International Journal of Approximate Reasoning 53 (2012) 1124–1136

Contents lists available at SciVerse ScienceDirect

International Journal of Approximate Reasoning

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

Incoherence correction strategies in statistical matching

Alessandro Brozzi [a], Andrea Capotorti [a,*], Barbara Vantaggi [b]

[a] *Dip. Matematica e Informatica, Università di Perugia, Italy*
[b] *Dip. Scienze di Base e Applicate per l'Ingegneria, Università "La Sapienza" Roma, Italy*

ARTICLE INFO

Article history:
Available online 27 July 2012

Keywords:
Statistical matching
Incoherence
Inference
Specialized discrepancy measure

ABSTRACT

Several economic applications require to consider different data sources and to integrate the information coming from them. This paper focuses on statistical matching, in particular we deal with incoherences. In fact, when logical constraints among the variables are present incoherencies on the probability evaluations can arise. The aim of this paper is to remove such incoherences by using different methods based on distances minimization or least commitment imprecise probabilities extensions. An illustrative example shows peculiarities of the different correction methods. Finally, limited to pseudo distance minimization, we performed a systematic comparison through a simulation study.

1. Introduction

The integration problem of knowledge coming from several separate data bases, which have some variables in common as well as some variables recorded only in one data base, occurs in several economic applications, some examples are marketing research [19] and microsimulation modeling [26,27,29,33,39].

In particular, we deal with the so called statistical matching problem for categorical variables, that can be represented by the following simple situation: there are two different sources, A and B, with some overlapping variables and some variables collected only in one source. Let X represent the common variables, Y denotes the variables collected only in A, and Z those only in B. Thus, the data consist of two samples, one on (X, Y) and the other one on (X, Z). In this context data are missing by design since they have been already collected separately, and to get joint data on Y and Z would be expensive and time-consuming. Traditionally, to cope with these problems, the available data are combined with assumptions, such as conditional independence between Y and Z given X, which are strong enough to assure a unique compatible distribution. Actually, since there are many distributions on (X, Y, Z) compatible with the available partial information on (X, Y) and (X, Z), it is too restrictive to consider just one of the compatible distributions (as already noted in [15,18,32] and for missing data problem in [12,25,34]).

This problem has been already faced in a coherent conditional probability setting in [36,37]: coherence allows to check the compatibility of partial (conditional) assessments, to manage further available knowledge and to make inference on the variables of interest. In these quoted papers it is supposed that the two samples are drawn randomly from the same population, here we allow that the two samples from the same population can be drawn according to different sample schemes. This leads to interval probabilities and so to coherent lower and upper conditional probabilities.

In this paper firstly we extend a result given in [37] to interval probabilities by showing that when there is no logical constraint among the variables, coherence is always satisfied but that, whenever logical constraints are present, it is necessary to check global coherence of the relevant partial assessments drawn from the different sources. If coherence is not satisfied, we need to remove incoherences. For precise assessments this has been already done in [36] adjusting the "minimal" incoherent assessments through minimization of norm $L1$.

[*] Corresponding author.
E-mail addresses: alessandro.brozzi@dmi.unipg.it (A. Brozzi), capot@dmi.unipg.it (A. Capotorti), barbara.vantaggi@sbai.uniroma1.it (B. Vantaggi).

0888-613X/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ijar.2012.06.009

The aim of this paper is to consider different correction methods: to perform a constrained maximum likelihood estimation; to find the "closer" coherent adjustment with respect to some specific pseudo-distance; or to extend the maximum coherent sub-assessment to the least commitment imprecise probabilities. The main effort is performed with the second aforementioned method, where the choice of the pseudo-distance to minimize is crucial. In fact, pseudo-distances need to be suitably adapted for partial conditional assessments, since they were introduced mainly for unconditional assessments. In particular, to properly deal with statistical matching, we introduce a specific adjustment of a discrepancy introduced in [7], that permits an unsupervised localization of the sub-domains where incoherence must be removed.

Once coherence is restored, it is possible to draw inference by computing analytically the lower and upper bounds for the quantities of interest. Actually, our aim is in the same line of those based on multiple imputation [32] and its extension [31], which yields an approximation of the lower and upper bounds in the case of a multivariate normal distribution. A similar approximation for these bounds is carried out in [15] on the base of maximum likelihood approach for categorical variables.

The paper is organized as follows: in Section 2 we briefly recall the main notions about coherent precise and imprecise conditionally probability assessments and we reformulate the statistical matching problem inside this framework. In Section 3 the three aforementioned correction methods are introduced, with a particular emphasis on the minimization of pseudo-distances and the specification of the discrepancy. In Section 4 we introduce an example built from data taken from [15] to better show advantages and drawbacks of the different methods. Finally, in Section 5 we compare the different pseudo-distances minimization performances through a simulation of 1000 random couples of samples drawn from a finite population according to the same sample scheme: A with cardinality $n_A = 1148$ and B with $n_B = 1165$. We perform 1000 unconstrained maximum likelihood estimations of the marginal distribution for the common random variable X and of the conditional distributions for $Y|X$ and $Z|X$ and we obtain 565 assessments which are not coherent. Corrected estimates induce credal sets [38], that means convex sets of probability distributions compatible with estimates. These estimates are compared with the original coherent ones through "goodness-of-fit" tests between the probability distribution of the population and the joint distributions in the credal sets. In Section 6 some conclusions and remarks are given.

2. Preliminaries about coherent conditional probability

We consider coherent (conditional) probability as framework of reference: an assessment $\mathbf{p}$ on a set $\mathcal{E}$ of conditional events is coherent if it is compatible with a conditional probability P (in the sense of de Finetti [13], see also [16,20]), on $\mathcal{A} \times \mathcal{A} \setminus \{\emptyset\}$, where $\mathcal{A}$ is an algebra. This means that the restriction $P_{|\mathcal{E}}$ of P on $\mathcal{E}$ coincides with $\mathbf{p}$. We recall that any conditional probability P on $\mathcal{A} \times \mathcal{A} \setminus \{\emptyset\}$ gives rise to a suitable class of probability distributions $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_l$ agreeing with P (for more details refer to [10,11]).

Coherence has an important role also for inference, that means extension of the given assessment to any new conditional event (see [13]):

Theorem 1. *Let* $\mathbf{p}$ *be an assessment on an arbitrary family* $\mathcal{E}$ *of conditional events; then there exists a (possibly not unique) coherent extension of* $\mathbf{p}$ *to any family* $\mathcal{K} \supset \mathcal{E}$ *if and only if* $\mathbf{p}$ *is a coherent conditional probability on* $\mathcal{E}$*. Moreover, if* $\mathbf{p}$ *is a coherent conditional probability on* $\mathcal{E}$*, then the coherent probability values for any conditional event* $F|K \in \mathcal{K} \setminus \mathcal{E}$ *belong to a closed interval* $[\underline{p_{F|K}}, \overline{p_{F|K}}]$.

The coherence notion has been given also for imprecise conditional probability assessments, i.e., whenever the numerical part of the assessment is elicited through interval values

$$\mathbf{lub} = ([lb_1, ub_1], \ldots, [lb_n, ub_n]). \tag{1}$$

For assessments like $(\mathcal{E}, \mathbf{lub})$, although defined on finite spaces, there could be different kinds of coherence requirements (for a detailed exposition, among others, refer to [28]). In this paper we focus on the most stringent one: (strong) coherence. By taking into account a Bayesian sensitivity analysis interpretation, coherent lower-upper conditional probability assessments $(\mathcal{E}, \mathbf{lub})$ are such that the numerical function defined by the lower (upper) bounds of the intervals $\mathbf{lub}$ can be obtained as lower (upper) envelope of a set of coherent precise conditional probability assessments on $\mathcal{E}$. It follows that each lower (lb_i) and upper (ub_i) bound is attained through at least one of these conditional probability.

