AStA Adv Stat Anal (2009) 93: 175–186
DOI 10.1007/s10182-008-0081-2

ORIGINAL PAPER

Semiparametric predictive mean matching

Marco Di Zio · Ugo Guarnera

Received: 1 April 2008 / Accepted: 4 September 2008 / Published online: 8 October 2008

Abstract Predictive mean matching is an imputation method that combines parametric and nonparametric techniques. It imputes missing values by means of the Nearest Neighbor Donor with distance based on the expected values of the missing variables conditional on the observed covariates, instead of computing the distance directly on the values of the covariates. In ordinary predictive mean matching the expected values are computed through a linear regression model. In this paper a generalization of the original predictive mean matching is studied. Here the expected values used for computing the distance are estimated through an approach based on Gaussian mixture models. This approach includes as a special case the original predictive mean matching but allows one to deal also with nonlinear relationships among the variables. In order to assess its performance, an empirical evaluation based on simulations is carried out.

Keywords Incomplete data · Imputation · Nearest neighbor donor · Gaussian mixture models

1 Introduction

The presence of partially incomplete data is one of the main issues to deal with in the context of Official Statistics. The most common way to manage missing values consists in compensating for nonresponse by imputing artificial data. A variety of imputation techniques have been introduced in literature and used by practitioners. They can be roughly divided into parametric and nonparametric techniques. Parametric methods are generally parsimonious, but, being based on explicit models, they fail

M. Di Zio (✉) · U. Guarnera
Istituto Nazionale di Statistica, via Cesare Balbo 16, 00184 Roma, Italy
e-mail: dizio@istat.it

when the model assumptions are not suitable for the data to be analyzed. On the contrary, nonparametric techniques do not rely on explicit model assumptions but require high amount of observations in order to be satisfactorily applied. One of the most popular nonparametric imputation methods is the Nearest Neighbor Donor (NND) that consists in matching completely observed units (donors) with incomplete units (recipients), based on some distance function, and transferring values from donors to recipients.

In order to overcome the difficulties of parametric and nonparametric methods, some techniques have been developed that could be considered in the middle of the two previous approaches. Among them, Predictive Mean Matching (PMM) (Little 1988) is one of the most commonly used (see, for instance, Durrant and Skinner 2006). PMM makes use of an explicit parametric model only to define a suitable criterion for matching complete and incomplete units. In a quite broad sense, PMM could be considered as a particular NND method with a suitable distance function. On the other hand, the function used in PMM is not a real distance function in the covariate space since the distance between distinct points can be zero. Thus, the asymptotic properties of the NND are no longer guaranteed, and the results of imputation via PMM still depend on the model. Nevertheless, the method is probably more robust than a fully model based approach, with respect to departures from the model assumptions. PMM is also appealing because it imputes "live" values, i.e., values that are really observed.

In a multivariate context, when the variables are continuous and in presence of arbitrary patterns of missing items, a typical application of the PMM is the following.

1. The parameters of a multivariate Gaussian distribution are estimated through the EM algorithm (Dempster et al. 1977) using all the available data (complete and incomplete).
2. Based on the estimates from EM, for each incomplete unit (recipient), predictions of the missing items conditional on the observed ones are computed. The same predictive means (i.e., corresponding to the same missing pattern) are computed for all the complete observations (donors).
3. Each recipient is matched to the donor having the closest predictive mean with respect to the Mahalanobis distance defined through the residual covariance matrix from the regression of the missing items on the observed ones.
4. Missing items are imputed in each recipient by transferring the corresponding values from its closest donor.

Although the previous procedure should be more robust than imputation based on standard linear regression, some degree of linearity is still assumed in the relations among variables. Thus, if this assumption is not appropriate, poor performances are expected. In this paper, this difficulty is overcome through a more flexible version of PMM that includes as a particular case the ordinary PMM. In the proposed method, data are modeled by means of a Gaussian mixture instead of a simple normal model. The idea is to exploit the flexibility of Gaussian mixture models for approximating more general data distributions (Marron and Wand 1992; Fraley and Raftery 2002). As in the ordinary PMM, the role of the model is only to provide a suitable distance function to be used for nearest neighbor imputation. This approach, which could be

defined "semiparametric," allows handling data that are far from normality, keeping the advantage of imputing "live" values. The last characteristic ensures univariate plausibility. For instance, missing items for nonnegative variables are guaranteed to be imputed with nonnegative values. As it will be clarified in the following, this semiparametric predictive mean matching (SPMM) is a generalized version of the standard predictive mean matching.

