Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2007) 5305–5316
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Imputation through finite Gaussian mixture models

Marco Di Zio*, Ugo Guarnera, Orietta Luzi

Istituto Nazionale di Statistica, via Cesare Balbo 16, 00184 Roma, Italy

Available online 30 October 2006

Abstract

Imputation is a widely used method for handling missing data. It consists in the replacement of missing values with plausible ones. Parametric and nonparametric techniques are generally adopted for modelling incomplete data. Both of them have advantages and drawbacks. Parametric techniques are parsimonious but depend on the model assumed, while nonparametric techniques are more flexible but require a high amount of observations. The use of finite mixture of multivariate Gaussian distributions for handling missing data is proposed. The main reason is that it allows to control the trade-off between parsimony and flexibility. An experimental comparison with the widely used imputation nearest neighbour donor is illustrated.

Keywords: Incomplete data; Imputation; Nearest neighbour donor

1. Introduction

The presence of missing values in statistical survey data is an important issue to deal with (Little and Rubin, 2002). Imputation is commonly used for the treatment of missing items: it consists in the replacement of the missing values with plausible ones. As stated by Marker et al. (2002), the main challenges in the field of imputation are: (1) to maximise the use of available data in order to minimise the mean square error for univariate statistics and to preserve covariance structures in multivariate data sets; (2) to include in the variance estimates the uncertainty caused by the use of imputed data, i.e. synthetic (not really observed) data (Rubin, 1987). In fact, imputation may seriously affect the statistical properties of both univariate and joint data distributions and lead to severe underestimation of the variance of the target estimates, if standard methods are used for variance estimation considering imputed data as they were really observed.

In the area of imputation, several parametric and nonparametric techniques have been proposed for modelling incomplete data and compensating for nonresponse. Both approaches have advantages and drawbacks. Parametric techniques are parsimonious but depend on the model assumed, while nonparametric techniques are more flexible since they do not require model fitting, but, being generally justified by asymptotic arguments, they require a high amount of observations.

These reasons suggest exploring semiparametric imputation methods. One possible approach is based on the use of finite Gaussian mixture models (McLachlan and Peel, 2000). Finite mixtures of Gaussian distributions are a powerful tool for statistical modelling in a wide variety of situations, as shown in Fraley and Raftery (2002) and in Marron and Wand (1992) where the authors show that many probability distributions may be well approximated by finite mixture models. The mixtures combine the advantages of both parametric and nonparametric methods. As already mentioned,

* Corresponding author. Tel.: +39 4673 2871; fax: +39 06 4673 2955.
E-mail address: dizio@istat.it (M. Di Zio).

0167-9473/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.10.002

they do not restrict to a specific functional form, allowing to model a large class of distributions. However, by contrast to the nonparametric case, the complexity of the model grows only with the complexity of the data structure, instead of merely of the data set size. For instance, in Priebe (1994) it is shown how, with 10 000 observations, a lognormal density may be well approximated by a mixture of 30 Gaussian components.

In this paper, the use of finite mixtures of multivariate Gaussian distributions for imputation is proposed. In particular, we investigate their performance in terms of preservation of mean and variance–covariance structure of the observed data. The evaluation is performed through a comparison with one of the most commonly used nonparametric imputation techniques, the nearest neighbour donor method (NND). The NND belongs to the class of hot-deck methods, which have long been used at statistical agencies to impute missing data in statistical surveys (Kalton and Kasprzyk, 1986). The comparative evaluation is carried out through iterative simulation experiments on both artificial and real data.

The paper is structured as follows. In Section 2 the finite mixture models and the algorithm used to estimate the parameters in the presence of missing data are introduced. The use of finite mixture models for imputation is illustrated in Section 3. In Section 4 the simulation experiments on artificial data are described, while the results are discussed in Section 5. Section 6 is devoted to the description of the results of an application to real data. Finally, concluding remarks about benefits, limits, and open problems are provided in Section 7.

2. Finite mixture models for missing data

The use of mixture models for imputation requires the estimation of model parameters in presence of missing data. The algorithm proposed for the estimation is that of Hunt and Jorgensen (2003), and is based on the maximum-likelihood estimates (MLE) via EM algorithm. The algorithm is detailed in the following.

Let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ be a random sample from the p-dimensional r.v. $\mathbf{Y}$ distributed as a finite mixture of K Gaussian distributions

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

where $\sum_k \pi_k = 1$, $\pi_k \geqslant 0$ for $k = 1, \ldots, K$, and $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. Note that $\boldsymbol{\Phi}$ denotes the full set of parameters of the mixture model: $\boldsymbol{\Phi} = (\pi_1, \ldots \pi_K; \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$. Let us introduce the vector of indicator variables $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})'$ where z_{ik} is 1 if the ith individual belongs to group k, and 0 otherwise.

In case of partially incomplete data, for each unit i it results $\mathbf{y}_i = (\mathbf{y}_{\text{mis},i}, \mathbf{y}_{\text{obs},i})$, where $(\mathbf{y}_{\text{obs},i})$ are the observed variables, and $(\mathbf{y}_{\text{mis},i})$ the missing ones.

The EM algorithm consists in defining some initial guess for the parameters to be estimated, and iteratively applying until convergence the Expectation step (E-step) and the Maximisation step (M-step) described in the following.