Also starting from a coherent lower-upper assessment $(\mathcal{E}, \mathbf{lub})$ it is possible to infer coherent bounds $[\underline{p_{F|K}}, \overline{p_{F|K}}]$ for the coherent values of any conditional event $F|K$ of interest through specific sequences of linear optimization problems or satisfiability of some logical configurations (for details refer to [5]).

2.1. Statistical matching in a coherent setting

Denote by $(X_1, Y_1), \ldots, (X_{n_A}, Y_{n_A})$ and by $(X_{n_A+1}, Z_{n_A+1}), \ldots, (X_{n_A+n_B}, Z_{n_A+n_B})$ two random samples (on categorical variables) related to two sources A and B. We suppose that the two samples are related to the same population of interest. In [37] the two samples are supposed to be drawn according to the same sampling scheme, here we remove this hypothesis in order to reinterpret some of the results given in [32].

Let S_s (with $s = 1, 2$) be the event "the unit is drawn according to the s-th sampling scheme" and data in file A are drawn according to the first sampling scheme, while those in B are drawn according to the second one.

The relevant population parameters, representing (conditional) probability values, can be estimated from the two files: from file A the probability that the next unit has $Y = y_j$ conditional on $(X = x_i)$ (for any $i \in I$)

$$\mathbf{y}_{j|i} = P_{Y|(X=x_i)}(y_j), \tag{2}$$

and analogously from file B

$$\mathbf{z}_{k|i} = P_{Z|(X=x_i)}(z_k). \tag{3}$$

Moreover, from file A the probability that the next unit has $X = x_j$ can be evaluated

$$\mathbf{x}_i^{S_1} = P_X(x_i|S_1), \tag{4}$$

and analogously from file B

$$\mathbf{x}_i^{S_2} = P_X(x_i|S_2), \tag{5}$$

and, by supposing that a unit can be selected by just one sampling scheme with known sampling probabilities $P(S_s)$ with $s = 1, 2$, we get

$$\mathbf{x}_i = P_X(x_i) = \mathbf{x}_i^{S_1} P(S_1) + \mathbf{x}_i^{S_2} P(S_2). \tag{6}$$

Usually, under the hypothesis that the samples are drawn according to the same sample scheme, such estimations are performed through the (unconstrained) partial maximum likelihood evaluations, which coincide with the following frequencies

$$\mathbf{y}_{j|i} = \frac{n_A^{ij}}{n_A^{i\cdot}}, \quad \mathbf{z}_{k|i} = \frac{n_B^{ik}}{n_B^{i\cdot}}, \quad \mathbf{x}_i = \frac{n_A^{i\cdot} + n_B^{i\cdot}}{n_A + n_B}, \tag{7}$$

where $n_A^{i\cdot}$ and $n_B^{i\cdot}$ represent the number of units expressing $(X = x_i)$ in samples A and B, respectively, while n_A^{ij} stands for the number of units in A with $(X = x_i, Y = y_j)$ and n_B^{ik} the number of units in B with $(X = x_i, Z = z_k)$.

Note that when $n_A^{i\cdot}$ (equivalently for $n_B^{i\cdot}$) is 0 (i.e., no observation in A is such that $(X = x_i)$) the value $\mathbf{y}_{j|i}$ ($\mathbf{z}_{k|i}$) is not defined and no estimation is given for this specific parameter.

Now, we should deal with the whole assessment $(\mathcal{E}, \mathbf{p})$ with

$$\mathcal{E} = \left\{ \begin{array}{r} (X = x_i),\ (Y = y_j)|(X = x_i),\ (Z = z_k)|(X = x_i) \\ \text{for any } x_i, y_j, z_k \end{array} \right\}, \tag{8}$$

$$\mathbf{p} = \{\mathbf{x}_i, \mathbf{y}_{j|i}, \mathbf{z}_{k|i}\}_{i,j,k}.$$

Then, first of all we need to check its coherence (see [11]). In the particular context of statistical matching (see [37]) the check of coherence reduces to the compatibility of the following linear system

$$\begin{cases} \mathbf{y}_{j|i} \sum_{j,k} \alpha_{ijk} = \sum_k \alpha_{ijk} & \text{for any } \mathbf{y}_{j|i} \\ \mathbf{z}_{k|i} \sum_{j,k} \alpha_{ijk} = \sum_j \alpha_{ijk} & \text{for any } \mathbf{z}_{k|i} \\ \mathbf{x}_i = \sum_{j,k} \alpha_{ijk} & \text{for any } \mathbf{x}_i \\ \sum_{i,j,k} \alpha_{ijk} = 1 & \\ \alpha_{ijk} \geq 0 & \end{cases} \tag{9}$$

with unknowns α_{ijk} related to event $(X = x_i, Y = y_j, Z = z_k)$ different from the impossible one.

We recall here the result, proved in [37], that under logical independence, coherence is assured:

Theorem 2. *Let X, Y, Z be three finite random variables and $\mathcal{E}_X, \mathcal{E}_Y, \mathcal{E}_Z$ the associated partition generated by X, Y, Z. Consider the following three separately coherent conditional probability assessments $\{\mathbf{x}_i\}_i$, $\{\mathbf{y}_{j|i}\}_j$ and $\{\mathbf{z}_{k|i}\}_k$.*

Table 1
Samples A and B of Example 1.

A

X	Y	
	0	**1**
0	21	27
1	4	12

B

X	Z	
	0	**1**
0	33	15
1	12	4

Table 2
Marginal and conditional probabilities based on samples A and B of Example 1.

$\mathcal{E}$	**p**
$X = 0$	3/4
$X = 1$	1/4
$Y = 0 \vert X = 0$	7/16
$Y = 1 \vert X = 0$	9/16
$Y = 0 \vert X = 1$	1/4
$Y = 1 \vert X = 1$	3/4
$Z = 0 \vert X = 0$	11/16
$Z = 1 \vert X = 0$	5/16
$Z = 0 \vert X = 1$	3/4
$Z = 1 \vert X = 1$	1/4

Then, the assessment

$$\{\mathbf{x}_i\,,\,\mathbf{y}_{j|i}\}_{i,j}$$

is coherent (analogously for $\{\mathbf{x}_i\,,\,\mathbf{z}_{k|i}\}_{i,k}$).

Moreover, if the partitions $\mathcal{E}_Y, \mathcal{E}_Z$ are logically independent with respect to $\mathcal{E}_X$ (i.e., $(X = x_i, Y = y_j, Z = z_k)$ is possible for any value x_i of X, y_j of Y, z_k of Z s.t. the events $(X = x_i, Y = y_j)$ and $(X = x_i, Z = z_k)$ are possible), then the whole assessment

$$\mathbf{p} = \{\mathbf{x}_i\,,\,\mathbf{y}_{j|i}\,,\,\mathbf{z}_{k|i}\}_{i,j,k} \text{ on } \mathcal{E}$$

is coherent.

Note that, in general, if $(\mathcal{E}, \mathbf{p})$ is coherent then system (9) has more than one single solution and the set of all such solutions constitutes a convex set usually named credal set [38], as the following example shows:

Example 1. Let X, Y, Z be three binary variables with the only constraint that the event $(X = 0, Y = 1, Z = 1)$ is impossible. Two random samples A on (X, Y) and B on (X, Z) are drawn from a common population obtaining the data given in Table 1, where bold values denote the variables modalities and the values inside the joint samples counts.

From these observations we obtain as maximum likelihood estimations the assessment **p** reported in Table 2.