The semiparametric predictive mean matching is compared to the nearest neighbor donor method and to model based imputations obtained via Gaussian mixture models as described in Di Zio et al. (2007). The experiments are performed on both simulated and real data.

The paper is organized as follows. In Sect. 2, general concepts and basic definitions on finite mixtures of Gaussian distributions are given. Section 3 illustrates the use of mixture models for imputation via PMM. Finally, simulations and results are described in Sect. 4.

2 Estimation of Gaussian mixtures models in presence of incomplete data

Let $\mathbf{Y}$ be a p-dimensional random vector (r.v.) with probability distribution (density) $f(\mathbf{y})$. Let us suppose that f can be represented in the form

$$f(\mathbf{y};\boldsymbol{\Phi}) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{y};\boldsymbol{\theta}_k), \tag{1}$$

where the densities f_k with parameters $\boldsymbol{\theta}_k$ belong to the same parametric family, and the parameters π_k are positive and subject to the constraint $\sum_{k=1}^{K} \pi_k = 1$. Model (1) is said to be a mixture of the distributions $f_1, \ldots, f_K$ with mixing proportions $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$. The functions $f_1, \ldots, f_K$ are generally named mixture components (McLachlan and Peel 2000). In formula (1), $\boldsymbol{\Phi}$ denotes the full set of parameters $(\pi_1, \ldots, \pi_K, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$. The distribution $f(\mathbf{y};\boldsymbol{\Phi})$ is a Gaussian mixture if the functions f_k are Gaussian densities: $f_k = N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, where $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the normal density function with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

The log-likelihood of a Gaussian mixture based on n observations $\mathbf{y}_1, \ldots, \mathbf{y}_n$ is

$$L(\boldsymbol{\Phi}) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k f_k(\mathbf{y}_i;\boldsymbol{\theta}_k)\right) \tag{2}$$

and cannot be analytically maximized. The maximum likelihood estimates (MLE) are usually determined by recasting the problem as an incomplete data problem and by using the EM algorithm (Dempster et al. 1977). To this aim, each unit i $(i = 1, \ldots, n)$ is supposed to belong to one of the K groups corresponding to the K mixture components, and each group k $(k = 1, \ldots, K)$ is given an unobserved indicator variable Z_{ik}, where Z_{ik} is 1 or 0 depending on whether unit i belongs or not to group k.

The random vector $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iK})$ is multinomially distributed: $\text{Mult}_K(1, \boldsymbol{\pi})$ so that $\text{Prob}\{Z_{ik} = 1\} = \pi_k$. The mixing proportion π_k $(k = 1, \ldots, K)$ can be interpreted as the *a priori* probability of belonging to the group k. Furthermore, by the

Bayes formula, the probability

$$\tau_{ik} = E(Z_{ik}|\mathbf{y}_i, \boldsymbol{\Phi}) = \frac{\pi_k f_k(\mathbf{y}_i; \boldsymbol{\theta}_k)}{\sum_t \pi_t f_t(\mathbf{y}_i; \boldsymbol{\theta}_t)}, \quad i = 1, \ldots, n,\ k = 1, \ldots, K,$$

is the corresponding posterior probability given the observation $\mathbf{y}_i$, and its estimate can be expressed in terms of the estimates of parameters $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$. Following this setting, the complete data log-likelihood can be written as

$$L_c(\boldsymbol{\Phi}) = \sum_{k=1}^{K} \sum_{i=1}^{n} z_{ik} \big[\log(\pi_k) + \log f_k(\mathbf{y}_i; \boldsymbol{\theta}_k)\big], \tag{3}$$

where z_{ik} is the realized value of the r.v. Z_{ik}. Actually, the values z_{ik} are not observed, and if we replace them by the corresponding r.v. Z_{ik} in formula (3), we obtain the r.v. which we still call $L_c(\boldsymbol{\Phi})$. The E-step of the EM algorithm consists in calculating, at each iteration, the expected value of $L_c(\boldsymbol{\Phi})$ conditional on $\mathbf{y}_i$ and the current estimates of the parameters. This reduces to compute the expectation of Z_{ik} given $\mathbf{y}_i$, i.e., τ_{ik}, for $i = 1, \ldots, n$ and $k = 1, \ldots, K$. In the normal case, the M-step can also be performed in closed form, providing recursive equations for the parameters $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$.