2.1. E-step

This step consists, at each iteration t, in computing the expectation of the complete data likelihood conditional on the observed data, using the current estimates of the parameters $\hat{\boldsymbol{\Phi}}^{(t)}$. This requires the computation of the following expected values:

$$\hat{\tau}_{ik}^{(t)} = E(z_{ik}|\mathbf{y}_{\text{obs},i}; \hat{\boldsymbol{\Phi}}^{(t)}) = \frac{\hat{\pi}_k^{(t)} N_k(\mathbf{y}_{\text{obs},i}; \hat{\boldsymbol{\theta}}_k^{(t)})}{\sum_{k=1}^{K} \hat{\pi}_k^{(t)} N_k(\mathbf{y}_{\text{obs},i}; \hat{\boldsymbol{\theta}}_k^{(t)})};$$

$$E(z_{ik} y_{ij}|\mathbf{y}_{\text{obs},i}; \hat{\boldsymbol{\theta}}_k^{(t)}),\ E(z_{ik} y_{ij} y_{ij'}|\mathbf{y}_{\text{obs},i}; \hat{\boldsymbol{\theta}}_k^{(t)}).$$

Note that the first expected value $\hat{\tau}_{ik}^{(t)}$ is the estimated posterior probability (at iteration t) that the ith observation belongs to the kth group. It corresponds to the usual E-step in the EM algorithm for Gaussian mixture models with complete data, provided that the full vector $\mathbf{y}_i$ is replaced by the observed data $\mathbf{y}_{\text{obs},i}$. The other two expectations are analogous to those required in the standard EM algorithm for incomplete normal data and can be easily computed using the sweep operator (Schafer, 1997).

2.2. M-step

In this step new parameter estimates $\hat{\boldsymbol{\Phi}}^{(t+1)}$ are obtained by maximizing the expected likelihood estimated in the E-step.

The mixing proportions are given by

$$\hat{\pi}_k^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} \hat{\tau}_{ik}^{(t)}, \quad k = 1, \ldots, K.$$

The multivariate normal parameters are obtained through

$$\hat{\mu}_{kj}^{(t+1)} = \frac{1}{n\hat{\pi}_k^{(t+1)}} E\left(\sum_{i=1}^{n} \hat{\tau}_{ik}^{(t)} y_{ij} | \mathbf{y}_{\text{obs},i}, \hat{\boldsymbol{\theta}}_k^{(t)}\right),$$

$$\hat{\Sigma}_{kjj'}^{(t+1)} = \frac{1}{n\hat{\pi}_k^{(t+1)}} E\left(\sum_{i=1}^{n} \hat{\tau}_{ik}^{(t)} y_{ij} y_{ij'} | \mathbf{y}_{\text{obs},i}, \hat{\boldsymbol{\theta}}_k^{(t)}\right) - \hat{\mu}_{kj}^{(t+1)} \hat{\mu}_{kj'}^{(t+1)}.$$

The algorithm so far described concerns the estimation of an heteroscedastic mixture model with a given number K of components. This is the most complex model in the set of Gaussian mixture models. Simpler models (e.g., homoscedastic mixture models) could be obtained by introducing suitable constraints on the covariance structure. There is a trade off between the number of components K and the complexity of the model: the more complex is the model, the lower is the number of components that are needed to provide a good representation of the data. In this paper, only heteroscedastic mixture models are considered, thus the model selection reduces to the choice of the number of components. Our approach to this problem is based on the maximisation of the Bayesian posterior model probability through the BIC (Bayesian Information Criterion) approximation (Schwarz, 1978). Given the MLE of the parameters of a K-component model, BIC is defined as $2L(\hat{\boldsymbol{\Phi}}_K) - v_K \log n$, where L is the log-likelihood function based on n observations, $\hat{\boldsymbol{\Phi}}_K$ are the MLEs for the K-component model, and v_K is the number of independent parameters to be estimated.

3. Imputation

Various ways of using models for imputation are described in Little and Rubin (2002). Among them, the most relevant are *Conditional mean imputation* and *Random draws imputation*.

Conditional mean imputation consists in imputing predictions from a very general regression on observed values. For instance, this group of methods includes the linear regression imputation, and imputation with means within cells. In the latter case, the dummy indicator variables for the imputation cells can be considered as regressor variables. Concerning random draws, missing values are replaced by predicted values drawn from the probability distribution of the missing items given the observed ones. Conditional mean imputation techniques have been developed in both parametric and nonparametric contexts. Among the others, a nonparametric conditional mean imputation method is proposed in Nielsen (2001). Despite of its good performances when the target of the survey is the estimation of linear statistics such as means or totals, conditional mean imputation can determine serious bias in estimating non-linear quantities and, even more, multivariate distributions. In order to preserve the distributional features, imputation by random draws from the estimated distribution is commonly used.

According to these general practices, we propose the following imputation strategies.

First, estimate the mixture model parameters obtaining

$$f(\mathbf{y}_i; \hat{\boldsymbol{\Phi}}) = \sum_{k=1}^{K} \hat{\pi}_k N_k(\mathbf{y}_i; \hat{\boldsymbol{\theta}}_k).$$

Then, impute missing data by means of the two alternatives described in the following.

- *Conditional mean imputation*. Impute each missing vector $\mathbf{y}_{\text{mis},i}$ with the conditional expectation of the r.v. $\mathbf{Y}_{\text{mis},i}|\mathbf{Y}_{\text{obs},i}$, for $i = 1, \ldots, n$ w.r.t. $f(\mathbf{y}_{\text{mis},i}|\mathbf{y}_{\text{obs},i}, \hat{\boldsymbol{\Phi}}) = \sum_{k=1}^{K} \hat{\tau}_{ik} N_k(\mathbf{y}_{\text{mis},i}|\mathbf{y}_{\text{obs},i}, \hat{\boldsymbol{\theta}}_k)$, i.e., with the weighted

sum $\sum_{k=1}^{K} \hat{\tau}_{ik} E_k(\mathbf{y}_{\text{mis},i}|\mathbf{y}_{\text{obs},i}, \hat{\boldsymbol{\theta}}_k)$ of the predictions $E_k(\mathbf{y}_{\text{mis},i}|\mathbf{y}_{\text{obs},i}, \hat{\boldsymbol{\theta}}_k)$ from each multivariate conditional normal distribution $N_k(\mathbf{y}_{\text{mis},i}|\mathbf{y}_{\text{obs},i}, \hat{\boldsymbol{\theta}}_k)$.

