Adv Data Anal Classif
DOI 10.1007/s11634-016-0243-0

REGULAR ARTICLE

A sequential distance-based approach for imputing missing data: Forward Imputation

Nadia Solaro[1] · **Alessandro Barbiero**[2] · **Giancarlo Manzi**[2] · **Pier Alda Ferrari**[2]

Received: 7 July 2014 / Revised: 28 February 2016 / Accepted: 7 March 2016

Abstract Missing data recurrently affect datasets in almost every field of quantitative research. The subject is vast and complex and has originated a literature rich in very different approaches to the problem. Within an exploratory framework, distance-based methods such as nearest-neighbour imputation (NNI), or procedures involving multivariate data analysis (MVDA) techniques seem to treat the problem properly. In NNI, the metric and the number of donors can be chosen at will. MVDA-based procedures expressly account for variable associations. The new approach proposed here, called Forward Imputation, ideally meets these features. It is designed as a sequential procedure that imputes missing data in a step-by-step process involving subsets of units according to their "completeness rate". Two methods within this context are developed for the imputation of quantitative data. One applies NNI with the Mahalanobis distance, the other combines NNI and principal component analysis. Statistical properties of the two methods are discussed, and their performance is assessed, also in comparison with alternative imputation methods. To this purpose, a simulation study in the presence of different data patterns along with an application to real data are carried out, and practical hints for users are also provided.

✉ Nadia Solaro
nadia.solaro@unimib.it

Alessandro Barbiero
alessandro.barbiero@unimi.it

Giancarlo Manzi
giancarlo.manzi@unimi.it

Pier Alda Ferrari
pieralda.ferrari@unimi.it

1 Department of Statistics and Quantitative Methods, Università degli Studi di Milano-Bicocca, Via Bicocca degli Arcimboldi, 8, 20126 Milan, Italy

2 Department of Economics, Management and Quantitative Methods, Università degli Studi di Milano, Milan, Italy

Published online: 25 March 2016

Keywords Iterative PCA · Mahalanobis distance · MCAR missing data · MissForest · Nearest-neighbour imputation · Skew data

Mathematics Subject Classification 62H25 · 62-07 · 62-04 · 62H99

1 Introduction

Oscar Wilde's Lady Bracknell said: *Ignorance is like a delicate exotic fruit: Touch it and the bloom is gone*. With these famous lines, Wilde aimed at attacking the English aristocracy of his time whose interest was to hinder the spread of education to the whole society. Wilde's subtle message was to stress the extreme importance of knowledge and the need to defeat ignorance. Similarly, in statistical analysis, missing data are an important obstacle to disclosing knowledge.

Whatever the statistical transposition of Wilde's message, however, missing data are a type of ignorance that must be handled carefully, because an inappropriate treatment of missingness may affect findings and lead to unreliable results. An overwhelming amount of approaches have been proposed for this purpose. A comprehensive and definitive taxonomy of these approaches is difficult to implement since multiple features should be taken into account, such as statistical or mathematical settings and properties, their specific purpose, the fields of application, and so on. Nonetheless, classifications recurrent in literature distinguish between parametric and nonparametric methods, single or multiple imputation, imputation-based procedures (e.g. nearest-neighbour imputation) and model-based procedures (e.g. likelihood-based or Bayesian methods). Reference can be made to Little and Rubin (2002) and Rässler et al. (2013) for a general review of missing data approaches and related discussion. In more specific contexts, Schafer (1997) treats the problem of imputing multivariate missing data within a model-based imputation framework, while Groves et al. (2002) collect many contributions dealing with the problem of nonresponse for different typologies of surveys.

Within a nonparametric framework, distance-based methods such as nearest-neighbour imputation (NNI), or procedures involving multivariate data analysis (MVDA) techniques seem to treat the problem properly in a data analysis perspective, where no a priori distribution assumption is made, and the main concern should be to preserve the observed features of data. In particular, the NNI method allows subsets of complete units to be selected on the basis of a specific metric and used as "donors" for incomplete units. MVDA-based techniques can recover values for incomplete units by taking into account associations between variables. Both the approaches have attractive features. This motivated us to develop a new approach that ideally meets the potentialities of the two methods, in spite of an already vast literature. In particular, our approach to the problem of imputation is distance-based and takes advantage of the underlying data structure by exploiting the level of interconnections among variables, as provided by covariance/correlation structures in the case of quantitative data.

We call this approach "Forward Imputation" (*ForImp*) because it is a sequential method that alternates NNI with an MVDA technique to impute missing data in a

step-by-step process involving subsets of units according to their "completeness rate". It is an exploratory-descriptive method that does not need distribution assumptions on the data structure. In this sense, it can be classified within the nonparametric framework and regarded as one of the *learning-from-data* procedures, which have been greatly developed over these last years (Hastie et al. 2009).

This work is therefore addressed to presenting the methodological and algorithmic aspects of this new approach focusing on imputation for quantitative missing data. It has been extensively tested through a wide range of many different scenarios concerning a multitude of simulated and real data patterns, together with benchmark comparisons with other two relevant imputation methods. A selection of the obtained results is provided here, whereas the extensive simulation study is presented in Solaro et al. (2015a), and a specific aspect is treated in Solaro et al. (2014).

The rest of the paper is organized as follows. Section 2 describes the *ForImp* approach in detail, along with two possible variants specifically proposed for quantitative data: The one based on Principal Component Analysis (PCA), the other on the Mahalanobis distance. Section 3 presents a selection of the simulation study results obtained under some representative scenarios pertaining to specific data structures. Section 4 proposes a case study concerning real data. Section 5 concludes the paper and also discusses further developments.

2 The proposed approach: Forward Imputation

The main idea of the *ForImp* approach is to estimate missing data by exclusively exploiting the complete part of the data, which becomes larger and larger at every step of an iterative procedure. No initialization of missing entries is required. The approach is grounded on three basic ingredients: the forward, sequential progression of the imputation task, the use of the NNI method, and the *possible* recourse to an MVDA technique. These elements together create a general "architecture" that must be specified differently according to the nature of the data, in particular the measurement level of the variables.

Here, focusing on the problem of imputation for quantitative data, we propose two variants within the *ForImp* approach. The simplest variant exclusively relies on the NNI method, which is applied directly to the original variables. The other *ForImp* variant introduces PCA as an MVDA technique to obtain synthesis statistics from the complete part of the data in a way that is instrumental in the set-up of so-called "pseudo-indicators". These are artificial, missing-data-free variables, which gather the main relevant information intrinsic to the complete part of the data.

We indicate the two methods developed according to these lines: (A) *ForImp* with PCA (*ForImpPCA-FIP*), which involves both an MVDA technique (i.e. PCA) and NNI (Sect. 2.1); (B) *ForImp* with the Mahalanobis distance (*ForImpMahalanobis-FIM*), which involves only NNI applied to the original variables (Sect. 2.2). Both *FIP* and *FIM* fulfil the imputation task by taking advantage of the covariance/correlation structure of the complete part of the data, which is sequentially updated at every step of the procedure. They work through distinct stages, as described in detail below.

2.1 *ForImp* with PCA

The *FIP* procedure exploits the PCA method for setting up the pseudo-indicators we term pseudo-principal components. These are computed for all the units (i.e. both complete and incomplete) in order to synthesize the main relevant information in the covariance/correlation matrix of the complete part of the data, and then transfer them to incomplete units in a way that is instrumental in the subsequent application of the NNI method. All is performed in four stages:

– Stage 0: *Data Preparation*. Let $\mathbf{X} = [x_{ij}]$ be an $n \times p$ data matrix of p quantitative variables X_j and n units ($i = 1, \ldots, n$; $j = 1, \ldots, p$). Assume that at least p rows of $\mathbf{X}$ are free of missing values and the other $n - p$ rows contain at most $p - 1$ missing values ($n > p$, $p \geq 2$), (the assumption $n > p$ can indeed be relaxed. See Remark 4 further on). Then, matrix $\mathbf{X}$ is split into $K + 1$ disjoint submatrices where the first, denoted by $\mathbf{X}_0$, is missing-data-free of dimension $n_0 \times p$, $p \leq n_0 < n$, while the other K submatrices $\mathbf{X}_k$ of dimension $n_k \times p$ contain exactly k missing values in each of their rows, with $k = 1, \ldots, K < p$ and $n_0 + \sum_{k=1}^{K} n_k = n$. Such submatrices are then ordered increasingly according to k.

→ Set $k = 1$. It is not necessary for index k to assume all the integers in $\{1, 2, \ldots, K\}$. It may occur that $n_k = 0$ for any k. In such a case, jump to the first $k' > k$ such that $n_{k'} > 0$.

