Algorithmic-Type Imputation Techniques with Different Data Structures: Alternative Approaches in Comparison

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

Abstract In recent years, with the spread availability of large datasets from multiple sources, increasing attention has been devoted to the treatment of missing information. Recent approaches have paved the way to the development of new powerful algorithmic techniques, in which imputation is performed through computer-intensive procedures. Although most of these approaches are attractive for many reasons, less attention has been paid to the problem of which method should be preferred according to the data structure at hand. This work addresses the problem by comparing the two methods *missForest* and *IPCA* with a new method we developed within the forward imputation approach. We carried out comparisons by considering different data patterns with varying skewness and correlation of variables, in order to ascertain in which situations a given method produces more satisfying results.

Keywords Forward imputation • Iterative PCA • missForest • Missing data

1 Missing Data Treatment

Missing data treatment is frequently invoked when performing data analysis. There exists no field of quantitative research where missing information is not a problem, and an optimal choice of an imputation procedure should be a guarantee of

N. Solaro (✉)
Department of Statistics and Quantitative Methods, Università di Milano-Bicocca, Milan, Italy
e-mail: nadia.solaro@unimib.it

A. Barbiero • G. Manzi • P.A. Ferrari
Department of Economics, Management and Quantitative Methods, Università di Milano, Milan, Italy
e-mail: alessandro.barbiero@unimi.it; giancarlo.manzi@unimi.it; pieralda.ferrari@unimi.it

D. Vicari et al. (eds.), *Analysis and Modeling of Complex Data in Behavioral and Social Sciences*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-319-06692-9_27,

reliable statistical analyses. In modern missing data handling, two broad taxonomies dominate recent literature: (1) parametric and nonparametric methods; (2) single and multiple imputation (Little and Rubin 2002). In parametric methods, likelihood-based procedures (e.g. the EM algorithm) are applied starting from a distributional assumption on the missing part of data in order to obtain estimates of missing values according to their generating model. Nonparametric missing data procedures are model-free methods that do not require distributional assumptions on the data. Imputation is thus performed by learning from the data structure at hand. While single imputation is concerned with the problem of assigning a single value to each missing datum, multiple imputation aims at accounting for the uncertainty implicit in the fact that the imputed values are not the actual values. This is achieved by deliberately adding sources of error during the imputation process, thus giving rise to a multitude of estimates for each missing datum from which standard errors and confidence intervals can be computed.

Among nonparametric single imputation techniques, methods based on computer-intensive iterative statistical procedures seem the most promising in producing reliable imputations. In this work, attention is specifically drawn to three different logics of imputing, based on the use of random forest (Stekhoven and Bühlmann 2012), iterative PCA (Nora-Chouteau 1974) and the forward (Ferrari et al. 2011) procedures respectively. In particular, Stekhoven and Bühlmann's method (*missForest*, Stekhoven and Bühlmann 2012) is an iterative technique for the imputation of continuous and/or categorical data based on a random forest, which is a random classifier introduced in the context of machine learning (Breiman 2001). The Iterative Principal Component Analysis (*IPCA*) (Greenacre 1984; Nora-Chouteau 1974) imputes missing values simultaneously by an iterative use of the principal component analysis. It has recently been subject to renewed interest as it is at the core of the multiple imputation technique with PCA, a component of a more general methodology (*missMDA*) introduced by Josse et al. (2011) for imputing missing data with multivariate data analysis techniques. The Forward Imputation (*ForImp*) by Ferrari et al. (2011) is a sequential procedure designed for extracting a latent dimension from ordinal variables in the presence of missing data. The nonlinear PCA (NLPCA) and the nearest-neighbour imputation (NNI) method are alternated in a step-by-step process that recovers the missing ordinal categories and then extracts the latent dimension.

Although grounded on distinct logics, *IPCA* and *ForImp* both depend on factorial methods, which are widely used also in contexts where the incompleteness of information requires a different approach from a purely imputation perspective. This is the case of data fusion and data grafting procedures which, allowing databases from different sources to be combined together by recovering mismatches of variables and/or units, can be regarded as special cases of missing data imputation (Aluja-Banet et al. 2007; Saporta 2002).

This work has two objectives. The first is to re-formulate *ForImp* as an imputation technique for quantitative variables. Indeed, in its original version *ForImp* was not expressly developed as an imputation method, but rather as a method for missing data handling in NLPCA in alternative to commonly used standard options, such as

passive treatment (Ferrari et al. 2011). The second is to offer a critical comparison of the thus revised *ForImp* with *missForest* and *IPCA* based on various configurations of quantitative data as given by different patterns of skewness and correlation of variables.

2 The Forward Imputation for Quantitative Variables

Since our goal is to re-design the *ForImp* method as a pure imputation technique, we specifically focused on missing data handling in the case of quantitative variables. Accordingly, we relied on the traditional linear PCA to build up the new version of the method, which will be termed Forward Imputation with the PCA (*ForImpPCA*). Although the logic behind *ForImpPCA* is very similar to the original *ForImp* (Ferrari et al. 2011), it is characterized by several features. Since the dimensionality reduction problem is not the primary concern, the PCA method is merely involved as a tool functional to the imputation exercise. In particular, the same number of principal components are extracted as the number of variables in the starting data matrix, in order to produce convenient synthesis indicators that are more or less related to the original variables.