- *Random draw*. Draw a value $\mathbf{y}_{\text{mis},i}$ from the distribution of $f(\mathbf{y}_{\text{mis},i}|\mathbf{y}_{\text{obs},i}, \hat{\boldsymbol{\Phi}}) = \sum_{k=1}^{K} \hat{\tau}_{ik} N_k(\mathbf{y}_{\text{mis},i}|\mathbf{y}_{\text{obs},i}, \hat{\boldsymbol{\theta}}_k)$, for $i = 1, \ldots, n$.

In practice this is accomplished by drawing a value k from the multinomial distribution $Mult_K(1; \hat{\tau}_{i1}, \ldots, \hat{\tau}_{iK})$ and then, given k, generating from the multivariate conditional normal distribution $N_k(\mathbf{y}_{\text{mis},i}|\mathbf{y}_{\text{obs},i}, \hat{\boldsymbol{\theta}}_k)$.Hereafter, for the sake of simplicity, the imputation by conditional expectation and random draw through mixture models will be referred to as MCM and MRD, respectively.

Imputation method based on mixtures is compared with the NND method which is the most frequently used technique in the practice of Statistical Institutes. The NND technique belongs to the wide family of the so-called hot-deck methods. These methods consist in matching completely observed units (donors) with units having some missing items (recipients), and transferring values from donors to recipients.

The match can be either random or based on some distance function computed on some set of covariates (matching variables). In the latter case the method is the NND. Hot-deck methods are generally preferred to other imputation techniques because of low operational cost, reduced nonresponse bias on univariate statistics, univariate plausibility (i.e., use of 'live' values). On the other hand, donor-based imputation can produce attenuation of associations (Kalton and Kasprzyk, 1986).

In order to assess the performance of MCM, MRD and NND, an empirical evaluation has been carried out analysing their behaviour in simulative contexts based on artificial data and on a real data set.

4. Experiments on artificial data

In this section the simulation study carried out in order to evaluate the performance of finite Gaussian mixture models for dealing with partially incomplete data is described.

The objective of the experiments is to evaluate the performance of the mixture model approach in terms of preservation of univariate statistics and covariance structure of the data. To this aim a comparison between imputation based on mixture models and NND is performed. The donors are chosen among observations without any missing items, and the similarity is evaluated according to a distance measure computed considering only the variables that are observed in the recipient. The adopted distance is the Euclidean one, and the variables are standardised in order to avoid scale problems.

The considered imputation methods are evaluated in several simulative contexts differing with respect to data generating distribution (Gaussian and non-Gaussian), sample size, and missing data mechanism, i.e., missing completely at random (MCAR) and missing at random (MAR) (Little and Rubin, 2002).

For each experimental setting, 1000 simulations have been performed consisting of the following steps: (1) artificial generation of a sample from a given multivariate data distribution; (2) artificial generation of missing values in the sample; (3) estimation and imputation; (4) evaluation of the imputation by comparing the imputed data with the original ones. All the experiments have been developed using SAS/IML software, Version 8.2 of the SAS System for Windows.

As far as the Gaussian experiments are concerned, two different cases are considered. In the first, data are characterised by medium and high correlation, while in the second an extreme situation is considered where the variables are grouped in two blocks, independent one of each other, but with high correlation within the blocks. Since in the MAR experiment the missing values are generated depending on an always observed variable, the second framework allows to have in the same data set some variables with MCAR mechanism and other affected by MAR mechanism. In both settings, the numerical values of the parameters are obtained by means of real data sets.

As far as the non-Gaussian experiments are concerned, the generating distribution is a multivariate Gamma with correlation structure similar to the one used in the second Gaussian framework, i.e., two blocks (almost uncorrelated) with high correlation within them.

In the following, it is given a detailed description of the four steps referring to the sample data generation, the nonresponse simulation, the estimation and imputation, and the evaluation.

(1) *Sample data generation:*

(i) Normal case (first experiment—G1). Data are drawn from a 5-variate Gaussian random vector $(Y_1, \ldots, Y_5)$ with mean vector

$$\boldsymbol{\mu} = (6, 3, 2, 2, 6)',$$

and covariance matrix

$$\Sigma = \begin{pmatrix} 8.0 & 4.5 & 3.5 & 3.5 & 4.0 \\ 4.5 & 8.4 & 4.6 & 4.0 & 4.5 \\ 3.5 & 4.6 & 3.7 & 3.0 & 3.5 \\ 3.5 & 4.0 & 3.0 & 6.7 & 3.2 \\ 4.0 & 4.5 & 3.5 & 3.2 & 4.0 \end{pmatrix}.$$

(ii) Normal case (second experiment—G2). Data are drawn from a 5-variate Gaussian random vector $(Y_1, \ldots, Y_5)$ having the sub-vectors of the first two components (Y_1, Y_2) and the last three (Y_3, Y_4, Y_5) independent and normally distributed with parameters $(\boldsymbol{\mu}_{(1)}, \Sigma_{(1)})$ and $(\boldsymbol{\mu}_{(2)}, \Sigma_{(2)})$ respectively, where

$$\boldsymbol{\mu}_{(1)} = (-2.5, -2.6)', \quad \boldsymbol{\mu}_{(2)} = (-5.5, -7.6, -6.0)',$$

$$\Sigma_{(1)} = \begin{pmatrix} 3.1 & 2.7 \\ 2.7 & 2.8 \end{pmatrix}, \quad \Sigma_{(2)} = \begin{pmatrix} 3.1 & 2.4 & 2.4 \\ 2.4 & 3.0 & 2.1 \\ 2.4 & 2.1 & 3.0 \end{pmatrix}.$$