– Stage 1: This stage consists of two phases. (1) *Running PCA*. PCA is run on the complete $\mathbf{X}_{k-1}$ available up to the $(k-1)$-th step to synthesize the complete part of the data. Accordingly, a number p of principal components (as many as there are variables) is extracted from either variance-covariance matrix $\mathbf{\Sigma}_{k-1}$ or correlation matrix $\mathbf{R}_{k-1}$ of the complete $n_{k-1} \times p$ matrix $\mathbf{X}_{k-1}$, in order to obtain the eigenvalues $\lambda_s^{(k-1)}$ and the normalized eigenvectors $\boldsymbol{\omega}_s^{(k-1)}$ having the loadings $\omega_{js}^{(k-1)}$ as generic elements ($j, s = 1, \ldots, p$).
(2) *PPC score computation*. Pseudo-indicators called Pseudo-Principal Components (PPCs) are set up from such quantities for both complete units in $\mathbf{X}_{k-1}$ and incomplete units in $\mathbf{X}_k$ involving only common variables without missing values. In all, as many PPCs are built as there are variables. Let ι_k be the set formed by the k-combinations of the p variable indices having missing values in the rows of $\mathbf{X}_k$. The total number of k-tuples $\iota_{\tau,k}$ in ι_k is thus given by $T_k \leq \binom{p}{k}$, $\tau = 1, \ldots, T_k$, $k \leq p-1$. For instance, if $p = 5$ and $k = 2$, then $\binom{5}{2} = 10$ is the maximum number of potential pairs of variables with missing values in $\mathbf{X}_2$, but if: $\{X_1, X_2\}$, $\{X_4, X_5\}$ were the actual pairs with missing values in $\mathbf{X}_2$, then the ι_2 set would contain $T_2 = 2$ elements given by: $\iota_2 = \{\iota_{1,2} = \{1, 2\}, \iota_{2,2} = \{4, 5\}\}$. Next, let $\mathcal{S}_{\tau,k}$ be the subset of incomplete units $u_i^{(k)}$ having missing values just on the combination of variables with indices in $\iota_{\tau,k}$, ($i = 1, \ldots, n_{\tau,k}$; $\tau = 1, \ldots, T_k$; $\sum_{\tau=1}^{T_k} n_{\tau,k} = n_k$). Then, for each $\iota_{\tau,k}$ the PPCs, denoted by $\tilde{C}$, are given:

(i) for submatrix $\mathbf{X}_k$ and the incomplete units in $\mathcal{S}_{\tau,k}$, by the linear combinations:

$$\tilde{C}^{(k)}_{s(\iota_{\tau,k})}(\mathcal{S}_{\tau,k}) = \sum_{\substack{l=1 \\ l \notin \iota_{\tau,k}}}^{p} \omega^{(k-1)}_{ls} X^{(k)}_{l}(\mathcal{S}_{\tau,k}), \tag{1}$$

$(s = 1, \ldots, p, \ \forall\tau)$, where $X^{(k)}_{l}(\mathcal{S}_{\tau,k})$ denotes the l-th variable (column) in $\mathbf{X}_k$ with index l outside $\iota_{\tau,k}$, and with values referred to the incomplete units $u^{(k)}_i$ in $\mathcal{S}_{\tau,k}$. In particular, $(\iota_{\tau,k})$ at subscript of $\tilde{C}_s$ stresses that computations in (1) are performed only on variables X_l with index $l \notin \iota_{\tau,k}$, while (k) at superscript indicates that $\tilde{C}_s$ is computed on the incomplete $\mathbf{X}_k$ occurring at the k-th step;

(ii) for submatrix $\mathbf{X}_{k-1}$ and its complete units, by the linear combinations:

$$\tilde{C}^{(k-1)}_{s(\iota_{\tau,k})} = \sum_{\substack{l=1 \\ l \notin \iota_{\tau,k}}}^{p} \omega^{(k-1)}_{ls} X^{(k-1)}_{l}, \tag{2}$$

$(s = 1, \ldots, p, \ \forall\tau)$, where $X^{(k-1)}_{l}$ denotes the l-th variable in $\mathbf{X}_{k-1}$ with $l \notin \iota_{\tau,k}$ considered with respect to all the units $u^{(k-1)}_c$ therein present, which are either originally complete or imputed during the process run up to the $(k-1)$-th step $(c = 1, \ldots, n_{k-1})$.

– Stage 2: *Application of the NNI method*. In order to select a proper subset of donors for each unit in $\mathbf{X}_k$, the NNI method is applied to both complete and incomplete units using the PPC scores (1) and (2). The weighted Minkowski distance d_r of order r ($r \geq 1$ and, for $r \to \infty$, the Tchebycheff or Lagrange distance) is used as the basic metric for comparing unit $u^{(k)}_i$ in $\mathbf{X}_k$ with each unit $u^{(k-1)}_c$ in $\mathbf{X}_{k-1}$, for $c = 1, \ldots, n_{k-1}$:

$$d_r(u^{(k)}_i, u^{(k-1)}_c) = \left\{ \sum_{s=1}^{p} \left| \left(\tilde{c}^{(k)}_{s(\iota_{\tau,k}),i} - \tilde{c}^{(k-1)}_{s(\iota_{\tau,k}),c} \right) w^{(k-1)}_s \right|^r \right\}^{1/r}, \tag{3}$$

where the values $\tilde{c}^{(k)}_{s(\iota_{\tau,k}),i}$ and $\tilde{c}^{(k-1)}_{s(\iota_{\tau,k}),c}$ are the PPC scores computed according to formulas (1) and (2), respectively, while the weight $w^{(k-1)}_s$ is given by:

$$w^{(k-1)}_s = \sqrt{\lambda^{(k-1)}_s / \sum_{m=1}^{p} \lambda^{(k-1)}_m}, \quad s = 1, \ldots, p, \tag{4}$$

with the eigenvalues $\lambda^{(k-1)}_s$ computed at Stage 1, phase (1). The weight (4) is introduced to strengthen (or weaken) the role of the PPCs that derive from the principal components with higher (or smaller) variance.

Donors $u_{\delta,i}^{(k)}$ for unit $u_i^{(k)}$ are then given by the first $q100$ % of complete units $u_c^{(k-1)}$ corresponding to the q-th quantile $d_{q,i}$ of Minkowski distances d_r in (3), $(0 < q < 1; \delta = 1, \ldots, n_\delta; i = 1, \ldots, n_k)$.

– Stage 3: *Imputation*. For each incomplete unit $u_i^{(k)}$, the missing value on variable X_j is imputed by computing a weighted average of its donors' values:

$$\tilde{x}_{ij}^{(k)} = \frac{\sum_{\delta=1}^{n_\delta} x_{\delta j}^{(k-1)} \frac{1}{d_{\delta i}}}{\sum_{\delta=1}^{n_\delta} \frac{1}{d_{\delta i}}}, \quad \forall j \in \iota_{\tau,k},\ d_{\delta i} \neq 0\ \forall \delta, \tag{5}$$

for $i = 1, \ldots, n_k$, where n_δ is the total number of donors for $u_i^{(k)}$, and $d_{\delta i}$ is the distance between the δ-th donor and unit $u_i^{(k)}$ as computed in Stage 2. The reciprocal of the distances in (5) allows more (less) emphasis to be put on nearer (more distant) donors. Should an incomplete unit coincide with one or more donors (i.e. $d_{\delta' i} = 0$ for any δ'), we would replace formula (5) with: $\tilde{x}_{ij}^{(k)} = \frac{1}{n_{\delta'}} \sum_{\delta'=1}^{n_{\delta'}} x_{\delta' j}^{(k-1)}$, thus exclusively taking up the $n_{\delta'}$ donors coincident with the incomplete $u_i^{(k)}$ and discarding the others.

The imputed submatrix $\tilde{\mathbf{X}}_k$ is thus obtained by replacing the missing values in $\mathbf{X}_k$ with the corresponding imputed values $\tilde{x}_{ij}^{(k)}$ derived from (5). The new complete matrix $\mathbf{X}_k$ is finally given by stacking the previous $\mathbf{X}_{k-1}$ with the imputed $\tilde{\mathbf{X}}_k$.

Stages 1 to 3 are sequentially repeated with the increasing of index k, thus making the complete part of the data larger and larger after the completion of each iteration. The *FIP* procedure then stops when matrix $\mathbf{X}$ is completely imputed.

Several aspects concerning the *FIP* algorithm are pointed out in the following.

