Accounting for Non-ignorable Sampling and Nonresponse In Statistical Matching

Daniela Marella[a] and Danny Pfeffermann[b,c]

Summary

Data for statistical analysis is often available from different samples, with each sample containing measurements on only some of the variables of interest. Statistical matching attempts to generate a fused database containing matched measurements on all the target variables. In this article, we consider the use of statistical matching when the samples are drawn by informative sampling designs and are subject to not missing at random nonresponse. The problem with ignoring the sampling process and nonresponse is that the distribution of the data observed for the responding units can be very different from the distribution holding for the population data, which may distort the inference process and result in a matched database that misrepresents the joint distribution in the population. Our proposed methodology employs the empirical likelihood approach and is shown to perform well in a simulation experiment and when applied to real sample data.

Keywords: empirical likelihood; fusion; IPF algorithm; matching uncertainty; NMAR nonresponse, sample and respondents distributions.

[a]*Dipartimento di Scienze Sociali ed Economiche, Sapienza Università di Roma, Roma, Italy;*

[b] Department of *Statistics, Hebrew University of Jerusalem, Israel;*

[c]*Southampton Statistical Sciences Research Institute* (S3RI)*, University of Southampton,UK.*

Corresponding author: Daniela Marella; daniela.marella@uniroma1.it

1. Introduction

Statistical matching has become popular in recent years. Information on a set of variables of interest is often available in different micro databases, with each database containing only some of the variables, but with no joint observations on all the variables. For example, in Italy reliable information on households income is provided by SHIW (Survey on Household Income and Wealth) conducted by Banca d'Italia. On the other hand, information on consumption expenses is provided by the HBS (Household Budget Survey), run by ISTAT (Italian National Institute of Statistics). Cf. Conti *et al.* (2017). This constitutes a serious problem since household data on income and expenditure are used by policy makers for analyzing the impact of policy strategies. Statistical matching attempts to combine the data obtained from different, non-overlapping samples, drawn from the same target population. At a micro level, the main objective is to construct a synthetic (fused) data set, with joint observations on all the variables of interest. At a macro level, the main objective is the estimation of the joint population distribution of all the variables of interest.

Let A and B be two independent samples of size n_A and n_B respectively, selected from a population of N independent and identically distributed (*i.i.d.*) records, generated from some joint probability (density) function (*pdf*), $f_p(x,y,z;\theta)$ of variables (X,Y,Z) indexed by a vector parameter θ, where p signifies the population model (the model holding for the population values). We suppose that the population is large, such that the samples A and B can be assumed to have no units in common. The statistical matching problem is that (X,Y,Z) are not jointly observed in the two samples: only (X,Y) are observed for the units in sample A, and only (X,Z) are observed for the units in sample B; see Rässler (2002) and D'Orazio *et al*. (2006b). Thus, the units in A have missing Z values while the units in B have missing Y values. Because of the lack of joint information on all the three variables, the joint *pdf* $f_p(x,y,z;\theta)$ is not directly identifiable, unless under strong assumptions, which are generally hard to confirm. Several alternative approaches have been proposed in the literature to overcome the identification problem. The first (common) approach assumes conditional independence (CIA) between Y and Z given X, see, *e.g.*, Okner (1972). A second approach assumes the existence of external information. Relevant external information may be available in one of the following forms: **i)** a sample C with joint observations on (X,Y,Z); (Singh *et al.*, 1993); **(ii)** proxy variables for Y, Z as in Zhang (2015), where a range of statistical matching techniques are reviewed and developed for estimating the joint population *pdf* of categorical variables. Proxy variables, if sufficiently associated with Y or Z, can help studying the relationship between Y and Z and in particular, help verifying or refuting the CIA. Empirical results in Zhang (2015) demonstrate that the use of proxy variables not only reduces the uncertainty associated with data fusion, but also provides more accurate estimates of the target joint distribution. Notice, however, that the CIA cannot be tested from the samples A and B alone, and external information is often not available. (As discussed and illustrated in subsequent sections, the CIA can be tested indirectly by use of the estimated respondents' distribution resulting from this assumption.)

A third approach proposed in the literature consists therefore of analyzing the uncertainty regarding the joint distribution of (X,Y,Z). Under this approach, several alternative models for the joint distribution of (X,Y,Z), compatible with the distributions of (X,Y) and (X,Z) in the samples A and B are considered, resulting in "uncertainty intervals" for the joint *pdf* of all the three variables, and the target estimators derived from them. See, *e.g.*, Moriarity and Scheuren (2001), Rässler (2002) and D'Orazio *et al.* (2006a). Uncertainty in statistical matching in a nonparametric setting is considered in Conti *et al.* (2015). Zhang and Chambers (2019) describe a general approach for inference based on incomplete 2X2 tables (including the case of statistical matching and nonresponse), when assumptions required for validating a likelihood-based approach cannot be supported by the available data. The authors develop the concept of corroboration, as a measure of the statistical evidence in the observed data for the unknown parameter values, which is not based on likelihoods. For this, the authors compute intervals for each of the parameter values (rather than point estimates), without relying on any additional assumptions that can lead to pointwise identification of the joint distribution. The interval corresponding to a maximum corroboration value identifies the parameter value that is the hardest to refute based on the observed data.

In practice, the independence assumption between sample measurements pertaining to different units in the sample is itself questionable when dealing with sample survey data. Often, the sample selection employs complex sampling designs that involve different inclusion probabilities, which could be related to the survey variables of interest, known in the statistical literature as informative sampling. This can distort the independence assumption and result in a different distribution of the observed data from the distribution holding in the population from which the sample is drawn. See Pfeffermann and Sverchkov (2009) for discussion of the notion of informative sampling and review of methods to handle this problem.

Statistical matching of complex sample surveys is studied by Rubin (1986), Renssen (1998) and Conti *et al.* (2016). Marella and Pfeffermann (2019) considered statistical matching under informative sampling designs, assuming complete response. However, in practice, not all the sampled units respond, and as well known, the response rates are steadily decreasing all over the world. Most of the approaches dealing with nonresponse assume that the missing data are missing at random (MAR, Little and Rubin, 1987). By this assumption, the response probabilities do not depend on the unobserved data, after conditioning on the observed data. In reality, the MAR assumption is often violated and when the response probabilities are correlated with the target outcomes even after conditioning on the observed data (often, the model covariates), the missing data are not missing at random (NMAR nonresponse).

In this article, we consider the case where the sampling designs used to select the samples A and B are informative, and the nonresponse in the two samples is NMAR. As studied theoretically and illustrated in many articles, even informative sampling with complete response, or noninformative sampling but with NMAR nonresponse, already results in a different joint distribution of the observed data in the sample from the distribution of the same variables in the population from which the sample is taken. See, *e.g.*, Pfeffermann (2017). Not surprisingly and as illustrated later, ignoring the sampling and response processes in statistical matching (the focus of the present article) can result in severely biased estimators and a misrepresentative fused dataset. To the best of our knowledge, no other article has been published so far, considering the dual effects of informative sampling and NMAR nonresponse in statistical matching. Our proposed methodology utilizes the empirical likelihood (EL) approach.

In Section 2 we define more formally the statistical framework under consideration. Section 3 develops the EL in the statistical matching context under informative sampling designs and NMAR nonresponse, assuming the CIA. The proposed approach combines the EL with a parametric model for the response probabilities. In Section 4 the CIA is dropped, the uncertainty in statistical matching is introduced and a procedure for choosing a population distribution from the class of plausible *pdf*s is described. Section 5 presents the results of a simulation study, aimed for assessing the performance of our proposed methodology. In Section 6 we apply the methodology to the SHIW and HBS samples mentioned in the introduction. Section 7 contains a brief summary.

2. Statistical matching under nonignorable sampling and nonresponse. Notation and assumptions

Consider a finite population of N units $\{i=1,...,N\}$. Associated with unit i are values of three study variables, (X,Y,Z). Suppose that the population values $D_p=\{(x_1,y_1,z_1),...,(x_N,y_N,z_N)\}$ are independent realizations from a distribution with *pdf* $f_p(x,y,z;\theta)$. Let $V_{p,A}$, $V_{p,B}$ be sets of population values of design variables used for selecting two nonoverlapping samples, A and B respectively. Some or all of the variables (X,Y,Z) might be included among the design variables. We assume that $D_p,V_{p,A},V_{p,B}$ are realizations of a random process, implying that the first order inclusion probabilities $\left\{\pi_{i,A},\,\pi_{i,B}\right\}$ may be viewed as random as well. Denote by $w_{i,A}=1/\pi_{i,A}$, $w_{i,B}=1/\pi_{i,B}$ the (base) sampling weights. We assume that under complete response, the data available to the analyst consist of the samples $A=(x_i,y_i,w_{i,A})$ of size n_A and $B=(x_i,z_i,w_{i,B})$ of size n_B, but not the population values of the design variables, which are known to the persons drawing the samples but generally not to the persons analysing the sample data. Following Marella and Pfeffermann (2019), we assume that the sampling designs for selecting the two samples are informative for the corresponding joint population distribution, in the sense that the sample selection probabilities are correlated with at least some of the variables (X,Y,Z), implying that even if all the three variables had been observed in the two samples, the joint sample *pdf* $f_S(x,y,z)$ of the sample data is different from the corresponding population *pdf*, $f_p(x,y,z)$, for $S=A,B$.

In this article, we assume that in addition to the use of informative sampling designs, the samples A and B are subject to not missing at random (NMAR) unit nonresponse, in the sense that the probability to respond depends on the study variables. The data available to the analyst consist therefore of the sets of responding units in $A(R_A)$ and $B(R_B)$, respectively. Consequently, the joint *pdf* of the observed data, $f_{R_S}(x,y,z)$, differs from the sample *pdf* $f_S(x,y,z)$ under complete response, and from the population *pdf* $f_p(x,y,z)$, $S=A,B$. Here, for convenience, we omit the parameters indexing the three distributions. Notice that while the sampling probabilities are generally known, and thus can be used to account for informative sampling, the response probabilities are practically unknown and need to be modelled.

Pfeffermann and Sikov (2011) review approaches proposed in the literature to deal with NMAR nonresponse.

Accounting for informative sampling but with complete response for statistical matching is considered in Marella and Pfeffermann (2019). The authors applied a parametric approach, basing the inference on the sample distribution, *i.e.*, the distribution holding for the observed sample data. However, as discussed in Pfeffermann and Landsman (2011), the maximization of *sample likelihoods* can be complicated numerically and result in unstable estimates, depending on the population model and the model assumed for the sample selection probabilities, given the observed data. For this reason, and in order to account also for NMAR nonresponse, we propose in Section 3 the use of the empirical likelihood (EL), which enables estimating the parameters governing the sampling and response models, without specifying the corresponding population model.

3. Statistical matching under nonignorable sampling and nonresponse by EL

In Section 3.1, the statistical framework under the EL approach is briefly described. The EL under informative sampling is introduced in Section 3.2 and extended to NMAR nonresponse in Section 3.4. The generation of a fused data set under the EL approach is described in Section 3.3.