(iii) Non-normal case (multivariate Gamma distribution—NG). Data are drawn through a slight modification of 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 such a distribution the following procedure is adopted. First, 7 independent random variables 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 $(Y_1, \ldots, Y_5)$ is obtained combining the X_i in the following way:

$$Y_1 = X_1 + X_3; \quad Y_2 = X_1 + X_4; \quad Y_3 = X_1 + X_2 + X_5;$$

$$Y_4 = X_2 + X_6; \quad Y_5 = X_2 + X_7.$$

Following Kotz et al. (2000), it is easy to compute the expected value and the correlation matrix of the r.v.s Y_i. The parameters θ_i are chosen to obtain a correlation structure similar to that of the first Gaussian experiment, i.e., two independent or 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 1000 observations from this distribution is shown in Fig. 1.

Finally, samples of 300 and 1000 units have been generated for both the normal and non-normal settings.

(2) *Nonresponse simulation*. Once a sample of complete data is generated, item nonresponse is simulated according to both MCAR and MAR mechanisms. In the MCAR simulations, sub-samples of values are randomly selected and dropped for the variables (Y_1, Y_2, Y_3, Y_4) according to the percentages 20%, 25%, 30%, 40% respectively.

In the MAR experiments, missing items are introduced for the variables (Y_1, Y_2, Y_3, Y_4) depending on the observed values y_5 of the variable Y_5 under the assumption that the higher is the value of Y_5, the higher is the nonresponse propensity. More in detail, denoting by q_i the ith quartile of the empirical distribution of Y_5, the nonresponse probabilities for (Y_2, Y_3, Y_4) are 0.1 if $y_5 < q_1$, 0.2 if $y_5 \in [q_1, q_2)$, 0.5 if $y_5 \in [q_2, q_3)$ and 0.6 if $y_5 \geqslant q_3$. For the variable Y_1 a more critical situation in terms of response rate has been chosen. Nonresponse probabilities are 0.1 if $y_5 < q_1$, 0.2 if $y_5 \in [q_1, q_2)$, 0.4 if $y_5 \in [q_2, q_3)$ and 0.9 if $y_5 \geqslant q_3$.

(3) *Estimation and imputation*. The incomplete sample of data is imputed by using MCM, MRD and NND. Concerning finite mixtures, models with different number of components have been estimated and used for imputation following

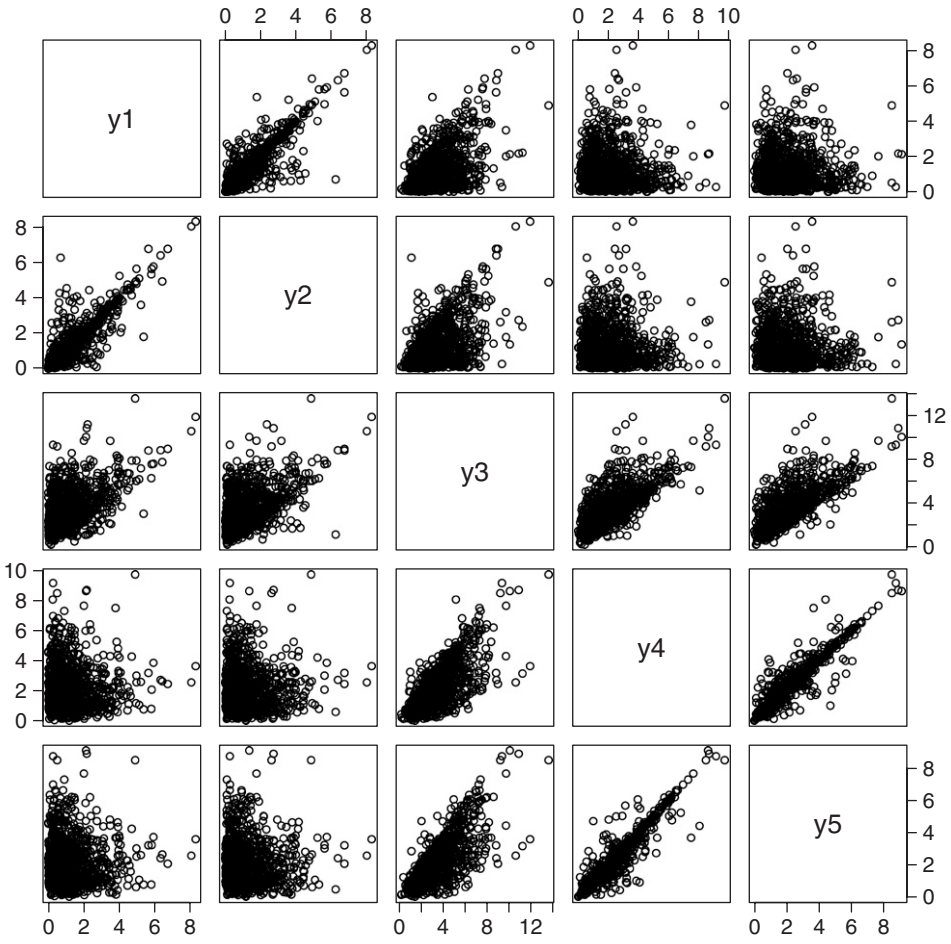

Fig. 1. The scatter-plot matrix of a sample of 1000 observations drawn from the multivariate Gamma NG.

the algorithm described in Section 2. Starting points for the EM algorithm have been identified by clustering data by means of the k-means algorithm, and considering as starting parameters the sample means and the sample covariance matrix within the clusters. For the parameter $\hat{\pi}_k$, the starting point is the relative frequency of the kth cluster.

