Behavior of L_1-based probabilistic correction applied to statistical matching with misclassification information

Giulia Buonumori **Andrea Capotorti**
Dip. Matematica e Informatica
University of Perugia - Italy
andrea.capotorti@unipg.it giulia.buonumori@studenti.unipg.it

Abstract

We illustrate the use of a recently proposed efficient procedure, based on L_1 distance minimization, for correcting inconsistent (i.e. incoherent) probability assessments for the so named statistical matching problem. Albeit the statistical matching problem is based on conditional probabilities estimates, inconsistencies can appear only among assessments given on the same conditioning values, hence a correction instance can be splitted in a finite set of unconditional correction instances where the L_1-based correction can efficiently operate. The statistical matching problem has been recently enriched with the possibility of a misclassification setting, breaking the aforementioned segmentation possibility. Anyhow the L_1-based procedure can be applied by a straightforward translation in a MIP problem, albeit the set of consistent solutions turns out to be not convex and hence potential disconnected solutions can appear.

1 Introduction

In recent contributions [1, 2] it has been proposed an efficient procedure for correcting inconsistent (i.e. incoherent) probability assessments based on L_1 distance minimization and encoded in mixed integer programming (MIP) problems. The procedure is particular apt to deal with assessments stemming from different sources of information, and the so named statistical matching problem is one of those cases (see e.g. [11]). Albeit the statistical matching problem is based on conditional probabilities estimates, always in [11] it has been proven that inconsistencies can appear only among assessments given on the same conditioning values, hence a correction instance can be splitted in a finite set of unconditional correction instances where the L_1-based correction can efficiently operate.

The problem has been recently enriched with the possibility of a misclassification setting [8], breaking the aforementioned segmentation possibility. If marginal assessments on the conditioning variable are taken for good, the only possible correction are the closest Fréchet-Hoeffding bounds for the misclassification probabilities. On the contrary, if also the marginal probabilities are allowed to be modified or the assessment is partial, the L_1-based procedure can be applied by a straightforward translation in a MIP problem, albeit the set of consistent solutions turns out to be not convex and hence potential disconnected solutions can appear. It is eventually notable that in the case the L_1-based correction would induce some marginal probability to be null, that could happen whenever the initial assessment would be based on rare or scarce observations, it will not be needed to proceed to further corrections on deeper zero layers (see [5]).

In the next sections we will briefly illustrate the general statistical matching (Sec.2), the merging and correction procedures for general unconditional probability assessments (Sec.3) and consequently their specific application to the statistical matching problem (Sec.4). Finally, in Sec.5 we will give a rough preliminary idea of the correction of incoherent evaluations when also a missclassification mechanism is assessed.

2 The statistical matching problem

As already stated, we propose to adopt a correction procedure applied to a merging operation for a specific practical problem named "statistical matching". Let us briefly recall what it means and which are its main peculiarities. A detailed description of such a problem can be found, e.g., in [9, 10].

Denote by $(\mathcal{X}_1, \mathcal{Y}_1), \ldots, (\mathcal{X}_{n_A}, \mathcal{Y}_{n_A})$ and by $(\mathcal{X}_{n_A+1}, \mathcal{Z}_{n_A+1}), \ldots, (\mathcal{X}_{n_A+n_B}, \mathcal{Z}_{n_A+n_B})$ two random samples, related to two sources A and B, of dimensions n_A and n_B. Samples observe three categorical variables $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ with modalities mx_i, $i \in I$, my_j, $j \in J$ and mz_k, $k \in K$, respectively. Hence in the sequel we will adopt the following notation for the possible observations:

$$X_i \equiv (\mathcal{X} = mx_i), i \in I, \quad Y_j \equiv (\mathcal{Y} = my_j), j \in J, \quad Z_k \equiv (\mathcal{Z} = mz_k), k \in K, \quad (1)$$

that will constitute our propositional variables (i.e. events).

Let S_s (with $s = 1, 2$) be the two, possibly different, sampling schemes. From them, relevant parameters, represented by (conditional) probabilities, can be estimated : from A the probability to observe Y_j conditional on X_i (for any $i \in I$)

$$\mathbf{y}_{j|i} = P_{\mathcal{Y}|(X_i)}(Y_j), \quad (2)$$

and analogously from B the probability to observe Z_k conditional on X_i (for any $i \in I$)

$$\mathbf{z}_{k|i} = P_{\mathcal{Z}|X_i}(Z_k). \quad (3)$$

Moreover, from A we can estimate the probability to observe X_i by following the first sampling scheme

$$\mathbf{x}_i^{S_1} = P_{\mathcal{X}}(X_i|S_1), \tag{4}$$

while from file B by following the second one

$$\mathbf{x}_i^{S_2} = P_{\mathcal{X}}(X_i|S_2), \tag{5}$$

and, by supposing that an observation can be obtained through one single sampling scheme S_s, with $s \in \{1, 2\}$ and probability $P(S_s)$, we get

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

Under the assumption of a common sampling scheme, estimations are obtained through partial maximum likelihood method, and the result brings to the frequencies

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

with $n_A^{i\cdot}$ and $n_B^{i\cdot}$ cardinalities of elements with X_i in samples A and B, respectively, while n_A^{ij} is the cardinality of elements in A with (X_i, Y_j) and n_B^{ik} is the cardinality of elements in B with (X_i, Z_k).