3.1. Statistical framework under the empirical likelihood approach

The use of the EL for analyzing complex survey data has its origins in the pioneering article by Hartley and Rao (1968), where an estimator based on the multinomial function under simple random sampling is proposed. The use of EL gained increasing interest in general statistical contexts, following the work of Owen (1990, 1991, 2001, 2013). See also Qin and Lawless (1994) and the review article by Chen and Van Keilegom (2009). The EL combines the robustness of nonparametric methods with the efficiency of the likelihood approach. It is essentially the likelihood of the multinomial distribution employed by Hartley and Rao (1968), where the unknown parameters are the point masses assigned to the distinct sample values. Chen and Qin (1993) proposed an EL approach for using auxiliary information in simple random sampling without replacement. Chen and Sitter (1999) extended the method to unequal probability sampling, applying a “pseudo-empirical likelihood approach”. Wu (2004) used pseudo-empirical likelihood methods to combine information from two independent surveys, and obtained an estimator for a mean, which is asymptotically equivalent to a GREGtype estimator. The EL approach facilitates the use of calibration constraints. See Remark 4 for the form of the constraints and Chaudhuri *et al*. (2010) for details of the constrained estimation procedure and the asymptotic properties of the resulting estimators. Most importantly, the use of this approach does not require specifying the population model, and is thus more robust and often easier to implement.

In the present section, we assume the CIA but this assumption is dropped in the following sections. We also assume that X can take K distinct values with probabilities $p_k^X = \mathrm{P}(X = x_k), \sum_{k=1}^{K} p_k^X = 1$, while Y and Z are continuous. The "matching variable", X, measured in both samples, might be a stratification variable and/or a socio-demographic variable. Socio-demographic characteristics are often related to other variables of interest.

The basic idea of the EL approach is to approximate the population distribution by a multinomial model, which support is given by the empirical observations. Let $(x_i, y_i.z_i)$ define the values associated with unit i and denote by $p_i^X = \Pr(X = x_i)$, $p_i^{Y|X} = \Pr(Y = y_i \mid X = x_i)$, $p_i^{Z|X} = \Pr(Z = z_i \mid X = x_i)$, each with the support observed in the samples. Then, under the CIA, the joint population multinomial probability of unit i is given by $p_i^{XYZ} = p_i^X p_i^{Y|X} p_i^{Z|X}$. Finally, let $A_k = \{i \in A : x_i = x_k\}$ be the set of sampled units in A with $X = x_k$, such that for $i \in A_k$, $p_i^X = P(X = x_k) = p_k^X$, $k = 1,..,K$.

3.2. The empirical likelihood approach under informative sampling

In what follows we define the EL in the statistical matching context under informative sampling. Let I_i^A be the sample indicator taking the value 1 if unit i is drawn to the sample A and 0 otherwise. For $i \in A_k$ denote,

$$\tau_{i,A}^{XY} = P(I_i^A = 1 \mid x_i, y_i),\ p_i^{Y|X} = P(y_i \mid x_i),\ \tau_{i,A}^{X} = P(I_i^A = 1 \mid x_i) = \sum_{j \in A_k} \tau_{j,A}^{XY} p_j^{Y|X} = \tau_{k,A}^{X}. \qquad (3.1)$$

It follows that,

$$p_{i,A}^{Y|X} = P(y_i \mid x_i, I_i^A = 1) = \frac{P(I_i^A = 1 \mid x_i, y_i)}{P(I_i^A = 1 \mid x_i)} p_i^{Y|X} = \frac{\tau_{i,A}^{XY} p_i^{Y|X}}{\sum_{j \in A_k} \tau_{j,A}^{XY} p_j^{Y|X}}. \qquad (3.2)$$

Similarly, for $i \in A_k$

$$p_{k,A}^X = P(x_k \mid I_i^A = 1) = \frac{P(I_i^A = 1 \mid x_k)}{P(I_i^A = 1)} p_k^X = \frac{\tau_{k,A}^X p_k^X}{\sum_{j=1}^K \tau_{j,A}^X p_j^X} . \tag{3.3}$$

Under informative sampling, the observed outcomes are no longer representative of the population outcomes and the sample models (3.2), (3.3) are different from the corresponding population models $p_i^{Y|X}, p_k^X$. Nonetheless, as shown and illustrated in Pfeffermann *et al*. (1998), if the population values are independent under the population model (see beginning of Section 2), then under mild conditions they are also asymptotically independent under the sample model, when the sample size remains fixed but the population size increases. This permits approximating the sample likelihood by the product of the sample *pdf*s over the corresponding sample observations. Hence, for sufficiently large populations, the sample EL (ESL), based on the observed data in A is,

$$ESL_{Obs}^A = \prod_{k=1}^K (p_{k,A}^X)^{n_{k,A}^X} \prod_{i \in A_k} p_{i,A}^{Y|X} , \tag{3.4}$$

where $n_{k,A}^X$ is the size of A_k. An analogous expression to (3.4) holds for the ESL based on the observed data in B. Hence, the ESL based on the sample $A \cup B$ is,

$$ESL_{Obs}^{A \cup B} = \left(\prod_{i \in A} p_{i,A}^X p_{i,A}^{Y|X} \right) \left(\prod_{i \in B} p_{i,B}^X p_{i,B}^{Z|X} \right) = \prod_{k=1}^K (p_{k,A}^X)^{n_{k,A}^X} \prod_{i \in A_k} p_{i,A}^{Y|X} \prod_{k=1}^K (p_{k,B}^X)^{n_{k,B}^X} \prod_{i \in B_k} p_{i,B}^{Z|X} , \tag{3.5}$$

where $p_{i,B}^{Z|X} = P(z_i \mid x_i, I_i^B = 1)$. By (3.2), (3.3) (with analogue expressions for the sample *B*), and (3.5), the log-likelihood based on $A \cup B$ is,

$$\begin{aligned} \log(ESL_{Obs}^{A \cup B}) = {} & \sum_{i \in A_k} \log(\tau_{i,A}^{XY} p_i^{Y|X}) - n_{k,A}^X \log\left(\sum_{i \in A_k} \tau_{i,A}^{XY} p_i^{Y|X} \right) + \sum_{k=1}^K n_{k,A}^X \log(\tau_{k,A}^X p_k^X) - \\ & + \sum_{k=1}^K n_{k,A}^X \log\left(\sum_{j=1}^K \tau_{j,A}^X p_j^X \right) + \sum_{i \in B_k} \log(\tau_{i,B}^{XZ} p_i^{Z|X}) - n_{k,B}^X \log\left(\sum_{i \in B_k} \tau_{i,B}^{XZ} p_i^{Z|X} \right) + \\ & + \sum_{k=1}^K n_{k,B}^X \log(\tau_{k,B}^X p_k^X) - \sum_{k=1}^K n_{k,B}^X \log\left(\sum_{j=1}^K \tau_{j,B}^X p_j^X \right). \end{aligned} \tag{3.6}$$

Notice that the sampling probabilities in A and B may depend on many unobserved variables and yet, by definition of the sample *pdf*, one only needs to model the probabilities $P(I_i^A = 1 \mid x_i, y_i)$ [$P(I_i^B = 1 \mid x_i, z_i)$]. Furthermore, following Pfeffermann and Sverchkov (1999), the probabilities $\tau_{i,A}^{XY} = P(I_i^A = 1 \mid x_i, y_i) = 1 / E_A(w_{i,A} \mid x_i, y_i)$ and $\tau_{i,B}^{XZ} = 1 / E_B(w_{i,B} \mid x_i, z_i)$ can be estimated outside the likelihood by regressing the sample weights $w_{i,A}\,(w_{i,B})$ against $(x_i\,, y_i)\,[(x_i\,, z_i)]$, using the observed data in A and B respectively, or non-parametrically, as considered in Feder and Pfeffermann (2019). The resulting estimates can then be inserted into the expressions for $\tau_{i,A}^{XY}$ and $\tau_{i,B}^{XZ}$, with $\tau_{k,A}^{X}$ and $\tau_{k,B}^{X}$ defined by (3.1). Moreover, as discussed and illustrated in Pfeffermann (2011), the resulting sample models can be tested. The unknown parameters in (3.6) are thus the probabilities $\{p_k^X\,,\, p_i^{Y|X}\,,\, p_i^{Z|X}\}$. In the statistical matching context, different approaches can be used for maximization of the likelihood. For convenience, we describe the approaches for the case where no variables with known sample values and corresponding population means exist, which as noted before can be used for calibration. When such variables exist, the calibration equations are imposed to constrain the maximization process. See Remark 4 below.

Remark 1. In practice, the covariates contained in the population model need not be the same as the covariates contained in the model of the conditional sample inclusion probabilities $P(I_i^A = 1 \,|\, x_i, y_i)$. However, to simplify the presentation, we assume for convenience that the same covariates appear in the two models or alternatively, that x_i defines the union of the two sets of covariates.

3.2.1. Estimating the unknown probabilities separately from the samples A and B

Noting that the likelihood (3.5) can be factorized into a likelihood based only on the sample A, and a likelihood based only on the sample B, the unknown probabilities $\{p_k^X\,,\, p_i^{Y|X}\,,\, p_i^{Z|X}\}$ can be estimated separately from the two samples. This implies two sets of estimates for the probabilities $\{p_k^X\}$, which need to be harmonized. See Renssen (1998) and below. The EL estimators of the unknown probabilities are obtained by maximizing the loglikelihood (3.6), subject to the constraints,

$$p_k^X \ge 0,\; p_i^{Y|X} \ge 0,\; p_i^{Z|X} \ge 0,\; \sum_{k=1}^{K} p_k^X = 1,\; \sum_{j \in A_k} p_j^{Y|X} = 1,\; \sum_{j \in B_k} p_j^{Z|X} = 1\,. \tag{3.7}$$

Following Kim (2009), Chaudhuri *et al*. (2010), and Marella and Pfeffermann (2019), the estimators are:

$$\hat{p}_{k,A}^{X} = [n_{k,A}^{X}(\tau_{k,A}^{X})^{-1}] / \sum_{j=1}^{K}[n_{j,A}^{X}(\tau_{j,A}^{X})^{-1}],\ \hat{p}_{k,B}^{X} = [n_{k,B}^{X}(\tau_{k,B}^{X})^{-1}] / \sum_{j=1}^{K}[n_{j,B}^{X}(\tau_{j,B}^{X})^{-1}]$$
$$\hat{p}_{i}^{Y|X} = (\tau_{i,A}^{XY})^{-1} / \sum_{j \in A_k}(\tau_{j,A}^{XY})^{-1},\ \hat{p}_{i}^{Z|X} = (\tau_{i,B}^{XZ})^{-1} / \sum_{j \in B_k}(\tau_{j,B}^{XZ})^{-1}, \tag{3.8}$$

where $\hat{p}_{k.A}^{X}, \hat{p}_{k,B}^{X}$ are the estimates of p_k^X obtained from the samples A and B, respectively. Harmonization of the estimates $\hat{p}_{k,A}^{X}$, $\hat{p}_{k.B}^{X}$ into a unique estimate p_k^X can be achieved by use of a linear combination of the two estimates, *i.e.*,

$$\hat{p}_{k}^{X} = \lambda \hat{p}_{k,A}^{X} + (1-\lambda)\hat{p}_{k,B}^{X},\ \lambda \in [0,1]. \tag{3.9}$$

A plausible choice is $\lambda = n_A / (n_A + n_B)$. Alternatively, one may choose the value λ minimizing the variance of (3.9). To this end, variance estimates of $\hat{p}_{k,A}^{X}, \hat{p}_{k,B}^{X}$ can be computed by resampling methods for finite populations, as proposed by Conti *et al.* (2020). The methods use a two-stage procedure. In the first stage, a pseudo-population, which can be viewed as a prediction of the target finite population is constructed, using the sampling weights. In the second stage, samples are drawn from the pseudo-population using the same sampling designs used for drawing the original samples. The procedure is also applicable for the case of informative sampling designs.