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 mixture 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. Once the parameters have been estimated for all the models, the 'best model' is chosen based on BIC, and the selected model is used to impute missing values following the two methods described in Section 2.

(4) *Evaluation*. As already stated, the performance of the different imputation methods is measured in terms of preservation of means and covariance structure. The evaluation is based on the comparison of the original sample data and the imputed ones. In order to perform this comparison, for each iteration of the simulation experiments, sample means and sample covariance matrices are computed for both original and imputed data. Differences between these statistics are then used to build the performance indicators described in the following.

Let $y_{i1}, \ldots, y_{ip}$ $(i = 1, \ldots, n)$ be the original 'true' values of the p-dimensional r.v. $\mathbf{Y}$ in the ith unit. Let $y^*_{i1}, \ldots, y^*_{ip}$ be the corresponding imputed values, i.e., the values of variables after imputation. Also, let $m_j^{(t)}$, $s_{jk}^{(t)}$ be the sample means and the elements of the sample covariance matrix respectively, computed on original data at the tth simulation, $(j, k = 1, \ldots, p)$. Finally, denote by $m_j^{*(t)}$, $s_{jk}^{*(t)}$ the corresponding quantities referred to the imputed data set.

The preservation of the mean is evaluated in terms of both relative bias (B_{m_j}) and relative root mean squared error (R_{m_j}) defined as follows:

$$B_{m_j} = \frac{1}{1000}\sum_{t=1}^{1000}\frac{m_j^{*(t)} - m_j^{(t)}}{|m_j^{(t)}|}, \quad j = 1, \ldots, p,$$

$$R_{m_j} = \sqrt{\frac{1}{1000}\sum_{t=1}^{1000}\frac{(m_j^{*(t)} - m_j^{(t)})^2}{m_j^{(t)2}}}, \quad j = 1, \ldots, p.$$

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

$$d_{jk} = \sqrt{\frac{1}{1000}\sum_{t=1}^{1000}\frac{(s_{jk}^{(t)} - s_{jk}^{*(t)})^2}{s_{jk}^{(t)2}}}, \quad j = 1, \ldots, p, \quad k = 1, \ldots, p$$

and building the overall evaluation indices

$$D_V = \sum_{j=1}^{p} d_{jj}, \quad D_C = \sum_{j=1}^{p}\sum_{k=j>k}^{p} d_{jk}.$$

The last indices provide measures for the variance and covariance preservation, respectively.

A further index for the evaluation of covariance structure is obtained by counting, over the 1000 simulations, the number of times that each method gives the best (the lowest) value of $D_V + D_C$. The corresponding index is denoted by $\%(D_S)$.

5. Results

The experiments described in the previous section allow us to analyse different aspects concerning the use of mixture models for imputation. An important issue is the comparison of mixture models with NND. Moreover, the two alternative ways of using mixture for imputation, MCM and MRD, can be comparatively evaluated. The comparison is done by assessing the performances with respect to the preservation of means and covariance matrix. The evaluation of these aspects is enriched by the fact that the experiments have been performed varying also the sample size and the missing mechanism. The results are shown in Tables 1–6.

Each table shows the results for the two different missing mechanism, MCAR and MAR.

Table 1 refers to the Gaussian experiment G1 with sample size 300. Table 2 is the same as the previous but with sample size 1000. Tables 3 and 4 concern the Gaussian experiment G2 with sample size 300 and 1000, respectively.

Table 1
Results for the experiment G1, sample size 300

Imp	B_{m_1}	B_{m_2}	B_{m_3}	B_{m_4}	R_{m_1}	R_{m_2}	R_{m_3}	R_{m_4}	D_V	D_C	$\%(D_S)$
MCAR											
NND	−0.001	0.000	0.000	0.000	0.018	0.019	0.010	0.010	0.776	1.514	18.5
MCM	0.000	0.000	0.000	0.000	0.012	0.011	0.006	0.007	1.269	0.922	3.3
MRD	−0.001	−0.001	0.000	0.000	0.016	0.016	0.008	0.008	0.584	1.059	78.2
MAR											
NND	−0.004	−0.004	0.026	0.015	0.105	0.082	0.037	0.026	2.243	4.920	0.5
MCM	−0.002	−0.001	0.000	0.000	0.029	0.024	0.008	0.007	2.152	1.687	0.2
MRD	−0.003	−0.002	0.000	0.000	0.036	0.030	0.010	0.009	0.744	1.565	99.3

Table 2
Results for the experiment G1, sample size 1000

Imp	B_{m_1}	B_{m_2}	B_{m_3}	B_{m_4}	R_{m_1}	R_{m_2}	R_{m_3}	R_{m_4}	D_V	D_C	$\%(D_S)$
MCAR											
NND	−0.001	0.000	0.000	0.000	0.010	0.010	0.005	0.005	0.429	0.790	21.7
MCM	0.000	0.000	0.000	0.000	0.006	0.006	0.003	0.004	1.216	0.584	0.0
MRD	0.000	0.000	0.000	0.000	0.009	0.009	0.004	0.005	0.318	0.578	78.3
MAR											
NND	−0.001	0.000	0.017	0.009	0.071	0.057	0.025	0.018	1.600	3.578	0.1
MCM	−0.001	−0.001	0.000	0.000	0.016	0.013	0.004	0.004	2.114	1.192	0.0
MRD	−0.001	−0.001	0.000	0.000	0.019	0.016	0.005	0.005	0.414	0.844	99.9

