Algorithmic imputation techniques for missing data: performance comparisons and development perspectives

Nadia Solaro, Alessandro Barbiero, Giancarlo Manzi and Pier Alda Ferrari

Abstract In recent years, much research has been devoted to solve the problem of missing data imputation. Although most of the novel proposals look attractive for some reason, less attention has been paid to the problem of when and why a particular method should be chosen while discarding the others. This matter is far crucial in applications, given that unsuitable solutions could heavily affect the reliability of statistical analyses. Starting from this, this work is addressed to study how well several algorithmic-type imputation methods perform in the case of quantitative data. We focus on three different logics of imputing, based respectively on the use of random forests, iterative PCA, and the forward procedure. In particular, the latter, having initially been introduced for ordinal data, has required us to develop an original adaptation so that it handles missing quantitative values.

Key words: multivariate exponential power distribution, multivariate skew-normal distribution, nearest neighbour, principal component analysis, random forest

1 Introduction

Missing data have always represented a hard-to-solve problem for researchers from every field. Unsuitable solutions could heavily affect the reliability of statistical results and lead to wrong conclusions. The increasing availability of data often characterized by missing values has paved the way for the development of new alternative

Nadia Solaro
Department of Statistics, Università degli Studi di Milano-Bicocca, Milan, Italy
e-mail: nadia.solaro@unimib.it

Alessandro Barbiero, Giancarlo Manzi, Pier Alda Ferrari
Department of Economics, Management and Quantitative Methods,
Università degli Studi di Milano, Milan, Italy
e-mail: alessandro.barbiero@unimi.it, giancarlo.manzi@unimi.it, pieralda.ferrari@unimi.it

methods for handling missing data. Users are therefore often faced with the dilemma of having to choose among many different imputation techniques, and, moreover, one is not always confident about the adequacy of the imputation exercise. It would therefore be important to find, for different situations and missing data distributions, the best algorithm to be used, as well as to possibly detect turning points where a given technique should be abandoned in favour of others.

This paper intends to offer a first inspection to these issues. We focus on comparing the performance of three specific methodologies which, although founded on very different logics, seem most promising in assigning "good" values to missing data: (i) Stekhoven and Bühlmann's method (*missForest* [9]), which uses an iterative imputation technique based on a "random forest", a random classifier introduced in the context of machine learning [3]; (ii) iterative imputation performed by means of multivariate data analysis techniques, such as the Iterative Principal Component Analysis (*IPCA*) ([8, 6, 7]), which permits to simultaneously estimate missing values and all the parameters connected with the chosen data analysis method; (iii) Ferrari, Annoni, Barbiero, and Manzi's forward imputation method (*ForImp* [4]), a sequential procedure that imputes missing values "forwards" by alternating the nonlinear PCA, carried out step-by-step on each updated complete part of data, and the nearest-neighbour imputation (NNI) method.

Given the wide scope of the subject, this comparative study is here confined to the context of quantitative data. *ForImp* having been designed for ordinal data, it has required us to develop a brand-new version capable to handle missing values for quantitative variables as well (*ForImpPCA*). The three methods *missForest*, *IPCA* and *ForImpPCA* are then compared through an extensive simulation study.

2 Adapting *ForImp* to quantitative data: *ForImpPCA*

The adaptation of *ForImp* [4] to quantitative data that we propose is based on the sequential use of PCA and the NNI method to detect subsets of donors and then impute missing values through opportune weighted averages of donors' values. Let $\mathbf{X}$ be an initial data matrix with x_{ij} values referred to n units (rows) and p quantitative variables (columns), with $n > p$. Assume that at least p rows in $\mathbf{X}$ are without missing values and the other rows contain at most $p-1$ missing values. Then, imputation is performed through the following procedure:

0. *Preliminary step*: split $\mathbf{X}$ into a $(n_0^{(0)} \times p)$-dimensional matrix $\mathbf{X}_0^{(0)}$ free of missing data ($p \leq n_0^{(0)} < n$), and K submatrices $\mathbf{X}_k$ of dimension $(n_k \times p)$, with $k = 1, \ldots, K < p$ expressing the number of missing values in each row. Note that it is not necessary that $n_k > 0$ for all k.
1. *Running PCA*: for k fixed, extract p principal components from either variance-covariance matrix $\mathbf{\Sigma}_0^{(k-1)}$ or correlation matrix $\mathbf{R}_0^{(k-1)}$ of the complete $\mathbf{X}_0^{(k-1)}$ of dimension $(n_0^{(k-1)} \times p)$ to obtain the eigenvalues $\lambda_s^{(k-1)}$ and the eigenvectors $\boldsymbol{\omega}_s^{(k-1)}$ with generic element $\omega_{js}^{(k-1)}$, $j, s = 1, \ldots, p$.

2. *PPC computation*: compute so-called Pseudo Principal Components (PPC) for both submatrices $\mathbf{X}_k$ and $\mathbf{X}_0^{(k-1)}$ by involving only common variables without missing values. Let ι be the set formed by those among the k-combinations of the p indices of variables which have missing values on the rows of $\mathbf{X}_k$. Then PPCs, denoted by $\tilde{C}$, are given by: $\tilde{C}_{s(\iota)}^{(k)} = \sum_{\substack{l=1 \\ l\notin\iota}}^{p} \omega_{ls}^{(k-1)} X_l^{(k)}$ for submatrix $\mathbf{X}_k$, and: $\tilde{C}_{s(\iota)}^{(k-1)} = \sum_{\substack{l=1 \\ l\notin\iota}}^{p} \omega_{ls}^{(k-1)} X_l^{(k-1)}$ for submatrix $\mathbf{X}_0^{(k-1)}$, $s = 1,\ldots,p$.
3. *Donors' selection*: compute the Minkowski distance d_r of order r, $(r \geq 1)$ between each incomplete unit $u_i^{(k)}$ in $\mathbf{X}_k$ and each complete unit $u_c^{(k-1)}$ in $\mathbf{X}_0^{(k-1)}$:

$$d_r(u_i^{(k)}, u_c^{(k-1)}) = \left\{ \sum_{s=1}^{p} \left| (\tilde{c}_{s(\iota),i}^{(k)} - \tilde{c}_{s(\iota),c}^{(k-1)}) w_s^{(k-1)} \right|^r \right\}^{1/r}, \quad c = 1,\ldots,n_0^{(k-1)},$$

where the weight $w_s^{(k-1)}$ is given by: $w_s^{(k-1)} = \sqrt{\lambda_s^{(k-1)} / \sum_{m=1}^{p} \lambda_m^{(k-1)}}$. Then, donors $u_{\delta,i}^{(k)}$ for unit $u_i^{(k)}$ are given by the first $q100\%$ complete units $u_c^{(k-1)}$ corresponding to the q-th quantile $d_{q,i}$ of the distances d_r, $(0 < q < 1; i = 1,\ldots,n_k)$.
4. *Imputation*: for each unit $u_i^{(k)}$, the missing value on variable X_j is imputed with the weighted average:

$$\tilde{x}_{ij}^{(k)} = \frac{\sum_{\delta=1}^{n_\delta} x_{\delta j}^{(k-1)} \frac{1}{d_{\delta i}}}{\sum_{\delta=1}^{n_\delta} \frac{1}{d_{\delta i}}}, \qquad \forall j \in \iota,$$

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 step 3. Next, set up $\mathbf{X}_0^{(k)}$ by row-stacking $\mathbf{X}_0^{(k)}$ with the imputed $\tilde{\mathbf{X}}_k$ and set $k = k+1$.

Steps 1–4 are then iterated until matrix $\mathbf{X}$ is completely imputed.

3 Simulation study

A Monte Carlo simulation study is performed in order to compare the performance of the three imputation techniques. Complete data matrices are generated according to different multivariate distributions. Along with the multivariate normal, two other families of multivariate distributions are considered: the skew-normal [1, 2] and the multivariate exponential power [5].