Remark 2. Another approach consists of replacing $p_{k,A}^{X}$ and $p_{k,B}^{X}$ in (3.5) by $\lambda p_{k,A}^{X} + (1-\lambda) p_{k,B}^{X}$, and maximizing the sample EL with respect to $\{p_k^X,\ p_i^{Y|X},\ p_i^{Z|X}\}$ and λ.

Remark 3. Chen and Sitter (1999) consider a pseudo empirical likelihood (PEL) approach, which in the context of statistical matching implies the following likelihood,

$$EL_{PEL}^{A\cup B} = \prod_{k=1}^{K}(p_k^X)^{\left(\sum_{i\in A_k} w_{i,A}\right)} \prod_{i\in A_k}\left(p_i^{Y|X}\right)^{w_{i,A}} \prod_{k=1}^{K}(p_k^X)^{\left(\sum_{i\in B_k} w_{i,B}\right)} \prod_{i\in B_k}\left(p_i^{Z|X}\right)^{w_{i,B}}. \tag{3.10}$$

Notice that in (3.10), the two samples are not considered separately. It follows from Chen and Sitter (1999) that in the absence of calibration constraints, the estimates maximizing the likelihood (3.10) are,

$$\hat{p}_{k,PEL}^{X} = \frac{\sum_{j\in A_k} w_{j,A} + \sum_{j\in B_k} w_{j,B}}{\sum_{j=1}^{n_A} w_{j,A} + \sum_{j=1}^{n_B} w_{j,B}},\ \ \hat{p}_{i,PEL}^{Y|X} = w_{i,A} / \sum_{j\in A_k} w_{j,A},\ \hat{p}_{i,PEL}^{Z|X} = w_{i,B} / \sum_{j\in B_k} w_{j,B}. \tag{3.11}$$

The PEL estimators of $\{ p_i^{Y|X},\ p_i^{Z|X} \}$ in (3.11) have the same form as in (3.8), but with the base sampling weights $\{w_{j,A}, w_{j,B}\}$, instead of the weights $\{(\tau_{j,A}^{XY})^{-1}, (\tau_{j,B}^{XZ})^{-1}\}$. The basic difference between the two approaches is that in Chen and Sitter (1999), the likelihood is with respect to the population distribution, while the likelihood in (3.5) is with respect to the sample distribution.

3.2.2. File concatenation for estimation of the probabilities p_k^X

Rubin (1986) proposed to estimate the population probability distribution of X by computing concatenated weights for the sample $A \cup B$ as follows:

$$p_{k,A\cup B}^X = P(x_k \mid I_i^A = 1 \cup I_i^B = 1) = \frac{(\tau_{k,A}^X + \tau_{k,B}^X)\, p_k^X}{\sum_{j=1}^K \tau_{j,A}^X p_j^X + \sum_{j=1}^K \tau_{j,B}^X p_j^X} = \frac{\tau_{k,A\cup B}^X p_k^X}{\sum_{j=1}^K \tau_{j,A\cup B}^X p_j^X}, \tag{3.12}$$

where $\tau_{k,A\cup B}^X = \tau_{k,A}^X + \tau_{k,B}^X$. The basic assumption underlying (3.12) is that the probability of a unit to be drawn to both samples is negligible, such that $P[(I_i^A = 1 \cap I_i^B = 1) \mid x_k] \cong 0$. This is generally true when the two samples are independent, with small sampling fractions. Define, $n_{k,A\cup B}^X = n_{k,A}^X + n_{k,B}^X$. With this notation,

$$ESL_{Obs}^{AUB} = \prod_{k=1}^K (p_{k,A\cup B}^X)^{n_{k,A\cup B}^X} \prod_{i\in A_k} p_{i,A}^{Y|X} \prod_{i\in B_k} p_{i,B}^{Z|X}. \tag{3.13}$$

The ESL (3.13) is maximized under the constraints (3.7), yielding the estimators,

$$\begin{aligned} &\hat{p}_k^X = [n_{k,A\cup B}^X (\tau_{k,A\cup B}^X)^{-1}] / \sum_{j=1}^K [n_{j,A\cup B}^X (\tau_{j,A\cup B}^X)^{-1}],\ \hat{p}_i^{Y|X} = (\tau_{i,A}^{XY})^{-1} / \sum_{j\in A_k} (\tau_{j,A}^{XY})^{-1}, \\ &\hat{p}_i^{Z|X} = (\tau_{i,B}^{XZ})^{-1} / \sum_{j\in B_k} (\tau_{j,B}^{XZ})^{-1}. \end{aligned} \tag{3.14}$$

Remark 4. When population means of variables measured in the sample A and/or in the sample B are known, they can be added to the constraints of the ESL. The following calibration constraints may be added, depending on data availability: $\sum_{k=1}^K p_k^X x_k = \mu_X$, $\sum_{k=1}^K p_k^X \sum_{i\in A_k} p_i^{Y|X} y_i = \mu_Y$, $\sum_{k=1}^K p_k^X \sum_{i\in B_k} p_i^{Z|X} z_i = \mu_Z$, where μ_X, μ_Y, μ_Z are the population means of X, Y, Z, respectively. In the simulation study of Section 5 and the application to real sample data in Section 6, we added the constraint $\sum_{k=1}^{K} p_k^X x_k = \mu_X$.

3.3. Generation of a fused data set

Once the probabilities $\{p_k^X,\ p_i^{Y|X},\ p_i^{Z|X}\}$ governing the population multinomial model have been estimated, a fused data set with joint observations (x, y, z) are constructed as follows:

(i) Generate $\tilde{n}$ observations taking values $(x_1, x_2, ..., x_K)$ with probabilities $(\hat{p}_1^X, \hat{p}_2^X, ..., \hat{p}_K^X)$;

(ii) For $i = 1, ..., \tilde{n}$ and $k = 1, ..., K$, draw at random a value $\tilde{y}_i$ from the estimated probability function $\hat{p}_i^{Y|X}$, taking the values $(y_1^k, y_2^k, ..., y_{n_{k,A}^X}^k)$ with probabilities $(\hat{p}_1^{Y|x_k}, \hat{p}_2^{Y|x_k}, ..., \hat{p}_{n_{k,A}^X}^{Y|x_k})$, where $n_{k,A}^X = \#\{i \in A : x_i = x_k\}$.

(iii) Apply a similar procedure for drawing values $\tilde{z}_i$ from the estimated probability function $\hat{p}_i^{Z|X}$.

The consistency of the estimators of the model parameters guarantees that for sufficiently large sample sizes n_A and n_B, the fused data set can be considered as being generated from the joint population *pdf*.

Remark 5. It is not correct to only impute the missing z-values in A, and only the missing y-values in B, and then consider the union of the two samples as the fused data set. This is so because although in the sample A the missing z-values could be imputed using the estimated probabilities $\hat{p}_i^{Z|X}$, under informative sampling the observed (x, y) values in A are not representative of the population (x, y) values. The same holds for the sample B.

3.4. Use of the EL under nonignorable sampling and nonresponse

In what follows we assume that additionally to informative sampling, the samples A and B are subject to NMAR nonresponse, by which the response probabilities depend in some stochastic way on the study variables of interest. Let R_i^A define the response indicator, taking the value 1 if sample unit $i \in A$ responds and 0 otherwise. Let R_A denote the set of responding units in A and r_A, the size of R_A. The response process is assumed to be independent between units. This way, the set of respondents can be viewed as the result of a two phase sampling process: **(i)** a sample A is selected from the finite population with known inclusion probabilities $\pi_{i,A}$; **(ii)** the response set R_A is selected from A with unknown response probabilities $P(R_i^A = 1 | I_i^A = 1)$. Let $\rho_{i,A}^X = P(R_i^A = 1 | x_i, I_i^A = 1)$. By Bayes rule, for $i \in A_k$

$$p_{k,R_A}^X = P(x_k \mid I_i^A = 1, R_i^A = 1) = \frac{P(R_i^A = 1 | x_k, I_i^A = 1)}{P(R_i^A = 1 | I_i^A = 1)} \ p_{k,A}^X = \frac{\tau_{k,A}^X \rho_{k,A}^X p_k^X}{\sum_{j=1}^{K} \tau_{j,A}^X \rho_{j,A}^X p_j^X}, \tag{3.15}$$

$$p_{i,R_A}^{Y|X} = P(y_i \mid x_k, I_i^A = 1, R_i^A = 1) = \frac{P(R_i^A = 1 | x_k, y_i, I_i^A = 1)}{P(R_i^A = 1 | x_k, I_i^A = 1)} \ p_{i,A}^{Y|X} = \frac{\tau_{i,A}^{XY} \rho_{i,A}^{XY} p_i^{Y|X}}{\sum_{i \in R_{A,k}} \tau_{i,A}^{XY} \rho_{i,A}^{XY} p_i^{Y|X}}, \tag{3.16}$$

where $\tau_{k,A}^X$ and $\tau_{i,A}^{XY}$ are defined in (3.1), $R_{A,k} = \{i \in R_A : x_i = x_k\}$ defines the group of respondents in A with $X = x_k$ of size $r_{k,A}^X$ and

$$\rho_{k,A}^X = P(R_i^A = 1 | x_k, I_i^A = 1) = E_A(R_i^A \mid x_k, I_i^A = 1) = \sum_{i \in R_{A,k}} \rho_{i,A}^{XY} p_{i,A}^{Y|X} , \tag{3.17}$$

$$\rho_{i,A}^{XY} = P(R_i^A = 1 | x_k, y_i, I_i^A = 1) = E_A(R_i^A \mid x_k, y_i, I_i^A = 1) . \tag{3.18}$$

In (3.15), the sample model $p_{i,A}^{Y|X}$ and the model assumed for the response probabilities define the model holding for the outcomes of the responding units. Notice that unless $P(R_i^A = 1 | x_k, y_i, I_i^A = 1) \ = P(R_i^A = 1 | x_k, I_i^A = 1)$ for all (x_k, y_i), the model (3.16) is different from the sample model $p_{i,A}^{Y|X}$ defined by (3.2), which is different from the corresponding population model under informative sampling. Specifically, the respondents model is a function of the corresponding population model, the conditional expectations of the sampling weights, $\tau_{i,A}^{XY} = P(I_i^A = 1 | x_i, y_i) = 1 / E_A(w_{i,A} \mid x_i, y_i)$, and the response probabilities $\rho_{i,A}^{XY} = P(R_i^A = 1 | x_k, y_i, I_i^A = 1)$. Assuming that the response is independent of the sample selection, $E_A(w_{i,A} \mid x_i, y_i) = E_{R_A}(w_{i,A} \mid x_i, y_i)$, in which case the probabilities $P(I_i^A = 1 | x_i, y_i)$ can be estimated by regressing $w_{i,A}$ against (x_i, y_i), using the observed data in A, and similarly for the sample B. See Section 3.2. Clearly, if the response probabilities depend in some way on the sample selection, say, higher nonresponse rates for units with higher sampling probabilities, the expectations $E_A(w_{i,A} \mid x_i, y_i)$ need to be estimated in some more elaborated manner. See also the concluding remarks in Section 7.

Remark 6. Under MAR nonresponse, the response probability does not depend on the target outcome variable after accounting for the model covariates, such that in (3.16), $p_{i,R_A}^{Y|X} = p_{i,A}^{Y|X}$. However, a nonresponse bias may still exist if the probabilities $\{p_k^X\}$ are not estimated properly. Recall that the covariates are only assumed to be known for the responding units.