In case of partially incomplete data the algorithm so far described has to be slightly modified in order to take into account the missing items. Now, the quantity to be maximized is not anymore the log-likelihood (2) but the observed-data log-likelihood

$$L_{\text{obs}}(\boldsymbol{\Phi}) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k f_k(\mathbf{y}_{\text{obs},i}; \boldsymbol{\theta}_k)\right),$$

where, according to the usual notation, $\mathbf{y}_{\text{obs},i}$ is the observed part of the vector $\mathbf{y}_i$ in the decomposition $\mathbf{y}_i = (\mathbf{y}_{\text{obs},i}, \mathbf{y}_{\text{mis},i})$. The modified EM algorithm is described by Hunt and Jorgensen (2003) and basically combines the standard EM algorithm for Gaussian mixtures, with the EM algorithm for incomplete normal data (Schafer 1997).

In our proposal, in order to initialize the EM, first a k-means algorithm is used to cluster data into as many groups as the number of the mixture components. Then, the proportions of units belonging to different clusters is taken as starting values of the parameters $\boldsymbol{\pi}$, while the parameters $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ are initialized with the sample mean and the sample covariance matrix of each cluster.

3 PMM via Gaussian mixtures models

The algorithm described in the previous section refers to the MLEs of a Gaussian mixture with a fixed number K of components. The problem then arises of how to choose the optimum value of K. The approach followed in this paper is based on the use of the Bayesian Information Criterion (BIC). In many problems of model selection, the BIC score can well approximate the Bayesian posterior model probability (Schwarz 1978). Moreover, as Roeder and Wasserman (1997) have shown, when a

normal mixture model is used to estimate a density "nonparametrically," the density estimate that uses BIC to select the number of mixture components is consistent.

For a given number K of mixture components, BIC is defined as $2L_{\text{obs}}(\hat{\boldsymbol{\Phi}}_K) - \nu_K \log(n)$, where $\hat{\boldsymbol{\Phi}}_K$ are the MLEs, ν_K is the number of independent parameters to be estimated, and n is the number of available observations. The proposed strategy consists in estimating different models with different number of components and choosing the model with the highest BIC.

Once the model that best fits data is selected and its parameters are estimated, for each incomplete observation $\mathbf{y}_i = (\mathbf{y}_{\text{obs},i}, \mathbf{y}_{\text{mis},i})$, the conditional distribution $f(\mathbf{y}_{\text{mis},i}|\mathbf{y}_{\text{obs},i}; \boldsymbol{\Phi})$ can be estimated as

$$f\left(\mathbf{y}_{\text{mis},i}|\mathbf{y}_{\text{obs},i}; \hat{\boldsymbol{\Phi}}\right) = \sum_{k=1}^{K} \hat{\tau}_{ik} N\left(\mathbf{y}_{\text{mis},i}|\mathbf{y}_{\text{obs},i}; \hat{\boldsymbol{\theta}}_k\right),$$

and this probability distribution can be used for imputing missing values via its expected value (hereafter MCM) or through a random draw (MRD), as described in Di Zio et al. (2007).

In the SPMM, analogously to the ordinary predictive mean matching, the conditional mean from the distribution is only used to find a nearest neighbor for the ith unit. More in detail, for each incomplete unit i, the donor j is the closest unit to i in terms of predictive mean. An important issue to deal with is the choice of the distance function to be used. As already mentioned in Sect. 1, in the ordinary PMM, a natural choice is the Mahalanobis distance based on the residual covariance matrix of the regression of $\mathbf{Y}_{\text{mis}}$ on $\mathbf{Y}_{\text{obs}}$. In fact, this choice gives a sort of standardization, where, roughly speaking, the contribution of the different variables to the global distance function is "weighted" with the inverse of the corresponding prediction error (Little 1988). Unlike standard PMM, in SPMM the residual covariance matrix depends also on the observations through the posterior probabilities τ_{ik}, hence a generalization must take into account the residual covariance matrices for both the recipient and the donor. In order to simplify the notation, let $\mathbf{x}_r = \mathbf{y}_{\text{obs},r}$ be the observed part of the incomplete record (recipient) r, and $\mathbf{y}_{m,r}$ the missing part. Correspondingly, for each complete record (possible donor) d, let $\mathbf{x}_d$ and $\mathbf{y}_{m,d}$ correspond to the missing and the observed subvectors, respectively, in the (recipient) record r.