The assessment **p** is coherent since the associated linear system *(9)* admits solution, as for example

$$\alpha_{111} = \frac{2}{64},\quad \alpha_{110} = \frac{10}{64},\quad \alpha_{101} = \frac{2}{64},\quad \alpha_{100} = \frac{2}{64},\quad \alpha_{010} = \frac{27}{64},\quad \alpha_{001} = \frac{15}{64},\quad \alpha_{000} = \frac{6}{64}.$$

Indeed, there is more than one joint distribution on (X, Y, Z) compatible with **p**. The corresponding credal set is in fact composed by all the possible solutions of *(9)* which can be obtained as convex combinations of the following two extreme distributions α^1 and α^2:

$$\alpha^1_{111} = \frac{1}{16},\quad \alpha^1_{110} = \frac{1}{8},\quad \alpha^1_{101} = 0,\quad \alpha^1_{100} = \frac{1}{16},\quad \alpha^1_{010} = \frac{27}{64},\quad \alpha^1_{001} = \frac{15}{64},\quad \alpha^1_{000} = \frac{6}{64},$$

$$\alpha^2_{111} = 0,\quad \alpha^2_{110} = \frac{3}{16},\quad \alpha^2_{101} = \frac{1}{16},\quad \alpha^2_{100} = 0,\quad \alpha^2_{010} = \frac{27}{64},\quad \alpha^2_{001} = \frac{15}{64},\quad \alpha^2_{000} = \frac{6}{64}.$$

The distributions α^1 and α^2 are computed by linear optimizations with constraints expressed through system (9).

When there are logical constraints among the variables Y and Z, the coherence of the whole assessment $(\mathcal{E}, \mathbf{p})$ in (8) is not in general assured by the separate coherence of the single assessments (2), (3), (6) and moreover incoherences, whenever present, are localized among conditional events with the same conditioning event $(X = x_i)$ (for the proofs and an example see again [37]). Note that the need to manage logical constraints arises from practical applications [15].

We consider now a more general situation: the two samples can be drawn according to different sample schemes with unknown probabilities $P(S_s)$. Obviously, in this case the estimate $\mathbf{x}_i$ is not univocally determined, in fact the coherent values for $\mathbf{x}_i$ are those in the interval

$$\mathbf{lub}_i = \left[\min\left\{\mathbf{x}_i^{S_1}, \mathbf{x}_i^{S_2}\right\}, \max\left\{\mathbf{x}_i^{S_1}, \mathbf{x}_i^{S_2}\right\}\right].$$

Then, it is necessary to work with the imprecise assessments

$$\mathbf{lub} = \{\mathbf{lub}_i, \mathbf{y}_{j|i}, \mathbf{z}_{k|i}\}_{i,j,k}. \tag{10}$$

on $\mathcal{E}$. These situations arise also in practical applications, see for instance [1,35].

First of all we need to check the coherence of **lub** on $\mathcal{E}$, that means to check the coherence as a lower probability assessment of

$$\underline{\mathbf{lub}} = \{\underline{\mathbf{x}}_i, \mathbf{y}_{j|i}, \mathbf{z}_{k|i}\}_{i,j,k} \tag{11}$$

with $\underline{\mathbf{x}}_i$ the left bound of $\mathbf{lub}_i$, and as an upper probability assessment of

$$\overline{\mathbf{lub}} = \{\overline{\mathbf{x}}_i, \mathbf{y}_{j|i}, \mathbf{z}_{k|i}\}_{i,j,k} \tag{12}$$

with $\overline{\mathbf{x}}_i$ the right bound of $\mathbf{lub}_i$.

Also in this case, the logical independence assumption guarantees coherence. In fact the following result holds:

Theorem 3. *Let X, Y, Z be three finite random variables and $\mathcal{E}_X, \mathcal{E}_Y, \mathcal{E}_Z$ the associated partition generated by X, Y, Z. Consider the separately coherent lower $\{\underline{\mathbf{x}}_i\}_i$, upper $\{\overline{\mathbf{x}}_i\}_i$ and conditional $\{\mathbf{y}_{j|i}\}_j$, $\{\mathbf{z}_{k|i}\}_k$ probability assessments.*

Then, the assessments

$$\{\underline{\mathbf{x}}_i\,,\mathbf{y}_{j|i}\}_{i,j}$$

and

$$\{\overline{\mathbf{x}}_i\,,\mathbf{y}_{j|i}\}_{i,j}$$

are separately coherent conditional lower and upper probabilities (analogously for $\{\underline{\mathbf{x}}_i\,,\mathbf{z}_{k|i}\}_{i,k}$ and $\{\overline{\mathbf{x}}_i\,,\mathbf{z}_{k|i}\}_{i,k}$).

Moreover, if the partitions $\mathcal{E}_Y, \mathcal{E}_Z$ are logically independent with respect to $\mathcal{E}_X$, then the whole assessments

$$\underline{\mathbf{p}} = \{\underline{\mathbf{x}}_i\,,\mathbf{y}_{j|i}\,,\mathbf{z}_{k|i}\}_{i,j,k}$$

and

$$\overline{\mathbf{p}} = \{\overline{\mathbf{x}}_i\,,\mathbf{y}_{j|i}\,,\mathbf{z}_{k|i}\}_{i,j,k}$$

are coherent conditional lower and upper probabilities, respectively.

Proof. Since $\{\underline{\mathbf{x}}_i\}_i$ is a coherent lower probability on $\mathcal{E}_X$, there is a class $\mathcal{P}$ of probabilities on $\mathcal{E}_X$ such that

$$\underline{\mathbf{x}}_i = \inf_{\mathcal{P}} P(X = x_i). \tag{13}$$

From Theorem 2 we have that for any $P \in \mathcal{P}$ the assessment $\{P(X = x_i)\,,\mathbf{y}_{j|i}\}_{i,j}$ is a coherent conditional probability. Hence, by taking the lower envelope of all the coherent conditional probability assessments in $\mathcal{P}' = \{P(X = x_i)\,,\mathbf{y}_{j|i} \,:\, P \in \mathcal{P}\}_{i,j}$ we get (through (13)) exactly $\{\underline{\mathbf{x}}_i\,,\mathbf{y}_{j|i}\}_{i,j}$, so coherence as lower conditional probability follows.

Moreover, if the partitions $\mathcal{E}_Y, \mathcal{E}_Z$ are logically independent with respect to $\mathcal{E}_X$, again from Theorem 2 we have that for any $P \in \mathcal{P}$ the assessment $\{P(X = x_i)\,,\mathbf{y}_{j|i}\,,\mathbf{z}_{k|i}\}_{i,j,k}$ is a coherent conditional probability, hence the lower envelope of the class of this kind of coherent conditional probabilities gives rise to $\{\underline{\mathbf{x}}_i\,,\mathbf{y}_{j|i}\,,\mathbf{z}_{k|i}\}_{i,j}$, implying its coherence as lower conditional probability.

The proof for $\{\overline{\mathbf{x}}_i\,,\mathbf{y}_{j|i}\}_{i,j}$ and $\overline{\mathbf{p}}$ goes along the same line. □

Also for imprecise evaluations $(\mathcal{E}, \mathbf{lub})$, if there is some logical constraint among the variables Y and Z, the coherence of the whole assessments $\underline{\mathbf{p}}$ and $\overline{\mathbf{p}}$ is not assured by separate coherence of the single assessments and, whenever present, incoherences localize among conditional events with the same conditioning event $(X = x_i)$ (the proof goes along the same line as that for precise evaluations $(\mathcal{E}, \mathbf{p})$, see [37]).

3. Removing incoherences in statistical matching

Estimate correction has been already studied (e.g., see [23]), but this approach does not seem suitable in the context of statistical matching because of the lack of information due to the fact that Y and Z are not jointly observed, so the prior distribution cannot be updated and the likelihood function has a flat ridge (as already noted in [32]).

In the following we show some methods, which could be reasonable to correct incoherent assessments.