With straightforward modification of the notation, similar expressions to (3.15)-(3.18) are obtained for the model holding for the responding units in B. Thus, the *empirical respondents' likelihood* (ERL) for the sample $A \cup B$ is given by,

$$ERL_{Obs}^{A\cup B} = \prod_{k=1}^{K} (p_{k,R_A}^X)^{r_{k,A}^X} \prod_{i \in R_{A,k}} p_{i,R_A}^{Y|X} \prod_{k=1}^{K} (p_{k,R_B}^X)^{r_{k,B}^X} \prod_{i \in R_{B,k}} p_{i,R_B}^{Z|X} . \tag{3.19}$$

Remark 7. The likelihood (3.19) only depends on the observed data for the responding units.

The response probabilities in (3.15)-(3.16), defining the probabilities in (3.19) are unknown and need to be estimated from the available data. Since no "response weights" are known, parametric models for the response probabilities in the two samples need to be postulated. For example,

$$P(R_i^A = 1 | x_i, y_i, I_i^A = 1) = g_A(\gamma_{0,A} + \gamma_{x,A} x_i + \gamma_{y,A} y_i), \tag{3.20}$$

$$P(R_i^B = 1 | x_i, z_i, I_i^B = 1) = g_B(\gamma_{0,B} + \gamma_{x,B} x_i + \gamma_{z,B} z_i), \tag{3.21}$$

for some functions g_A, g_B, with unknown parameters $\gamma_A = (\gamma_{0,A}, \gamma_{x,A}, \gamma_{y,A})$, $\gamma_B = (\gamma_{0,B}, \gamma_{x,B}, \gamma_{z,B})$. Here again, we assume for convenience that the response probabilities depend on the same covariates as in the sample model. See Remark 1. Modelling the response probabilities by the logit or probit functions is common, but notice that in our case the probabilities depend also on the study variables, which is different from the familiar "propensity scores" approach, under which the response probabilities only depend on the observed covariates, which are in common use under MAR nonresponse. The unknown vector parameters, γ_A, γ_B, indexing the response models in the two samples are then estimated as part of the maximization of the likelihood. Thus, one needs to maximize the likelihood (3.19) with respect to a larger set of parameters $[\{p_k^X, p_i^{Y|X}, p_i^{Z|X}\}, \gamma_A, \gamma_B]$, satisfying the constraints,

$$p_k^X \geq 0, p_i^{Y|X} \geq 0, p_i^{Z|X} \geq 0, \sum_{k=1}^{K} p_k^X = 1, \sum_{j \in R_{A,k}} p_j^{Y|X} = 1, \sum_{j \in R_{B,k}} p_j^{Z|X} = 1 . \tag{3.22}$$

for all k and i. Notice the difference from the constraints in (3.7) under full response.

For subsequent inference in the statistical matching context, one only needs estimates of the probabilities $\{p_k^X,\ p_i^{Y|X},\ p_i^{Z|X}\}$, suggesting considering the coefficients γ_A, γ_B as nuisance parameters. In order to write the likelihood (3.19) as only a function of the three sets of probabilities, we adopt the profile likelihood approach. Suppose that the three sets of probabilities are "known". (In practice, we use some initial estimates, see Remark 8 below). The profile likelihood function is defined as $G(\gamma_A,\gamma_B) = ERL_{Obs}^{A\cup B}(\ \gamma_A,\gamma_B \mid p_k^X, p_i^{Y|X}, p_i^{Z|X})$ and it is maximized with respect to (γ_A, γ_B), yielding the estimators,

$$(\hat{\gamma}_A,\hat{\gamma}_B) = \underset{(\gamma_A,\gamma_B)}{\arg\max} \quad ERL_{Obs}^{A\cup B}(\ \gamma_A,\gamma_B \mid p_k^X, p_i^{Y|X}, p_i^{Z|X})\,. \tag{3.23}$$

Next we substitute the estimates (3.23) into the likelihood (3.19) and maximize the resulting likelihood with respect to the unknown sets of probabilities, yielding

$$(\hat{p}_k^X, \hat{p}_i^{Y|X}, \hat{p}_i^{Z|X}) = \underset{(p_k^X, p_i^{Y|X}, p_i^{Z|X})}{\arg\max} ERL_{Obs}^{A\cup B}(\ p_k^X, p_i^{Y|X}, p_i^{Z|X}, \hat{\gamma}_A, \hat{\gamma}_B)\,. \tag{3.24}$$

This completes the first iteration in the estimation process. In the second iteration, we consider the estimates in (3.24) as "known", re-estimate the parameters (γ_A, γ_B), and then the unknown probabilities. The iterations continue until convergence. See Feder and Pfeffermann (2019) for conditions guaranteeing the convergence of the maximization process.

As noted before, the model for the response probabilities can be tested by testing the estimated models, $\hat{p}_{i,R_A}^{Y|X}$ and $\hat{p}_{i,R_B}^{Z|X}$ for the observed data, using standard goodness of fit tests. See Pfeffermann and Landsman (2011) and Feder and Pfeffermann (2019) for examples of relevant test procedures. Once the probabilities of the population multinomial models have been estimated, a fused data set with observations (x, y, z) is constructed, following the procedure in Section 3.3.

Remark 8. In the simulation study (Section 5), initial estimates of $\{p_k^X,\ p_i^{Y|X},\ p_i^{Z|X}\}$ are computed by the relative frequency of the observed values in the samples A and B. For example, for $X = x_k$ and $Y = y_i$, the initial value of $p_i^{Y|X}$ is the ratio between the number of units in $R_{A,k}$ with $X = x_k$ and $Y = y_i$, and $r_{k,A}^X$. If Y is a continuous variable, all the observed values are different and the initial estimates are $1/r_{k,A}^X$. We maximized the ERL (3.19) by using the R function *emplik*. See Owen (2013) for related theory and further details.

Remark 9. One of the reviewers of the present article proposed an EM algorithm for maximization of the ERL. We hope to investigate the properties of this algorithm in the future. See also the concluding remarks in Section 7.

4. Uncertainty in statistical matching under informative sampling and NMAR nonresponse

So far, we assumed that the joint population *pdf* satisfies the CIA. Clearly, the CIA may not hold in practice and having no joint measurements for the variables of interest, disallows distinguishing between different plausible distributions. In Section 4.1 we drop the CIA and define instead a class of plausible joint *pdf*s for the outcome variables of interest. Some measures quantifying the size of the class are introduced. In Section 4.2, a procedure for choosing a *pdf* from the class of plausible *pdf*s is described.

4.1. Measuring uncertainty in statistical matching

In statistical matching, estimation of the joint *pdf* of (X,Y,Z) requires the estimation of **(i)** the marginal *pdf* of X and **(ii)** the joint conditional *pdf* of (Y,Z) given X. Denote by $F_p(y,z\,|\,x_k)$ the joint cumulative population distribution function (*cdf*) of (Y,Z) given $X=x_k$, and by $F_p(y\,|\,x_k)$, $G_p(z\,|\,x_k)$ the corresponding marginal *cdf*s.

Notice that unless under additional assumptions, the only valid statement regarding $F_p(y,z\,|\,x_k)$ is that it lies in the set Ω_p^k of all joint distributions having marginal *cdf*s $F_p(y\,|\,x_k)$, $G_p(z\,|\,x_k)$, *i.e.*, $\Omega_p^k=\{F_p(y,z\,|\,x_k): F_p(y,\infty\,|\,x_k)=F_p(y\,|\,x_k); F_p(\infty,z\,|\,x_k)=G_p(z\,|\,x_k)\}$. For known $F_p(y\,|\,x_k)$, $G_p(z\,|\,x_k)$, $L[F_p(y\,|\,x_k),G_p(z\,|\,x_k)]\le F_p(y,z\,|\,x_k)\le U[F_p(y\,|\,x_k),G_p(z\,|\,x_k)]$, where

$$U[F_p(y\,|\,x_k),G_p(z\,|\,x_k)]=\min[F_p(y\,|\,x_k),G_p(z\,|\,x_k)]\,, \tag{4.1}$$

$$L[F_p(y\,|\,x_k),G_p(z\,|\,x_k)]=\max[0,F_p(y\,|\,x_k)+G_p(z\,|\,x_k)-1]\,. \tag{4.2}$$

The bounds (4.1), (4.2) are the Fréchet bounds, see Nelsen (1999). A natural pointwise uncertainty measure is the length of the interval $\{L[...],U[...]\}$. For $X=x_k$, the measure is,

$$\Delta_p^k=\int_{\Re^2}\{U[F_p(y\,|\,x_k),G_p(z\,|\,x_k)]-L[F_p(y\,|\,x_k),G_p(z\,|\,x_k)]\}\,dF_p(y\,|\,x_k)dG_p(z\,|\,x_k)\,. \tag{4.3}$$

Weight functions different from $dF_p(y\,|\,x_k)dG_p(z\,|\,x_k)$ can be used instead. Our choice has a clear interpretation, with larger weights assigned to intervals with larger marginal densities. The measure in (4.3) is easily estimated from the sample data (Equation (4.5)).

An overall uncertainty measure is defined by the average of the conditional measures (4.3),

$$\Delta_p = \sum_{k=1}^{K} \Delta_p^k p_k^X \,. \tag{4.4}$$

As shown in Conti *et al.* (2012), the value $\Delta_p^k = 1/6$ of the conditional uncertainty measure (4.3) represents the maximum uncertainty when no external information beyond knowledge of the marginal *cdf*s $F_p(y \mid x_k)$ and $G_p(z \mid x_k)$ is available. Consequently, the maximum unconditional uncertainty measure (4.4) also equals 1/6. Denote, $\Upsilon_{k,R_A} = (y_1^k, y_2^k, ..., y_{r_{k,A}^X}^k)$, $\Gamma_{k,R_B} = (z_1^k, z_2^k, ..., z_{r_{k,B}^X}^k)$. The measure (4.3) can be estimated by averaging the $r_{k,A}^X r_{k,B}^X$ pointwise uncertainty measures,

$$\hat{\Delta}_p^k = \frac{1}{r_{k,A}^X r_{k,B}^X} \sum_{y \in \Upsilon_{k,R_A}} \sum_{z \in \Gamma_{k,R_B}} [U(\hat{F}_p(y \mid x_k), \hat{G}_p(z \mid x_k)) - L(\hat{F}_p(y \mid x_k), \hat{G}_p(z \mid x_k)] \,, \tag{4.5}$$

where $\hat{F}_p(y \mid x_k)$ and $\hat{G}_p(z \mid x_k)$ are the estimated *cdfs* of $F_p(y \mid x_k)$ and $G_p(z \mid x_k)$; $\hat{F}_p(y \mid x_k) = \sum_{i=1}^{r_{k,A}^X} \hat{p}_i^{Y|x_k} I(y_i^k \le y)$, $\hat{G}_p(z \mid x_k) = \sum_{i=1}^{r_{k,B}^X} \hat{p}_i^{Z|x_k} I(z_i^k \le z)$. The overall uncertainty measure (4.4) is estimated as,

$$\hat{\Delta}_p = \sum_{k=1}^{K} \hat{\Delta}_p^k \hat{p}_k^X \,. \tag{4.6}$$

The bounds (4.1), (4.2) can be narrowed when additional information is available. The reduction in uncertainty due to the use of external information is investigated in Conti *et al.* (2015, 2016), where conditionally on $X = x_k$, constraints of the form $a_k \le c_k(y,z) \le b_k$ with $c_k(y,z)$ defining a monotone function of y (z) for each z (y), are added. The class of plausible *pdf*s is now,

$$\Omega_{p,c}^k = \{F_p(y,z \mid x_k) : F_p(y,\infty \mid x_k) = F_p(y \mid x_k), F_p(\infty,z \mid x_k) = G_p(z \mid x_k),\ a_k \le c_k(y,z) \le b_k\} \,. \tag{4.7}$$

Hereafter, each bivariate *pdf* in the class (4.7) is referred to as a *plausible matching pdf* for (Y,Z), conditionally on $X = x_k$. For example, Okner (1972) imposed the constraint $Y \le Z$. With this constraint, the Fréchet bounds (4.1)-(4.2) become (see Conti *et al.* 2015),

$$U_c[F_p(y \mid x_k), G_p(z \mid x_k)] = \min[F_p(y \mid x_k), F_p(z \mid x_k), G_p(z \mid x_k)] \tag{4.8}$$

$$L_c[F_p(y \mid x_k), G_p(z \mid x_k)] = \max[0, F_p(y \mid x_k) + G_p(z \mid x_k) - 1, \\ \min(F_p(y \mid x_k), F_p(z \mid x_k)) + G_p(z \mid x_k) - 1] \quad . \tag{4.9}$$

Notice the difference from (4.1) and (4.2), when no additional information is available. The corresponding uncertainty measures, $\Delta_{p,c}^{k}, \Delta_{p,c}$, are defined similarly to (4.3), (4.4) but with respect to the bounds (4.8), (4.9). By choosing a *matching distribution* from the class (4.7), the uncertainty measure $\Delta_{p,c}$ provides an upper bound for the matching error. The statistical matching problem consists therefore of choosing a *matching distribution* from the class (4.7).