Remark 1 Loadings for computing PPC scores. For comparability reasons, in formulas (1) and (2) the PPC scores at step k are computed using only non-missing variables common to both complete and incomplete units along with the loadings derived from PCA carried out on the complete matrix $\mathbf{X}_{k-1}$. In this sense, we call them "pseudo" principal component scores. We do not compute the linear combinations for the complete units in $\mathbf{X}_{k-1}$ using *all* the variables. The intent is to synthesize the information through a criterion common to all the units—both complete and incomplete—and based on common elements (i.e. same variables and same loadings).

For the same comparability reasons, we do not re-compute the loadings for each possible combination of common complete variables occurring for a given k. Apart from an additional computational burden, this procedure would actually change the weighting system for the variables. The loadings would be normalized each time on different subsets of complete variables according to the positions where the missing values occurred.

All this is especially crucial when the difference between the number p of overall variables and the number k of variables having missing values is small, since PCA would be performed on sub-matrices related to the complete part with a very small number, $p - k$, of columns, in the extreme case equal to one.

Remark 2 Use of quantiles for the set of donors. The problem of how many donors should be selected in an NNI approach is widely debated in literature (Little and Rubin 2002; Tarsitano and Falcone 2010). No conclusive answer has been given, since solutions for the best number of donors are often found for a specific framework. In our algorithm, for each incomplete unit we use the first $q100$ % of units in the complete matrix $\mathbf{X}_{k-1}$ as donors ($0 < q < 1$), a natural choice since we are dealing with sub-matrices of different sizes at each k. We have performed an extensive simulation study (Solaro et al. 2015a) elsewhere in order to inspect the problem of how to choose the proportion q of donors. It showed that the ideal proportion of donors ultimately depends on the pattern of data at hand. However, since the number of donors grows as k increases, it is possible to conceive a mixed approach in which the number of donors is limited by a maximum threshold (e.g. fixing an upper bound).

Remark 3 Full-rank condition. Matrices $\mathbf{\Sigma}_{k-1}$ or $\mathbf{R}_{k-1}$ do not need to be of full rank for each k. Denoting with $p' < p$ the rank of $\mathbf{\Sigma}_{k-1}$ or $\mathbf{R}_{k-1}$ for a given k, one has: $\lambda_{p'+1}^{(k-1)} = \lambda_{p'+2}^{(k-1)} = \cdots = \lambda_{p}^{(k-1)} = 0$. Formulas (1) and (2) for PPC score computation thus still hold for every $s = 1, \ldots, p', \ldots, p$, with certain loadings $\omega_{ls}^{(k-1)}$ possibly equal to zero for the last indices $s > p'$. Nevertheless, the contribution of such PPCs to the Minkowski distance (3) vanishes because of the corresponding zero eigenvalues $\lambda_{s}^{(k-1)}$ in the weight (4).

Remark 4 Number of units less than variables. The requirement that at least p rows in $\mathbf{X}$ are free of missing values, i.e. $n_0 \geq p$, is not binding. For the arguments advanced in the Appendix, which descend from the well-known relationship between classical multidimensional scaling (also known as Principal Coordinates Analysis-PCO) and PCA (Cox and Cox 2001; Gower 2005), in case of $n_0 < p$ the first step of *FIP* is liable to be modified as follows:

- Stage 1: extract the first non-null $n_0' \leq n_0 - 1$ eigenvalues $\lambda_s^{(0)}$ and p-dimensional eigenvectors $\boldsymbol{\omega}_s^{(0)}$ from $\mathbf{R}_0$ (or $\mathbf{\Sigma}_0$), (with index s ranging over 1 to $n_0' < p$);
- Stage 2: compute formulas (1) and (2) to set up all the possible n_0' PPCs regarding only the variables outside the various ι sets, then apply formulas (3) and (4) for selecting donors;
- Stage 3: carry out imputation by formula (5).

For the procedure to work, a minimum number n_0 of units equal to 3 is required.

In general, the number p of variables might exceed the total number of complete units not only at Stage 0, but in one or more of the subsequent steps k, i.e. $n_{k-1} < p$ for any $k > 1$. In such a case, the above sketched procedure still applies by extracting the first non-null $n_{k-1}' \leq n_{k-1} - 1$ eigenvalues and corresponding eigenvectors at Stage 1, and then all the possible n_{k-1}' PPCs at Stage 2. Finally, given the relationship between PCA and PCO, this procedure can be applied to the extreme, though not unusual, situation in which the starting matrix $\mathbf{X}$ has a number of units smaller than variables, i.e. $n < p$. In such a case, one obviously has: $n_{k-1} < p$ at every step k, ($k = 1, \ldots, K$). We indicate this latter variant of the method as *ForImp with PCO*, to stress the substantial difference from the $n > p$ case.

Remark 5 Choice between the variance-covariance matrix and correlation matrix. Since *FIP* does not aim at dimensionality reduction problems, the input matrix from which the PCs are extracted should be chosen consistently with the imputation problem. For example, if a variable had a greater variability and a smaller number of missing values than the others in the set, extracting the PCs from matrix $\mathbf{\Sigma}_{k-1}$ rather than $\mathbf{R}_{k-1}$ would be preferable even if variables are not directly comparable. This would put more weight on the most complete variable, and the subsequent selection of donors would depend more on it than the others. This feature would turn out to be attractive if this variable were highly correlated with the others.

Remark 6 Ordinal data. In Ferrari et al. (2011), an alternative treatment method for missing data is proposed in the context of nonlinear PCA when the objective is to measure a single latent dimension from ordinal data with missing categories. Since the proposal belongs to the *ForImp* approach, it can be readily extended to a pure imputation perspective by extracting more than one latent dimension and carrying out a procedure similar to *FIP*.

2.2 *ForImp* with the Mahalanobis distance

The *FIM* variant starts from Stage 0 as in *FIP*, but does not embrace Stage 1 because *FIM* works directly on the original variables. Regarding Stage 2, the basic metric for selecting donors is the Mahalanobis distance d_M:

$$d_M(u_i^{(k)}, u_c^{(k-1)}) = \left\{ \left(\mathbf{x}^{(k)}_{(\iota_{\tau,k}),i} - \mathbf{x}^{(k-1)}_{(\iota_{\tau,k}),c} \right)^t \mathbf{\Sigma}^{-1}_{(\iota_{\tau,k}),ic} \left(\mathbf{x}^{(k)}_{(\iota_{\tau,k}),i} - \mathbf{x}^{(k-1)}_{(\iota_{\tau,k}),c} \right) \right\}^{1/2}, \quad (6)$$

where $\mathbf{x}^{(k)t}_{(\iota_{\tau,k}),i}$ and $\mathbf{x}^{(k-1)t}_{(\iota_{\tau,k}),c}$ are the $(p-k)$-dimensional row-vectors of unit $u_i^{(k)}$ in $\mathbf{X}_k$ and unit $u_c^{(k-1)}$ in $\mathbf{X}_{k-1}$, respectively, with values given by the variables X_l outside the $\iota_{\tau,k}$ set. The variance-covariance matrix $\mathbf{\Sigma}_{(\iota_{\tau,k}),ic}$ of order $(p-k)$ is computed on the variables X_l outside the $\iota_{\tau,k}$ set using values of all complete units $u_c^{(k-1)}$ plus the unit $u_i^{(k)}$ (this set of units is indicated with "ic" at the subscript of $\mathbf{\Sigma}_{(\iota_{\tau,k}),ic}$). If $k = p-1$, then d_M reduces to: $d_M(u_i^{(k)}, u_c^{(k-1)}) = \frac{\left| x_{l,i}^{(k)} - x_{l,c}^{(k-1)} \right|}{\sqrt{\mathrm{Var}(X_{l,ic})}}$, $c = 1, \ldots, n_{k-1}$, where $X_{l,ic}$ denotes the unique, common complete variable outside the $\iota_{\tau,k}$ set with values restricted to the units $u_i^{(k)}$ and $u_c^{(k-1)}$.

Similarly to *FIP*, donors are then given by the first $q100$ % of complete units corresponding to the q-th quantile $d_{q,i}$ of Mahalanobis distances d_M ($0 < q < 1$; $i = 1, \ldots, n_k$). For the rest, the routine applies in the same way as described above.

Unlike *FIP*, *FIM* requires the invertibility of matrix $\mathbf{\Sigma}_{(\iota_{\tau,k}),ic}$ in the Mahalanobis distance (6), which must then be of full rank for all k, and whichever the set ic of units is. A necessary condition for this to happen is: $n_{k-1} + 1 \geq p$ for every k. Clearly, the full-rank requirement rules out the possibility that *FIM* can be applied when $n < p$. In such a case, one could resort to the *FIP* method as explained in Remark 4.