3.1. Maximum likelihood estimates

As just mentioned, a first attempt could be based on the constrained maximum likelihood criterion ([24]): as estimates of $\theta = (P(X = x_i), P_{Y|(X=x_i)}(Y = y_j), P_{Z|(X=x_i)}(Z = z_k))_{i,j,k}$ are taken the values of the parameters $\left(\widehat{\theta}_i, \widehat{\theta}_{j|i}, \widehat{\theta}_{z|i}\right)_{i,j,k}$, derived as solution of the program

$$\max_{\theta} l(\theta|n_A, n_B) \tag{14}$$

under the constraint that θ is a coherent conditional probability assessment over $\mathcal{E}$, where $l(\theta|n_A, n_B)$ is the log likelihood

$$\begin{aligned} l(\theta|n_A, n_B) &= \log(L(\theta|n_A, n_B)) = \log\left(\prod_{i,j}(\theta_{j|i}\theta_i)^{n_A^{ij}}\prod_{i,k}(\theta_{k|i}\theta_i)^{n_B^{ik}}\right) \\ &= \sum_{i,j} n_A^{ij}(\log(\theta_{j|i}) + \log(\theta_i)) + \sum_{i,k} n_B^{ik}(\log(\theta_{k|i}) + \log(\theta_i)). \end{aligned} \tag{15}$$

Hence, we need to deal with an optimization problem with the observed data log likelihood $l(\theta|n_A, n_B)$ as non-linear objective function and a set of linear constraints, for the unknowns α_{ijk}, induced by the coherence requirement on θ (see system (9)).

3.2. Pseudo-distances minimization

Another possible correction method is to find coherent estimates as "close" as possible to the available information represented by the whole assessment (8). This approach implies the choice of some pseudo-distance such as Euclidean distance, Kullback–Leibler divergence, Csiszár f-divergences. Some of these can be applied only among unconditional probabilities; while others could be applied also for partial conditional probability assessments.

Given two conditional probability estimates $\mathbf{p} = [p_1, \ldots, p_n]$ and $\mathbf{q} = [q_1, \ldots, q_n]$ on $\mathcal{E}$, the most widely adopted divergencies among them are

(1) $L1(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} |q_i - p_i|$;

(2) $L2(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} (q_i - p_i)^2$;

(3) $KL(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} (q_i \ln(q_i/p_i) - q_i + p_i)$.

$L1$ and $L2$ are usual metric distances, endowed with all their geometric properties, but until now remain without an intuitive probabilistic interpretation for conditional assessments. Moreover, their use in conditional context could lead to numerical troubles due to non-convexity of coherent assessments (see e.g. [3]).

$KL(\mathbf{p}, \mathbf{q})$ coincides with the so-called logarithmic Bregman divergence $d_f(\mathbf{q}, \mathbf{p})$ and, in the unconditional case, it is widely used for its information theoretic properties. In fact, such divergence generalizes the well known Kullback-Leibler divergence [21] to partial assessments, however in some cases it presents some unpleasant situation since it is based on a scoring rule which takes into account only the events which occur and not those which do not occur.

To overcome this characteristic and to encompass the need of considering the conditional framework where the assessment is given, recently in [6,7] for partial conditional assessments $\mathbf{v} = [v_1, \ldots, v_n] \in (0, 1)^n$ over $\mathcal{E} = [E_1|H_1, \ldots, E_n|H_n]$, the following random variable has been proposed as scoring rule:

$$S(\mathbf{v}) := \sum_{i=1}^{n} I_{E_i H_i} \ln v_i + \sum_{i=1}^{n} I_{E_i^c H_i} \ln(1 - v_i) \tag{16}$$

with $I_{\cdot}$ the indicator function of unconditional events.

Note that the scoring rule $S(\cdot)$ is a not positive function, in fact the motivation for its introduction is that the assessor "loses less" the higher the probabilities are of occurring events, and at the same time, the lower the probabilities of events which do not occur. The values assessed on events that turn out to be undetermined do not influence the score. Such a score $S(\cdot)$ is an extension to partial and conditional probability assessments of the "total-log proper scoring rule" for probability distributions proposed by Lad in [22, p. 355] and it has been considered in [17] as an example of equivalence between the coherence of a partial conditional probability $\mathbf{p}$ and its admissibility with respect to a proper scoring rule. By considering as a distance the difference between expected scores related to the initial evaluation $\mathbf{p}$ and to the evaluation $\mathbf{q}_{\boldsymbol{\alpha}}$ induced by a probability distribution $\boldsymbol{\alpha}$, it is possible to define the following "discrepancy":

$$\Delta(\mathbf{p},\boldsymbol{\alpha}) = \sum_{i|\alpha(H_i)>0} \alpha(H_i)\left(q_i \ln \frac{q_i}{p_i} + (1-q_i)\ln\frac{(1-q_i)}{(1-p_i)}\right) \tag{17}$$

taking the convention $0\ln(0) = 0$. Note that in $\Delta(\mathbf{p},\boldsymbol{\alpha})$ each term is weighted by $\alpha(H_i)$, which reflects the "relevance" of each context H_i with respect to all the assessments.

The main idea is to take as coherent correction of $\mathbf{p}$ the assessment $\mathbf{q}_{\mathbf{p}} \equiv \mathbf{q}_{\widetilde{\boldsymbol{\alpha}}}$ generated by the distribution $\widetilde{\boldsymbol{\alpha}}$ solution of the nonlinear optimization program

$$\min_{\boldsymbol{\alpha}} \Delta(\mathbf{p},\boldsymbol{\alpha}). \tag{18}$$

The motivation for this choice is that (intuitively) the assessor of $\mathbf{p}$ would expect to suffer the penalty $S(\mathbf{p})$, hence we select the coherent assessment $\mathbf{q}_{\mathbf{p}}$ that has a (probabilistic) expected score as close as possible to it. In [7] it is formally proved that $\Delta(\mathbf{p},\boldsymbol{\alpha})$ has all the usual divergencies properties.

Now, we stress that $L1$, $L2$, KL and discrepancy Δ apply on the whole domain $\mathcal{E}$, so their minimizations would induce changes also on the marginal estimate $\{\mathbf{x}_i\}_{i\in I}$, which is coherent and it would be better to avoid any change on it. Hence, to encompass the need to keep the marginal probabilities $\{\mathbf{x}_i\}_{i\in I}$ unchanged, recently in [8] the following modified discrepancy (19) has been introduced:

$$\begin{aligned}\Delta_{mix}(\mathbf{p},\{\boldsymbol{\alpha}_i\}_i) = \sum_i \mathbf{x}_i \Bigg[&\sum_j \left(q^{\alpha_i}_{j|i} \ln \frac{q^{\alpha_i}_{j|i}}{\mathbf{y}_{j|i}} + \left(1-q^{\alpha_i}_{j|i}\right)\ln\frac{\left(1-q^{\alpha_i}_{j|i}\right)}{(1-\mathbf{y}_{j|i})}\right)\\ &+\sum_k \left(q^{\alpha_i}_{k|i} \ln \frac{q^{\alpha_i}_{k|i}}{\mathbf{z}_{k|i}} + \left(1-q^{\alpha_i}_{k|i}\right)\ln\frac{\left(1-q^{\alpha_i}_{k|i}\right)}{(1-\mathbf{z}_{k|i})}\right)\Bigg].\end{aligned} \tag{19}$$

For any $i \in I$, each distribution α_i on the sample space spanned by $(Y = y_j)|(X = x_i)$ and $(Z = z_k)|(X = x_i)$ should fulfill the normalizing condition $\alpha_i(X = x_i) = \mathbf{x}_i$, and generates the conditional probabilities

$$q^{\alpha_i}_{j|i} = \frac{\alpha_i(Y = y_j)}{\alpha_i(X = x_i)} \quad q^{\alpha_i}_{k|i} = \frac{\alpha_i(Z = z_k)}{\alpha_i(X = x_i)}. \tag{20}$$