4.2. Choosing a matching distribution

Conti *et al.* (2016) proposed a procedure for choosing a *pdf* in the class (4.7), based on Iterative Proportional Fitting (IPF, Bishop *et al.*1975). The procedure consists of the following steps:

Step 1: Discretize Y and Z by grouping their ascending values in pre-defined classes. Conditionally on $X = x_k$, the range of Y is divided into h_k adjacent intervals $I_1^{Y|x_k}, .., I_h^{Y|x_k}, .., I_{h_k}^{Y|x_k}$, where $I_h^{Y|x_k} = [y_{h-1}, y_h]$, $h = 1,.., h_k$ with $y_0 = min\, y_i$, $y_h = max\, y_i$. Similar notation applies to the variable Z; $I_g^{Z|x_k} = [z_{g-1}, z_g]$ for $g = 1,.., g_k$. For $X = x_k$, denote by $Y_{d,k}$ ($Z_{d,k}$) the discretized variable corresponding to Y (Z), taking h_k (g_k) values defined by the midpoints $y_{d,h}$ ($z_{d,g}$) of each interval. Let $\{C^k\}$ be the contingency table defined by the $h_k g_k$ values $\Upsilon^{YZ|x_k} = [(y_{d,1}, z_{d,1}),..,(y_{d,h}, z_{d,g}),..,(y_{d,h_k}, z_{d,g_k})]$, with cell probabilities $(p_{11}^{Y_{d,k} Z_{d,k}|x_k},.., p_{hg}^{Y_{d,k} Z_{d,k}|x_k},..., p_{h_k g_k}^{Y_{d,k} Z_{d,k}|x_k})$. Initial values $\{p_{hg}^{0, Y_{d,k} Z_{d,k}|x_k}\}$ of the cell probabilities when applying the IPF are defined in Step 3 below. Note that a separate contingency table $\{C^k\}$ is defined for each value x_k. As also explained in Step 3, the constraint $a_k \le c_k(y,z) \le b_k$ on the support of $(Y,Z) \mid x_k$ is applied to the values $(Y_{d,k}, Z_{d,k})$, resulting in cells with structural zeroes.

Step 2: For $X = x_k$, the marginal probabilities $p_{h.}^{Y_{d,k}|x_k}$, $p_{.g}^{Z_{d,k}|x_k}$ in $\{C^k\}$, *i.e.*, the probabilities that $Y_{d,k}$ and $Z_{d,k}$ take the values $y_{d,h}$, $z_{d,g}$, are estimated as, $\hat{p}_{h.}^{Y_{d,k}|x_k} = \sum_{i=1}^{r_{k,A}^X} \hat{p}_i^{Y|x_k} I(y_i^k \in I_h^{Y|x_k})$, $\hat{p}_{.g}^{Z_{d,k}|x_k} = \sum_{i=1}^{r_{k,B}^X} \hat{p}_i^{Z|x_k} I(z_i^k \in I_g^{Z|x_k})$, where $\hat{p}_i^{Y|x_k}, \hat{p}_i^{Z|x_k}$ are the MLE of the ERL (3.19).

Step 3: Once the contingency table $\{C^k\}$ has been defined, the midpoints $(y_{d,h}, z_{d,g})$ are checked to identify cells in $\{C^k\}$, which do not satisfy the constraint $a_k \le c_k(y_{d,h}, z_{d,g}) \le b_k$. These cells define structural zeroes in $\{C^k\}$. The IPF initial cell probabilities are defined as, $p_{hg}^{0,Y_{d,k}Z_{d,k}|x_k} = \delta_{hg} \hat{p}_{h.}^{Y_{d,k}|x_k} \hat{p}_{.g}^{Z_{d,k}|x_k}$, where $\delta_{hg} = 1$ for cells not containing structural zeroes and $\delta_{hg} = 0$ otherwise.

A fused data set for (X,Y,Z) is constructed from the estimated matching distribution obtained at the end of the iterations, as follows: **(i)** Generate $\tilde{n}$ observations $\tilde{x}_i$ from the estimated distribution of X, taking values $(x_1, x_2, ..., x_K)$ with probabilities $(\hat{p}_1^X, \hat{p}_2^X, ..., \hat{p}_K^X)$. Let $\tilde{n}_k^X$ be the number of observations with $\tilde{x}_i = x_k$; **(ii)** For each observation x_i, $i = 1,..,\tilde{n}_k^X$, draw independently $\tilde{n}_k^X$ pairs $[(y_{d,1}, z_{d,1}),..,(y_{d,h}, z_{d,g}),..,(y_{d,h_k}, z_{d,g_k})]$ with cell probabilities $(\hat{p}_{11}^{Y_{d,k}Z_{d,k}|x_k},.., \hat{p}_{hg}^{Y_{d,k}Z_{d,k}|x_k},..., \hat{p}_{h_k g_k}^{Y_{d,k}Z_{d,k}|x_k})$, computed by the IPF algorithm.

5. Simulation Study

5.1. Description of simulation experiment

In order to evaluate the performance of our proposed methodology, we performed a simulation experiment, consisting of the following steps:

*Step*1. Generate a population of $N = 10,000$ values x_i, taking the values $k = 1,2,3,4$ with probabilities $p^X = (p_1^X, p_2^X, p_3^X, p_4^X) = (0.4, 0.1, 0.3, 0.2)$. For each x_i, generate independently values y_i and z_i from the following distributions: **(i)** $y_i \mid x_i$ is normal with parameters $\theta_{Y|X} = (\beta_0 + \beta_1 x_i, \sigma_{Y|X}^2)$; $\beta_0 = 0.5$, $\beta_1 = 2$, $\sigma_{Y|X} = 4$, (ii) $z_i \mid x_i$ is normal with parameters $\theta_{Z|X} = (\alpha_0 + \alpha_1 x_i, \sigma_{Z|X}^2)$; $\alpha_0 = 2$, $\alpha_1 = 2$, $\sigma_{Z|X} = 4$.

Thus, the CIA holds in the population and $cor_{YZ}^{CIA} = cor_{XY} cor_{XZ} = 0.27$.

Remark 10. In Section 5.3 and in the application in Section 6 with real sample data, we no longer assume the CIA and illustrate the theory of Section 4.

Step 2. Draw independently samples A and B from the population generated in *Step* 1 by use of Poisson sampling with expected sample sizes $E(n_A) = E(n_B) = 3,000$ and selection probabilities,

$$\pi_{i,A} = n_A \frac{\exp(\kappa_{x,A} x_i + \kappa_{y,A} y_i)}{\sum_{j=1}^{N} \exp(\kappa_{x,A} x_j + \kappa_{y,A} y_j)}; \quad \pi_{i,B} = n_B \frac{\exp(\kappa_{x,B} x_i + \kappa_{z,B} z_i)}{\sum_{j=1}^{N} \exp(\kappa_{x,B} x_j + \kappa_{z,B} z_j)}, \tag{5.1}$$

where $\kappa_A = (\kappa_{x,A}, \kappa_{y,A})$ and $\kappa_B = (\kappa_{x,B}, \kappa_{z,B})$ denote the sampling model coefficients (specified later). Notice that for $\kappa_{y,A} \neq 0$, $\kappa_{z,B} \neq 0$, the two sampling designs are informative.

Step 3. Generate the samples of responding units in the two samples with response probabilities,

$$\rho_{i,A}^{XY}(\gamma_A) = logit^{-1}(\gamma_{x,A} x_i + \gamma_{y,A} y_i); \rho_{i,B}^{XZ}(\gamma_B) = logit^{-1}(\gamma_{x,B} x_i + \gamma_{z,B} z_i), \tag{5.2}$$

where $\gamma_A = (\gamma_{x,A}, \gamma_{y,A})$, $\gamma_B = (\gamma_{x,B}, \gamma_{z,B})$ govern the response models acting in the samples A and B respectively (specified later). Clearly, the nonresponse is NMAR.

In what follows we assume knowledge of the mean $\mu_X = \sum_{k=1}^{4} p_k^X k$ of X, hereafter the *calibration constraint*, abbreviated C-C. See Remark 4.

The probabilities $\{p_k^X, p_i^{Y|X}, p_i^{Z|X}\}$ are estimated under three scenarios:

Scenario 1: All the sampled units respond and the sampling designs used for selecting the samples A and B are ignored. The ESL is in this case, $ESL_{Obs}^{A\cup B} = \prod_{k=1}^{K} (p_k^X)^{n_{k,A}^X + n_{k,B}^X} \prod_{i\in A_k} p_i^{Y|X} \prod_{i\in B_k} p_i^{Z|X}$ and it is maximized under the constraints (3.7) and the C-C. The estimates of p_k^X obtained from the two samples are harmonized according to (3.9), with $\lambda = n_A/(n_A + n_B)$. Denote by $\{\hat{p}_{k,1}^X, \hat{p}_{i,1}^{Y|X}, \hat{p}_{i,1}^{Z|X}\}$ the estimated population *pdf*.

Scenario 2: All the sampled units respond, but the informative sampling designs are taken into account in the estimation process. The ESL (3.6) is maximized subject to the constraints (3.7) and the C-C. The expectations $E_A(w_{i,A} | x_i, y_i; \kappa_A)$ are estimated by regressing $w_{i,A}$ against (x_i, y_i), assuming the model, $E_A(w_{i,A} | x_i, y_i) = \exp\{ax + bx^2 + cy + dy^2\}$. A similar model is used for estimating the expectations $E_B(w_{i,B} | x_i, z_i)$. The use of these models guaranties positive expectations. The two estimates of p_k^X obtained from samples A and B are harmonized as under *Scenario* 1. We denote by $\{\hat{p}_{k,2}^X, \hat{p}_{i,2}^{Y|X}, \hat{p}_{i,2}^{Z|X}\}$ the estimated population *pdf* under this scenario.