Table 3
Results for the experiment G2, sample size 300

Imp	B_{m_1}	B_{m_2}	B_{m_3}	B_{m_4}	R_{m_1}	R_{m_2}	R_{m_3}	R_{m_4}	D_V	D_C	$\%(D_S)$
MCAR											
NND	0.000	0.000	0.000	0.002	0.015	0.029	0.023	0.075	1.744	3.049	30.5
MCM	0.000	0.000	0.000	0.000	0.010	0.019	0.014	0.049	3.432	1.789	0.8
MRD	0.000	0.001	0.000	0.002	0.012	0.025	0.018	0.061	1.421	2.340	68.7
MAR											
NND	−0.047	−0.069	−0.092	−0.064	0.077	0.113	0.112	0.155	5.729	11.594	0.9
MCM	0.001	−0.001	−0.001	0.003	0.021	0.028	0.018	0.049	4.443	2.501	3.5
MRD	0.001	0.000	−0.001	0.001	0.024	0.034	0.022	0.060	1.877	3.018	95.6

Table 4
Results for the experiment G2, sample size 1000

Imp	B_{m_1}	B_{m_2}	B_{m_3}	B_{m_4}	R_{m_1}	R_{m_2}	R_{m_3}	R_{m_4}	D_V	D_C	$\%(D_S)$
MCAR											
NND	0.000	0.000	0.000	0.002	0.008	0.016	0.011	0.038	0.930	1.619	31.3
MCM	0.000	0.000	0.000	0.001	0.005	0.010	0.007	0.025	3.312	1.002	0.0
MRD	0.000	0.001	0.000	0.000	0.007	0.014	0.010	0.033	0.775	1.274	68.7
MAR											
NND	−0.034	−0.042	−0.056	−0.044	0.058	0.073	0.071	0.114	4.444	8.724	0.3
MCM	0.000	0.000	0.000	0.000	0.011	0.015	0.010	0.026	4.327	1.414	0.0
MRD	0.000	0.000	0.000	0.000	0.013	0.019	0.012	0.032	1.43	1.680	99.7

Table 5
Results for the experiment NG, sample size 300

Imp	B_{m_1}	B_{m_2}	B_{m_3}	B_{m_4}	R_{m_1}	R_{m_2}	R_{m_3}	R_{m_4}	D_V	D_C	$\%(D_S)$
MCAR											
NND	−0.004	−0.002	0.000	−0.002	0.025	0.027	0.014	0.019	0.678	1.091	13.7
MCM	0.001	0.003	0.001	0.001	0.017	0.019	0.009	0.014	0.566	0.598	31.4
MRD	0.002	0.004	0.001	0.002	0.022	0.024	0.011	0.016	0.469	0.701	54.9
MAR											
NND	0.005	0.006	−0.054	−0.084	0.165	0.126	0.077	0.099	2.491	4.700	2.9
MCM	0.008	0.001	0.002	0.000	0.065	0.036	0.016	0.015	1.074	1.615	29.7
MRD	0.009	0.000	0.002	0.000	0.069	0.043	0.018	0.018	0.878	1.638	67.4

Table 6
Results for the experiment NG, sample size 1000

Imp	B_{m_1}	B_{m_2}	B_{m_3}	B_{m_4}	R_{m_1}	R_{m_2}	R_{m_3}	R_{m_4}	D_V	D_C	$\%(D_S)$
MCAR											
NND	0.000	0.000	0.002	0.001	0.013	0.015	0.007	0.010	0.341	0.522	24.1
MCM	0.000	0.000	0.000	0.000	0.009	0.010	0.005	0.007	0.482	0.294	2.4
MRD	0.000	0.000	0.000	0.000	0.012	0.013	0.007	0.008	0.236	0.349	73.5
MAR											
NND	0.006	0.002	−0.031	−0.052	0.119	0.088	0.052	0.063	1.892	3.641	0.7
MCM	0.006	0.001	0.001	0.000	0.029	0.018	0.008	0.008	0.952	0.893	8.3
MRD	0.006	0.002	0.001	0.000	0.032	0.022	0.010	0.009	0.441	0.820	91.0

Finally Tables 5 and 6 show results concerning the experiment in a non-Gaussian case (NG) described in detail in Section 4. Also in this case the two tables refer to the two different sample sizes.

As far as the preservation of the sample mean is concerned, the B_{m_j} and R_{m_j} indicators show that the methods provide similar results in the MCAR setting. All the methods are unbiased, and show similar values for the R_{m_j}. Despite the small differences, it is possible to observe that the behaviour of the MCM is always preferable to the others, while the worst is that of NND.

Under the MAR mechanism, the gain obtained by using the mixtures is apparent in particular for MCM. Analysing Tables 1, 2, 5 and 6 MCM and MRD outperform NND especially on the variables (Y_3 and Y_4) that are correlated with the variable used to build the MAR mechanism (Y_5). In fact, the other variables (Y_1 and Y_2) could be considered as affected only by an MCAR mechanism (see Section 4). Also in this case, MCM has the best performances while NND is the worst method.

The other results are interesting in order to analyse a characteristic referring to bivariate distributions, and in particular to the covariance matrix.

Analysing the $\%(D_S)$ indicator, MRD is always the best method, and in particular, differences are more appreciable when the mechanism is MAR. It is interesting to note that in the non-Gaussian experiment under the MAR setting, NND is almost never chosen as the best method.

Hence, the results show that under the MCAR mechanism there is a small advantage in the use of mixture models with respect to the NND. However, when MAR mechanism affects data, the imputation through mixture models provide remarkably better results than NND.

Among the mixture models methods, when the main objective of the survey is the estimation of mean or total, MCM is the most appropriate. On the other hand, when also covariance structure must be preserved, MRD seems to be preferable. In fact it outperforms MCM for the preservation of the covariance matrix while its performances, in terms of mean preservation, are close to those of MCM.