Whenever $n_A^{i\cdot}$ (the same for $n_B^{i\cdot}$) is equal to zero (i.e. no observation in A has X_i) the value $\mathbf{y}_{j|i}$ ($\mathbf{z}_{k|i}$) is undefined and this specific parameter has not any estimation.

If the probabilities $P(S_s), s = 1, 2$, can be elicited, we get a precise conditional probability assessment $(V, \mathcal{E}, \mathbf{p}, \mathfrak{C})$ with

$$V = \{X_i,\, Y_j,\, Z_k\}\,,\ \mathcal{E} = \{X_i,\, Y_j|X_i,\, Z_k|X_i\}\,,\ \mathbf{p} = \{\mathbf{x}_i, \mathbf{y}_{j|i}, \mathbf{z}_{k|i}\},\ i \in I, j \in J, k \in K, \tag{8}$$

while $\mathfrak{C}$ is a set of logical constraints, in this field named as "structural zeroes", among elements of V.

Usually, the first step is to check the coherence of $(V, \mathcal{E}, \mathbf{p}, \mathfrak{C})$, that should resort to check the satisfiability of a sequence of linear systems (see, e.g., [5]) but that in the particular context of the statistical matching can be reduced to the solvability of a unique linear system (see [11]). Generally, whenever $(V, \mathcal{E}, \mathbf{p}, \mathfrak{C})$ is coherent there is more than one solution and the set of all of them forms a so called "credal set".

In the trivial case of logical independence, coherence is automatically ensured (see [11]). In the more worthwhile case of structural zeroes among random variables $\mathcal{Y}$ and $\mathcal{Z}$ (for real applications where these are present refer, e.g., to [9]), coherence of the entire assessment $(V, \mathcal{E}, \mathbf{p}, \mathfrak{C})$ in (8) is not directly ensured by the separate coherence of the distinct assessments with numerical parts (2), (3), (6). The problem is hence to find a coherent assessment that solves inconsistencies.

Anyhow, whenever present, inconsistencies focus on conditional events with the same conditioning X_i (proofs and examples again in [11]).

This result will permit to split the problem of the merging of the two estimates into separate subproblems: one for the unconditional values $\mathbf{x}_i, i \in I$, and one for each conditioning X_i about the conditional quantities $\{\mathbf{y}_{j|i}, \mathbf{z}_{k|i}\}$, $j \in J$ and $k \in K$. In each of these subproblems the merging and correction procedure can be applied, even being the statistical matching a conditional problem, by fixing in each subproblem the conditioning event, that could be the sure event $\top$ or X_i, and dealing with actually unconditional problems. To see how this could be possible, let us formalize in the next Sections the merging and correction procedure, starting with the formal definition of the unconditional probability assessments.

3 Correction of probability assessments

A probability assessment on a finite domain is a quadruple $\pi = (V, U, p, \mathfrak{C})$, where $V = \{X_1, \ldots, X_k\}$ is a finite set of propositional variables, representing any potential event of interest, U is a subset of V that contains the effective events taken into consideration, $p : U \to [0, 1]$ is a function which assigns a probability value to each variable in U, and $\mathfrak{C}$ is a finite set of logical constraints which lie among all the variables in V.

With such framework, the user provides a probability value for the elements of set U, but logical constraints can also be written in terms of all the existing events V. This feature allows to extend an initial assessment to a larger domain without redefining the whole model.

The constraints in $\mathfrak{C}$ are written with the usual logical notation, where $\neg$, $\wedge$ and $\vee$ denote the negation, disjunction and conjunction connectives, respectively; $\Rightarrow$ the material implication; $=$ the logical equivalence; $\top$ and $\bot$ the universal tautology and contradiction (sure and impossible events), respectively. These constraints can be used to represent any kind of compound event, for instance that an event is the conjunction of other two events, or denote the implications or incompatibilities among the elements of V. Without loss of generality, we suppose that $\mathfrak{C}$ is expressed in conjunctive normal form (CNF) that will help in the implementation part of the correction procedure. Hence $\mathfrak{C} = \{c_1, \ldots, c_m\}$ where each element c_i of $\mathfrak{C}$ is a disjunctive clause, i.e. $c_i = \left(\bigvee_{h \in H_i} X_h \right) \vee \left(\bigvee_{l \in L_i} \neg X_l \right)$ for some $H_i, L_i \subseteq \{1, \ldots, n\}$. Since we will require that all the logical constraint present in $\mathfrak{C}$ must be satisfied, $\mathfrak{C}$ can be seen as the conjunction of $c_1, \ldots, c_m$.