Scenario 3: The sampled units respond with probabilities defined by (5.2) and we account for both the informative sampling designs and the NMAR nonresponse. For this, we maximized the ERL (3.19) with respect $\{p_k^X, p_i^{Y|X}, p_i^{Z|X}\}$, under the constraints (3.22) and the C-C. The response is independent of the sample selection, such that $E_A(w_{i,A} | x_i, y_i) = E_{R_A}(w_{i,A} | x_i, y_i)$, and the probabilities $P(I_i^A = 1 | x_i, y_i)$ are estimated by regressing $w_{i,A}$ against (x_i, y_i), using the observed data. A similar procedure is applied for the sample B. As in *Scenario* 2, we used exponential regression models. Denote by $\{\hat{p}_{k,3}^X, \hat{p}_{i,3}^{Y|X}, \hat{p}_{i,3}^{Z|X}\}$ the estimated population *pdf* under this scenario. The two estimates of p_k^X are harmonized as under *Scenario* 1.

Different sampling parameters κ_A, κ_B and response parameters γ_A, γ_B are considered, thus distinguishing between informative and noninformative samples and different NMAR nonresponse models. We repeated Steps 2-3 for each scenario and each combination of the parameters κ_A, κ_B, γ_A, γ_B, $M = 400$ times.

5.2 Simulation results when the population distribution satisfies the CIA

We begin by studying the effect of ignoring the informative sampling mechanisms used for drawing the samples A and B. To this end, we estimated for each of the 400 samples the probabilities $\{p_k^X\}$ under the scenarios 1 and 2 ($h = 1,2$). Next, we computed the mean $\bar{\hat{p}}_{k,h}^X$ and their variance-covariance matrix, but only for $k = 1,2,3$, since the sum of the probabilities and their estimates equals 1. In order to evaluate the overall performance of the estimators, we use the Hotelling T^2 statistic $(\hat{p} - p)'\hat{V}^{-1}(\hat{p} - p)$, where $\hat{p}$ is the mean vector of the estimated probabilities over the 400 samples and $\hat{V}$ is the empirical V-C matrix of $\hat{p}$.

Table 1 displays the p-values (pv_h) of the test for different choices of the vectors κ_A, κ_B, defining the sampling probabilities (Eq. 5.1).

Table 1. P-values for different choices of the vectors κ_A, κ_B defining the sampling probabilities.

$\kappa_A = \kappa_B$	pv_1	pv_2
(0,0)	0.614	0.614
(0.25,0.25)	<0.0001	0.727
(0.5,0.5)	<0.0001	0.824

As can be seen, when $\kappa_A = \kappa_B = (0,0)$, the sampling designs generating the samples A and B are not informative, and the null hypothesis of no sampling effects is not rejected. However, for $\kappa_A = \kappa_B = (0.25, 0.25)$ and $\kappa_A = \kappa_B = (0.5, 0.5)$, when the sampling processes are ignored under *Scenario* 1, the null hypothesis is rejected with extremely small p-values. When the sampling processes are accounted for under *Scenario* 2, the null hypothesis is not rejected.

So far we focused on the estimation of the probabilities $\{p_k^X\}$. Next we turn our attention to the estimation of the population model $F_p(y \mid x_k)$. For each $X = x_k$, we used the estimated probabilities $\{\hat{p}_i^{Y|X}\}$ to generate a fused data set of size $\tilde{n} = 10,000$ (Section 3.3) and computed the Kolmogorov-Smirnov (*KS*) distance $KS_{p,h}^{Y|x_k} = \sup_{-\infty<y<\infty} \left| F_p(y \mid x_k) - \hat{F}_{p,h}(y \mid x_k) \right|$ between the normal *pdf* $F_p(y \mid x_k)$ used to generate the population values (Step 1 in Section 5.1) and the estimated *pdf*, $\hat{F}_{p,h}(y \mid x_k)$ in the fused data set, with the index $h = 1,2$ labelling the scenario. Table 2 shows the average of the 400 *KS* values, denoted $KSd_{p,h}^{Y|x_k}$, $x_k = 1,2,3,4$.

Table 2. Distance measures $KSd_{p,h}^{Y|x_k}$, for $x_k = 1,2,3,4$, $h = 1,2$ with different choices of the vector coefficients κ_A, κ_B defining the sample selection probabilities (Eq. 5.1).

$\kappa_A = \kappa_B$	$KSd_{p,1}^{Y\|1}$	$KSd_{p,2}^{Y\|1}$	$KSd_{p,1}^{Y\|2}$	$KSd_{p,2}^{Y\|2}$	$KSd_{p,1}^{Y\|3}$	$KSd_{p,2}^{Y\|3}$	$KSd_{p,1}^{Y\|4}$	$KSd_{p,2}^{Y\|4}$
(0,0)	0.033	0.033	0.060	0.060	0.020	0.020	0.021	0.021
(0.25,0.25)	0.380	0.181	0.381	0.164	0.334	0.071	0.251	0.043
(0.5, 0.5)	0.624	0.258	0.566	0.222	0.438	0.123	0.271	0.069

The conclusions from Table 2 are similar to those drawn from Table 1. When $\kappa_A = \kappa_B = (0,0)$, $KSd_{p,1}^{Y|x_k} = KSd_{p,2}^{Y|x_k}$ for $x_k = 1,2,3,4$ and all the distances are very small. When $\kappa_A = \kappa_B = (0.25, 0.25)$, the distance measures are much larger, and they increase further when $\kappa_A = \kappa_B = (0.5, 0.5)$. Notice that for each $x_k = 1,2,3,4$, $KSd_{p,1}^{Y|x_k}$ is much larger than $KSd_{p,2}^{Y|x_k}$, because under *Scenario* 2, we account for the informative sampling designs. We also observe that the $KSd_{p,h}^{Y|x_k}$ distances for $x_k = 1,2$ are much larger than the corresponding distances for $x_k = 3,4$. This result is explained by the fact that the mean of the inclusion probabilities increases as x_k increases, changing from 0.10 for $x_k = 1$, 0.20 for $x_k = 2$, 0.39 for $x_k = 3$ and 0.62 for $x_k = 4$ when $\kappa_A = \kappa_B = (0.25, 0.25)$, with similar means for $\kappa_A = \kappa_B = (0.5, 0.5)$. Thus, the informativeness of the sampling design reduces, as X increases. Similar results (not reported) are obtained when estimating the population *cdf* of $Z \mid X$.

Next consider *Scenario* 3, by which in addition to informative sampling, the samples A and B are subject to NMAR nonresponse. Figure 1 exhibits the population *pdf* and the kernel density estimates of the sample *pdf* with full response**,** the respondents *pdf* and the estimated population *pdf* of $Y \mid x_k = 2$, for one of the 400 samples A, for the case $\kappa_A = (0.5, 0.5)$, $\gamma_A = \gamma_B = (0.05, 0.1)$. For selecting the bandwidth for the kernel estimates, we followed Sheather and Jones (1991). Evidently, the sample *pdf* is different from the population *pdf* due to informative sampling, and the respondents' *pdf* is different from the sample *pdf* because of the nonresponse. Notice that the estimated population *pdf* is the closest to the population *pdf*. Similar results (not reported) are obtained for the *pdf*s of $Y \mid x_k$, $x_k = 1,3,4$ and $Z \mid x_k$, $x_k = 1,2,3,4$.

Figure 1. Population *pdf* and kernel density estimates of the sample *pdf*, the respondents *pdf* and the estimated *pdf* of $Y \mid x_k = 2$**,** $\kappa_A = (0.5, 0.5)$**,** $\gamma_A = \gamma_B = (0.05, 0.1)$**.**

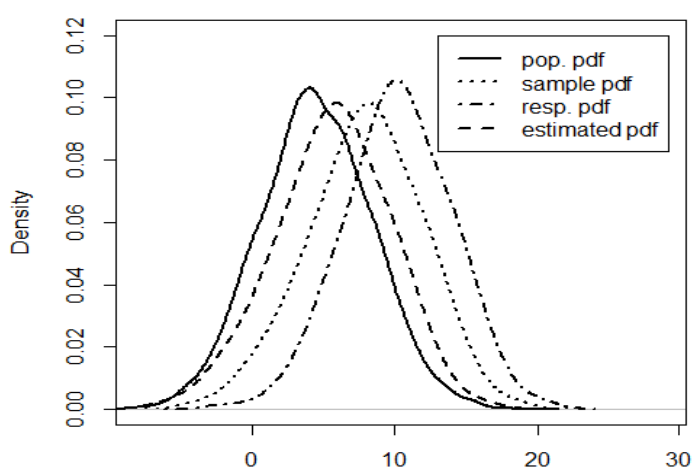


Table 3 shows how by accounting for the sampling and response effects under *Scenario* 3, we are able to fit the population model, using the same sample used for Figure 1. For this, we use the *KS* test statistic with critical values computed by parametric bootstrap, as established theoretically by Babu and Rao (2004) and applied by Pfeffermann and Landsman (2011). Specifically, we generated $B = 500$ bootstrap samples from the estimated model, re-estimated for each sample the unknown model parameters and computed the *KS* statistic with the estimated parameters, and then obtained the critical value at the $\alpha = 0.05$ level from the resulting empirical distribution of the *KS* statistics. Table 3 reports the *KS* statistic of the estimated *pdf* of $Y \mid x_k$ ($x_k = 1,2,3,4$) for the sample in Figure 1 and the corresponding critical value computed by the parametric bootstrap.

Table 3. Kolmogorov-Smirnov test statistic and critical values for $\alpha = 0.05$.

Distribution	*KS statistic*	*Critical value*
$\mathrm{Y} \mid X = 1$	0.14	0.18
$\mathrm{Y} \mid X = 2$	0.11	0.16
$\mathrm{Y} \mid X = 3$	0.04	0.11
$\mathrm{Y} \mid X = 4$	0.04	0.07

We also applied the Hotteling test based on all the 400 samples as in Table 1, with $\gamma_A = \gamma_B = (0.05, 0.1)$ and $\gamma_A = \gamma_B = (0.1, 0.1)$, and obtained extremely high p-values for all the three choices of the vectors κ_A, κ_B defining the sample selection probabilities, thus verifying that the model which accounts for the sampling and response processes fits well the population distribution of X. Table 4 shows the $KSd_{p,3}^{Y|x_k}$ distances for the estimated *cdf* $\hat{F}_p(y \mid x_k)$, computed as in Table 2 by constructing a fused data set. See Section 3.3.

Table 4. Distance measures $KSd_{p,3}^{Y|x_k}$ **for different choices of** κ_A, κ_B**, with** $\gamma_A = \gamma_B = (0.05, 0.1)$.

$\kappa_A = \kappa_B$	$KSd_{p,3}^{Y\|1}$	$KSd_{p,3}^{Y\|2}$	$KSd_{p,3}^{Y\|3}$	$KSd_{p,3}^{Y\|4}$
(0,0)	0.088	0.073	0.043	0.062
(0.25,0.25)	0.223	0.197	0.096	0.057
(0.5,0.5)	0.281	0.231	0.143	0.081

It appears from Table 4 that the distortion in the estimation of the *pdf*s $\{p_i^{Y|X}\}$ worsens under the combination of informative sampling and NMAR nonresponse, particularly for $x_k = 1,2$. Note, however, that the measures $KSd_{p,3}^{Y|x_k}$ are always much smaller than the corresponding measures $KSd_{p,1}^{Y|x_k}$ reported in Table 2, and only mildly larger than the measures $KSd_{p,2}^{Y|x_k}$.