2.3 Further characteristics of the *FIP* and *FIM* methods

We will now discuss some further characteristics of *FIP* and *FIM* concerning: (a) role of units and (b) weight of variables, both in the selection of donors, and (c) admissibility of imputed values. Specifically:

(a) The forward and sequential nature of the procedure, common to *FIP* and *FIM*, implies that units are included into the imputation process according to the extent of their completeness. Accordingly, the more missing values the units have, the later they enter the process, and the fewer chances to become donors throughout the procedure they have.

(b) Common complete variables are weighted differently or not according to which of the two *ForImp* variants is considered. In *FIP*, common variables are weighted through their loadings, which derive from running PCA on the complete part of the data and then appear in PPC computation formulas (1) and (2). Common variables having the highest loadings in (1) and (2) and taking part in the computation of the PPCs that are associated with the highest eigenvalues, and then the highest weights (4), play a more important role in the subsequent selection of donors, because this evaluation is based on the Minkowski distance (3) just computed using the PPCs and their weight (4). On the other hand, in *FIM* the use of the Mahalanobis distance (6) implicitly assigns a same weight to the variables, adjusting for their correlation. As a result, selection of donors is based on variables that are put on the same footing and thus play the same role.

(c) Imputations are always fulfilled in the range of observed values as a result of the well-known Cauchy internality property held by power means of order r, ($r \neq 0$, and with $r \to 0$ for the geometric mean), with (5) being an average. This property, which seems obvious, may not be possessed by other imputation methods, and turns out to be desirable in many applications where it is known that values of variables cannot be outside a range of admissible values. *FIP* and *FIM*, imputing in the range of observed values, certainly produce admissible values.

Moreover, as the *ForImp* approach is based on the NNI method, complete units that are very far from the incomplete ones—at least with reference to the common observed part—in principle should not take on the role of donors. *ForImp* is thus expected to be intrinsically robust to the presence of outlying units or groups of units more or less separated from the main core of the data. However, the actual resistance of *FIP* and *FIM* to contamination greatly depends on the structure and pattern of anomalies, which are likely to assume more varied and complex shapes in a multidimensional space, and should then be carefully inspected before proceeding to imputation. A cautious approach could consist of removing the outlying units detected by some convenient methods [e.g. Forward Search (Atkinson et al. 2004)] from the complete part of the data and then proceeding to imputation on the outlier-free dataset. In this regard, Sect. 4 proposes an example of imputation in the presence of multivariate outliers, which uses both the full and the outlier-free datasets. Alternatively, one can replace $\mathbf{\Sigma}$ (in the case of *FIM* and *FIP*) or $\mathbf{R}$ (as for *FIP*) with a robust estimate [e.g. minimum covariance determinant, based on the "good" part of the data (Rousseeuw and Leroy 1987)] to produce robust PPC scores or robust Mahalanobis distances.

All the routines concerning *FIP* and *FIM* were implemented in the R environment (R Core Team 2015). These routines are part of the R library "GenForImp" (Solaro et al. 2015b), available on the CRAN package repository of the R web site.

3 Performance assessment of *FIP* and *FIM*: simulation study

The performance of *FIM* and *FIP* was assessed and compared through an extensive Monte Carlo simulation study in the presence of data having different patterns. The extended version of the study is reported in Solaro et al. (2015a). Results discussed here concern a subset of the simulation scenarios regarded as more representative since they closely interpret common real situations.

Alternative imputation techniques were also considered in the simulation study as a benchmark for comparisons, namely: (i) Stekhoven and Bühlmann's *missForest* method (2012), a nonparametric imputation technique for continuous and/or categorical data based on a random forest [i.e. a random classifier introduced in the context of machine learning by Breiman (2001)]; (ii) Iterative PCA [*IPCA*, Nora-Chouteau (1974), Greenacre (1984)], an algorithmic-type technique that imputes missing values simultaneously by the iterative use of PCA, and recently reassessed by Josse et al. (2011) to be part of a more general methodology with principal component methods (*missMDA*).

Of course, many other imputation methods could have been considered for benchmarking, such as the extensions introduced by Wasito and Mirkin (2005) of the least-squares imputation algorithms using the NNI approach. However, besides comparing *FIM* and *FIP* with either an imputation method similarly involving PCA, like *IPCA*, or based on a far different logic, such is *missForest*, our choice in favour of these two methods was also secondary to the availability of implemented routines in the R environment (R Core Team 2015). In particular, *IPCA* is implemented in the R library "missMDA" by Husson and Josse (2015), and *missForest* in the homonymous R library "missForest" by Stekhoven (2013).

3.1 Simulation study

The main objective of the simulation study was to assess the performance of *FIP* and *FIM*, also in comparison with *missForest* and *IPCA*, in the presence of data patterns often encountered in applications, i.e. heavy/thin-tailed symmetric or skew shapes having specific correlation structures between variables. On this point, besides considering "dimensionality of data" (number of units and variables) and "seriousness of missingness" (i.e. percentages of missing values), correlations of variables, kurtosis and skewness of the data distribution were more closely taken into account, as they were expected to greatly affect the imputation performance. In Solaro et al. (2015a), the simulation study is extensively described along its main steps, i.e. simulation design, multivariate distributions chosen for data generation, relationship between input parameters and output parameters of the involved multivariate distributions, set-up of data patterns having specific characteristics in terms of kurtosis or skewness and correlation structures, simulation procedure and summary of simulation results, and finally, descriptive and inferential analyses of the achieved outcomes.

Table 1 Formal definitions of *MSN* family ofdistributions

MSN family of distributions: $\boldsymbol{X} \sim \mathrm{MSN}_p(\boldsymbol{\Omega}, \boldsymbol{\alpha})$	
Density function	$f(\mathbf{x}; \boldsymbol{\Omega}, \boldsymbol{\alpha}) = 2\phi_p(\mathbf{x}; \boldsymbol{\Omega})\Phi(\boldsymbol{\alpha}^t\mathbf{x})$, where:
	– $\phi_p(\mathbf{x}; \boldsymbol{\Omega})$ is the $N_p(\mathbf{0}, \boldsymbol{\Omega})$ d.f. with correlation matrix $\boldsymbol{\Omega}$
	– $\Phi(\cdot)$ is the $N(0, 1)$ distribution function, and $\boldsymbol{\alpha} \in \mathbb{R}^p$
Meaning of parameters	– Parameter related to the skewness: $\boldsymbol{\alpha} \in \mathbb{R}^p$. If: $\boldsymbol{\alpha} = \mathbf{0}$, then: $\boldsymbol{X} \sim N_p(\mathbf{0}, \boldsymbol{\Omega})$
	– Mean vector: $\mathrm{E}(\boldsymbol{X}) = \boldsymbol{\mu} = \sqrt{2/\pi}\boldsymbol{\delta}$, with: $\boldsymbol{\delta} = \frac{\boldsymbol{\Omega}\boldsymbol{\alpha}}{\sqrt{1+\boldsymbol{\alpha}^t\boldsymbol{\Omega}\boldsymbol{\alpha}}}$
	– Variance–covariance matrix: $\mathrm{V}(\boldsymbol{X}) = \boldsymbol{\Sigma} = \boldsymbol{\Omega} - \boldsymbol{\mu}\boldsymbol{\mu}^t$
	– Correlation matrix: $\mathbf{R} = \mathbf{D}^{-1}\boldsymbol{\Sigma}\mathbf{D}^{-1}$, with: $\mathbf{D} = \mathrm{diag}\left\{\sqrt{1 - 2\pi^{-1}\delta_j^2}\right\}_{j=1,\ldots,p}$
	– $\boldsymbol{\Omega}$ input correlation matrix, $\mathbf{R}$ output correlation matrix
	– Multivariate (MV) skewness index: $\gamma_{1\mathrm{MV}} = \left(\frac{4-\pi}{2}\right)^2 (\boldsymbol{\mu}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^3 \in (-0.9905, +0.9905)$

As an instance of the study carried out, we focus here on simulations performed under a pattern of data very common in real situations, which is given by a mix of distribution skewness and different correlations between variables. It will be described in the next Sect. 3.2. Data inherent to this pattern were randomly generated by relying on the Multivariate Skew-Normal (*MSN*) family of distributions (Azzalini and Dalla Valle 1996; Azzalini and Capitanio 1998), the description of which is provided in Table 1, along with the method implemented by Azzalini in the R library "sn" (Azzalini 2015).

