$I_2 \xrightarrow{w} \int_{\mathbf{R}^2} \Big\{ \sqrt{\alpha} \left(I_{((y,z)\in T_1^x)} W_2^x(z) + I_{((y,z)\in T_2^x)} W_2^x(\gamma_y(a_x)) \right) + \sqrt{1-\alpha} \left(I_{((y,z)\in T_3^x)} W_1^x(y) + I_{((y,z)\in T_4^x)} W_1^x(\delta_z(b_x)) \right) \Big\} \, d[F(y\,|x)G(z\,|x)] \tag{A.10}$$

Define

$$I_{2j} = \sqrt{\frac{n_{A,x} n_{B,x}}{n_{A,x} + n_{B,x}}} \int_{\mathbf{R}^2} I_{((y,z)\in T_j^x)} \left(\widehat{K}_+^x(y,\, z) - K_+^x(y,\, z) \right) d[F(y\,|x)G(z\,|x)], \qquad j = 1,\, \ldots,\, 4 \tag{A.11}$$

As $(y,\, z)$ is in T_1^x, we have

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

Due to the consistency of the e.d.f., the terms $\widehat{F}_{n_A}(\delta_z(b_x)\,|x) - G(z\,|x)$ and $\widehat{G}_{n_B}(y\,|x) - G(z\,|x)$, $\widehat{G}_{n_B}(\gamma_y(a_x)\,|x) - G(z\,|x)$ converge a.s. to positive constants for every $(x,\, y)$ in T_1^x. When multiplied either by $\sqrt{n_{A,x}}$ or by $\sqrt{n_{B,x}}$, they tend to infinity a.s. Hence, 670

$$\sqrt{\frac{n_{A,x} n_{B,x}}{n_{A,x} + n_{B,x}}} \left(\widehat{K}_+^x(y,\, z) - K_+^x(y,\, z) \right) I_{((y,z)\in T_1^x)} \xrightarrow{w} \sqrt{\alpha} W_2^x(z) I_{((y,z)\in T_1^x)} \tag{A.12}$$

as n_A and n_B tend to infinity. From this result, it follows that

$$I_{21} \xrightarrow{w} \sqrt{\alpha} \int_{\mathbf{R}^2} W_2^x(z) I_{((y,z)\in T_1^x)} \, d[F(y\,|x)G(z\,|x)] \tag{A.13}$$

as n_A and n_B go to infinity. In the same way, it can be shown that

$$I_{22} \xrightarrow{w} \sqrt{\alpha} \int_{\mathbf{R}^2} W_2^x(\gamma_y(a_x)) I_{((y,z)\in T_2^x)} \, d[F(y\,|x)G(z\,|x)] \tag{A.14}$$

$$I_{23} \xrightarrow{w} \sqrt{1-\alpha} \int_{\mathbf{R}^2} W_1^x(y) I_{((y,z)\in T_3^x)} \, d[F(y\,|x)G(z\,|x)] \tag{A.15}$$

$$I_{24} \xrightarrow{w} \sqrt{1-\alpha} \int_{\mathbf{R}^2} W_1^x(\delta_z(b_x)) I_{((y,z)\in T_4^x)} \, d[F(y\,|x)G(z\,|x)] \tag{A.16}$$

as n_A and n_B increase. From (A.14) to (A.16), the conclusion (A.10) easily follows. 675

Claim 3. As n_A and n_B tend to infinity, we have

$$I_3 \xrightarrow{w} \sqrt{1-\alpha}\int_{\boldsymbol{R}} W_1^x(y)\,dF(y\,|x) + \sqrt{\alpha}\int_{\boldsymbol{R}} W_2^x(z)\,dG(z\,|x) - \sqrt{1-\alpha}\int_{\boldsymbol{R}^2} G(z\,|x)W_1^x(y)\,dK_-^x(y,\,z) - \sqrt{\alpha}\int_{\boldsymbol{R}^2} F(y\,|x)W_2^x(z)\,dK_-^x(y,\,z) \tag{A.17}$$