Since a probabilistic assessment π is partial, it may or not be coherent, i.e. consistent with a probability distribution.

The problem of checking the coherence of a probability assessment, called **CPA**, has been already studied (see [3, 4] among the many), albeit in a slightly different form, showing that it is a NP-complete problem, even when the constraints in $\mathfrak{C}$ are binary (i.e., each of them involves only two variables).

There exist several approach to solve **CPA**. Among those, the Mixed Integer Programming (MIP) based approach has proved to be very effective as reported in

Table 1: Variables of $\mathcal{P}1$

name	size	type
a_{ij}, for $i = 1, \ldots, n$ and $j = 1, \ldots, n+1$	$n(n+1)$	binary
b_{ij}, for $i = 1, \ldots, n$ and $j = 1, \ldots, n+1$	$n(n+1)$	real
q_j, for $j = 1, \ldots, n+1$	$n+1$	real
r_i for $i = 1, \ldots, n$	n	real
s_i for $i = 1, \ldots, n$	n	real

[6, 7], where their implementation was able to handle coherence testing instances up to 1000 variables and 1000 disjunctive clauses in average time ranging from some seconds to some minutes.

When a probability assessment $\pi = (V, U, p, \mathfrak{C})$ is not coherent, then it is possible to "correct" it in order to obtain a coherent probability assessment π' which is as close as possible to π, according to a distance or a pseudo-distance function between probability assessments.

Definition 1 *Given a distance d, a d-correction of a probability assessment $\pi = (V, U, p, \mathfrak{C})$ is a vector p' such that the probability assessment $\pi' = (V, U, p', \mathfrak{C})$ is coherent and $d(p, p')$ is minimized. We denote $\mathcal{C}_d(\pi)$ the sets of all the d-correction of π.*

It is important to notice that for certain choices of d, $\mathcal{C}_d(\pi)$ has just one element, for instance when d is the Euclidean distance. On the other hand, for some other choices of d, $\mathcal{C}_d(\pi)$ has more than one element for some probability assessments π. In this case, the operation of correcting a probability assessment leads to an imprecise probability model, called "credal set". Clearly if π is coherent, then $\mathcal{C}_d(\pi) = \{p\}$, for any distance d of $\mathbb{R}^n$.

In this paper we focus on the L_1 distance defined as $d_1(p, p') = \sum_{i=1}^{n} |p(X_i) - p'(X_i)|$ and we denote $\mathcal{C}_{d_1}(\pi)$ as $\mathcal{C}(\pi)$.

This distance has two important properties. First of all, the correction can be easily interpreted as a cost of changing the probability values, in terms of the sum of the displacements $|p(X_i) - p'(X_i)|$. Minimization of such displacements obeys to the basic principle of minimal change in a numerical uncertainty setting. Secondly, the resulting minimization problem with L_1 distance can be solved by using linear programming with both integer and real variables and this represents a clear computational advantage compared to other distances which require non linear (quadratic, logarithmic, etc.) optimizations tools.

In [1] the details of the MIP-based program $\mathcal{P}1$ implementation have been give. Here we just recall the basic quantities involved in it.

It is well known that if a probability assessmentis coherent, there exists a sparse probability distribution μ so that p' can be written as a convex combination of at most $n+1$ atoms. Let us call $\alpha^{(1)}, \ldots, \alpha^{(n+1)}$ these atoms.

The variables of $\mathcal{P}1$ are summarized in Table 3, while its linear constraints are

$$\sum_{h \in H_i} a_{h,j} + \sum_{l \in L_i} (1 - a_{l,j}) \geq 1 \qquad i = 1, \ldots, m \quad j = 1, \ldots, n+1 \tag{9}$$

$$\sum_{j=1}^{n+1} b_{ij} = p(X_i) + (r_i - s_i) \qquad i = 1, \ldots, n \tag{10}$$

$$0 \leq b_{ij} \leq a_{ij}, \quad a_{ij} - 1 + q_j \leq b_{ij} \leq q_j \qquad i = 1, \ldots, n \quad j = 1, \ldots, n+1 \tag{11}$$

$$\sum_{i=1}^{n+1} q_j = 1 \tag{12}$$

$$r_i \leq 1, \quad s_i \leq 1 \qquad i = 1, \ldots, n \tag{13}$$

The implicit constraint is that all of the variables must be non-negative, as usual in linear programming.