Table 2 reports the simulation settings considered for the study. Combined together, they gave rise to a total number of 216 simulation scenarios. Table 2 also shows the correspondence between some relevant values of parameters of the generating *MSN* distribution (*input parameters*) and the values of some indices (*output parameters*) that can be of interest for the statistical analysis of the generated data. This correspondence is essential for interpreting simulation results in the presence of such a complex variation structure among all the involved parameters, which is intrinsic to *MSN* distributions. Input parameters stand for correlation coefficients ω_{lj} in matrix $\boldsymbol{\Omega}$ ($l \neq j$), skewness parameters α in vector $\boldsymbol{\alpha}$, and coefficient c regulating the unbalance of correlations in $\boldsymbol{\Omega}$, while output parameters are correlation coefficients ρ_{lj} in matrix $\mathbf{R}$ ($l \neq j$) and multivariate (MV) skewness index $\gamma_{1\mathrm{MV}}$.

Under each scenario a complete $n \times p$ data matrix $\mathbf{X}^*$ ($n > p$) was first generated according to a specific *MSN* distribution. Then, 1000 incomplete matrices $\mathbf{X}_t$ were derived from it by deleting 5, 10, or 20 %, respectively, of Missing Completely At Random (MCAR) values ($t = 1, \ldots, 1000$). Subsequently, *missForest*, *IPCA*, *FIP*, and *FIM* were applied to each $\mathbf{X}_t$ with their options fixed as reported in Table 2. For each method, simulation results were synthesized by means of the Relative Mean Square Error (RMSE):

Table 2 Experimental conditions, correspondence of input–output correlations and skewness, and options fixed for the applicability of the imputation methods

Data dimensionality and seriousness of missingness:	
– Number of variables in $\mathbf{X}^*$	$p = 5$
– Number of units in $\mathbf{X}^*$	$n = 500; 1000$
– Percentage of MCAR values	$5\%; 10\%; 20\%$

Data generation from $\text{MSN}_p(\boldsymbol{\Omega}, \boldsymbol{\alpha})$ with $\boldsymbol{\Omega} = \left[\omega_{lj}\right]_{l \neq j = 1,\ldots,p}$ and $\boldsymbol{\alpha} = [\alpha]_{j=1,\ldots,p}$:

→ *Input parameters:*	→ *Output parameters:*
– Skewness parameter: $\alpha = 1; 4; 10; 30, \forall j$	– Output correlation coefficients in $\mathbf{R}$:
– Input correlation coefficients in $\boldsymbol{\Omega}$: $\omega_{1j} = \omega_{j1} = -\omega$, $0 < \omega < 1$, $j \neq 1$; $\omega_{lj} = \omega / c$, for $l, j \neq 1$, $l \neq j$, with $\omega = 0.2; 0.5; 0.8$ and $c = 1; 1.25; 1.5$	$\rho_{1j} = \rho_1$, $j \neq 1$; $\rho_{lj} = \rho_2$, for $l, j \neq 1$, $l \neq j$

Correspondence between input and output correlations (with α varying):

Input correlations in $\boldsymbol{\Omega}$:	Output correlations in $\mathbf{R}$:	Correlation structure*:
$\omega = 0.2$, $c = 1; 1.25; 1.5$	$\rightarrow \rho_1 \approx -0.25$, $\rho_2 \approx 0$	neg. low and nearly null ρs
$\omega = 0.5$, $c = 1; 1.25$	$\rightarrow \rho_1 \approx -0.4$, $\rho_2 \approx 0.2$	neg. moderate and pos. low ρs
$\omega = 0.5$, $c = 1.5$	$\rightarrow \rho_1 \approx -0.4$, $\rho_2 \approx 0.1$	neg. moderate and nearly null ρs
$\omega = 0.8$, $c = 1$	$\rightarrow \rho_1 \approx -0.7$, $\rho_2 \approx 0.6$	neg. high and pos. high ρs
$\omega = 0.8$, $c = 1.25$	$\rightarrow \rho_1 \approx -0.7$, $\rho_2 \approx 0.35$	neg. high and pos. moderate ρs
$\omega = 0.8$, $c = 1.5$	$\rightarrow \rho_1 \approx -0.7$, $\rho_2 \approx 0.25$	neg. high and pos. low ρs

* neg.: negative, pos.: positive

Correspondence between input and output skewness:

Input skewness:	Output MV skewness index:	Strength:
$\alpha = 1$ (with: $\omega \geq 0.2$, $c = 1; 1.25; 1.5$)	$\rightarrow \gamma_{1\text{MV}} \in (0.27, 0.42)$	moderate skewness
$\alpha \geq 4$ (with: $\omega \geq 0.2$, $c = 1; 1.25; 1.5$)	$\rightarrow \gamma_{1\text{MV}} \in (0.89, 0.99)$	strong skewness

Options fixed for each imputation method:	
– *missForest*	maximum number of iterations: 50
– *IPCA*	default option "Regularized method" in the function "imputePCA" maximum number of iterations: 5000 number of extracted PCs: $p - 2$, $(p \geq 3)$
– *FIP*	extraction from the var-cov matrix (option "cor=False") distance: city-block ($r = 1$); Euclidean ($r = 2$); Tchebycheff ($r = \infty$)
– *FIP* and *FIM*	donor quantile: $q = 0.05; 0.1; 0.15; 0.2$

$$\mathsf{m}\text{RMSE}_t = \sum_{j=1}^{p} \frac{1}{n\sigma_j^2} (\mathbf{x}_j^* - \mathsf{m}\tilde{\mathbf{x}}_{j,t})^t (\mathbf{x}_j^* - \mathsf{m}\tilde{\mathbf{x}}_{j,t}), \tag{7}$$

($t = 1, \ldots, 1000$), where $\mathbf{x}_j^*$ is the j-th column vector of the complete matrix $\mathbf{X}^*$, $\mathsf{m}\tilde{\mathbf{x}}_{j,t}$ is the j-th column vector of the matrix $\mathsf{m}\tilde{\mathbf{X}}_t$ imputed with method m at the t-th simulation run, and σ_j^2 is the variance of the j-th variable in $\mathbf{X}^*$.

After that, we carried out descriptive and inferential analyses of RMSE values. Descriptive analyses were performed by computing usual synthesis measures (mean, standard deviation, and quartiles) and providing graphical representations (e.g. dot

plots of RMSE median values). Inferential analyses were carried out to test the null hypothesis of equality of RMSE distributions of the examined methods against different ordered alternative hypotheses under a same simulation scenario. To this end, the Jonckheere–Terpstra (J–T) test was preferred to Page's test (Hollander and Wolfe 1999) because we regarded comparisons among the various RMSE distributions, as fixed in our systems of hypothesis, as overall (i.e. on all runs) rather than punctual (i.e. run-by-run), as Page's test would have indicated instead.

3.2 Simulation results under skewness and unbalanced correlations

Data within the pattern of skewness and unbalanced correlations were generated to be skew-distributed, with the first variable being negatively correlated with all the others by a same value $-\omega$, and the other variables positively correlated to each other by a same magnitude ω/c, with $0 < \omega < 1, c \geq 1$ (Table 2). In order to produce unbalanced correlations, this magnitude was progressively reduced by increasing coefficient c from 1 (absence of unbalance) to 1.25 and 1.5 (Table 2). Unbalance was thus defined on the absolute values of input correlations ω_{lj} in matrix $\mathbf{\Omega}$ without taking into account their sign ($l \neq j = 1, \ldots, p$). The unbalance extent among absolute correlations in the input $\mathbf{\Omega}$ then reflects in the output $\mathbf{R}$ correlation matrix though not linearly, as sketched in the input-output correspondences in Table 2. To help the reader interpret these correspondences, a short description of the types of correlation structures is therefore proposed and later referred to in the description of simulation findings.

It can be immediately seen from Table 2 that the patterns of variations among input-output parameters are quite complex. The crucial point in interpreting simulation results is understanding how changes in input ω, c and α parameters affect the output ρ and $\gamma_{1\text{MV}}$ values. In Solaro et al. (2015a), there is a wide discussion on this point. Within the scopes of this study, it is helpful to consider that, *ceteris paribus*, the skewness index $\gamma_{1\text{MV}}$ increases as α increases—in particular index $\gamma_{1\text{MV}}$ assumes moderate values with $\alpha = 1$ and approaches its maximum possible value equal to $+0.9905$ (Table 1) as α increases—, while the absolute values of the two distinct correlation coefficients ρ_1 and ρ_2 in matrix $\mathbf{R}$ (Table 2) increase as ω increases and decrease as c increases.

