WP.25
ENGLISH ONLY

**STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Data Editing
(25-27 September 2006, Bonn, Germany)

Topic (v): New and emerging methods

A SEMIPARAMETRIC PREDICTIVE MEAN MATCHING: AN EMPIRICAL EVALUATION

Invited Paper

Prepared by Marco Di Zio and Ugo Guarnera, ISTAT, Italy

Abstract

Predictive mean matching is an imputation method that combines parametric and nonparametric techniques. It imputes missing values by means of the Nearest Neighbour Donor where the distance is computed on the expected values of the missing variables conditional on the observed covariates, instead of 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 allows to deal also with non linear relationships among the variables, and includes as a special case the original predictive mean matching. In order to assess its performance, an empirical evaluation based on simulations is carried out.

I. INTRODUCTION

1. 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 non-response by imputing artificial data. A big variety of imputation techniques have been introduced in literature and used by practitioners. They can be roughly divided in parametric and nonparametric. Parametric methods are generally parsimonious, but, being based on explicit models, they fail 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.

2. 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 us 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 a particular NND method with a suitable distance function. On the other hand, the function used in PMM is not a real distance function and the asymptotic properties of the NND are no longer guaranteed. Thus, the results of imputation via PMM still depend on the model, though the method is probably more robust than a fully model based approach, with respect to departures from the model assumptions.

3. In multivariate contexts, when the variables are numerical and with arbitrary patterns of missing items, a typical application of the PMM is the following.

a) 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).
b) 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).
c) 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.
d) Missing items are imputed in each recipient by transferring the corresponding values from its closest donor.

4. Although the previous procedure should be more robust than imputation based on standard linear regression, in the conditional distribution of the missing items given the observed ones, 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. 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 et al. 2002). As in the ordinary PMM however, the role of the model is only that of providing a suitable distance function to be used for nearest neighbor imputation. This approach, that could be defined as "semiparametric", allows handling data that are far from normality, keeping the advantage of imputing "live" values, that is, values that are really observed. The last characteristic ensures univariate plausibility. For instance, missing items for nonnegative variables are guaranteed to be imputed with nonnegative values. As will be clarified in the following, this semiparametric predictive mean matching is a generalized version of the standard Predictive Mean Matching. In fact, the latter is a special case of the proposed method.

5. The paper is organized as follows. In the next section general concepts and basic definitions are given on finite mixtures of Gaussian distributions, and the EM algorithm for the estimation of Gaussian mixtures in presence of missing values is briefly described. Section III illustrates the use of mixture models for imputation via PMM. Simulations are described in Section IV and conclusions follow in Section V.

II. ESTIMATION OF GAUSSIAN MIXTURES MODELS IN PRESENCE OF INCOMPLETE DATA

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

$$(1) \qquad f(\ ;\ ) = \sum_{k=1}^{K} \pi_k f_k(\ _k)$$

where the densities f_k belong to the same parametric family and the parameters π_k are positive and subject to the constraint $\sum_{k=1}^{K} \pi_k = 1$. Then, the model 1) is said to be a *mixture* of the distributions $f_{1,..}, f_K$ with *mixing proportions* $\boldsymbol{\pi} = (\pi_1,..,\pi_k)$ and the functions $f_{1,..}, f_K$ are called the *mixture components* (McLachlan and Peel, 2000). In formula (1) $\boldsymbol{\phi}$ denotes the full set of parameters $(\pi_1,..,\pi_k;\theta_1,..\theta_K)$. The distribution $f(\ ;\ )$ is a *Gaussian mixture* if the functions f_k are Gaussian densities: $f_k = N_k = N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, where $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$ denotes the normal density function with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