Note that the generated estimate $\mathbf{q} = \{\mathbf{x}_i, q^{\alpha_i}_{j|i}, q^{\alpha_i}_{k|i}\}_{i,j,k}$ is coherent (see e.g., [37]) and it leaves the marginal probabilities $\{\mathbf{x}_i\}_{i\in I}$ unchanged. This characteristic differentiates the specialized discrepancy (19) from the discrepancy (17), as the following simple example shows:

Example 2. Let X, Y, Z be three random variables and $\mathcal{E}_X = \{A_1, A_2\}$, $\mathcal{E}_Y = \{E_1, E_2, E_3\}$, $\mathcal{E}_Z = \{S_1, S_2, S_3\}$ be the three corresponding partitions such that

$$S_1 \wedge (E_1 \vee E_2) = \emptyset \text{ and } S_2 \wedge E_1 = \emptyset. \tag{21}$$

Consider the following conditional assessments

$$P(A_1) = \frac{1}{3}, \qquad P(A_2) = \frac{2}{3};$$

$$P(S_1|A_1) = \frac{179}{1108}, \qquad P(S_2|A_1) = \frac{443}{1108}, \qquad P(S_3|A_1) = \frac{486}{1108},$$
$$P(S_1|A_2) = \frac{2}{3}, \qquad P(S_2|A_2) = \frac{1}{9}, \qquad P(S_3|A_2) = \frac{2}{9};$$

Table 3
A comparison of corrections for Example 2, bold and script values highlight differences between the two different corrections.

$\mathcal{E}$	P	Δ	Δ_{mix}
$A1$	0.3333	**0.3726**	–
$A2$	0.6667	**0.6274**	–
$S1\|A1$	0.1616	0.1616	0.1616
$S2\|A1$	0.3998	0.3998	0.3998
$S3\|A1$	0.4386	0.4386	0.4386
$E1\|A1$	0.3483	0.3483	0.3483
$E2\|A1$	0.0918	0.0918	0.0918
$E3\|A1$	0.5599	0.5599	0.5599
$S1\|A2$	0.6667	*0.4848*	*0.4848*
$S2\|A2$	0.1111	*0.0996*	*0.0996*
$S3\|A2$	0.2222	*0.4156*	*0.4156*
$E1\|A2$	0.6667	*0.4156*	*0.4156*
$E2\|A2$	0.00	*0.0996*	*0.0996*
$E3\|A2$	0.3333	*0.4848*	*0.4848*

$$P(E_1|A_1) = \frac{387}{1111}, \qquad P(E_2|A_1) = \frac{102}{1111}, \qquad P(E_3|A_1) = \frac{622}{1111},$$

$$P(E_1|A_2) = \frac{2}{3}, \qquad P(E_2|A_2) = 0, \qquad P(E_3|A_2) = \frac{1}{3}.$$

It is easy to check that P on events A_i is coherent, as well as $P(S_j|A_i)$ (and analogously $P(E_k|A_i)$) for any A_i. However, the whole assessment is not coherent: incoherence is localized on events conditioned to A_2 since $P(S_1|A_2) + P(E_1|A_2) > 1$ is in contradiction with the first constraint in (21) .

By minimizing either $\Delta(\mathbf{p}, \boldsymbol{\alpha})$ or $\Delta_{mix}(\mathbf{p}, \{\boldsymbol{\alpha}_j\})$ we obtain the same correction limited to the incoherent part (see Table 3), whereas with the former also the unconditional values for $P(A_i)$ are modified, even being coherent.

We stress that for the statistical matching problem Δ_{mix} seems to be more appropriate than Δ. Note that, with such specialized discrepancy, the sub-domains, where incoherence must be removed, are implicitly detected, without the need for a preliminary inspection of the assessment $(\mathcal{E}, \mathbf{p})$.

Going back to a general approach, it is possible to proceed in two ways: a supervised procedure, apt to correct incoherent sub-assessments which were previously detected; or an unsupervised approach, which usually adjusts the whole assessment (8). In any case, in order to correct an estimation $\mathbf{p}$ we need to look for the assessment $\mathbf{q_p}$ that is a solution of the following nonlinear optimization program:

$$\min_{\mathbf{q}} \delta(\mathbf{p}, \mathbf{q}), \tag{22}$$

with $\delta(\mathbf{p}, \mathbf{q})$ any pseudo-distance (if δ is Δ or Δ_{mix} then $\mathbf{q}$ are those induced by $\boldsymbol{\alpha}$ or $\{\boldsymbol{\alpha}_i\}_i$, respectively).

Note that the discrepancy $\Delta_{mix}(\mathbf{p}, \{\boldsymbol{\alpha}_i\}_i)$ is based on the already mentioned segmentation of the possible incoherences. In fact, it separately applies on scenarios $(X = x_i)$ and its use in an optimization program like (22) allows to adjust only the values inside sub-domains where incoherences appear, without any other change. Hence, its application implies that also an unsupervised approach actually works as a supervised one.

3.3. Coherent dilation

A third possibility to adjust the initially incoherent assessment $(\mathcal{E}, \mathbf{p})$ is to determine a coherent sub-assessment $(\mathcal{G}, \mathbf{p}_{|\mathcal{G}})$ and coherently extend it to the rest $\mathcal{F} = \mathcal{E} \setminus \mathcal{G}$ as prescribed by the generalized Bayesian updating scheme (see e.g. [10,11,37] among others). Since, in general, coherent extension gives rise to an interval of plausible values, with this approach the whole assessment turns out to be imprecise due to the interval values $((\mathcal{F}, [\underline{\mathbf{p}_{\mathcal{F}}}, \overline{\mathbf{p}_{\mathcal{F}}}]))$. Also in such a situation, inference can be performed again through the generalized Bayesian updating scheme but applied to imprecise evaluations (see e.g. [2,5] among others). Whenever results of such inference are too vague since the intervals are very wide (close to [0,1]), they can be eventually reduced by a procedure proposed in [9] that enucleates coherent cores, i.e., surely coherent subintervals with highest degree of support.

The choice of the coherent sub-assessment $(\mathcal{G}, \mathbf{p}_{|\mathcal{G}})$ should follow some criterion, since it may not be determined uniquely. Anyhow, for the specific application to statistical matching such a choice comes quite naturally, since in [36] it has been shown that it is possible to detect an incoherent sub-assessment $(\mathcal{F}, \mathbf{p}_{|\mathcal{F}})$ with minimal cardinality.

4. A practical example

In order to compare the different proposed correction methods, we develop an example with data taken from [15] (see also [14,37]). The data are a subset of 2313 employees (people at least 15 years old) extracted from the pilot survey of the

Table 4
Distribution of age and professional status in file A.

Age	Prof. status			
	S_1	S_2	S_3	Total
A_1	–	–	9	9
A_2	–	5	17	22
A_3	179	443	486	1108
A_4	6	1	2	9
Total	185	449	514	1148

Table 5
Distribution of age and educational level in file B.

Age	Educ. level				
	E_1	E_2	E_3	E_4	Total
A_1	6	0	–	–	6
A_2	14	6	13	–	33
A_3	387	102	464	158	1111
A_4	10	0	3	2	15
Total	417	108	480	160	1165

Italian Population and Household Census in the year 2000. Three categorical variables have been analyzed: Age, Educational Level and Professional Status. In file A, containing 1148 units, the variables Age and Professional Status are observed, while file B, consisting of 1165 observations, the variables Age and Educational Level are considered. The variables are grouped in homogeneous response categories as follows: $A_1 = 15$–17 years old, $A_2 = 18$–22 years old, $A_3 = 23$–64 years old, $A_4 =$ more than 65; $E_1 =$ None or compulsory school, $E_2 =$ Vocational school, E_3=Secondary school, E_4=Degree; S_1=Manager, S_2=Clerk, S_3=Worker.