A natural metric for the Mahalanobis distance is the estimate $\hat{S}_{\mathbf{Y}|\mathbf{X}}$ of the covariance matrix $S_{\mathbf{Y}|\mathbf{X}} = \text{Cov}[(\mathbf{Y}_{m,r}|\mathbf{x}_r) - (\mathbf{Y}_{m,d}|\mathbf{x}_d)]$. In fact, since the variables $\mathbf{Y}_{m,r}|\mathbf{x}_r$ and $\mathbf{Y}_{m,d}|\mathbf{x}_d$ are independent one of each other, $S_{\mathbf{Y}|\mathbf{X}}$ is the sum of their covariance matrices $\text{Cov}(\mathbf{Y}|\mathbf{x}_r)$ and $\text{Cov}(\mathbf{Y}|\mathbf{x}_d)$, respectively. In order to provide an explicit formula for $S_{\mathbf{Y}|\mathbf{X}}$, we note that the covariance matrix $\text{Cov}(\mathbf{Y}|\mathbf{x})$ of the distribution of the random vector $\mathbf{Y}$ conditional on $\mathbf{X} = \mathbf{x}$ can be decomposed as $\text{Cov}(\mathbf{Y}|\mathbf{x}) = \text{Cov}^{(1)}(\mathbf{Y}|\mathbf{x}) + \text{Cov}^{(2)}(\mathbf{Y}|\mathbf{x}) = E_k[\text{Cov}(\mathbf{Y}|\mathbf{x}, k)] + \text{Cov}_k[E(\mathbf{Y}|\mathbf{x}, k)]$, where the covariance matrix $\text{Cov}(\mathbf{Y}|\mathbf{x}, k)$ and the expected value $E(\mathbf{Y}|\mathbf{x}, k)$ refer to the distribution of $\mathbf{Y}$ conditional on $\mathbf{X} = \mathbf{x}$ and a specific mixture component k, while E_k and Cov_k refer to the distribution of the indicator variable Z for the group labels $k = 1, \ldots, K$.

The first term $\text{Cov}^{(1)}(\mathbf{Y}|\mathbf{x})$ on the r.h.s. of the above decomposition is $\sum_{k=1}^{K} \tau_k(\mathbf{x}) \times \boldsymbol{\Sigma}^{(k)}_{\mathbf{Y}|\mathbf{X}}$, where $\tau_k(\mathbf{x})$ denotes the posterior probability of belonging to the group k for a

unit where $\mathbf{x}$ is observed, and $\mathbf{\Sigma}^{(k)}_{\mathbf{Y}|\mathbf{X}}$ is the residual covariance matrix of the regression of $\mathbf{Y}$ on $\mathbf{X}$ referring to the kth Gaussian distribution of the mixture.

As far as the second term $\mathrm{Cov}^{(2)}(\mathbf{Y}|\mathbf{x})$ is concerned, it can be shown that $\mathrm{Cov}^{(2)}(\mathbf{Y}|\mathbf{x}) = \sum_{k=1}^{K} \tau_k(\mathbf{x}) D^{(k)}(\mathbf{x})$, where

$$D^{(k)}(\mathbf{x}) = \big[E(\mathbf{Y}|\mathbf{x},k) - E(\mathbf{Y}|\mathbf{x})\big]\big[E(\mathbf{Y}|\mathbf{x},k) - E(\mathbf{Y}|\mathbf{x})\big]^{\mathrm{T}}.$$

The total covariance matrix is

$$\mathrm{Cov}(\mathbf{Y}|\mathbf{x}) = \mathrm{Cov}^{(1)}(\mathbf{Y}|\mathbf{x}) + \mathrm{Cov}^{(2)}(\mathbf{Y}|\mathbf{x}) = \sum_{k=1}^{K} \tau_k(\mathbf{x})\big(\mathbf{\Sigma}^{(k)}_{\mathbf{Y}|\mathbf{X}} + D^{(k)}(\mathbf{x})\big).$$

The final estimate of the metric of the Mahalanobis distance is given by $\hat{S}_{\mathbf{Y}|\mathbf{X}} = \widehat{\mathrm{Cov}}(\mathbf{Y}|\mathbf{x}_r) + \widehat{\mathrm{Cov}}(\mathbf{Y}|\mathbf{x}_d)$, where $\widehat{\mathrm{Cov}}(\mathbf{Y}|\mathbf{x})$ is obtained by using the MLE of the relevant parameters.

It is worthwhile to note that when the number of components of the mixture model is $K = 1$, the proposed method coincides with the original PMM; in fact, it reduces to a simple Gaussian model, and the distance is proportional, up to a constant, to the Mahalanobis metric.