The variables a_{ij} are binary , i.e. constrained in $\{0, 1\}$. Each value a_{ij} should correspond to the atom component $\alpha^{(j)}(X_i)$, for $i = 1, \ldots, n$ and $j = 1, \ldots, n+1$. Indeed, the constraint (9) forces each assignment $(a_{1j}, \ldots, a_{nj})$ to satisfy all the clauses $c_i \in \mathfrak{C}$. The values $q_1, \ldots, q_{n+1}$ represent the coefficient of the convex combination which generates p', which also correspond to the probabilities $\mu(\alpha^{(1)}), \ldots, \mu(\alpha^{(n+1)})$. The constraint (11) allows to express the equation

$$b_{ij} = a_{ij} \cdot q_j \text{ for } i = 1, \ldots, n \text{ and } j = 1, \ldots, n+1,$$

without using the multiplication, otherwise $\mathcal{P}1$ would not be a linear problem. Indeed, if $a_{ij} = 0$, then $b_{ij} = 0$ too. On the other hand, if $a_{ij} = 1$, then $a_{ij} - 1 + q_j \leq b_{ij} \leq q_j$ reduces to $q_j \leq b_{ij} \leq q_j$. In this way, for each $i = 1, \ldots, n$ the sum $\sum_{j=1}^{n+1} b_{ij}$ corresponds to $\sum_{j=1}^{n+1} a_{ij} \cdot q_j$. Since $a_{ij} = 1$ if and only if $\alpha^{(j)}$ satisfies X_i, the sum is also equal to $p'(X_i)$.

The variables r_i, s_i are slack variables, which represent, respectively, the positive and the negative difference between $p(X_i)$ and $p'(X_i)$, as implied by the constraint (10). Hence $(r_i - s_i)$ is the correction on the probability of X_i, for each $i = 1, \ldots, n$.

Finally, the objective function to be minimized is

$$\sum_{i=1}^{n} (r_i + s_i) \tag{14}$$

that, being the sum of these corrections, corresponds to the L_1-distance between p and p', i.e., $\sum_{i=1}^{n} |p(X_i) - p'(X_i)|$. Note that for each $i = 1, \ldots, n$, it is impossible that $r_i > 0$ and $s_i > 0$, otherwise the objective function would not be minimized.

It is easy to see that any solution of the linear program $\mathcal{P}1$ corresponds to a L_1-correction p' of p. And vice versa, any L_1-correction p' of p corresponds to a solution of $\mathcal{P}1$.

The optimal value δ for the objective function corresponds to the minimum possible correction on p and any coherent probability assessment $\pi' = (V, U, p', \mathfrak{C})$ such that $d_1(p, p') = \delta$ is a possible solution i.e., p' is an element of $\mathcal{C}(\pi)$. Note that p' can be simply obtained as $p'_i = p_i + r_i - s_i$ for $i = 1, \ldots, n$.

In many situations $\mathcal{C}(\pi)$ has more than one element and the MIP problem is able to find just one solution, which could not be a good representative of all the elements of $\mathcal{C}(\pi)$, as happens when it is an extreme value. Hence program $\mathcal{P}1$ must be associated with an other MIP program $\mathcal{P}2$ to generate all the elements of $\mathcal{C}(\pi)$. In $\mathcal{P}2$ all the constraints and the variables of $\mathcal{P}1$ are reported and it contains a new real variable z, which is subject to the constraints $r_i + s_i \leq z$, for $i = 1, \ldots, n$ (hence $z \geq \max_{i=1,\ldots,n}(r_i + s_i)$), and the new additional constraint $\sum_{i=1}^{n}(r_i + s_i) = \delta$. In this way, the $\mathcal{P}2$ objective function to be minimized is simply z.

The corrected assessment $\bar{\pi} = (V, U, \bar{p}, \mathfrak{C})$ tries to spread the difference δ as much as possible among all the dimensions, i.e. the variables of U. Hence $\bar{p}$ is, in some sense, the most "entropic" point of $\mathcal{C}(\pi)$.

Using $\bar{p}$, it is possible to find the face F_1 of the polytope $\mathcal{Q}$ where $\mathcal{C}(\pi)$ lies. The face F_1 is itself a convex set with at most $n + 1$ atoms as extremal points, which can be found as a part of the solutions of $\mathcal{P}2$ (i.e., the optimal values of a_{ij}).

By looking at the signs of $\bar{p}(X_i) - p(X_i)$, for $i = 1, \ldots, n$, it is also possible to determine the face F_2 of $\mathcal{B}_\pi(\delta)$ which contains $\mathcal{C}(\pi)$. Indeed, F_2 is a convex set with at most n extremal points of the form $p + sign(\bar{p}(X_j) - p(X_j)) \cdot \delta \cdot e_j$.