A final consideration concerns the use of the BIC score for choosing the best model. In the Gaussian experiments, BIC works satisfactorily, in fact the chosen model is always the multinormal one (1 component) for both the sample sizes. In the multivariate Gamma experiments the most frequently chosen model is the mixture of three components. When the sample size is 300, this model is chosen 82% of times in MCAR case, and 90% in MAR case. When sample size is 1000, the frequency is 99% of times in both the missing data mechanisms.

6. Application to real data

In order to illustrate the effectiveness and test the performance of our proposal, we carry out also an experiment on a subset of the *1997 Italian Labour Cost Survey* (LCS).

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. The survey is subject to a specific European Regulation requiring all the European Community Member States to collect every four years detailed information about the labour cost and employment structure in some specific Industries.

Our data set consists of 1000 units that belong to the metallurgic economic activity sector.

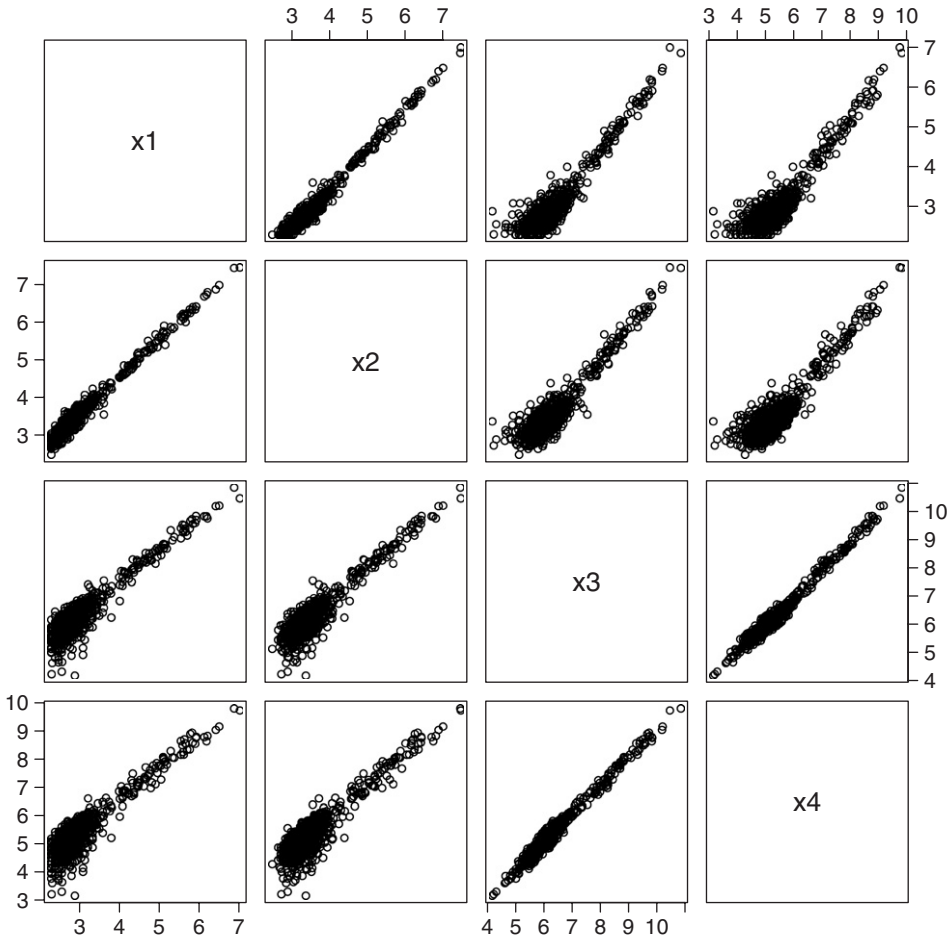

Fig. 2. The scatter-plot matrix of the 1000 units that belong to the metallurgic economic activity sector in the LCS.

In particular, we analyze four main variables measuring the *total number of employees* (X_1), the *total number of hours worked* (X_2), the *wages and salaries* (X_3), and the *total labour cost* (X_4). The values of the variables are obtained by means of a logarithmic transformation of the original data. Fig. 2 shows the scatter-plot matrix of the data used for the experiments .

In this situation, since the random generating mechanism of the r.v.s is unknown, a resampling approach has been adopted. The resampling scheme consists in sampling 1000 observations (through a simple random sampling with replacement) $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(1000)}$ (bootstrap sample) from the initial sample, where $\mathbf{x}^{(i)}$ represents the ith unit where the variables (X_1, X_2, X_3, X_4) are observed. The bootstrap sample can be thought of as generated from the estimated empirical distribution of (X_1, X_2, X_3, X_4). Similarly to the previous experiments (described in Section 4), missing values have been introduced only on the first three variables according to the MCAR and MAR mechanisms.

In particular, as far as the MCAR mechanism is concerned, the variables X_1, X_2, X_3 are affected by missing values with probabilities 0.25, 0.30 and 0.40, respectively. For the MAR mechanism, the missing items in the variables (X_1, X_2, X_3) are introduced according to the observed values x_4.

More in detail, denoting by q_i the ith quartile of the empirical distribution of X_4, the nonresponse probabilities for (X_2, X_3) are 0.1 if $x_4 < q_1$, 0.2 if $x_4 \in [q_1, q_2)$, 0.5 if $x_4 \in [q_2, q_3)$ and 0.6 if $x_4 \geqslant q_3$. For the variable X_1 a more critical situation in terms of response rate has been chosen. Nonresponse probabilities are 0.1 if $x_4 < q_1$, 0.2 if $x_4 \in [q_1, q_2)$, 0.4 if $x_4 \in [q_2, q_3)$ and 0.9 if $x_4 \geqslant q_3$. Once the missing values are introduced in the bootstrap sample, they are imputed by means of NND and the proposed methods based on mixtures.