Logical constraints between the variables Age and Educational level (Age and Professional Status) are denoted by the symbol "–" (to be distinguished from the zero frequencies) in Table 4 (Table 5): for example, in Italy a 17 years old person cannot have a University degree. Tables 4 and 5 show, respectively, the distribution of Age and Professional Status in file A, and in file B that related to Age and Educational level.

Additional logical constraints involving both the variables Professional Status and Educational level are:

$$S_1 \wedge (E_1 \vee E_2) = \emptyset \text{ and } S_2 \wedge E_1 = \emptyset.$$

By considering the maximum likelihood estimations as evaluation of the relevant conditional probabilities, we get the assessment for the variable Age:

$$P(A_1) = \frac{15}{2313}, \quad P(A_2) = \frac{55}{2313},$$

$$P(A_3) = \frac{2219}{2313}, \quad P(A_4) = \frac{24}{2313};$$

for the Professional Status given the Age:

$$P(S_2|A_2) = \frac{5}{22}, \quad P(S_3|A_2) = \frac{17}{22},$$

$$P(S_1|A_3) = \frac{179}{1108}, \quad P(S_2|A_3) = \frac{443}{1108}, \quad P(S_3|A_3) = \frac{486}{1108},$$

$$P(S_1|A_4) = \frac{2}{3}, \quad P(S_2|A_4) = \frac{1}{9}, \quad P(S_3|A_4) = \frac{2}{9};$$

for the Educational level given the Age:

$$P(E_1|A_1) = 1, \quad P(E_2|A_1) = 0, \quad P(E_1|A_2) = \frac{14}{33},$$

$$P(E_2|A_2) = \frac{6}{33}, \quad P(E_3|A_2) = \frac{13}{33}, \quad P(E_1|A_3) = \frac{387}{1111},$$

Table 6
Incoherence corrections with associated inference results for $S3|E4$.

	$S_1\|A_4$	$S_2\|A_4$	$S_3\|A_4$	$E_1\|A_4$	$E_2\|A_4$	$E_3\|A_4$	$E_4\|A_4$	$S_3\|E_4$
p	0.6667	0.1111	0.2222	0.6667	0	0.2000	0.1333	Ø
$L1_{\|\mathcal{F}}$	0.2222	–	0.6667	0.6667	–	–	–	[0,0.6285]
$L1_{\|A4}$	0.5266	0.0000	0.4734	0.4734	0.0000	0.2836	0.2431	[0,0.6234]
$L2_{\|A4}$	0.5333	0.0389	0.4278	0.4278	0.0389	0.3	0.2333	[0,0.6238]
$KL_{\|A4}$	0.4856	0.1179	0.3965	0.3965	0.1179	0.2914	0.1942	[0,0.6257]
Δ_{mix}	0.4985	0.0939	0.4077	0.4077	0.0939	0.2943	0.2042	[0,0.6252]
ML	0.4286	0.0714	0.5000	0.5000	0.0000	0.3000	0.2000	[0,0.6254]
$IP_{\mathcal{E}\setminus\mathcal{F}}$	[0 , 0.2222]	-	[0.6667 0.8889]	-	-	-	-	[0,0.6386]
core								[0.0017,0.6286]
$IP_{\mathcal{E}\setminus\{\cdot\|A_4\}}$	[0 , 1]	[0 , 1]	[0 , 1]	[0 , 1]	[0 , 1]	[0 , 1]	[0 , 1]	[0,0.6607]
core								[0,0.6349]

$$P(E_2|A_3) = \frac{102}{1111}, \qquad P(E_3|A_3) = \frac{464}{1111}, \qquad P(E_4|A_3) = \frac{158}{1111},$$

$$P(E_1|A_4) = \frac{2}{3}, \qquad P(E_2|A_4) = 0,$$

$$P(E_3|A_4) = \frac{1}{5}, \qquad P(E_4|A_4) = \frac{2}{15}.$$

The above assessment is not coherent as shown in [37], and in particular incoherence is localized in $P(\cdot|A_4)$ since from logical constraints between Educational Level and Professional Status it follows that $E_1 \wedge S_1 = \emptyset$ and $E_1 \subseteq S_3$, while we have $P(E_1|A_4) + P(S_1|A_4) > 1$ and $P(E_1|A_4) > P(S_3|A_4)$.

Then, we either focus on the minimal set of conditional events $\mathcal{F} = \{E_1|A_4, S_1|A_4, S_3|A_4\}$ involved in incoherencies as proposed in [37] and we correct the assessment only on it (supervised approach), or we adjust the whole distribution on Professional Status and Educational Level conditioned to A_4 (semi-supervised approach).

The given assessment **p** is therefore corrected with respect to the different aforementioned pseudo-distances.

Results are shown in Table 6, where

- $L1_{|\mathcal{F}}$ gives the solution proposed in [37] by minimizing $L1$ distance only on $\mathcal{F}$;
- $L1_{|A4}$, $L2_{|A4}$, $KL_{|A4}$ gives the solutions by minimizing distances only on the events conditioned on A_4;
- Δ_{mix} generates the solution obtained by minimizing the specific discrepancy (19);
- ML gives the maximum likelihood estimation;
- $IP_{\mathcal{E}\setminus\mathcal{F}}$ gives the coherent lower-upper extension induced by the given assessment on $\mathcal{E} \setminus \mathcal{F}$;
- $IP_{\mathcal{E}\setminus\{\cdot|A_4\}}$ gives the coherent lower-upper extension induced by the given assessment on $\mathcal{E}\setminus\{S_i|A_4, E_j|A_4 \ : \ i = 1, 2, 3\ ; \ j = 1, \ldots, 4\}$;
- the last column shows the extensions of the corrections on the conditional event $S_3|E_4$ with the respective "core" rows showing the coherent sub-interval extension with maximum support (see [9]).

Note that due to lack of space we restrict the assessment only to the values conditioned to A_4, that are just those involved in the incoherence.

Firstly, we compare the rows related to the minimal set of incoherence, and it seems that $L1_{|\mathcal{F}}$ and $IP_{\mathcal{E}\setminus\mathcal{F}}$ perform similarly. Quite reasonable inference bounds are obtained by removing not all the set of events conditioned on A_4, but just a subset (a minimal subset). However, we can observe a drastic change on the probability values. In particular, the imprecise adjustment $IP_{\mathcal{E}\setminus\mathcal{F}}$ performs quite well. In fact, it induces inference bounds for $S_3|E_4$ similar to the precise corrections with the additional possibility of focusing on the "core" sub-interval. This sub-interval, even remaining quite vague, presents the positive feature of bounding the lower probability away from zero.

Note that $L1_{|A4}$ and ML give similar results and in particular they take into consideration the absence of observations for $E_2|A_4$ in a way that the related value is not modified. Thus, the peculiarities of the maximum likelihood principle also show in this correction of the incoherence. On the other hand, by using the other distances, precise adjustments on the sub-family conditioned to A_4 have all quite similar behavior, and in particular they modify also the assessment related to $E_2|A_4$, where there is no observation.

The advantage of Δ_{mix} correction is its automatic localization of the scenarios (in this specific example A_4) where the adjustment can be performed and their relative importance expressed by the unconditional probabilities $\mathbf{x}_i$. Note that we apply Δ_{mix}, instead of Δ, in order to avoid any change on the probability distribution of X, that is coherent with any conditional probability on $Y|(X = x)$ (or equivalently $Z|(X = x)$), for any x, as shown in Theorem 2 and 3. In fact, Δ tends to change also the distribution of X (through the weights) in order to reduce the incoherences, as shown in Example 2.

Table 7
Finite population with (X, Y, Z) endowed with structural zeros $(-)$.

X	Y	Z		
		z_1	z_2	z_3
x_1	y_1	–	–	116
	y_2	–	26	5
	y_3	54	108	25
x_2	y_1	–	–	277
	y_2	–	65	1
	y_3	321	1	1

Table 8
Marginal and conditional probabilities based on the population of Table 7.