The whole set of corrections $\mathcal{C}(\pi)$ will result as $F_1 \cap F_2$.

These steps have been implemented in a procedure named Correct that, given in input any partial assessment π, returns the extremal points of the credal set $\mathcal{C}(\pi)$ (for details refer again to [1]).

In Sec.2 we have seen that an incoherent assessment could come by the merging of two separate assessments π_1 and π_2. Let us show how to produce a new coherent probability assessment π_3 which is a "compromise" between π_1 and π_2, keeping as much as possible the information from both.

Depending if the two assessments are compatible (i.e. they give the same values to common variables) or not (i.e. there is an explicit contradiction given by different probabilities to some common variable) there are two different way of defining the joining of them. We report here just the basic notions, referring again to [1] for all the details.

In case of compatibility, it is possible to join directly the two original assessments, so that the merging will result as $\pi_1 \oplus \pi_2 = \text{Correct}(\pi_1 + \pi_2)$. Note that, since such merging procedure is the result of our Correct procedure, its output could be a credal set, as already outlined in the previous Section.

When the probability assessments to be merged are non compatible it is not possible to join directly them into a unique assessment. Hence, in addition to possible initial incoherences present in the separate assessment, we have to tackle with a sure incoherence in the joint one. Anyhow two different correction procedures are possible: a "weighted combination" of the two assessments, or a "assignment to

duplicates". The first approach requires to create a non contradictory probability assessment derived from π_1 and π_2, by choosing a weighted average probability value for each variable in common.

The merging operation between π_1 and π_2 is then defined as the new assessment obtained as correction of the weighted average $\pi_1 \oplus_\omega \pi_2 = \text{Correct}(\pi_1 +_\omega \pi_2)$.

The second approach is to create a probability assessment which maintains both numerical values and to solve the apparent contradiction by adding a new logical variable X_i', for each variable X_i in common. Obviously the logical constraints $\neg X_i \vee X_i'$ and $X_i \vee \neg X_i'$ must be added to $\mathfrak{C} \cup \mathfrak{D}$ to represent the duplicated events $X_i = X_i'$.

Indeed, apart from separate initial incoherences of the two initial assessments π_1 and π_2, the new assessment so obtained $\pi_1 + \pi_2$ is obviously incoherent since the duplicated events with different associated values and the merging operation of π_1 and π_2 results as $\pi_1 \oplus_I \pi_2 = \text{Correct}(\pi_1 + \pi_2)$. Note that, whenever the two assessments π_1 and π_2 are compatible, this merging operator $\pi_1 \oplus_I \pi_2$ coincides with the previous $\pi_1 \oplus \pi_2$ since no duplication of variables is needed in such a case.

The main difference between the two merging of incompatible assessments just described is that $\oplus_I$ is an unsupervised approach since it tries to automatically solve the contradictions, while the operator $\oplus_\omega$ is a supervised approach since it needs an explicit and "exogenous" conciliation among explicit numerical contradictions through the choice of the weight ω. These differences can lead to very different final results. Anyway, the idea behind these two methods is the same, i.e., the merging of two information sources can be performed in two steps. First, put together all the information $\mathcal{I}$, and then find the smallest number of corrections on $\mathcal{I}$ such that the new information $\mathcal{I}'$ is consistent. The choice of which merging operator to adopt should be based on the availability or not of the weight ω representing the relevance, or better of the reliability, of the sources of information. If a reliability grade ω is available, or reasonably assessed, the $\oplus_\omega$ should be preferred, if not the $\oplus_I$ operator avoids the use of unrealistic assumptions.

4 Application of the merging and correction procedures to the statistical matching problem

We can now describe how the merging an correction procedures defined in the previous Section can be applied to the statistical matching problem described in Sec. 2. The preliminary operation is to merge the estimates coming from the two different sampling schemes S_1 and S_2. In particular, since incoherences could be focused only on events conditioned to the same event, we can split the domain $\mathcal{E}$ into sub-domains

$$\mathcal{E}_\Omega = \{X_i\}_{i \in I}\,; \tag{15}$$

$$\mathcal{E}_i = \{Y_j | X_i, Z_k | X_i\}_{j \in J, k \in K} \text{ for } i \in I \tag{16}$$

Since, as described in Section 2, variables $\mathcal{Y}$ and $\mathcal{Z}$ are not jointly observed, on the domains E_i the two sources of information do not overlap and hence the problem will be to, eventually, correct the estimates $\{\mathbf{y}_{j|i}, \mathbf{z}_{k|i}\}$ obtained through (2) and (3). A proper merging operation is needed for the estimates $\{\mathbf{x}_i^{S_1}\}_{i\in I}$ and $\{\mathbf{x}_i^{S_2}\}_{i\in I}$, both on elements of $\mathcal{E}_\Omega$.