Figure 1 presents dot plots of the RMSE median values obtained for each method under the settings of Table 2, with $p = 5$ variables, $n = 1000$ units, 20 % of MCAR values, and $q = 0.1$ donor quantile in the case of *FIM* and *FIP*. This situation represents the other omitted scenarios well and thus allows us to draw some general considerations. By taking into account the correspondence between input $\mathbf{\Omega}$ and output $\mathbf{R}$ provided in Table 2, it can be seen that: (a) When $\omega = 0.2$ (i.e. negative low and nearly null ρs, 1st row of panels in Fig. 1) *FIM* is the best imputation method, achieving the smallest RMSE values, followed by *FIP*. The performances of *IPCA* and *missForest* are worse; (b) when $\omega = 0.5$ (i.e. negative moderate and positive low/nearly null ρs, 2nd row in Fig. 1) *FIP* mostly proves to be the best method with few exceptions; (c) the case $\omega = 0.8$ (last row, Fig. 1) highlights the main differences among the methods. There is an inversion of trend moving from "negative high and positive high ρs" ($c = 1$), where *IPCA* is the best method and *FIM* the worst, to

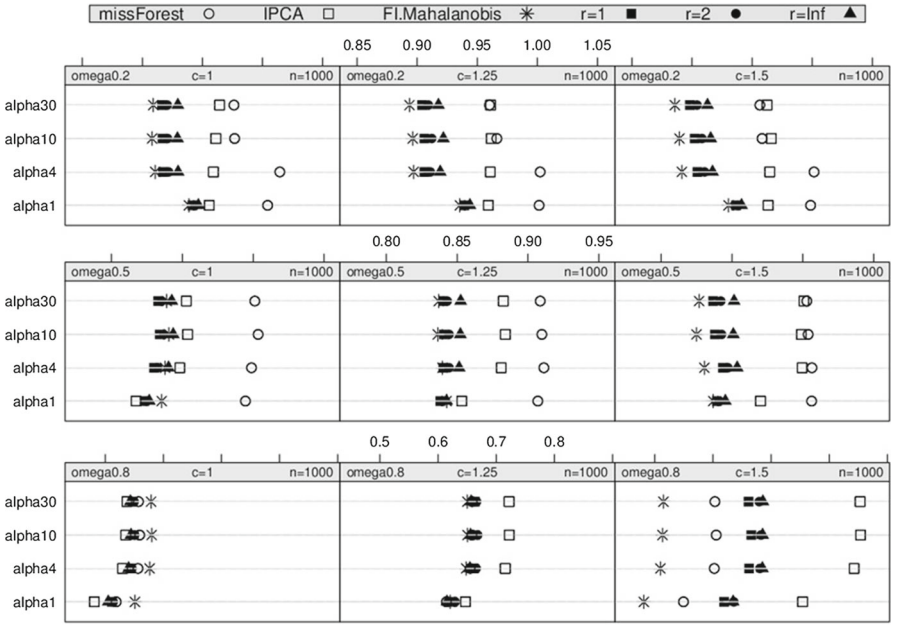

Fig. 1 *Dot plots* of RMSE median values of *missForest*, *IPCA*, *FIM* and *FIP* with $q = 0.1$ donor quantile, 20 % of MCAR values, $p = 5$ variables and $n = 1000$ units. Values of ω in *each row of panels*: $\omega = 0.2$ (*first row*), $\omega = 0.5$ (*second row*), $\omega = 0.8$ (*third row*). Values of c in *each column of panels*: $c = 1$ (*first column*), $c = 1.25$ (*second column*), $c = 1.5$ (*third column*)

"negative high and positive low ρs" ($c = 1.5$), where *FIM* is the best method and *IPCA* the worst. The panel related to the "negative high and positive moderate ρs" level ($c = 1.25$) displays an intermediate situation, where *missForest*, *FIM* and *FIP* have a very similar good performance, while *IPCA* performs less well.

Regarding the skewness effect, the dichotomy $\alpha = 1$ vs. $\alpha \geq 4$ is clearly visible. In particular, when $\omega = 0.2$, *missForest*, *FIM*, and *FIP* perform better for more skew data ($\alpha \geq 4$), while *IPCA* seems less sensitive to variations of α. On the other hand, when $\omega = 0.8$, all the methods tend to perform better for less skew data ($\alpha = 1$). The case $\omega = 0.5$ is intermediate, showing both the trends, in particular the first for $c = 1.5$ (better for more skew), and the second for $c = 1$ (better for less skew).

The one-sided J–T test results reported here concern the problem of detecting the best method from among *FIM*, *FIP* and *IPCA*. *missForest* was discarded from this analysis because *IPCA* proved to perform better than it in almost all the considered scenarios of the overall study (Solaro et al. 2015a). The aim is to test the null hypothesis:

$$H_0 : F_{FIM}(x) = F_{FIP}(x) = F_{IPCA}(x) \quad \forall x \geq 0, \tag{8}$$

where in (8) F_{FIM}, F_{FIP} and F_{IPCA} denote the RMSE distribution function of methods *FIM*, *FIP* and *IPCA*, respectively, against each of the following six ordered alternative hypotheses:

Table 3 J–T test for detection of the best imputation method among *FIM*, *FIP* and *IPCA*

		$p = 5$					
		$n = 500$			$n = 1000$		
	c	1	1.25	1.5	1	1.25	1.5
$\omega = 0.2$	$\alpha = 1$	6	6	6	1	1	1
	$\alpha = 4, 10, 30$	4	1	1	1	1	1
$\omega = 0.5$	$\alpha = 1$	3	3	6	3	2	1
	$\alpha = 4$	3	6	1	2	1	1
	$\alpha = 10$	3	4	1	2	1	1
	$\alpha = 30$	3	1	1	2	1	1
$\omega = 0.8$	$\alpha = 1, 4, 10, 30$	3	1	1	3	1	1

Legend:
1, 4 : *FIM* the best
2 : *FIP* the best
3, 6 : *IPCA* the best

$$
\begin{aligned}
&(1)\quad H_1 : F_{FIM}(x) \le F_{FIP}(x) \le F_{IPCA}(x)\\
&(2)\quad H_1 : F_{FIP}(x) \le F_{FIM}(x) \le F_{IPCA}(x)\\
&(3)\quad H_1 : F_{IPCA}(x) \le F_{FIP}(x) \le F_{FIM}(x)\\
&(4)\quad H_1 : F_{FIM}(x) \le F_{IPCA}(x) \le F_{FIP}(x)\\
&(5)\quad H_1 : F_{FIP}(x) \le F_{IPCA}(x) \le F_{FIM}(x)\\
&(6)\quad H_1 : F_{IPCA}(x) \le F_{FIM}(x) \le F_{FIP}(x),
\end{aligned}
\tag{9}
$$

with at least one strict inequality for any x, at the 0.05 nominal level. For each scenario the best method is then judged as the one that corresponds to the significant test result having the smallest *p-value* on the reference asymptotic normal distribution.

Results are displayed in Table 3. With a significant result, the number of the "most significant" ranking is given according to the numbering of system (9). Moreover, the cells are differently coloured depending on which method appears as the first in the significant ranking. Grey cells denote rankings 1 and 4 where *FIM* is the best. Light-grey cells stand for rankings 2 and 5 where *FIP* is the best. A blank background in the cells denotes rankings 3 and 6 with *IPCA* as the best. Comparisons reported there are made with *FIM* and *FIP* applied at their default options ($q = 0.1$ and $r = 2$). As is apparent, *FIM* in particular performs better than the others in almost all the considered scenarios. *IPCA* works better with balanced moderate or high ρs ($c = 1$ with $\omega = 0.5; 0.8$), with the only exceptions of $\omega = 0.5$, $c = 1$, $\alpha \ge 4$ and $n = 1000$, where *FIP*, followed by *FIM*, is better than *IPCA*.

3.3 Discussion and practical hints for users

A few general considerations about the main findings of the overall simulation study are worth making. As expected, the four methods share the fact that their RMSE tends to increase with the dimensionality of data, especially the number of variables, and to decrease as the value of the correlation parameters increases. This is consistent with the fact that if variables are medium/highly correlated then imputation is generally subject to smaller errors. In addition, the most important factor that seems to discriminate, on the whole, between a "good" and a "less good" imputation method reveals itself to be the type of correlation structure along with the magnitude of correlation coefficients, whose impact can become even stronger if data are skew (see also Solaro et al. 2015a).

Regarding the examined data patterns, the main impressive findings of the work can be given in the form of practical hints for users. In particular:

1. *FIM* works well, especially in the presence of data with small or negative correlations of a same magnitude, or a mix of negative and positive correlations provided that such correlations are strongly unbalanced towards lower absolute values. Such considerations hold particularly for skew data;
2. *FIP* has characteristics similar to *FIM*, but tends to perform better with a slightly higher level of correlations, i.e. small/medium correlations, according to the findings of Sect. 3.2 (Table 2; Fig. 1).