5.3 *Simulation results when the CIA in the population model does not hold*

In this section, we study the performance of the methodology proposed in Section 4. For this, we consider the following *Scenario* 4, which consists of 3 parts:

(i) Generate a population of $N = 10,000$ values x_i, taking the values $k = 1,2,3,4$ with the same probabilities as before. Conditionally on $X = x_k$, generate (Y, Z)-values from a bivariate normal distribution with parameters as in Step 1 of Section 5.1 and $cor_{YZ|x} = 0.77$. The unconditional correlation is $cor_{YZ} = 0.83$.

(ii) Remove values (Y, Z) for which $Y > Z$. The resulting final population of joint (X, Y, Z) values consists of $N = 7,135$ observations, with empirical correlation $cor_{YZ} = 0.91$.

(iii) Select samples (A, B) similarly to Section 5.1, with $\kappa_A = \kappa_B = (0.25, 0.25)$. Select the responding units in the two samples according to Eq. (5.2), with $\gamma_A = \gamma_B = (0.05, 0.1)$.

We start by computing the overall (average) uncertainty measure (4.4), under the constraint $Y \le Z$. For this, we split the population data in (ii) into two datasets, the first containing the values (X, Y) and the second containing the values (X, Z). Under the constraint $Y \le Z$, the measure is $\Delta_{p,c} = 0.10$. When estimating the uncertainty measure but ignoring the sampling and response processes, the estimate is $\hat{\Delta}_{p,c} = \sum_{k=1}^{K} \hat{\Delta}_{p,c}^{k} \hat{p}_k^X = 0.15$. When accounting for the two processes, $\hat{\Delta}_{p,c} = 0.11$.

Next, we estimated the parameters defining the marginal distributions of $Y | x_k$ and $Z | x_k$ under *Scenario* 3 of Section 5.1, following the methodology of Section 3. We then used the estimates for choosing a *matching distribution* from the class (4.7) of plausible distributions under the constraint $Y \le Z$ by use of the IPF, as developed in Section 4.2. For each value x_k, the range of the variable Y (Z) has been divided into intervals of equal size, $\sqrt{r_{k,A}^X}$ $\left(\sqrt{r_{k,B}^X}\right)$, (Dougherty *et al.*, 1995). The IPF accuracy, measured by the maximum deviation between the final row and column marginal probabilities upon convergence and the target probabilities as estimated from the original samples, over all values x_k was found to be 0.02. Finally, we generated a fused data set of size $\tilde{n} = 10,000$, as described at the end of Section 4.2. The correlation between the imputed values of Y and Z obtained from the IPF *distribution* is 0.95, very close to the correlation, $cor_{YZ} = 0.91$ in **(ii)** above. For $k = (1,2,3,4)$, $cor_{(Y,Z)|X=x_k} = (0.87, 0.88, 0.88, 0.88)$ for the population values and $(0.90,\ 0.91, 0.91, 0.90)$ for the imputed values.

6. Application to real data: matching of household income and expenditure

6.1. *Sampling designs and choice of the matching variable*

In this section, we apply our proposed methodology to the SHIW and HBS samples mentioned in the introduction, and construct a fused data set with joint measurements of income and expenditure. SHIW is conducted by Banca d'Italia every two years. Its main goal is to study the economic status of Italian households, focusing on income and wealth. The SHIW questionnaire also contains a section on households expenditures (food consumption, expenses for housing, health, etc.), and some “recall questions” used for constructing an approximate measure of total expenditure. A main drawback of these questions is that they lead to “heaping and rounding”. For example, the concept of nondurable goods is too complex to be measured by a single question. It includes many diverse items and without specific instructions of which items to include, different respondents account for different items in their assessment of total expenditure. Consequently, SHIW suffers from significant underreporting of household expenditure (about 30%).

SHIW is drawn in two stages, with municipalities as the primary sampling units and households (HH) as the secondary sampling units. In the present application, we use the 2010 wave, which consists of 387 municipalities drawn with probabilities proportional to size (PPS) and 7,951 HH sampled by simple random sampling (SRS). The HH income is defined as the combined disposable annual income of all the people living in the HH. The HBS uses a similar sampling design and collects detailed information on socio-demographic characteristics and expenditures on a disaggregated set of commodities (durable and non-durable). Here again, we use the 2010 wave, which consists of 470 municipalities and 22,227 HHs.

As stated and illustrated throughout the article, statistical matching is usually based on a set of variables measured in all the data sources (the *X* variables). In our application, we considered three variables as plausible candidate matching variables, harmonized across the two samples: household size (*hsize*=1,2,3,4+), area of residence (*area*), and occupational status (*condlav*). The literature highlights three main criteria for selecting matching variables;

see, *e.g.*, D'Orazio *et al.* (2006). 1- the variables need to be comparable with regard to their statistical content and have a similar distribution in the two surveys. 2- the variables must have good prediction power in predicting the outcome variables. 3- the use of these variables should minimize the "maximum error" in matching the joint distribution of the outcome variables of interest.

Regarding the first criterion, a common method for comparing the distribution of variables in different data sets is by use of the Hellinger distance $HD = \frac{1}{\sqrt{2}}\sqrt{\sum_{k=1}^{K}\left(\sqrt{\hat{p}_{k,A}^{X}} - \sqrt{\hat{p}_{k,B}^{X}}\right)^2}$, where $\hat{p}_{k,S}^{X}$ are the estimates of the probabilities p_k^X, obtained from sample $S = A, B$. It is generally accepted that a value exceeding 0.05 should raise concerns about the similarity of the distributions. The values in our case are 0.027 for *hsize*, 0.024 for *area* and 0.055 for *condlav*. As for the second criterion, we modelled the log-expenditure (Y) based on the HBS data, and log-income (Z) based on the SHIW data, each time as a linear function of one of the candidate matching variables as the sole explanatory variable. The variables *hsize*, *area* and *condlav* are all statistically significant in explaining the variation of both the expenditure and income. However, *hsize* was found to have the best prediction power, with coefficients of determination $R^2 = 0.20$ in the expenditure model, and $R^2 = 0.11$ in the income model.

In order to examine the third criterion, we proceeded as follows: **(i)** compute for each pair (y_i^k, z_j^k), $i = 1,..,r_{k,A}^X$, $j = 1,..,r_{k,B}^X$ the pointwise uncertainty measure defined by the length of the Fréchet interval (L_c, U_c), with the bounds (4.8) and (4.9), where for $X = x_k$, (y_i^k, z_j^k) defines a pair composed by an observed value of Y and an observed value of Z.

(ii) compute the average of the $r_{k,A}^X r_{k,B}^X$ measures as an estimate of $\Delta_{p,c}^k$, defined in Eq. (4.5); **(iii)** compute the unconditional uncertainty measure $\hat{\Delta}_{p,c}$ defined in Eq. (4.6). We found that when *hsize* is used as the matching variable, the uncertainty measure is $\hat{\Delta}_{p,c} = 0.11$, and it remains approximately the same when including all three matching variables in the analysis. ($\hat{\Delta}_{p,c} = 0.107$). Based on these findings, we use *hsize* as our sole matching variable. For applying our proposed methodology, we added the calibration constraint $\sum_{k=1}^{K} p_k^X x_k = 2.4$, (hereafter C-C), where 2.4 is the average size of households in 2010, as published in the ISTAT site http://dati.istat.it/#.

6.2. *Results obtained when matching the two surveys*

SHIW and HBS suffer from low response rates, about 62% in both samples. It is quite evident that the nonresponse is explained, at least in part, by the size of the HH and the income (or expenditure). The larger the HH, the more possibilities exist to find a contact person for an interview. In addition, HH consisting of only one or two elder people, often tend not to participate in surveys. Furthermore, as often reported in the literature, the response probability tends to decrease as the HH income or expenditure increase (Korinek *et al.*, 2006). In order to obtain a response rate of about 62%, we computed the response probabilities in the two samples by use of the models defined by (5.2), with coefficients $(\gamma_{x,A}, \gamma_{y,A}) = (0.2, -0.002)$, $(\gamma_{x,B}, \gamma_{z,B}) = (0.2, -0.003)$.

Table 5 displays 4 different estimates of the probabilities $\{p_k^X\}$, when considering the 4 possible size values (*hsize*=1,2,3,4+). The first column headed p_k^X, shows the ISTAT's estimates of the household size distribution in Italy in 2010. These values are considered as the true probabilities, and serve as benchmarks for the performance of the other estimates. The estimates are defined as follows: $\hat{p}_{k,1}^X$ are the estimates obtained when ignoring the sampling design effects and assuming that all the units responded, and not imposing the C-C. The estimates are obtained by maximizing the likelihood as under *Scenario* 1 in Section 5.1, but only imposing the constraints (3.7); $\hat{p}_{k,1C}^X$ are the estimates obtained under the same setup, but imposing also the C-C; $\hat{p}_{k,2C}^X$ are the estimates obtained when accounting for the sampling effects (but still assuming full response) and imposing the C-C, obtained by maximizing the ESL (3.6), subject to the constraints (3.7) and the C-C; $\hat{p}_{k,2CM}^X$ are our proposed estimates, which account for the sampling designs and the nonresponse (*Scenario* 3 of Section 5.1), obtained by maximizing the ERL (3.19) under the constraints (3.22) and the C-C. We accounted for the sampling design effects by following the approach described in Section 3.2. The last four columns of Table 5 display the sample sizes and the numbers of respondents, with the index A defining the HBS and the index B the SHIW.

Table 5. Different Estimates of the probabilities p_k^X .

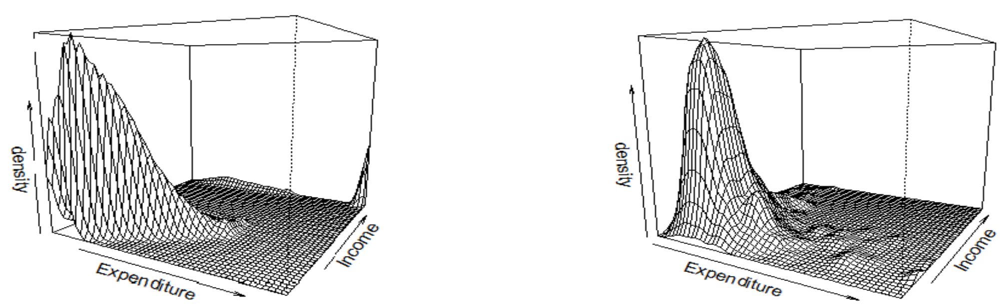
hsize	p_k^X	$\hat{p}_{k,1}^X$	$\hat{p}_{k,1C}^X$	$\hat{p}_{k,2C}^X$	$\hat{p}_{k,2CM}^X$	$n_{k,A}^X$	$n_{k,B}^X$	$r_{k,A}^X$	$r_{k,B}^X$
1	0.284	0.260	0.264	0.276	0.276	5851	1989	3194	1074
2	0.276	0.293	0.293	0.281	0.280	6292	2522	3783	1504
3	0.209	0.210	0.208	0.200	0.205	4758	1589	3069	1028
4	0.232	0.238	0.233	0.243	0.239	5326	1851	3730	1258

In order to compare the goodness of fit of the four sets of estimators in Table 5, we computed again the Hellinger distances, with the estimates compared to the true probabilities, p_k^X . For the estimates $\hat{p}_{k,1}^X$, the *HD* distance is 0.023. It reduces to 0.018 for $\hat{p}_{k,1C}^X$, to 0.012 for $\hat{p}_{k,2C}^X$ and to 0.009 for $\hat{p}_{k,2CM}^X$.