As described in Sec. 3, two different approaches can be used: the "supervised" procedure if we can assess the "weight" ω of the relevance or reliability of sources; or the "unsupervised" one that relies on the duplication of all events X_i and consequent addition of structural constraints that express such duplication.

Schematically, the first approach needs hence to compute at first a componentwise"weighted average"

$$\mathbf{x}^{S_1} +_\omega \mathbf{x}^{S_2} = \omega\{\mathbf{x}_i^{S_1}\}_{i\in I} + (1-\omega)\{\mathbf{x}_i^{S_2}\}_{i\in I} \tag{17}$$

for a chosen weight $\omega \in [0,1]$, and consequently apply the correct procedure to $(V, \mathcal{E}_\Omega, \mathbf{x}^{S_1} +_\omega \mathbf{x}^{S_2}, \mathfrak{C})$ obtaining for the numerical part

$$\mathbf{lub} = \mathbf{x}^{S_1} \oplus_\omega \mathbf{x}^{S_2} = \text{Correct}(\mathbf{x}^{S_1} +_\omega \mathbf{x}^{S_2}) \tag{18}$$

If there is some missing value for $\{\mathbf{x}_i^{S_1}\}_{i\in I}$ or for $\{\mathbf{x}_i^{S_2}\}_{i\in I}$ it must be put equal to 0 in (17). Remember that the correct procedure could lead to either a single solution or to a convex set of solutions, hence **lub** in (18) could be either an actually precise coherent assessment $\{\mathbf{x}_i\}_{i\in I}$ or a proper lower-upper assessment $\{\mathbf{lub}_i\}_{i\in I}$.

Note moreover that, if estimates are taken through frequencies in both samples, $\mathbf{x}^{S_1} +_\omega \mathbf{x}^{S_2}$ in (17) turns out to be directly coherent for any choice of $\omega \in [0,1]$ so that $\mathbf{lub} = \{\mathbf{x}_i\}_{i\in I} = \mathbf{x}^{S_1} +_\omega \mathbf{x}^{S_2}$. In particular, choosing $\omega = \frac{n_A}{n_A+n_B}$ we obtain exactly the $\mathbf{x}_i$ estimates already described in (7). So the common sampling scheme can be re-interpreted in our method as separate sampling schemes with weights proportional to the different sample dimensions.

The second approach is to let the correct procedure work without any exogenous weight of the sources and contemplating simultaneously the two different estimates $\{\mathbf{x}_i^{S_1}\}_{i\in I}$ and $\{\mathbf{x}_i^{S_2}\}_{i\in I}$. The obvious inconsistencies are solved by duplicating the events in $\mathcal{E}_\Omega$ as $\mathcal{E}'_\Omega = \{A_i \equiv X_i, B_i \equiv X_i\}_{i\in I}$ and by adding structural zeros induced by the duplicates $A_i = B_i$, for $i \in I$. Hence the correction procedure can be applied to the concatenated assessment $\mathbf{x}^{S_1} \biguplus \mathbf{x}^{S_2}$ that assigns $\mathbf{x}_i^{S_1}$ to A_i and $\mathbf{x}_i^{S_2}$ to B_i, for any $i \in I$, by obtaining a, generally imprecise, assessment $\mathbf{lub} = \mathbf{x}^{S_1} \oplus_I \mathbf{x}^{S_2} = \text{Correct}(\mathbf{x}^{S_1} \biguplus \mathbf{x}^{S_2})$.

As already mentioned, to the other conditioned "strata" $(\mathcal{E}_i, \{\mathbf{y}_{j|i}, \mathbf{z}_{k|i}\}_{j\in J, k\in K})$ the correction procedure can be straightly applied obtaining, generally imprecise, estimates $\{\mathbf{lub}_{j|i}, \mathbf{lub}_{k|i}\}_{j\in J, k\in K}$, for $i \in I$.

At the end, by collecting all the corrections we get a, generally imprecise, coherent assessment $(V, \mathcal{E}, \{\mathbf{lub}_i, \mathbf{lub}_{j|i}, \mathbf{lub}_{k|i}\}_{i\in I, j\in J, k\in K}, \mathfrak{C})$ as the merging of the separate estimates based on the two sample schemes S_1 and S_2.