7. The maximum likelihood estimates (MLE) of the parameters $\boldsymbol{\phi}$ of a Gaussian mixture based on n observations $(\boldsymbol{y}_1,\ldots,\mathbf{y}_n)$ cannot be performed in closed form. The usual way to calculate the MLEs is to recast the problem as an incomplete data problem and use the EM algorithm (Laird et al., 1977). To this aim, each unit i (i =1,..,n) is supposed to belong to one of K groups corresponding to the K mixture components, and each group k (k=1,..,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 $\boldsymbol{Z}_i = (Z_{i1},\ldots,Z_{iK})$ is multinomially distributed: $Z_i \sim \text{Mult}_K(1, \boldsymbol{\pi})$ so that $\text{Prob}\{Z_{ik}=1\} = \pi_k$. The mixing proportion π_k (k =1,...,K) can be interpreted as the *a priori* probability of belonging to group k. Furthermore, from the Bayes formula, the probability $\tau_{ik} = \dfrac{\pi_k f_k(\ _i;\ _k)}{\sum_t \pi_t f_t(\ _i;\ _t)}$, ($i$ =1,…,n ; k =1,...,K) is the corresponding *posterior probability* given the observation $\boldsymbol{y}_i$ and its estimate can be expressed in terms of the estimates of parameters $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$. The "complete data" log-likelihood can be written as: $L_c(\boldsymbol{\Phi}, \boldsymbol{z}) = \sum_{k=1}^{K} \sum_{i=1}^{n} z_{ik}\{\log \pi_k + \log f_k(\ _i;\ _k)\}$ where z_{ik} is the realized value of the variable Z_{ik}. The E-step of the EM algorithm consists in calculating, at each iteration, the expected value of $L_c(\boldsymbol{\Phi}, \boldsymbol{z})$ conditional on $\boldsymbol{y}_i$ and the current estimates of the parameters. This reduces to compute the expectation of Z_{ik} , i.e. τ_{ik} , for i =1,..,n and k = 1,...,K. In the normal case, the M-step can also be performed in closed form, providing recursive equations for the parameters $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$.

8. In case of partially incomplete data the algorithm so far described has to be slightly modified in order to taken into account the missing items. Now the quantity to be maximized is the "observed-data log-likelihood" $L(\ _{obs};\ ) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k f_k(\ _{obs,i}\ _k)$ where, according to the usual notation, $\boldsymbol{y}_{obs,i}$ is the observed part of the vector $\boldsymbol{y}_i$ in the

decomposition $\boldsymbol{y}_i = (\boldsymbol{y}_{obs,i}, \boldsymbol{y}_{mis,i})$. The modified 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 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 proportion of units belonging to different clusters is taken as starting values of the parameters $\boldsymbol{\pi}$, while the parameters $(\mu_k ; \boldsymbol{\Sigma}_k)$ are initialized with the sample mean and the sample covariance matrix of each cluster.

III. PMM VIA GAUSSIAN MIXTURES MODELS

9. 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 fact, in many problems of model selection, the BIC score can well approximate the Bayesian posterior model probability (Schwartz, 1978). For a given number *K* of mixture components, BIC is defined as $2L(\hat{\Phi}_K ; y_{obs}) - \nu_K \log n$ where $L(\hat{\Phi}_K ; y_{obs})$ is the observed log-likelihood based on all the observations (complete and incomplete), $\hat{\Phi}_K$ 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.

10. Once the model that best fits data is selected and its parameters are estimated, for each incomplete observation $\boldsymbol{y}_i = (\boldsymbol{y}_{obs,i}, \boldsymbol{y}_{mis,i})$, the conditional distributions $f(y_{mis,i} \mid y_{obs,i}, \boldsymbol{\Phi})$ can also been estimated as: $f(y_{mis,i} \mid y_{obs,i}, \hat{\ }) = \sum_{k=1}^{K} \hat{\tau}_{ik} N_k(y_{mis,i} \mid y_{obs,i}, \hat{\theta}_k)$. One possible approach (Di Zio et al. 2006a, Di Zio et al. 2006b) could be that of imputing $\boldsymbol{y}_{mis,i}$ with the expected value (hereafter MCM) or a random draw from the conditional distribution (MRD). In this paper, where the PMM strategy is followed, the conditional mean from the distribution $f(y_{mis,i} \mid y_{obs,i}, \hat{\Phi})$ is only used to find a nearest neighbour for the *i*th unit. More in detail, for each incomplete unit *i*, the donor *j* is the closest 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 Section I, in the ordinary PMM a natural choice is the Mahalanobis distance based on the residual covariance matrix of the regression of $\boldsymbol{Y}_{mis,}$ on $\boldsymbol{Y}_{obs}$. In fact, this choice gives a sort of standardization where, roughly speaking, the contribution of the different variables to the global distance function are "weighted" with the inverse of the corresponding prediction errors (Little, 1988).