Otherwise, in the presence of either symmetric or skew data with medium/high correlations, other imputation methods such as *IPCA* could give better results. To this regard, it is worth remarking that, while not showing satisfactory results in most of the considered scenarios, *missForest* (not inspected by the J–T test) was expressly designed for imputing mixed-type data. This could explain its lack of effectiveness for quantitative data in the comparisons that were carried out.

As for the choice of the "ideal" donor quantile (*FIM* and *FIP*) and the distance (*FIP*), the simulation study pointed out that:

(i) The choice concerning donor quantiles appeared to be strongly linked to the magnitude of correlation of variables. If correlations are low then selecting a higher proportion of donors (e.g. $q = 0.2$) leads to smaller errors, while if correlations are high having few donors (e.g. $q = 0.05$) implies better results. In general, a good choice seems to fix the percentage of donors equal to 10 % (which is set as default in *FIM* and *FIP*) or 15 %.

(ii) Concerning *FIP*, differences among the various Minkowski distances here considered (i.e. city-block, Euclidean and Tchebycheff distances) did not appear too substantial in the considered comparisons among the methods. The Euclidean distance could hence be used as a default metric. Nonetheless, the performance of *FIP* can improve by taking into account the data structure more carefully, for instance by considering that in the presence of higher levels of correlations the Tchebycheff distance ($r = \infty$) seems a better choice, while the city-block distance ($r = 1$) has proved to be fitting in most of the considered scenarios with low correlations.

4 Consistency of *FIP* and *FIM* with real situations: a case study

In what follows, a case study concerning real data is shown in order to highlight certain good properties of the *ForImp* approach, as outlined in Sect. 2.3. In particular, the imputation performance of the two *ForImp* methods is compared with *missForest* and *IPCA* in the presence of data with multivariate outliers. This case study is carried out using the Swiss Bank Note dataset, which is extensively considered by Atkinson et al. (2004) as an example of real data contaminated with multivariate outliers and underlying group structures. The authors apply the Forward Search (FS) methodology to detect such kinds of departures from the main core of the data. The dataset collects $p = 6$ types of readings (size of bank notes) made on $n = 200$ Swiss bank notes, of which 100 are genuine and 100 are forged. The six variables are: Y_1: length of bank

note near the top; Y_2: left-hand height of bank note; Y_3: right-hand height of bank note; Y_4: distance from bottom of bank note to beginning of the patterned border; Y_5: distance from top of bank note to beginning of the patterned border; Y_6: diagonal distance. A more complete description is provided in the cited reference (Atkinson et al. 2004, pp. 22–30). Imputation in the presence of mixed data is beyond the scope of this work. Accordingly, in the following analysis the group structure is not taken into account, while emphasis is put on the presence of multivariate outliers. A first application of FS on the full dataset (analyses here omitted) revealed that 19 bank notes (with id code: 1, 5, 16, 40, 111, 116, 138, 148, 160, 161, 162, 167, 168, 171, 180, 182, 187, 192, 194) could be considered as outlying with respect to the main core of the data.

To stress the impact of these outlying units, a description of the data pattern and subsequent analyses were performed twice, the first time involving all the units ("with outliers", $n = 200$) and the second time the non-outlying units only ("without outliers", $n = 181$). Regarding the shape of the data, Mardia's multivariate skewness index (Mardia 1970) is equal to 6.9826 ("with outliers") and 1.9891 ("without outliers"). As expected, skewness of the multivariate empirical distribution appears less marked when outliers are deleted. Moreover, the empirical correlation matrix $\mathbf{R}$ contains a mix of positive and negative correlations with $\rho_{\max} = 0.743$ and $\rho_{\min} = -0.623$ ("with outliers") and $\rho_{\max} = 0.743$ and $\rho_{\min} = -0.832$ ("without outliers"), while the mean absolute correlation, given by: $\bar{\rho}_{\text{abs}} = 0.374$ ("with outliers") and $\bar{\rho}_{\text{abs}} = 0.416$ ("without outliers"), shows a moderate average level of the (off-diagonal) correlations in absolute value. These features let us understand that the *ForImp* approach could perform effectively on both datasets, in particular for the moderate average magnitude of correlations along with marked unbalance among them characterizing both datasets. In addition, we expect *FIP* to perform better than *FIM* because of the medium average level of absolute correlation coefficients, and the Tchebycheff distance leading to better results than the city-block or the Euclidean distance given the presence of several high correlations in $\mathbf{R}$.

The original dataset does not contain any missing value. To appraise the performance of the considered methods we proceeded through a simulation study designed as follows. First, we removed the 19 outliers from the dataset. Second, we generated MCAR values at 5, 10, and 20 % on the remaining submatrix with 181 rows. Next, we imputed missing values according to the two approaches "with outliers" and "without outliers". In the with-outliers case, before proceeding to imputation, we added the previously excluded 19 rows to the complete part of the data. In the without-outliers case, we imputed missing values on the outlier-free dataset of $n = 181$ units. Finally, in both the approaches we applied the four imputation methods *FIP*, *FIM*, *IPCA*, and *missForest*. We repeated this procedure 1000 times. Imputation errors were assessed through the RMSE defined in (7).

Figure 2 displays box plots of the RMSE values obtained for *missForest*, *IPCA*, *FIP*, and *FIM* (the latter two with $q = 0.1$) in the case of 20 % missing values. The other two cases of 5 and 10 % reporting the same trend are here omitted. The *IPCA* method proves to be more sensitive to the kind of contamination present in the data, as shown by the two box plots of RMSE in the with- and without-outliers cases. Moreover, *IPCA* has the worst performance, as its errors turn out to be basically higher than those of

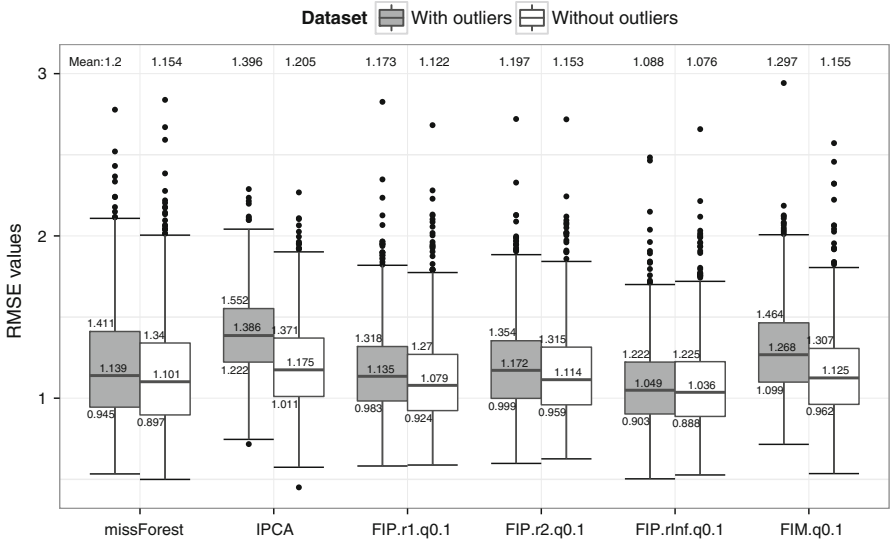

Fig. 2 Swiss Bank Note data imputation in the presence of outliers and without outliers: box plots of RMSE values of *missForest*, *IPCA*, *FIM* and *FIP* ($q = 0.1$) with 20 % of MCAR values

the other methods. *FIM* shows a similar trend though with slightly smaller errors. Its poor performance is clearly caused by the presence of medium/high correlations between the variables, in line with the simulation results of Sect. 3.2. It is also sensitive to the presence of outliers, as appears from the different heights of the two with- and without-outliers box plots. This reveals that the Mahalanobis distance, adjusting for correlations of variables, could cause inclusion of outlying units in the donor set depending on the type of data structure in the multidimensional space. On the other hand, *missForest* and *FIP* are only slightly or not at all sensitive to the presence of outliers. In particular, *FIP* with the Tchebycheff distance (second-last couple of box plots) has its two with- and without-outliers box plots almost perfectly overlapping, thus indicating that it is intrinsically robust to the presence of outliers. Moreover, *FIP* has the smallest errors basically, so that it performs best among the considered methods. Parenthetically, in the with-outliers case we observe that among the three metrics, the Euclidean distance tends to produce slightly higher RMSE values, given that, as known, it is more sensitive to outliers.

5 Conclusions

In this paper, we have introduced a sequential distance-based approach for missing data imputation. We have specified this approach, called Forward Imputation, to the case of quantitative data by proposing two different variants, the first alternating NNI with PCA (*ForImpPCA*), the second using the NNI method with the Mahalanobis distance (*ForImpMahalanobis*). We have evaluated the performance of the proposed methods, also in comparison with two popular competitors (*missForest* and *IPCA*), by

means of an extensive MC simulation study undertaken in the presence of different data patterns. In particular, we have reported and discussed results obtained under skewness and unbalanced correlations, which are features frequently encountered in real applications. We have also considered analyses and comparisons on real data affected by the presence of outliers.