5 Correction of a statistical matching with missclassification

In [8] it is described a variation of the usual statistical matching problem by introducing a missclassification mechanism that could be summarized by saying that the common variable $\mathcal{X}$ is biasedly observed in source A (e.g. if its values are assessed by not experts in the field) giving rise to a new variable $\mathcal{X}^*$ with the same modalities $mx_i, i \in I$, while $\mathcal{X}$ remains properly observed in the second source B.

In addition, a missclassification mechanism, specified by conditional probabilities $P_{\mathcal{X}|X^*_{i^*}}(X_i)$, can be fully or partially assessed. Hence the whole assessment that results from the joining of all the available information will be of the form $\pi = (V^*, \mathcal{E}^*, \mathbf{p}^*, \mathfrak{C}^*)$ with

$$\begin{aligned} V^* = \{X_i,\, X^*_{i^*}\, Y_j,\, Z_k\} \quad &, \quad \mathcal{E}^* = \{X_i,\, X^*_{i^*}, Y_j | X^*_{i^*},\, Z_k | X_i,\, X_i | X^*_{i^*}\}\,, \\ \mathbf{p}^* = \{\mathbf{x}_i, \mathbf{x}^*_{i^*}, \mathbf{y}_{j|i^*}, \mathbf{z}_{k|i}, \mathbf{x}_{i|i^*}\} \quad &, \quad (i, i^*) \in \mathcal{I} \subseteq I \times I, j \in J, k \in K, \end{aligned} \tag{19}$$

while $\mathfrak{C}^*$ incorporates the structural zeroes among elements of V^*.

This brakes the division in the subdomains (15,16) and the possibility to correct incoherence of the whole assessments with a finite set of corrections on the subdomains. Anyhow, always in [8], it has been proven that the coherence of the whole assessments is basically due to the coherence of the subassessment involving only $\mathcal{X}^*$ and $\mathcal{X}$, hence with numerical part $\mathbf{p}^*_{|\mathcal{I}} = \{\mathbf{x}_i, \mathbf{x}_{i^*}, \mathbf{x}_{i|i^*}\}_{(i,i^*)\in\mathcal{I}}$, and that, in the case of absence of structural zeroes between $\mathcal{X}^*$ and $\mathcal{X}$ (i.e. $\mathcal{I} = I \times I$), the conditional probabilities $\mathbf{x}_{i|i^*}$, $i, i^* \in I$, are constrained by coherence to lay inside the so called Fréchet-Hoeffding bounds:

$$\frac{\max(0, \mathbf{x}_i + \mathbf{x}^*_{i^*} - 1)}{\mathbf{x}^*_{i^*}} \leq \mathbf{x}_{i|i^*} \leq \frac{\min(\mathbf{x}_i, \mathbf{x}^*_{i^*})}{\mathbf{x}^*_{i^*}}. \tag{20}$$

Such bounds imply that the set of coherent values for $\{\mathbf{x}_i, \mathbf{x}_{i^*}, \mathbf{x}_{i|i^*}\}_{(i,i^*)\in\mathcal{I}}$ is not convex in general, hence the credal set of a correction of an incoherent assessments could result not connected and hardly computable. Hence we cannot expect a procedure that produces the whole credal set of correction $\mathcal{C}(\pi)$. Anyhow, we can find just one element of such credal set by a particular setting of linear constraints in a new MIP-based optimization.

More precisely, we change a little bit the notation with respect the MIP program $\mathcal{P}1$ described in Sec.3. In fact now the atoms are characterized by the simple possibility of having the conjunction $X_i \wedge X^*_{i^*}$, so that the set of constraints associated to the subassessment can be simply represented by set of couples of indexes $\mathfrak{C}^*_{|\mathcal{I}} = \{(i, i^*) \in I \times I : X_i \wedge X^*_{i^*} = \bot\}$ (in the sequel we will denote with c^* the cardinality of $\mathfrak{C}^*_{|\mathcal{I}}$). Consequently the binary variables can be denoted with a_{ii^*}, while the real variables with b_{ii^*} and q_{ii^*}. About the slack variables, we need them for the potential modification of both the marginal and conditional probabilities, hence we denote them with $r_i, s_i, r_{i^*}, s_{i^*}, r_{i|i^*}, s_{i|i^*}$, respectively. With such a choice

the constraints of a new MIP program $\mathcal{P}3$ become:

$$\sum_{(i,i^*)\in\mathfrak{C}^*_{|\mathcal{I}}} (1 - a_{ii^*}) \geq c^* \tag{21}$$

$$0 \leq b_{ii^*} \leq a_{ii^*} \qquad a_{ii^*} - 1 + q_{ii^*} \leq b_{ii^*} \leq q_{ii^*} \tag{22}$$