11. A natural generalization of this metric in the case of PMM via mixture model is obtained through the matrix defined as follows. In order to simplify the notations, let $\boldsymbol{x}_r = \boldsymbol{y}_{obs,r}$ be the observed part of the incomplete record (recipient) *r*, and $\boldsymbol{y}_{m,r}$ the missing part. Correspondingly, for each complete record (possible donor) *d*, let $\boldsymbol{x}_d, \boldsymbol{y}_{m,d}$ be the values of the variables $\boldsymbol{X}$, $\boldsymbol{Y}_m$ that are observed and missing, respectively, in the record *r*. The matrix $\hat{S}_{Y|X}$ used in the Mahalanobis distance is given by the estimate of the expectation $S_{Y|X} = E(\boldsymbol{w}\boldsymbol{w}^T)$ where $\boldsymbol{w} = [(\boldsymbol{Y}_{m,r} \mid \boldsymbol{x}_r) - E(\boldsymbol{Y}_{m,r} \mid \boldsymbol{x}_r)] - [(\boldsymbol{Y}_{m,d} \mid \boldsymbol{x}_d) - E(\boldsymbol{Y}_{m,d} \mid \boldsymbol{x}_d)]$. In fact, since the variables $\boldsymbol{Y}_{m,r} \mid \boldsymbol{x}_r$ and $\boldsymbol{Y}_{m,d} \mid \boldsymbol{x}_d$ are independent one of each other, $S_{Y|X}$ is the sum of their covariance matrices

$V_{\mathbf{Y}|\mathbf{X}}(\mathbf{x}_r)$, $V_{\mathbf{Y}|\mathbf{X}}(\mathbf{x}_d)$ respectively. In order to provide an explicit formula for $S_{Y|X}$, we note that the covariance matrix $V_{\mathbf{Y}|\mathbf{X}}(\mathbf{x})$ of the distribution of the random vector $\mathbf{Y}$ conditional on $\mathbf{X}=\mathbf{x}$, can be decomposed as $V_{\mathbf{Y}|\mathbf{X}}(\mathbf{x})= V^{(1)}(\mathbf{x})+V^{(2)}(\mathbf{x}) =E_k[V(\mathbf{Y} \mid \mathbf{x},k)] + V_k[E(\mathbf{Y} \mid \mathbf{x},k)]$, where the variance $V(\mathbf{Y} \mid \mathbf{x},k)$ and the expected value $E(\mathbf{Y} \mid \mathbf{x},k)$ refer to the distribution of $\mathbf{Y}$ conditional on $\mathbf{X}=\mathbf{x}$ and a specific mixture component k, while E_k and V_k refer to the distribution of the indicator variable Z for the group labels k = *1,..,K*. The first term $V^{(1)}(\mathbf{x})$ on the r.h.s. of the above decomposition is $\sum_{k=1}^{K}\tau_k(\ )\sum_{Y|X}^{(k)}$ where $\tau_k(\mathbf{x})$ denotes the posterior probability for a unit where $\mathbf{x}$ is observed, of belonging to the group k, and $\sum_{Y|X}^{(k)}$ is the residual covariance matrix of the regression of Y on X from the kth Gaussian distribution of the mixture. As far as the second term $V^{(2)}$ is concerned, it can be shown that $V^{(2)}(\mathbf{x})=\sum_{k=1}^{K}\tau_k(\ )D^{(k)}(\ )$ where $D^{(k)}(\ )$ is the matrix $[E(\mathbf{Y} \mid \mathbf{x},k) - E(\mathbf{Y} \mid \mathbf{x})]\,[E_k(\mathbf{Y} \mid \mathbf{x},k) - E(\mathbf{Y} \mid \mathbf{x})]^T$. Thus, the total covariance matrix is $V_{\mathbf{Y}|\mathbf{X}}(\mathbf{x}) = V^{(1)}(\mathbf{x})+V^{(2)}(\mathbf{x}) = \sum_{k=1}^{K}\tau_k(\ )\left(\sum_{Y|X}^{(k)}+D^{(k)}(\ )\right)$. The final estimate of S is given by $\hat{V}_{Y|X}(\ _r)+\hat{V}_{Y|X}(\ _d)$ by using the estimates from the EM algorithm.