Our approach has turned out to be effective compared to *IPCA* and *missForest* under data structures in general represented by skew-distributed data and low/medium correlations of variables, as also confirmed by the analyses carried out on real data. Further inspections in the presence of different mechanisms generating missing data, or potential data contaminations such as multivariate outliers, or data other than all-quantitative, should be considered in the future in order to better appraise the potentialities of the *ForImp* approach in such different situations. Moreover, the role of its principal constituents, i.e. metrics and donors in NNI and the choice of the MVDA technique, should be examined more carefully for supporting users with additional clues.

Acknowledgments N. Solaro's work was partly funded by the MIUR PRIN "MISURA—Multivariate models for risk assessment" project. The authors would like to thank the Coordinating Editor, Associate Editor and the two anonymous referees for their valuable comments and suggestions, which greatly improved the paper.

Appendix

In *FIP*, the requirement: $n_k \geq p$ for each $k = 0, 1, \ldots, K$ is not binding. For a given k, suppose that $n_k < p$, and consider the correlation matrix $\mathbf{R}_k$ computed from the complete matrix $\mathbf{X}_k$ (similar arguments would hold for the variance-covariance matrix $\mathbf{\Sigma}_k$). Let $\mathbf{Z}_k$ be the standardized matrix derived from $\mathbf{X}_k$. It follows that: $\mathbf{R}_k = \frac{1}{n_k}\mathbf{Z}_k^t\mathbf{Z}_k$, for which: $\text{rank}(\mathbf{R}_k) \leq n_k < p$. Now, consider the $n_k \times n_k$ product-matrix: $\mathbf{P}_k = \mathbf{Z}_k\mathbf{Z}_k^t$. Standard results of matrix algebra ensure that $\mathbf{P}_k$ and $\mathbf{R}_k$ have the same rank. In particular, as $\mathbf{Z}_k$ has zero-mean columns, it always holds that: $\mathbf{Z}_k\mathbf{Z}_k^t\mathbf{1} = 0 \cdot \mathbf{1} = \mathbf{0}$, where $\mathbf{1}$ is a vector of n_k ones, so that $\mathbf{P}_k$ always admits a zero eigenvalue. Therefore, $\text{rank}(\mathbf{P}_k) = \text{rank}(\mathbf{R}_k) \leq n_k - 1$. Moreover, $n_k\mathbf{R}_k$ has the same non-null eigenvalues of $\mathbf{P}_k$, with (at least) extra $p - n_k + 1$ eigenvalues equal to zero.

Let $\eta_s^{(k)}$ be a non-null eigenvalue of $\mathbf{P}_k$, and $\mathbf{v}_s^{(k)}$ the corresponding normalized eigenvector ($s = 1, \ldots, n'_k \leq n_k - 1$, where: $n'_k = \text{rank}(\mathbf{P}_k)$). By definition, $\mathbf{P}_k\mathbf{v}_s^{(k)} = \eta_s^{(k)}\mathbf{v}_s^{(k)}$. Pre-multiplying both members by $\mathbf{Z}_k^t$ leads to: $n_k\mathbf{R}_k\mathbf{Z}_k^t\mathbf{v}_s^{(k)} = \eta_s^{(k)}\mathbf{Z}_k^t\mathbf{v}_s^{(k)}$, from which it is apparent that: $\mathbf{Z}_k^t\mathbf{v}_s^{(k)} = \boldsymbol{\xi}_s^{(k)}$ is the s-th p-dimensional eigenvector of $n_k\mathbf{R}_k$. Finally, from the fact that: $\boldsymbol{\xi}_s^{(k)t}\boldsymbol{\xi}_s^{(k)} = \mathbf{v}_s^{(k)t}\mathbf{P}_k\mathbf{v}_s^{(k)} = \eta_s^{(k)}\mathbf{v}_s^{(k)t}\mathbf{v}_s^{(k)} = \eta_s^{(k)}$, it derives that the s-th normalized eigenvector of $n_k\mathbf{R}_k$, which is also an eigenvector for $\mathbf{R}_k$, is given by: $\boldsymbol{\omega}_s^{(k)} = \boldsymbol{\xi}_s^{(k)}/\sqrt{\eta_s^{(k)}}$, while the s-th eigenvalue of $\mathbf{R}_k$ is: $\lambda_s^{(k)} = \eta_s^{(k)}/n_k$, for $s = 1, \ldots, n'_k \leq n_k - 1 < p$.

References

Atkinson AC, Riani M, Cerioli A (2004) Exploring multivariate data with the Forward Search. Springer, New York

Azzalini A (2015) R package "sn": the skew-normal and skew-t distributions (version 1.2-4). http://azzalini.stat.unipd.it/SN
Azzalini A, Capitanio A (1999) Statistical applications of the multivariate skew normal distribution. J R Stat Soc B 61(3):579–602
Azzalini A, Dalla Valle A (1996) The multivariate skew-normal distribution. Biometrika 83(4):715–726
Breiman L (2001) Random forests. Mach Learn 45:5–32
Cox TF, Cox MAA (2001) Multidimensional scaling, 2nd edn. Chapman & Hall/CRC, Boca Raton
Ferrari PA, Annoni P, Barbiero A, Manzi G (2011) An imputation method for categorical variables with application to nonlinear principal component analysis. Comput Stat Data Anal 55:2410–2420
Gower JC (2005) Principal coordinates analysis. In: Armitage P, Colton T (eds) Encyclopedia of biostatistics. Wiley, New York
Greenacre M (1984) Theory and applications of correspondence analysis. Academic Press, London
Groves RM, Dillman DA, Eltinge JL, Little RJA (2002) Survey nonresponse. Wiley, New York
Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning. Data mining, inference and prediction, 2nd edn. Springer, New York
Hollander M, Wolfe DA (1999) Nonparametric statistical methods, 2nd edn. Wiley-Interscience, New York
Husson F, Josse J (2015) missMDA: Handling missing values with/in multivariate data analysis (principal component methods). R package version 1.8.2. http://CRAN.R-project.org/package=missMDA
Josse J, Pagès J, Husson F (2011) Multiple imputation in principal component analysis. Adv Data Anal Classif 5:231–246
Little RJA, Rubin DB (2002) Statistical analysis with missing data, 2nd edn. Wiley, New York
Mardia KV (1970) Measures of multivariate skewness and kurtosis with applications. Biometrika 57(3):519–530
Nora-Chouteau C (1974) Une méthode de reconstitution et d'analyse de données incomplètes. PhD thesis, Université Pierre et Marie Curie
R Core Team (2015) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org
Rässler S, Rubin DB, Zell ER (2013) Imputation. Wiley Interdiscip Rev Comput Stat 5(1):20–29. doi:10.1002/wics.1240
Rousseeuw PJ, Leroy AM (1987) Robust regression and outlier detection. Wiley, New York
Schafer JL (1997) Analysis of incomplete multivariate data. Chapman and Hall/CRC, London
Solaro N, Barbiero A, Manzi G, Ferrari PA (2014) Algorithmic-type imputation techniques with different data structures: alternative approaches in comparison. In: Vicari D, Okada A, Ragozini G, Weihs C (eds) Analysis and modeling of complex data in behavioural and social sciences. Studies in classification, data analysis, and knowledge organization. Springer International Publishing, Cham, pp 253–261
Solaro N, Barbiero A, Manzi G, Ferrari PA (2015a) A comprehensive simulation study on the Forward Imputation. Working Paper 2015_4, Università degli Studi di Milano, Italy. https://ideas.repec.org/p/mil/wpdepa/2015-04.html
Solaro N, Barbiero A, Manzi G, Ferrari PA (2015b) GenForImp: a sequential distance-based approach for imputing missing data. R package version 1.0.0. http://CRAN.R-project.org/package=GenForImp
Stekhoven DJ (2013). missForest: nonparametric missing value imputation using random forest. R package version 1.4. http://CRAN.R-project.org/package=missForest
Stekhoven DJ, Bühlmann P (2012) MissForest—nonparametric missing value imputation for mixed-type data. Bioinformatics 28(1):112–118
Tarsitano A, Falcone M (2010) Missing values adjustment for mixed-type data. Working Paper n. 15-2010, Università della Calabria, Italy. https://ideas.repec.org/p/clb/wpaper/201015.html
Wasito I, Mirkin B (2005) Nearest neighbour approach in the least-squares data imputation algorithms. Inf Sci 169(1):1–25