$$\sum_{i^*} b_{ii^*} = \mathbf{x}_i + (r_i - s_i) \tag{23}$$

$$\sum_{i} b_{ii^*} = \mathbf{x}^*_{i^*} + (r_{i^*} - s_{i^*}) \tag{24}$$

$$b_{ii^*} = \mathbf{x}_{i|i^*}\mathbf{x}_{i^*} + \mathbf{x}_{i|i^*}(r_{i^*} - s_{i^*}) + \mathbf{x}_{i^*}(r_{i|i^*} - s_{i|i^*}) \tag{25}$$

$$\sum_{i,i^*\in I} b_{ii^*} = 1 \tag{26}$$

$$r_i \leq 1, \quad s_i \leq 1, \quad r_{i^*} \leq 1 \quad , \quad s_{i^*} \leq 1, \quad r_{i|i^*} \leq 1, \quad s_{i|i^*} \leq 1, \tag{27}$$

where the constraint (21) induces the binary variables a_{ii^*} to be 0 for the couples of indexes in $\mathfrak{C}^*_{|\mathcal{I}}$; constraints like (22) are set for all $i, i^* \in I$ and induce equalities $b_{ii^*} = a_{ii^*}q_{ii^*}$ that otherwise will not be linear; constraints like (23) and (24) are set for all the assessed marginal probabilities $\mathbf{x}_i$ and $\mathbf{x}^*_{i^*}$ and permit their correction through the slack variables; constraints like (25) are set for all assessed conditional probabilities $\mathbf{x}_{i|i^*}$ and constraint the joint distribution with corrected conditional and marginal values. Note that these last type of constraints are equivalent to set

$$b_{ii^*} = (\mathbf{x}_{i|i^*} + r_{i|i^*} - s_{i|i^*})(\mathbf{x}_{i^*} + r_{i^*} - s_{i^*}) \tag{28}$$

but without developing the cross products among the slack variables since they will constitute corrections of the joint distribution that, not being assessed, does not need any correction. This permits us to remain in a linear program.

The objective function to minimize is again the sum of the slack variables

$$\sum_{i,i^*} r_i + s_i + r_{i^*} + s_{i^*} + r_{i|i^*} + s_{i|i^*} \tag{29}$$

that obviously represents the L_1 distance between the assess probability values and the coherent ones.

Note that whenever the corrected assessment would present some marginal probability to be zero, all the new probabilities conditioned on such $X^*_{i^*}$ will result automatically coherent since, for the structure of the assessment, the various zero layers (for such a notion refer to [5]) will involve only one such conditioning event per time, so that the $P_{\mathcal{X}|X^*_{i^*}}(X_i)$ do not have any particular constraint to satisfy.

At the moment we have developed only the theoretical part of this section, leaving its practical application to future developments.

References

[1] M. Baioletti, A. Capotorti. Efficient L_1-Based Probability Assessments Correction: Algorithms and Applications to Belief Merging and Revision. in: *ISIPTA'15 Proc. of the 9th International Symposium on Imprecise Probability: Theories and Applications*. Pescara (IT), 37–46, ARACNE, 2015.

[2] M. Baioletti, A. Capotorti. An efficient probabilistic merging procedure applied to statistical matching. *Lecture Notes in Computer Science*, 10351 LNCS, 65–74, 2017.

[3] Baioletti, M., Capotorti, A., Tulipani, S., Vantaggi B. Simplification Rules for the Coherent Probability Assessment Problem. *Ann. of Math. and Artif. Intell.*, 35:11–28, 2002.

[4] Baioletti, M., Capotorti, A., Tulipani, S. An empirical complexity study for a 2CPA solver. In: *Modern Information Processing: From Theory to Applications*. B. Bouchon-Meunier, G. Coletti and R.R. Yager, Eds. 1–12, 2005.

[5] G. Coletti, R. Scozzafava. *Probabilistic Logic in a Coherent Setting*, Dordrecht: Kluwer, Series "Trends in Logic", 2002.

[6] Cozman, F.G., Fargonidi Ianni, L. Probabilistic Satisfiability and Coherence Checking through Integer Programming. Lecture Notes in Computer Science, 7958, 145–156, 2013.

[7] Cozman, F.G., Fargonidi Ianni, L. Probabilistic satisfiability and coherence checking through integer programming. International Journal of Approximate Reasoning, 58, 57–70, 2015.

[8] M. Di Zio, B. Vantaggi. Partial identification in statistical matching with misclassification, *International Journal of Approximate Reasoning*, 82: 227–241, 2017.

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

[10] D'Orazio, M., Di Zio, M., Scanu, M. Statistical Matching for Categorical Data: displaying uncertainty and using logical constraints. Journal of Official Statistics 22, 137–157, 2006.

[11] B. Vantaggi. Statistical matching of multiple sources: A look through coherence. *International Journal of Approximate Reasoning*, 49(3): 701–711, 2008.