4 Empirical evaluation

In this section we describe the simulation study carried out to evaluate the performance of the proposed semiparametric predictive mean matching. The experiments rely on data artificially generated from different probability distributions and on a subset of data obtained from a real survey. The assessment is made in terms of preservation of means and of covariance structure of the data. To this aim a comparison between SPMM, NND, MCM, and MRD is performed. The conditional distribution used for MCM, MRD, and SPMM is estimated through a finite mixture of Gaussian distributions as described in Sect. 2. The considered imputation methods are evaluated in different simulation frameworks.

For each experimental setting, 100 simulations have been performed consisting of the following steps:

1. artificial generation of a sample from a given multivariate probability distribution;
2. introduction of missing values in the sample;
3. estimation of the mixture model used for SPMM, MCM, MRD, and imputation;
4. comparison of the imputed dataset with the original one through appropriate indices.

All the experiments are developed using SAS/IML software, Version 9.1 of the SAS System for Windows.

The previous 4 steps are detailed in the following subsection.

4.1 Sample data generation

In this section, the probability distributions and the set of real data used for the empirical evaluation are described.

4.1.1 Lognornal distribution, LN

A first experiment has been performed by drawing data from a multivariate lognormal distribution. In practice, this is accomplished by drawing a sample of a 5-dimensional random vector $(X_1, \ldots, X_5)$ from a 5-variate Gaussian distribution and then defining new variables $(Y_1, \ldots, Y_5)$ through the transformation: $Y_i = \exp(X_i)$ for $i = 1, \ldots, 5$. The normal random vector $(X_1, \ldots, X_5)$ is obtained by merging two independent random vectors (X_1, X_2) and (X_3, X_4, X_5) having normal distributions characterized by parameters $(\boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma}^{(2)})$ and $(\boldsymbol{\mu}^{(3)}, \boldsymbol{\Sigma}^{(3)})$, respectively. The values of the parameters for the normal distributions are:

$$\boldsymbol{\mu}^{(2)} = (-2.5, -2.6)', \qquad \boldsymbol{\mu}^{(3)} = (-2.5, -2.6, -2.6)',$$
$$\Sigma^{(2)} = \begin{pmatrix} 3.1 & 2.7 \\ 2.7 & 2.8 \end{pmatrix}, \qquad \Sigma^{(3)} = \begin{pmatrix} 3.1 & 2.4 & 2.4 \\ 2.4 & 3.0 & 2.1 \\ 2.4 & 2.1 & 3.0 \end{pmatrix}.$$

The parameters are obtained by a real survey.

4.1.2 Multivariate Gamma distribution, MG

Data are drawn from the Cheriyan and Ramabhadran's multivariate Gamma distribution described in Kotz et al. (2000, pp. 454–456). In order to draw a sample of a 5-variate random vector $(Y_1, \ldots, Y_5)$ from this distribution, the following procedure is adopted. First, samples are drawn from 6 independent random variables $X_1, \ldots, X_6$ distributed according to Gamma distributions with parameters θ_i $(i = 1, \ldots, 6)$. Then, samples from $(Y_1, \ldots, Y_5)$ are obtained through the transformations

$$Y_1 = X_1 + X_2; \qquad Y_2 = X_1 + X_3;$$
$$Y_3 = X_1 + X_4; \qquad Y_4 = X_1 + X_5; \qquad Y_5 = X_1 + X_6.$$

The values of the parameters are

$$\boldsymbol{\theta} = (1.0, 0.2, 0.3, 0.4, 0.5)'.$$

Following Kotz et al. (2000), it is easy to compute the expected value and the correlation matrix of the random variables Y_i. The values of $\boldsymbol{\theta}$ are chosen so that the variables Y_i are characterized by high correlations. This experiment will be denoted as MGH.

Another experimental setting (hereafter MGL) is obtained through the following slight modification of the Cheriyan and Ramabhadran's procedure. First, 7 independent r.v.s. X_i for $i = 1, \ldots, 7$ are considered distributed according to Gamma distributions characterised by different parameters θ_i. Then, the 5-variate random vector is obtained combining the X_i in the following way:

$$Y_1 = X_1 + X_3; \qquad Y_2 = X_1 + X_4;$$
$$Y_3 = X_1 + X_2 + X_5; \qquad Y_4 = X_2 + X_6; \qquad Y_5 = X_2 + X_7.$$