12. 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.

IV. EMPIRICAL EVALUATION

13. In this section we describe the simulation study carried out to evaluate the performance of the proposed semiparametric Predictive Mean Matching (SPMM).
The assessment is in terms of preservation of means and covariance structure of the data. To this aim a comparison between SPMM, NND and also with respect to imputations obtained directly via mixture models is performed. In the latter, imputations are obtained in two ways. By means of a random draw from the relevant conditional distribution of the missing items given the observed ones (MRD), and by means of their conditional expectations (MCM). The conditional distribution is estimated through a finite mixture of Gaussian distributions as described in Section II, and is the same used for SPMM. The considered imputation methods are evaluated in different simulation frameworks.

14. For each experimental setting, 100 simulations have been performed consisting of the following steps:
i) artificial generation of a sample from a given multivariate probability distribution;
ii) introduction of missing values in the sample;
iii) estimation of mixture model used for SPMM, MCM, MRD and imputation;
iv) comparison of the imputed dataset with the original ones through appropriate indices.

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

The previous 4 steps are detailed in the following.

Sample data generation

For the sample data generation, three different probability distributions are chosen.

15. 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 ($X_1,...,X_5$) from a 5-variate Gaussian distribution, and then defining new variables ($Y_1,...,Y_5$) through the transformation: $Y_i = exp(\alpha X_i)$ for i =1,..5. The normal random vector ($X_1,...,X_5$) 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:

$\boldsymbol{\mu}^{(2)} = (2.5, 2.6)$; $\boldsymbol{\mu}^{(3)} = (2.5, 2.6)$;

$$\boldsymbol{\Sigma}^{(2)} = \begin{pmatrix} 3.1 & 2.7 \\ 2.7 & 2.8 \end{pmatrix}; \quad \boldsymbol{\Sigma}^{(3)} = \begin{pmatrix} 3.0 & 2.4 & 2.4 \\ 2.4 & 3.0 & 2.1 \\ 2.4 & 2.1 & 3.0 \end{pmatrix}.$$

The value of the parameter α is 0.2. The parameters are obtained by a real survey.

16. Multivariate Gamma distribution - MG

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

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

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.

17. Another experimental setting (hereafter MGL) is obtained through the following slight modification of the Cheriyan and Ramabhadran's procedure. First, 7 independent random variables (r.v.s) X_i for i=1,…,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:

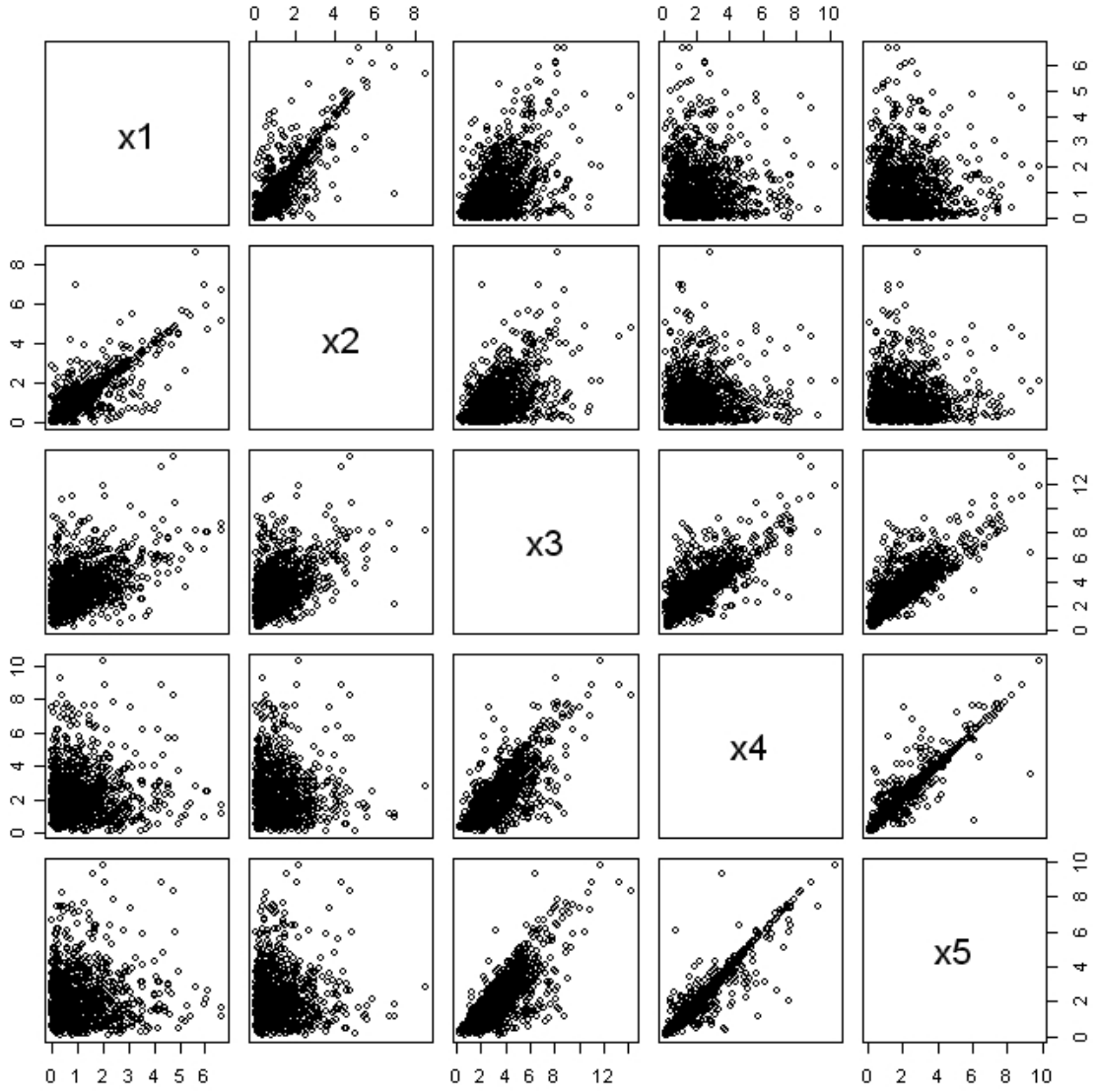
$Y_1=X_1+X_3$; $Y_2=X_1+X_4$; $Y_3=X_1+X_2+X_5$; $Y_4=X_2+X_6$; $Y_5=X_2+X_7$.

The parameters θ_i are chosen to obtain a correlation structure characterised 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 1000 observations from this distribution is shown in Figure 1.

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



Samples of 500 and 1000 units have been generated for all the settings.

Non-response simulation

18. Once a sample of complete data is generated, item non-response is simulated according to Missing at Random (MAR) mechanism (Little and Rubin, 2002). Items for (Y_1, Y_2, Y_3, Y_4) have response-rates depending 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 non-response probabilities for the variables are the following: 0.10 if $y_5 < q_1$, 0.20 if $y_5 \in [q_1, q_3)$, 0.30 if $y_5 \geq q_3$. No missing values are introduced in the variable Y_5.