In addition to estimating the probabilities p_k^X, we estimated the *pdf*s $\{p_i^{Y|X}, p_i^{Z|X}\}$, both when ignoring the sampling designs and nonresponse and when accounting for them, imposing the calibration constraint C-C in both cases. Next, we generated a fused data set of size $\tilde{n} = 10{,}000$ by assuming the CIA, as described in Section 3.3. The (weighted) correlations cor_{XY}, cor_{XZ} in the original samples are 0.38 and 0.31, respectively. In the fused data sets, the correlations are 0.34 and 0.28 when ignoring the sampling designs and nonresponse, and {0.38, 0.32} when accounting for them. The correlation between the imputed values of Y and Z when ignoring the sampling designs and nonresponse in the estimation of the probabilities $\{p_k^X, p_i^{Y|X}, p_i^{Z|X}\}$ is 0.08. The correlation increases to 0.13 when both processes are accounted for. Notice that when assuming the CIA, the correlation computed from the original samples is $cor_{YZ}^{CIA} = cor_{XY} cor_{XZ} = 0.12$.

As mentioned in Section 6.1, SHIW contains also some recall questions, aimed for constructing an approximate measure of total expenditure. The correlation in the SHIW sample between income and expenditure is 0.65. Thus, the fused data set constructed under the CIA seems to misrepresent the joint population distribution of (Y, Z). Consequently, we no longer assume the CIA and estimate instead a *matching distribution* for income and expenditure by assuming the class (4.7) of plausible distributions, with the added constraints $Y \le Z$ and the C-C, and applying the IPF. (Section 4.2.) The IPF accuracy was found to be 7×10^{-4}, much smaller than in the simulation study. Next, we used the estimated joint distribution for generating $\tilde{n} = 10{,}000$ values (x_i, y_i, z_i), as described at the end of Section 4.2. Figure 2 shows the bivariate density estimates obtained by application of the IPF and under the CIA, for households of size 3. Similar figures (not shown) have been produced for HH of size 1, 2 and 4+. Evidently, the two estimated densities are different. As noted above, the correlation between the imputed values of Y and Z under the CIA is 0.12. The correlation increases to 0.55 by use of the IPF. The correlation in the SHIW sample is 0.65, but recall that expenditure is not directly observed in SHIW. See Section 6.1.

Figure 2. Estimation of *pdf* of (Y,Z) under the constraint $\boldsymbol{Y \leq Z}$ for $hsize = 3$. Estimate obtained by IPF (left) and under the CIA (right).



Rässler (2002) proposes four validation measures of decreasing importance in a statistical matching problem, which in our case are as follows: **(1)** preserving the true household values; **(2)** preserving the true joint distribution; **(3)** preserving correlation structures; **(4)** preserving marginal distributions. We cannot assess the first measure since the true incomes and expenditures at the HH level are unknown. The second measure requires knowledge of the true joint population distribution of (X,Y,Z), which is likewise unknown, but an uncertainty measure of the kind introduced in Section 4.1 can be used to assess how far the matching distribution is from the true joint distribution. When accounting for the sampling and nonresponse effects and imposing the constraint $Y \leq Z$, the estimated uncertainty measure $\hat{\Delta}_{p,c}$ decreases from 0.16, (its maximum value with no constraint) to 0.11. The uncertainty measure increases to 0.13 when the sampling and nonresponse processes are ignored. Regarding the third measure, we note that the correlation between the imputed values of expenditure and income is 0.55 when applying the IPF. Thus, our proposed methodology seems to recover pretty well the "approximate" correlation of 0.65 between income and expenditure in the SHIW sample. Regarding the fourth measure, the constructed fused data set preserves by construction, the marginal distributions of the income and expenditure. This follows from the use of the IPF, which adjusts the initial cell probabilities to fit the marginal distributions of the two variables, as estimated from the two samples separately.

7. Concluding Remarks

In this paper, we propose a comprehensive approch to deal with statistical matching, when the samples containing the unmatched data are drawn by informative sampling designs and are subject to NMAR nonresponse. Our approach employs the EL to account for the sampling and response processes, thus enabling generating a fused data set, which represents sufficiently accurately the true joint population *pdf* of the target variables. We first consider the case where the target variables of interest are conditionally independent given the available matching variables (the CIA), and then the much more challenging problem when the CIA cannot be assumed. In order to deal with the latter case, we apply a procedure based on the IPF for choosing a *pdf* from a class of plausible *pdf*s, which satisfy available information regarding the relationship between the target variables and calibration constraints. An extensive simulation study and application to real datasets illustrate the good performance of our proposed methodology.

We obviously hope that other researchers will apply our proposed approach with appropriate modifications required for their data. New theoretical developments of the present work include the use of proxy variables for estimation of the conditional sample inclusion probabilities $P(I_i^A = 1 | x_i, y_i)$ when the response process is not independent of the sampling process (Section 3.4), possibly by adding them to the covariates of the sampling and/or the response models. Good proxy variables may also be used for initialization of the IPF algorithm. Finally, we mention the EM algorithm for maximization of the empirical respondents' likelihood (3.19), as proposed by one of the reviewers of the article. (See Remark 9.)

References

Babu, G.J. and Rao, C.R. (2004). Goodness-of-Fit Tests when Parameters are Estimated. *Sankhya, Series A*, **66**, 63-74.

Bishop, Y.M., Fienberg, S.E. and Holland, P.W. (1975). *Discrete Multivariate Analysis*. Springer, New-York.

Chaudhuri S., Handcock M.S and Rendall, M.S. (2010). A conditional empirical likelihood approach to combine sampling design and population level information. Technical report No.3/2010, National University of Singapore, Singapore, 117546.

Chen, J. and Qin, J. (1993). Empirical likelihood estimation for finite population and the effective usage of auxiliary information. *Biometrika,* **80**, 107-116.

Chen, J. and Sitter, R. R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica*, **9**, 385-406.

Chen, S.X. and Van Keilegom, I. (2009). A review on empirical likelihood methods for regression. *Test*, **18**, 415-447.

Conti, P.L., Marella, D. and Scanu, M. (2012). Uncertainty Analysis in Statistical Matching. *Journal of Official Statistics*. **28**, 1–21.

(2013). Uncertainty Analysis for Statistical Matching of ordered categorical variables. *Computational Statistics & Data Analysis*. **68**, 311–325.

(2015). How far from identifiability? A systematic overview of the statistical matching problem in a non-parametric framework. *Communications in Statistics-Theory and Methods*,**46**,967-994.

(2016). Statistical matching analysis for complex survey data with applications. *Journal of the American Statistical Association*, **111**, 1715-1725.

Conti, P.L., Marella, D. and Neri, A. (2017) Statistical matching and uncertainty analysis in combining household income and expenditure data. *Statistical Methods & Applications*, **26**, 3, 485-505.

Conti P.L., Marella D., Mecatti F. and Andreis F. (2020). A unified principled framework for resampling based on pseudo-populations: Asymptotic theory. *Bernoulli*, **26**, 2, 1044-1069.

D’Orazio, M., Di Zio, M., and Scanu, M. (2006a). Statistical matching for categorical data: displaying uncertainty and using logical constraints. *Journal of Official Statistics*, **22**, 137–157.

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

Dougherty, J., Kohavi, R. and Mehran, S. (1995). Supervised and unsupervised discretization of continuous features. *Machine Learning*: *Proocedings of the Twelfth International Conference*. Morgan Kaufmann Publishers, San Francisco, CA.

Feder, M. and Pfeffermann, D. (2019). Statistical inference under non-ignorable sampling and non-response-an empirical likelihood approach. http://eprints.soton.ac.uk/id/eprint/378245.

Hartley, H.O. and Rao, J.N.K. (1968). A new estimation theory for sample surveys. *Biometrika*, **55**, 547-557.

Kim, J.K. (2009). Calibration estimation using empirical likelihood in survey sampling. *Statistica Sinica*, **19**, 145-157.

Korinek, A., Mistiaen, J.A. and Ravallion, M. (2006). Survey nonresponse and the distribution of income. *The Journal of Economic Inequality*, **4**, 33-55.

Little, R.J.A. and Rubin, D.B. (1987), *Statistical Analysis with Missing Data*, New York: Wiley.

Marella, D. and Pfeffermann, D. (2019). Matching information from two independent informative sampling. *Journal of Statistical Planning and Inference*, **203**, 70-81.

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

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

Okner, B. (1972). Constructing a new data base from existing microdata sets: the 1966 merge file. *Annals of Economic and Social Measurement*, **1**, 325–342.

Owen, A. (1990). Empirical likelihood confidence regions. *The Annals of Statistics*,**18**,1,90-120.

(1991). Empirical likelihood for linear models. *The Annals of Statistics*, **19**, 1725-1747.

(2001). *Empirical likelihood*. Boca Raton: Chapman & Hall/CRC.

(2013). Self-concordance for empirical likelihood. *Canadian Journal of Statistics*, **41**, 387-397.

Pfeffermann, D., Krieger, A.M. and Rinott, Y. (1998). Parametric distribution of complex survey data under informative probability sampling. *Statistica Sinica*, 8, 1087-1114.

Pfeffermann, D. and Sverchkov, M. (1999). Parametric and and semi-parametric estimation of regression models fitted to survey data. *Sankhya*, Series B, **61**, 166-186.

Pfeffermann, D. and Sverchkov, M. (2009). Inference under informative sampling. In: *Handbook of Statistics* 29B; *Sample Surveys: Inference and Analysis*. (Eds, D. Pfeffermann and C.R. Rao). Amsterdam: North Holland, pp. 455-487.

Pfeffermann, D. (2011). Modelling of complex survey data: Why model? Why is it a problem? How can we approach it? *Survey Methodology*, **37**, 115-136.

Pfeffermann, D. and Sikov, A. (2011). Imputation and estimation under nonignorable nonresponse in household surveys with missing covariate information. *Journal of Official Statistics*, **27**, 181-209.

Pfeffermann, D. and Landsman, V.(2011).Are private schools really better than public schools? Assessment by methods for observational studies. *Annals of Applied Statistics*, **5**, 1726-1751.

Qin, J., and Lawless, J. (1994), Empirical likelihood and general estimating equations. *The Annals of Statistics*, **22**, 300-325.

Pfeffermann, D. (2017). Bayes-based Non-Bayesian Inference on Finite Populations from Non-representative Samples: A Unified Approach. *Calcutta Statistical Association Bulletin*, **69**, 35-63.

Rässler, S. (2002). *Statistical Matching*: *A Frequentist Theory, Practical Applications and Alternative Bayesian Approaches*. Springer, New York.

Renssen, R.H. (1998). Use of statistical matching techniques in calibration estimation. *Survey Methodology,* **24**, 171-183.

Rubin, D.B. (1986). Statistical Matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economics Statistics*, **4**, 87-94.

Sheather, S.J. and Jones, M.C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society*, *Series* B, **53**, 683-690.

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

Wu, C. (2004). Combining information from multipe surveys through the empirical likelihood method. *The Canadian Journal of Statistics*, **32**, 112, 1-12.

Zhang, L.-C. (2015). On proxy variables and categorical data fusion. *Journal of Official Statistics*, **31**, 783-807.

Zhang, L.-C. and Chambers, R.L. (2019). Minimal inference from incomplete 2 x 2- tables. In *Analysis of Integrated Data*, eds. L.-C. Zhang and R.L. Chambers. Chapter 6, 121-136. Chapman & Hall/CRC.