Table 7
Results for the experiment LCS with sample size 1000

Imp	B_{m_1}	B_{m_2}	B_{m_3}	R_{m_1}	R_{m_2}	R_{m_3}	D_V	D_C	$\%(D_S)$
MCAR									
NND	0.000	0.000	0.000	0.001	0.001	0.001	0.025	0.036	21.2
MCM	0.000	0.000	0.000	0.001	0.001	0.000	0.020	0.018	24.5
MRD	0.000	0.000	0.000	0.001	0.001	0.001	0.015	0.021	54.3
MAR									
NND	−0.020	−0.012	−0.008	0.029	0.017	0.011	0.563	1.206	3.0
MCM	−0.002	−0.002	0.000	0.005	0.003	0.003	0.092	0.222	37.0
MRD	−0.002	−0.002	0.000	0.005	0.003	0.003	0.101	0.223	60.0

The results of the imputations are evaluated using the indices illustrated in Section 4. This procedure has been repeated 1000 times, and the results are averaged over them. The results are shown in Table 7.

They confirm the results obtained via the experiments illustrated in Section 5. When the missing data are MCAR, the performances of the three methods are similar, with a slight preference for MRD. When the missing data are MAR, the mixture methods outperform NND, both in preservation of sample means and sample covariance matrix. In particular, MRD is preferable to MCM because of its performances in the preservation of the covariance structure, and because its behaviour concerning the mean preservation is very close to that of MCM (indeed almost equal in this case).

It is worthwhile noting that, the model chosen by the BIC is almost always the one with 2 components.

7. Concluding remarks

The results discussed in the previous sections show that mixture models are an appealing method for imputing missing data and a valuable alternative to the NND. In particular imputation by random drawing (MRD) from a finite Gaussian mixture model is the best choice because it preserves better than the others both sample mean and covariance.

Although the experiments have produced satisfactory results, a number of important problems still remains open.

The NND technique has some properties that make it still a valid competitor. In fact, it allows the researcher to deal, at least from an operational point of view, with semicontinuous variables, i.e., variables whose probability distribution has a mass concentration at some points, for instance zero. This is a frequent case in the surveys carried out by Statistical Institutes, and further studies should be devoted to this topic.

Another important remark is about the indicators used in the experiments. They focus on some aspects, as the mean and the covariance matrix, that are extremely important in statistics. Actually, they are the quantities suggested to be taken into account by Marker et al. (2002) when they discuss about the main objectives of an imputation procedure. In fact, in the context of official statistics, means or totals and linear relationships between variables are often the main target quantities to be estimated. However, analysts might also be interested in other population characteristics such as quantiles or non-linear relationships involving moments of higher order of the data distributions. Thus, especially in a non-Gaussian context, other measures should be adopted. Further studies should be done also for this topic.

The last consideration is about the goal of this paper and future research direction. This paper is devoted to the study of a new imputation method. The assessment of its performance is carried out having as target the first issue of the main two challenges stated by Marker et al. (2002) and introduced in Section 1: “to maximize the use of available data in order to minimise the mean square error for univariate statistics and to preserve covariance structures in multivariate data sets”. The second issue, “to include in the variance estimates the uncertainty caused by the use of imputed data”, still remains an open problem. In fact, if the variance of the estimator is computed on the imputed data set, considering the imputed values as they were really observed, the source of uncertainty due to the fact that the values are artificial and not really observed is neglected. This leads to underestimate the variance and thus to p-values that are too high. Different approaches have been introduced for dealing with this problem. The main are multiple imputation (Rubin, 1987), and methods based on resampling techniques (Rao, 1996). The study of how to adapt mixture models as used in this paper to include this further source of randomness in variance estimation is an important issue to consider in the next studies.

Acknowledgements

This work has been partially supported by MIUR Grant PRIN2003. We would like to thank the Associate Editor and referees for their constructive comments.

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

References

Fraley, C., Raftery, A.E., 2002. Model-based clustering, discriminant analysis, and density estimation. J. Amer. Statist. Assoc. 97, 611–631.
Hunt, L., Jorgensen, M., 2003. Mixture model clustering for mixed data with missing information. Comput. Statist. Data Anal. 41, 429–440.
Kalton, G., Kasprzyk, D., 1986. The treatment of missing survey data. Survey Methodology 12, 1–16.
Kotz, S., Balakrishnan, N., Johnson, N.L. 2000. Continuous Multivariate Distributions, vol. 1, second ed. Wiley, New York.
Little, J., Rubin, D., 2002. Statistical Analysis with Missing Data. Wiley, New York.
Marker, D.A., Judkins, D.R., Winglee, M., 2002. Large-Scale Imputation for Complex Surveys. In: Groves, R.M., Dillman, D.A., Eltinge, E.L., Little, R.J.A. (Eds.), Survey Nonresponse. Wiley, New York.
Marron, J.S., Wand, M.P., 1992. Exact mean integrated squared error. Ann. Statist. 20, 712–736.
McLachlan, G.J., Peel, D., 2000. Finite Mixture Models. Wiley, New York.
Nielsen, S.F., 2001. Nonparametric conditional mean imputation. J. Statist. Plann. Inference 99, 129–150.
Priebe, C.E., 1994. Adaptive mixtures. J. Amer. Statist. Assoc. 89, 796–806.
Rao, J.N.K., 1996. On variance estimation with imputed survey data. J. Amer. Statist. Assoc. 91, 499–506.
Rubin, D.B., 1987. Multiple Imputation for Non-Response in Surveys. Wiley, New York.
Schafer, J.L., 1997. Analysis of Incomplete Multivariate Data. Chapman & Hall, New York.
Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist. 6, 461–464.