$\mathcal{E}$	π
$X = x_1$	0.3407
$X = x_2$	0.6593
$Y = y_1 \| X = x_1$	0.1856
$Y = y_2 \| X = x_1$	0.3763
$Y = y_3 \| X = x_1$	0.4381
$Z = z_1 \| X = x_1$	0.4903
$Z = z_2 \| X = x_1$	0.0965
$Z = z_3 \| X = x_1$	0.4131
$Y = y_1 \| X = x_2$	0.3551
$Y = y_2 \| X = x_2$	0.0783
$Y = y_3 \| X = x_2$	0.5666
$Z = z_1 \| X = x_2$	0.4105
$Z = z_2 \| X = x_2$	0.0980
$Z = z_3 \| X = x_2$	0.4915

On the other side, the widest imprecise correction $IP_{\mathcal{E}\setminus\{\cdot|A_4\}}$, being the one with fewer assumptions, surely performs worst. Its vagueness on the values conditioned on A_4 is due to the freedom induced by the coherence characterization, and this reflects also on the inference performances.

Note that in Table 6 we report only the extension values for the conditional event $S_3|E_4$ as an example, however we could compute all the values for the (conditional) events of interest, as for example for the partition generated by the three random variables.

5. A systematic comparison of pseudo-distances minimizations

To have a finer discernment among the pseudo-distances to minimize in the second proposed method, we have performed a systematic comparison[1] by simulating 1000 couples of samples, with cardinality $n_A = 1148$ and $n_B = 1165$, respectively, drawn randomly from a finite population along the same lines of Example 2. Consider three categorical variables (X, Y, Z), with $I = \{1, 2\}, J = \{1, 2, 3\}, K = \{1, 2, 3\}$, distributed as described in Table 7, where the "—" represent the structural zeros implied by the logical constraints

$$(Z = z_1) \wedge ((Y = y_1) \vee (Y = y_2)) = \emptyset, \quad (Y = y_1, Z = z_2) = \emptyset. \tag{23}$$

From each couple of samples A and B, as described in Section 2.1, we can obtain an estimate $\mathbf{p}$ of the probabilities π of Table 8. Over the 1000 frequencies estimations (7) that we observed 565 were incoherent, as can be seen by computing $L1$ distances between the 1000 estimates $\mathbf{p}$ and the corresponding corrections $\mathbf{q_p}$ solutions of (22). In fact, $L1(\mathbf{p}, \mathbf{q_p}) = 0$ corresponds to original coherent frequencies $\mathbf{p}$ (see e.g., Fig. 1 about corrections obtained through $L2$ minimization). Note that for any choice of $\delta(\mathbf{p}, \mathbf{q})$ among those proposed, we obtain the same set of null distances, since all of them are proper pseudo-distances.

By means of the minimization (22) for pseudo-distances $L1, L2, KL, \Delta, \Delta_{mix}$ and for the constrained likelihood maximization (14) applied to the 565 incoherent estimates over the whole domain $\mathcal{E}$ (hence with unsupervised procedures), we obtain six different data-sets with coherent corrections. To compare the performances we evaluate, through chi-squared goodness-of-fit test, the adequacy of the (credal) set of joint probability distributions compatible with each estimate with respect to the joint distribution of the population. We use the minimal χ^2 statistics since in this way we look for the probability in the credal set "closer" to that of reference and in particular, when the credal set contains the joint distribution of the population, we get a zero distance.

Results are in Fig. 2: there are box-plots of minimal χ^2 statistics associated to the six data-sets of corrections and to the data-set of the 435 coherent estimates obtained directly by applying (7). Values are reported in logarithmic scale because

[1] Simulation and post-elaboration have been done through R package [30], non linear optimizations through GAMS software [4].

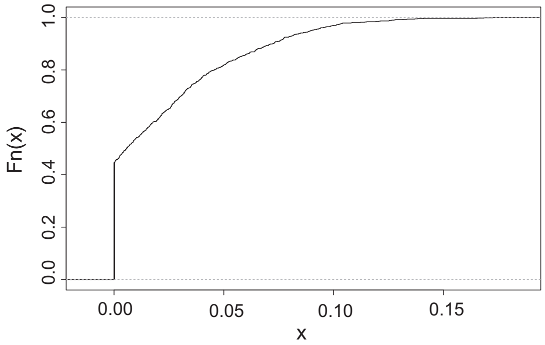

Fig. 1. Empirical cumulative distribution function of $L1$ distances between simulated estimates $\mathbf{p}$ and their corrections $\mathbf{q_p}$ through $L2$ minimization.

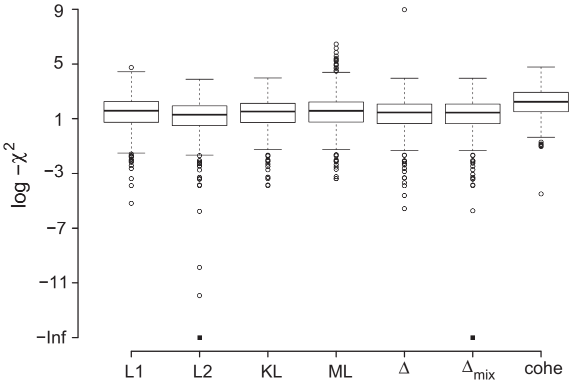

Fig. 2. Logs of minimal χ^2 "goodnes-of-fit" for credal sets induced by pseudo-distances minimizations (labels "$L1$", "$L2$", "KL", "Δ", "Δ_{mix}"), constrained maximum likelihoods (label "ML") and coherent frequencies (label "cohe") estimates. Black-boxes correspond to perfectly matched induced joint distributions.

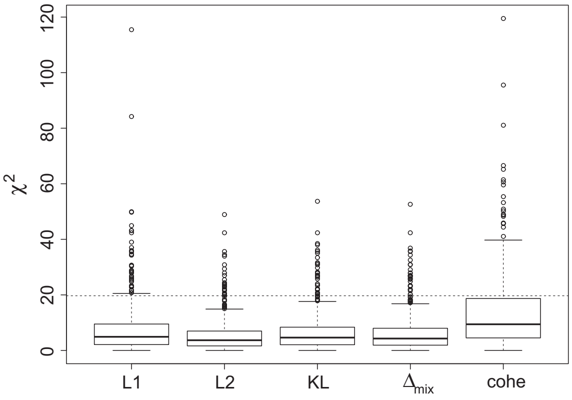

Fig. 3. Minimal χ^2 "goodnes-of-fit" for credal sets induced by pseudo-distances minimizations "$L1$", "$L2$", "KL", "Δ_{mix}" and coherent frequencies (label "cohe") estimates. The dashed line corresponds to the 95% confidence threshold.

correction through Δ minimization or constrained likelihood maximization (ML) presents several or severe upper outliers (in logarithmic scale the lower outliers actually correspond to credal-sets that better fit the real joint distribution). The presence of such anomalies induces to discard techniques based on Δ and *ML*. Without them we can compare the performance on the remaining five data-sets obtaining the results reported in Fig. 3.

Notice the better performance of the corrected estimates with respect to the not changed ones and the best behavior of the minimization of $L2$ and Δ_{mix} with respect to the other pseudo-distances. Then, it seems that the correction produces an information merging of the two samples A and B that the frequencies estimation does not meet. Among the different possible corrections, $L2$ and Δ_{mix} minimizations seems to better preserve the original information, moreover Δ_{mix} has the further feature of the automatic localization of the sub-domains of $\mathcal{E}$ where the changes are needed.