Claim 4. As n_A and n_B tend to infinity, we have

$$\begin{aligned} I_4 \xrightarrow{w} \int_{\boldsymbol{R}^2} \Big\{ & \sqrt{\alpha}\left(I_{((y,z)\in S_1^x)} + I_{((y,z)\in S_2^x)}\right) W_2^x(z) \\ & + \sqrt{\alpha}\left(I_{((y,z)\in S_3^x)} + I_{((y,z)\in S_4^x)}\right) W_2^x(\gamma_y(a_x)) \\ & + \sqrt{1-\alpha}\left(I_{((y,z)\in S_1^x)} + I_{((y,z)\in S_4^3)}\right) W_1^x(y) \\ & + \sqrt{1-\alpha}\left(I_{((y,z)\in S_2^x)} + I_{((y,z)\in S_4^x)}\right) W_2^x(\delta_z(b_x))\Big\}\; d[F(y\,|x)G(z\,|x)] \end{aligned} \tag{A.18}$$

Claims 3 and 4 are proved by the same technique used in Claims 1 and 2, respectively. Taking into account that

$$\sqrt{\frac{n_{A,x}\,n_{B,x}}{n_{A,x}+n_{B,x}}}\left(\widehat{\Delta}_c^x - \Delta_c^x(F,\,G)\right) = I_1 + I_2 - I_3 - I_4$$

and that linear functionals of Gaussian processes have normal distribution, from (A.7), (A.10), (A.17), (A.18), result (41) easily follows. □

Proof of Proposition 4. Let

$$J_1 = \sqrt{\frac{n_A\,n_B}{n_A+n_B}}\left(\sum_x \widehat{p}(x)\,\left(\widehat{\Delta}_c^x - \Delta_c^x(F,\,G)\right)\right) \tag{A.19}$$

$$J_2 = \sqrt{\frac{n_A\,n_B}{n_A+n_B}}\sum_x \Delta_c^x(F,\,G)\left(\widehat{p}(x) - p(x)\right) \tag{A.20}$$

so that

$$\sqrt{\frac{n_A\,n_B}{n_A+n_B}}\left(\widehat{\Delta}_c - \Delta_c(F,\,G)\right) = J_1 + J_2 \tag{A.21}$$

Now, it is not difficult to see that the two statistics J_1 and J_2 are asymptotically independent. Furthermore, the statistics

$$\sqrt{\frac{n_{A,x}\,n_{B,x}}{n_{A,x}+n_{B,x}}}\left(\widehat{\Delta}_c^x - \Delta_c^x(F,\,G)\right)$$

as x ranges in the support of X, are asymptotically independent, too. As a consequence of Proposition 3, and taking into account that

$$\widehat{p}(x) \xrightarrow{a.s.} p(x) \quad as\ n_A,\ n_B \to \infty$$

$$\sqrt{\frac{n_A\, n_B}{n_{A,x} n_{B,x}}} \sqrt{\frac{n_{A,x} + n_{B,x}}{n_A + n_B}} \xrightarrow{a.s.} p(x)^{-1/2} \quad as\ n_A,\ n_B \to \infty$$

it is not difficult to prove that

$$J_1 \xrightarrow{w} N\left(0,\ \sum_x p(x)\, V(F, G;\ x)\right) \quad as\ n_A,\ n_B \to \infty \tag{A.22}$$

In the second place, the random vector $\sqrt{n_A + n_B}(\widehat{\boldsymbol{p}} - \boldsymbol{p})$ tends in distribution to a (singular) multinormal variate, with null mean vector and covariance matrix $\boldsymbol{\Sigma}$. Hence,

$$J_2 \xrightarrow{w} N(0,\ \alpha(1-\alpha)\, \boldsymbol{\Delta}_c^x(F, G)^T\, \boldsymbol{\Sigma}\, \boldsymbol{\Delta}_c^x(F, G)) \tag{A.23}$$

From (A.22) and (A.23), result (45) follows. □ 695