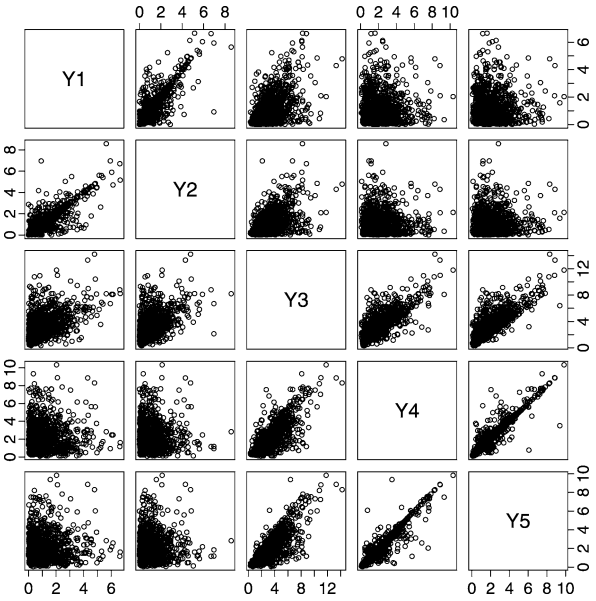

Fig. 1 Scatter-plot matrix of a sample drawn from the distribution used in MGL

The parameters θ_i are chosen to obtain a correlation structure characterized by two weakly correlated blocks of variables with high correlation within the blocks. The values of the parameters are

$$\boldsymbol{\theta} = (1, 2, 0.2, 0.2, 0.4, 0.2, 0.1)'.$$

A plot of a sample of 1,000 observations from this distribution is shown in Fig. 1.

For all the probability distributions (LN, MCH, MGL), samples of 500 and 1,000 units have been generated.

4.1.3 Experiments on a real data set

A subset of the 1997 Italian *Labour Cost Survey* (LCS) is also used for the evaluation of the procedure. The LCS is a periodic sample survey that collects information on employment, hours worked, wages, salaries, and labour cost on about 12,000 firms with more than 10 employees. Our dataset consists of 1,000 units that belong to the metallurgic economic activity sector. We analyze four main variables measuring the "*Total number of Employees*" (Y_1), the "*Total number of Hours Worked*" (Y_2), the "*Wages and Salaries*" (Y_3), and the "*Total Labour Cost*" (Y_4). The values of the variables are obtained by means of a logarithmic transformation of the original data. The experiment will be denoted by CLAV. Figure 2 shows the scatter-plot matrix of the data used for the experiments.

Since the underlying data distribution is unknown, a resampling approach has been adopted. The adopted resampling scheme consists in sampling 1,000 observations (through a simple random sampling with replacement) $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(1,000)}$ (bootstrap sample) from the initial sample, where $\mathbf{y}^{(i)}$ represents the ith unit whose observed variables are (Y_1, Y_2, Y_3, Y_4). The bootstrap sample can be thought of as generated from the empirical distribution of (Y_1, Y_2, Y_3, Y_4).

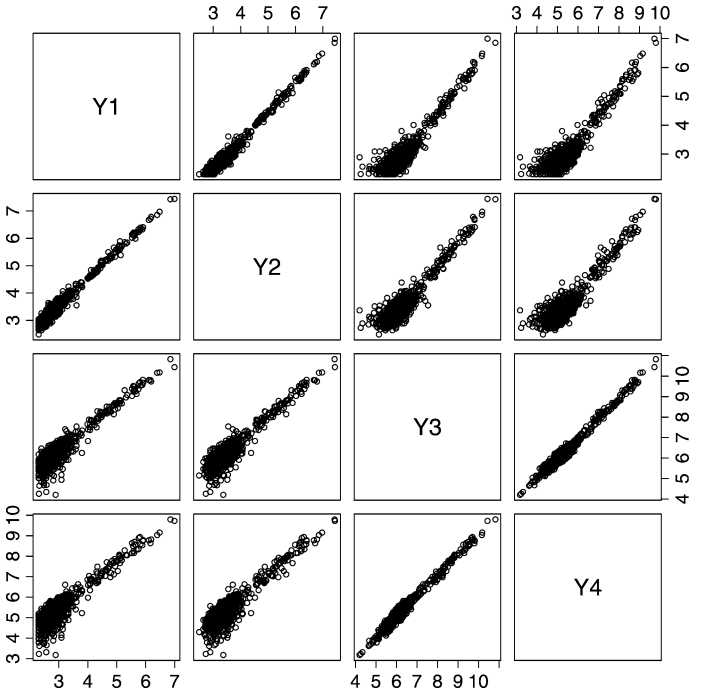

Fig. 2 The scatter-plot matrix of the dataset used for the experiment CLAV

4.2 Nonresponse simulation

Once a sample of complete data is generated, item nonresponse is simulated according to a Missing at Random (MAR) mechanism (Little and Rubin 2002).