6. Conclusion

Checking coherence and removing incoherences in the estimations is a long debated problem in the literature; we have studied it by focusing on statistical matching applications. In fact, in this kind of application the incoherence can arise when the variables are linked by logical relations. We have applied several incoherence adjustment procedures in this specific ambit: partial likelihood maximization, pseudo-distances minimization and coherent dilation. The study revealed some differences among these adjustments. We mainly focused on minimization of pseudo-distances and we have observed how a specialization of usual pseudo-distances performs better. This is due to integration of sources and lack of information on the variables not jointly observed, as is typical for the statistical matching problem. In particular, a specific adjustment of a discrepancy shows the advantage of an automatic and weighted localization of the sub-domains where incoherence must be removed. A comparison among different pseudo-distances based on simulated values has confirmed our expectations and has shown a surprisingly better performance of the corrected assessment with respect to the originally coherent ones.

We have also analyzed a very simple practical application and we have shown that better results are obtained not simply focusing on the minimal number of incoherent values, but involving all the elements conditioned to the same scenarios in which incoherence arises. On the other hand, coherent imprecise adjustment performs better with minimal number of changed values, with the counterpart of obviously vaguer inference conclusions that however could be improved by a "maximally supported" core detection.

References

[1] M. Ballin, M. D'Orazio, M. Di Zio, M. Scanu, N. Torelli, Statistical Matching of Two Surveys with a Common Subset, Tec. Report no. 124 of University of Trieste, Dept. Scienze Economiche e Statistiche, Italy, 2009, pp. 68–79.

[2] V. Biazzo, A. Gilio, A generalization of the fundamental theorem of de Finetti for imprecise conditional probability assessments, Int. J. Approx. Reason. 24 (2000) 251–272.

[3] V. Biazzo, A. Gilio, Some theoretical properties of conditional probability assessments, Lect. Notes Comput. Sci. LNAI 3571 (2005) 775–787.

[4] T. Brooke, D. Kendrick, A. Meeraus, GAMS: A User's Guide, The Scientific Press, Redwood City, California, 1988.

[5] A. Capotorti, L. Galli, B. Vantaggi, How to use locally strong coherence in an inferential process based on upper-lower probabilities, Soft Comput. 7 (5) (2003) 280–287.

[6] A. Capotorti, G. Regoli, Coherent correction of inconsistent conditional probability assessments, in: L. Magdalena, M. Ojeda-Aciego, J.L. Verdegay (Eds.), Proceeding of IPMU'08, Malaga (ES), 2008, pp. 891–898.

[7] A. Capotorti, G. Regoli, F. Vattari, Correction of incoherent conditional probability assessments, Int. J. Approx. Reason. 51 (6) (2010) 718–727.

[8] A. Capotorti, B. Vantaggi, Incoherence correction strategies in statistical matching, in: Proceedings of ISIPTA 2011, Innsbruck (Austria), 2011, pp. 109–118.

[9] A. Capotorti, M. Zagoraiou, Implicit degree of support for finite lower-upper conditional probabilities extensions, in: Proceedings of Information Processing and Management of Uncertainty in Knowledge-based Systems, vol. III, IPMU'06, EDK Paris, France, 2006, pp. 2331–2338.

[10] G. Coletti, Coherent numerical and ordinal probabilistic assessments, IEEE Transaction on Systems, Man, and Cybernetics 24 (1994) 1747–1754.

[11] G. Coletti, R. Scozzafava, Probabilistic logic in a coherent setting, Kluwer, Dordrecht, 2002., Series "Trends in Logic".

[12] G. De Cooman, M. Zaffalon, Updating beliefs with incomplete observations, Art. Intell. 159 (2004) 75–125.

[13] B. de Finetti, Sull'impostazione assiomatica del calcolo delle probabilità, Ann. Univ. Trieste 19 (1949) 3–55., Engl. transl. in: Ch. 5 of Probability, Induction, Statistics, Wiley, London, 1972.

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

[15] M. D'Orazio, M. Di Zio, M. Scanu, Statistical Matching: Theory and Practice, Wiley, New York, 2006.

[16] L.E. Dubins, Finitely additive conditional probabilities, conglomerability and disintegration, The Annals of Probability 3 (1975) 89–99.

[17] A. Gilio, G. Sanfilippo, Coherent conditional probabilities and proper scoring rules, in: Proceedings of ISIPTA 2011, Innsbruck (Austria), 2011, pp. 189–198.

[18] J.B. Kadane, Some statistical problems in merging data files, J. Off. Stat. 17 (2001) 423–433.

[19] W.A. Kamakura, M. Wedel, Statistical data fusion for cross-tabulation, J. Market. Res. 34 (1997) 485–498.

[20] P.H. Krauss, Representation of conditional probability measures on Boolean algebras, Acta Math. Acad. Scient. Hungar. 19 (1968) 229–241.

[21] S. Kullback, Information Theory and Statistics, Wiley, New York, 1957.

[22] F. Lad, Operational Subjective Statistical Methods: A Mathematical, Philosophical, and Historical Introduction, Wiley, New York, 1996.

[23] D.V. Lindley, A. Tversky, R.V. Brown, On the reconciliation of probability assessments, J. Roy. Stat. Soc. Ser. A 142 (2) (1979) 146–180.

[24] R.J.A. Little, D.B. Rubin, On jointly estimating parameters and missing data by maximising the complete-data likelihood, Am. Stat. 37 (1983) 218–220.

[25] C.F. Manski, Identification Problems in the Social Sciences, Harvard University Press, Cambridge, MA, 1995.

[26] B.A. Okner, Constructing a new data base from existing microdata sets: the 1966 merge file, Ann. Econ. Soc. Measure. 1 (3) (1972) 325–342.

[27] B.A. Okner, Data matching and merging: an overview, Ann. Econ. Soc. Measure. 3 (2) (1974) 347–352.

[28] E. Miranda, Updating coherent previsions on finite spaces, Fuzzy Sets Syst. 160 (9) (2009) 1286–1307.

[29] G. Paass, Statistical match: evaluation of existing procedures and improvements by using additional information, in: G.H. Orcutt, H. Quinke (Eds.), Microanalytic Simulation Models to Support Social and Financial Policy, Elsevier Science, Amsterdam, 1986, pp. 401–422.

[30] R Development Core Team R: a language and environment for statistical computing. R Foundation for Statistical Computing, http://www.R-project.org.

[31] S. Rässler, Statistical matching: a frequentist theory, practical applications and alternative Bayesian approaches, Lecture Notes in Statistics, Springer Verlag, 2002.

[32] D.B. Rubin, Statistical matching using file concatenation with adjusted weights and multiple imputations, J. Business Econ. Stat. 2 (1986) 87–94.

[33] P. Szivós, T. Rudas, I.G. Tóth, A tax-benefit microsimulation model for Hungary, in: Workshop on Microsimulation in the New Millennium: Challenges and Innovations, Cambridge, 1998.

[34] J.L. Schafer, Analysis of Incomplete Multivariate Data, Chapman & Hall, London, 1997.

[35] N. Torelli, M. Ballin, M. Di Zio, M. D'Orazio, M. Scanu, G. Corsetti, Statistical matching of two surveys with a non randomly selected common subset, in: Insights on Data Integration Methodologies, Proceedings of ESSnet-ISAD workshop, Vienna, 2008.

[36] B. Vantaggi, The role of coherence for the integration of different sources, in: Proceedings of 4th International Symposium on Imprecise Probabilities and their Applications ISIPTA'05, Pittsburgh, 2005, pp. 369–378.

[37] B. Vantaggi, Statistical matching of multiple sources: a look through coherence, Int. J. Approx. Reason. 49 (3) (2008) 701–711.

[38] P. Walley, Measures of uncertainty in expert systems, Artif. Intell. 83 (1) (1996) 1–58.

[39] M. Wolfson, S. Gribble, M. Bordt, B. Murphy, G. Rowe, The social policy simulation database and model: an example of survey and administrative data integration, Surv. Curr. Business 69 (1989) 36–41.