Estimation and imputation

19. The incomplete sample is imputed using the NND method, the SPMM and MCM, 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.

20. Concerning finite mixtures, they are estimated following the algorithm described in Section II. Models with different number of components have been estimated. Once the parameters have been estimated for all the models, the model with the highest BIC is chosen, and the selected model is used to impute missing values following the SPMM, MCM, MRD methods described in Section III.

21. 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.

Evaluation

22. 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 $y_{i1},...,y_{ip}$ $(i=1,...,n)$ be the original "true" values of the p-dimensional random variable $\boldsymbol{Y}$ in the ith unit. Let the ${}^{*}y_{i1},...,{}^{*}y_{ip}$ be the corresponding imputed values, i.e. the values of variables 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

$$Dm_j = \sqrt{\frac{1}{100}\sum_{t=1}^{100}\frac{(m_j^{(t)} - {}^{*}m_j^{(t)})^2}{(m_j^{(t)})^2}}\ , \quad j=1,...,p$$

where ${}^{*}m_j^{(t)}$ is the mean of variable Y_j computed on the imputed dataset in the tth experiment.

An overall evaluation index can be obtained by the indicator

$$Dm = \sum_{j=1}^{p} Dm_j\ .$$

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

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

where $s_{jk}^{(t)}$ and ${}^{*}s_{jk}^{(t)}$ are the corresponding elements of the sample covariance matrices S and ${}^{*}S$ 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 following index:

$$DS=\sum_{j=1}^{p}\sum_{k=j\geq k}^{p}d_{jk}$$

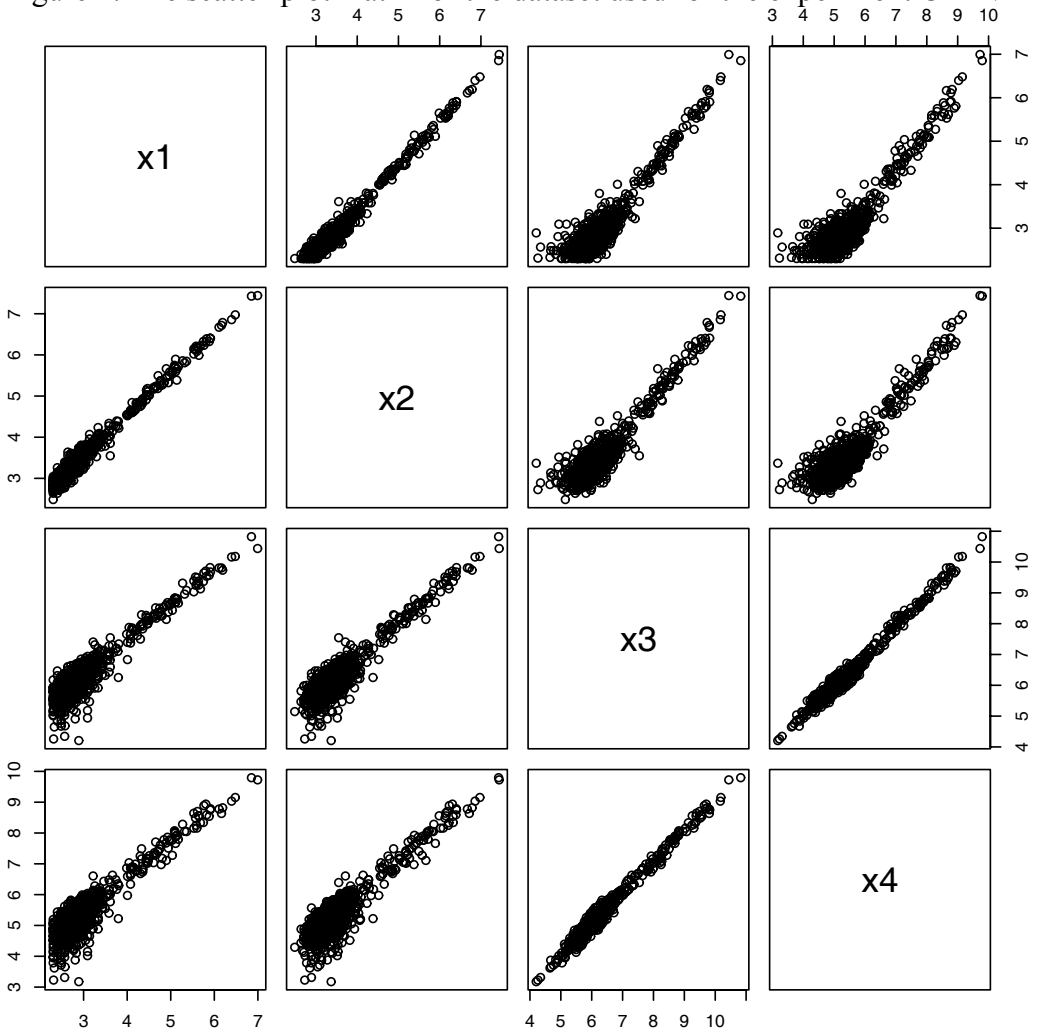
providing a measure for the variance and covariance structure preservation.