The simulation study is carried out under different settings defined by the number of variables, association/correlation structures, parameters related to skewness or kurtosis. Missing data in different percentages (5%, 10%, 20%) are then generated through a MCAR mechanism. For each scenario $1,000$ matrices $\mathbf{X}$ with missing data are produced; the three methods are then applied and compared through their

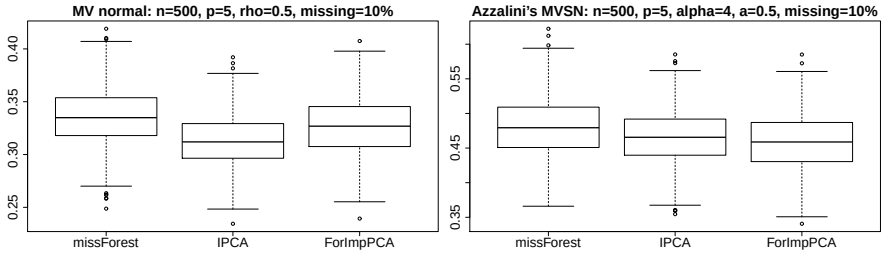

Fig. 1 Simulation results with MV normal (left panel) and Azzalini skew-normal (right panel)

RMSE computed with respect to the original complete data matrix $\mathbf{X}^*$: $RMSE = \sum_{j=1}^{p} \frac{1}{n\sigma_j^2}(\mathbf{x}_j^* - \tilde{\mathbf{x}}_j)^t(\mathbf{x}_j^* - \tilde{\mathbf{x}}_j)$, where $\mathbf{x}_j^*$ is the j-th column vector of $\mathbf{X}^*$, $\tilde{\mathbf{x}}_j$ is the column vector of the imputed data matrix $\tilde{\mathbf{X}}$, and σ_j^2 is the variance of the j-th variable in $\mathbf{X}^*$.

Results achieved up to now show that *IPCA* tends to work well in most of the scenarios considered, especially when distributions are symmetric. *ForImpPCA* tends to have its best performance with skewed distributions and variables not highly correlated (i.e. medium values of association parameter of Azzalini's skew-normal). *missForest* tends to produce the highest RMSE values, but further inspections are needed. Figure 1 reports an example of the typical results we found.

The work done so far seems to be susceptible of further developments. From a methodological point of view, we will investigate potential optimal properties of *ForImpPCA*. We will also consider more complex data structures in order to better highlight the aptitude of *ForImpPCA* to cope with differently skewed distributions.

References

1. Azzalini, A., Dalla Valle, A.: The multivariate skew-normal distribution. Biometrika **83**, 4, 715–726 (1996)
2. Azzalini, A., Capitanio, A.: Statistical applications of the multivariate skew normal distribution. J.R. Statist. Soc. B **61**, 3, 579–602 (1999)
3. Breiman, L.: Random forests. Mach. Learn. **45**, 1, 5–32 (2001)
4. Ferrari, P.A., Annoni, P., Barbiero, A., Manzi, G.: An imputation method for categorical variables with application to nonlinear principal component analysis. Comput. Stat. Data Anal. **55**, 7, 2410–2420 (2011)
5. Gómez, E., Gómez-Villegas, M.A., Marin, J.M.: A multivariate generalization of the power exponential family of distributions. Commun. Stat. Theory Meth. **27**, 3, 589–600 (1998)
6. Greenacre, M.J.: Theory and applications of correspondance analysis. Academic Press, London (1984)
7. Josse, J., Pagès, J., Husson, F.: Multiple imputation in principal component analysis. Adv. Data Anal. Classif. **5**, 3, 231–246 (2011)
8. Nora-Chouteau, C.: Une méthode de reconstitution et d'analyse de données incomplètes. Ph.D. thesis, Université Pierre et Marie Curie (1974)
9. Stekhoven, D.J., Bühlmann, P.: MissForest - nonparametric missing value imputation for mixed-type data. Bioinformatics **28**, 1, 112–118 (2012)