When data are generated from a probability distribution, nonresponse-rates for (Y_1, Y_2, Y_3, Y_4) depend on the observed values y_5 of the variable Y_5. More in detail, denoting by q_i the ith quartile of the empirical distribution of Y_5, the nonresponse probabilities for (Y_1, Y_2, Y_3, Y_4) are the following: 0.10 if $y_5 < q_1$, 0.20 if $y_5 \in [q_1, q_3)$, and 0.30 if $y_5 \geq q_3$. No missing values are introduced in the variable Y_5.

As far as real data is concerned, missing values have been introduced according to the previous described mechanism noting that the conditioning completely observed variable is Y_4.

4.3 Estimation and imputation

The incomplete sample is imputed using the NND, the SPMM, the MCM, and the MRD. In the NND method, the Euclidean distance is used, and the matching variables for a given incomplete unit are all those observed in that unit.

Concerning finite mixtures, estimates are performed following the algorithm described in Sect. 2. Models with different number of components have been estimated. Once the parameters have been estimated for all the models, the model with the high-

est BIC is used to impute missing values following the SPMM, MCM, MRD methods described in Sect. 3.

Starting points for the EM algorithm have been identified by the k-means algorithm. The stopping rule is based on a threshold for the relative increase of the likelihood in two consecutive iterations. In order to avoid singularities due to the unboundness of the likelihood function for heteroscedastic mixtures models (McLachlan and Peel 2000), the EM runs have been discarded whenever any matrix involved in the estimation algorithm had determinant below a prefixed threshold.

4.4 Evaluation

The process is replicated 100 times. For each iteration, the following indicators are computed based on the comparison of the original dataset with the imputed ones. Let $y_{i1}, \ldots, y_{ip}$ $(i = 1, \ldots, n)$ be the original "true" values of the p-dimensional random variable $\mathbf{Y}$ in the ith unit, and $\tilde{y}_{i1}, \ldots, \tilde{y}_{ip}$ the corresponding values after imputation. As already stated, the performance of an imputation method is measured in terms of preservation of means and of covariance matrix.

The preservation of the mean is measured through the relative root mean square error

$$D_{m_j} = \sqrt{\frac{1}{100}\sum_{t=1}^{100}\frac{(m_j^{(t)} - \tilde{m}_j^{(t)})^2}{(m_j^{(t)})^2}}, \quad j = 1, \ldots, p,$$

where $\tilde{m}_j^{(t)}$ is the mean of variable Y_j computed on the imputed dataset in the tth experiment, and $m_j^{(t)}$ is the mean computed on the original values.

An overall evaluation index can be obtained by the indicator

$$D_m = \sum_{j=1}^{p} D_{m_j}.$$

The preservation of the covariance structure is measured by computing for each pair of variables Y_j and Y_k the following quantities:

$$d_{jk} = \sqrt{\frac{1}{100}\sum_{t=1}^{100}\frac{(s_{jk}^{(t)} - \tilde{s}_{jk}^{(t)})^2}{(s_{jk}^{(t)})^2}}, \quad j = 1, \ldots, p,\ k = 1, \ldots, p,$$

where $s_{jk}^{(t)}$ and $\tilde{s}_{jk}^{(t)}$ are the corresponding elements of the sample covariance matrices S and $\tilde{S}$ computed on the original and imputed data, respectively, in the tth experiment. In order to provide an overall evaluation index, the quantities d_{jk} are summarized by the index

$$D_S = \sum_{j=1}^{p}\sum_{k=j}^{p} d_{jk},$$

providing a measure for the variance and covariance structure preservation.

Table 1 Results of the indices D_m and D_S computed on the simulations based on lognormal data with sample size 500 (LN 500) and 1000 (LN 1000)

LN 500	D_m	D_S	LN 1000	D_m	D_S
SPMM	0.0002	0.0037	SPMM	0.0001	0.0018
NND	0.0003	0.0045	NND	0.0001	0.0023
MCM	0.0001	0.0049	MCM	0.0001	0.0045
MRD	0.0002	0.0028	MRD	0.0001	0.0011

Table 2 Results of the indices Dm and DS computed on the simulations based on Multivariate Gamma MGL with sample size 500 (MGL 500) and 1000 (MGL 1000)