Experiments on a real data set

23. A subset of the 1997 Italian Labour Cost Survey (LCS) is 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 1000 units that belong to the metallurgic economic activity sector. 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 logaritmic 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

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



24. In this situation, since the underlying data distribution is unknown, a resampling approach has been adopted. The adopted resampling scheme consists in sampling 1000 observations (through a simple random sampling with replacement) $\boldsymbol{x}^{(1)},\ldots,\boldsymbol{x}^{(1000)}$ (bootstrap sample) from the initial sample, where $\boldsymbol{x}^{(i)}$ represents the ith unit whose observed variables are (X_1,X_2,X_3,X_4). The bootstrap sample can be thought of as generated from the empirical distribution of (X_1,X_2,X_3,X_4).

25. Missing values have been introduced according to the previous described mechanism noting that the conditioning variable is X_4.
Once the missing values are introduced in the bootstrap sample, they are imputed by means of NND and the methods based on mixtures. The results of the imputations are evaluated using the indices so far described.
This procedure has been repeated 100 times, and the results are averaged over them.
The results of the experiments are reported in the following tables.

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

LN 500

	Dm	DS
SPMM	0.0002	0.0037
NND	0.0003	0.0045
MCM	0.0001	0.0049
MRD	0.0002	0.0028

LN 1000

	Dm	DS
SPMM	0.0001	0.0018
NND	0.0001	0.0023
MDM	0.0000	0.0045
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

	Dm	DS
SPMM	0.0008	0.0813
NND	0.0007	0.0871
MCM	0.0004	0.0573
MRD	0.0006	0.0499

MGL 1000

	Dm	DS
SPMM	0.0004	0.0348
NND	0.0004	0.0434
MCM	0.0002	0.0422
MRD	0.0003	0.0190

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

MGH 500

	Dm	DS
SPMM	0.0014	0.0881
NND	0.0013	0.0715
MCM	0.0006	0.0545
MRD	0.0009	0.0434

MGH 1000

	Dm	DS
SPMM	0.0007	0.0332
NND	0.0007	0.0319
MCM	0.0003	0.0659
MRD	0.0005	0.0250

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

CLAV

	Dm	DS
SPMM	0.0000	0.0003
NND	0.0000	0.0003
MCM	0.0000	0.0001
MRD	0.0000	0.0001

26. 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 behaviour of MCM 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).

27. It is interesting to compare NND with SPMM, since the latter can be interpreted as a nearest neighbour donor with a particular distance. When there are some variables with low correlations the behaviour of the SPMM is better (Tables 1 and 2). When the variables are

correlated, the behaviour 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 considered also 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 the correlation among the variable is quite 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, as remarked in the next section.