MGL 500	D_m	D_S	MGL 1000	D_m	D_S
SPMM	0.0008	0.0813	SPMM	0.0004	0.0348
NND	0.0007	0.0871	NND	0.0004	0.0434
MCM	0.0004	0.0573	MCM	0.0002	0.0422
MRD	0.0006	0.0499	MRD	0.0003	0.0190

Table 3 Results of the indices D_m and D_S computed on the simulations based on Multivariate Gamma MGH with sample size 500 (MGH 500) and 1000 (MGH 1000)

MGH 500	D_m	D_S	MGH 1000	D_m	D_S
SPMM	0.0014	0.0881	SPMM	0.0007	0.0332
NND	0.0013	0.0715	NND	0.0007	0.0319
MCM	0.0006	0.0545	MCM	0.0003	0.0659
MRD	0.0009	0.0434	MRD	0.0005	0.0250

Table 4 Results of the indices Dm and DS computed on the experiment CLAV

CLAV	D_m	D_S
SPMM	0.0014	0.0881
NND	0.0013	0.0715
MCM	0.0006	0.0545
MRD	0.0009	0.0434

5 Results

The results of the experiments concerning the lognormal distribution are reported in Table 1, while Tables 2 and 3 contain the results referring to the multivariate Gamma distribution with low and high correlation, respectively. Finally, Table 4 shows the values of the indicators related to the experiment carried out on the Labour Cost data.

The results show that the preservation of the mean is similar in all the methods, although it can be noticed, as it was expected, a better behavior of MCM that is based on imputation of conditional means. Concerning the preservation of the covariance matrix, there is more difference among the methods. The best one is the imputation based on random draw from the estimated conditional probability distribution (MRD).

It is interesting to compare NND with SPMM, since the latter can be interpreted as a nearest neighbor donor with a particular distance. When there are some variables

with low correlations, the behavior of the SPMM is better (Tables 1 and 2). When the variables are correlated, the behavior is similar with a slight preference for NND. This can be explained by the fact that the distance used in SPMM is based on the conditional expected values estimated through an explicit model. Thus, SPMM takes into account the different influences of the covariates on the response variables, while NND treats all the covariates at the same way, unless different weights are assigned to different variables in the distance function. However, in the latter case, it is evident how difficult is to assign a weight to the variable. This difficulty is also increased by the fact that the weights should change according to the missing data pattern. In other words the SPMM can be broadly also considered as a distance computation assigning different weights to the covariates, where the weights vary according to the missing data pattern. On the other hand, when correlation among the variable is high, we can say that all the covariates explain the response variables, and imputation based on a real distance function, instead of predictive means, results in a better estimation of the conditional probability distribution.

The results suggest the random draw from the model as the best method to use. However, it is also worthwhile to remark an important characteristic of the SPMM. This method imputes only "live" values, thus avoiding strange "synthetic" values, for instance, the imputation of negative values when the variables are nonnegative. Hence, this method is particularly appealing whenever micro data must be released.

Acknowledgements The views expressed by the authors do not necessarily reflect the policy of Istituto Nazionale di Statistica.

References

Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. B **39**, 1–38 (1977)

Di Zio, M., Guarnera, U., Luzi, O.: Imputation through finite mixture models. Comput. Stat. Data Anal. **51**, 5305–5316 (2007)

Durrant, G.B., Skinner, C.: Using missing data methods to correct for measurement error in a distribution function. Surv. Methodol. **32**, 25–36 (2006)

Fraley, C., Raftery, E.: Model-based clustering, discriminant analysis, and density estimation. J. Am. Stat. Assoc. **97**, 611–629 (2002)

Hunt, L., Jorgensen, M.: Mixture model clustering for mixed data with missing information. Comput. Stat. Data Anal. **41**, 561–575 (2003)

Kotz, S., Balakrishnan, N., Johnson, N.L.: Continuous Multivariate Distributions, vol. 1, 2nd edn. Wiley, New York (2000)

Little, R.J.A.: Missing-data adjustments in large surveys. J. Bus. Econ. Stat. **6**, 287–296 (1988)

Little, J., Rubin, D.: Statistical Analysis with Missing Data. Wiley, New York (2002)

Marron, S., Wand, M.: Exact Mean Integrated Squared Error. Ann. Stat. **20**, 712–736 (1992)

McLachlan, G., Peel, D.: Finite Mixture Models. Wiley, New York (2000)

Roeder, K., Wasserman, L.: Practical density estimation using mixtures of normals. J. Am. Stat. Assoc. **92**, 894–902 (1997)

Schafer, J.L.: Analysis of Incomplete Multivariate Data. Chapman & Hall, London (1997)

Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**, 461–464 (1978)