28. 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 (and of course of the NND). Those methods impute only "live" values, thus avoiding strange "synthetic" values, for instance the imputation of negative values when the variables are nonnegative. Hence the choice of the best imputation method depends on the use of the data, whether they have to be released at some micro level detail, or only used to produce aggregated values.

V. FINAL CONSIDERATIONS

29. Final remarks are about general considerations on PMM. Let us introduce the concept for the univariate case and with not a non formal language, however those rough considerations can be useful to clarify the peculiarities of the PMM techniques. In general when an imputation of ith observation is performed (let us call the imputed value ${}^{*}y|x_i$), it would be desirable to have an imputed value ${}^{*}y|x_i$ that can be considered as generated from the conditional probability distribution of $Y|x_i$ (hereafter f_i). Hence, when the imputation is performed through a NND, the imputed value will be transferred from the jth observation which is the closest one with respect to a distance $d(x_i , x_j)$. Actually, this imputed value can be approximately thought of as a realization of the random variable $Y|x_i$ with probability distribution f_j provided that x_i and x_j are close enough. The underlying idea is that, at least for $n \to \infty$, $P(|X_i - X_j| < \varepsilon) \to 1$ (for all $\varepsilon > 0$), i.e., when a high number of observations is available the realized values of X are very close each other, and the imputed value ${}^{*}y|x_i = y|x_j$ can be considered as generated by the probability distribution f_i. As far as the PMM is concerned, the donor is chosen according to a distance computed on the expected values $E(Y|x_i)$. Thus the jth observation, chosen as donor, is the closest unit to the ith with respect to a certain distance $d(E(Y|x_i), E(Y|x_j))$. The desired property that the imputed value ${}^{*}y|x_i$ is a realization of ${}^{*}Y|x_i \sim f_i$, is achieved when the distributions f_i and f_j differ each other only by functions of their expected values. It is now evident the lack of normality in the request of the PMM, because it refers to all the distributions that depend on the conditioning variables only through their expected value. On the other hand this characterization refers to distributions that are not usually discussed in literature and more study should be devoted to this topic.

References

Dempster, A.P., Laird, N.M., Rubin, D.B., (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society*, B 39, pp. 1-38.

Di Zio, M., Guarnera, U. and Luzi, O. (2006a). Imputation Through Finite Mixture Modelling, *IASC-CSDA conference*, Cyprus, 28-31 October, 2005.

Di Zio, M., Guarnera, U. and Luzi, O. (2006b). Use of Finite Mixture Models in Editing and Imputation of Survey Data, *European Conference on Quality in Survey Statistics*, Cardiff, 24-26 April, 2006.

Durrant, G.B., and Skinner, C. (2006). Using Missing Data Methods to Correct for Measurement Error in a Distribution Function. *Survey Methodology,* Vol. 32, No. 1, 25-36.

Fraley, C. and Raftery, E. (2002). Model-Based Clustering, Discriminant Analysis, and Density Estimation, *Journal of the American Statistical Association*, 97, pp. 611-629.

Hunt, L. and Jorgensen, M. (2003). Mixture model clustering for mixed data with missing information, *Computational statistics and data analysis*, 41, pp. 561-575.

Kotz, S., Balakrishnan, N., Johnson, N.L. (2000). *Continuous Multivariate Distributions.* Vol.1, 2nd ed. Wiley, New York.

Little R.J.A. (1988). Missing-Data Adjustments in Large Surveys. *Journal of Business & Economic Statistics*, Vol. 6, No. 3, 287-296.

Little, J. and Rubin, D., (2002). *Statistical Analysis with Missing data.* Wiley, New York.

Marron, S. and Wand, M. (1992). Exact Mean Integrated Squared Error, *Annals of Statistics* 20, pp. 712–736.

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

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

Schwarz G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6, pp. 461-464.