The *ForImpPCA* method assumes an $n \times p$ quantitative data matrix $\mathbf{X}$ with x_{ij} values ($i = 1, \ldots, n$, $j = 1, \ldots, p$) with at least p rows free of missing values and the other $n - p$ rows with at most $p - 1$ missing values ($n > p, p \geq 2$). Then, in a preliminary phase, data are prepared by splitting $\mathbf{X}$ into a complete submatrix $\mathbf{X}_0$ and K submatrices $\mathbf{X}_k$, where index k denotes the number of missing values potentially contained in each row ($k = 1, \ldots, K \leq p - 1$). Should k identify a submatrix without elements, we would set: $\mathbf{X}_k = \mathbf{X}_{0 \times p}$, and then jump to the submatrix corresponding to the subsequent k. The core steps of the *ForImpPCA* algorithm are the following:

– Set $k = 1$.

1. *PCA step*: Perform a PCA on the complete $\mathbf{X}_{k-1}$ from either its own variance-covariance matrix or correlation matrix, assumed of full rank, and obtain eigenvalues $\lambda_s^{(k-1)}$ and eigenvectors $\boldsymbol{\omega}_s^{(k-1)}$ with generic element $\omega_{js}^{(k-1)}$ from it, ($j, s = 1, \ldots, p$).
2. *PPC step*: Compute so-called Pseudo Principal Components (PPC) for both the complete $\mathbf{X}_{k-1}$ and the incomplete $\mathbf{X}_k$ by involving only common variables without missing values and eigenvectors obtained at the previous step, in order to obtain artificial variables free of missing values for both complete and incomplete units. We denote by ι the set formed by those among the k-combinations of the p indices of variables containing missing values in the rows of $\mathbf{X}_k$. Then PPCs, denoted by $\tilde{C}$, are given by linear combinations of the original variables outside the ι set with coefficients given by the element in the corresponding eigenvectors:

$\tilde{C}^{(k)}_{s(\iota)} = \sum_{\substack{l=1 \\ l \notin \iota}}^{p} \omega^{(k-1)}_{ls} X^{(k)}_{l}$ for submatrix $\mathbf{X}_k$, and: $\tilde{C}^{(k-1)}_{s(\iota)} = \sum_{\substack{l=1 \\ l \notin \iota}}^{p} \omega^{(k-1)}_{ls} X^{(k-1)}_{l}$ for submatrix $\mathbf{X}_{k-1}$, $s = 1, \ldots, p$.

3. *Donors' selection step*: PPCs represent common, complete information for the comparison of complete and incomplete units. PPCs are accordingly used to compute the Minkowski distance d_r of order r, $(r \geq 1)$, between each incomplete unit $u^{(k)}_i$ in $\mathbf{X}_k$ and the complete units $u^{(k-1)}_c$ in $\mathbf{X}_{k-1}$:

$$d_r(u^{(k)}_i, u^{(k-1)}_c) = \left\{ \sum_{s=1}^{p} \left| (\tilde{c}^{(k)}_{s(\iota),i} - \tilde{c}^{(k-1)}_{s(\iota),c}) w^{(k-1)}_s \right|^r \right\}^{1/r}, \quad c = 1, \ldots, n_{k-1}, \quad (1)$$

where the weights: $w^{(k-1)}_s = \sqrt{\lambda^{(k-1)}_s / \sum_{m=1}^{p} \lambda^{(k-1)}_m}$, being the square root of normalized eigenvalues, are used to strengthen (weaken) the role of PPCs derived from principal components with higher (smaller) variances. Thereafter, donors are detected as an opportune percentage of the complete units nearest to a specific incomplete unit. Formally, donors $u^{(k)}_{\delta,i}$ for unit $u^{(k)}_i$ are given by the first $q100\,\%$ complete units $u^{(k-1)}_c$ corresponding to the q-th quantile $d_{q,i}$ of the distances d_r, $(0 < q < 1; i = 1, \ldots, n_k)$.

4. *Imputation step*: Once the donors have been identified, their values in the original data matrix are used for imputation by means of a weighted average. Weights are given by the reciprocals of the distances between donors and each specific incomplete unit in order to put more (less) emphasis on less (more) distant donors. For a missing value on variable X_j and unit $u^{(k)}_i$ the imputed value is therefore given by:

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

where n_δ is the total number of donors for $u^{(k)}_i$ and $d_{\delta i}$ is the distance between the δ-th donor and unit $u^{(k)}_i$ as computed in step 3.

– Set $k = k + 1$ and jump to the *PCA step* until $\mathbf{X}$ is completely imputed.

3 A Data Structure-Driven Simulation Study for Comparison

A Monte Carlo simulation study was carried out to assess the performance of the *ForImpPCA* method by comparing it with *missForest* and *IPCA* in the presence of different data patterns and Missing Completely At Random (MCAR) generated missing values (Little and Rubin 2002). In this study, attention was specifically addressed to skewed data structures, in order to verify whether and to what extent

Table 1 Experimental conditions in the simulation study (1,000 runs for each scenario)

Common set of experimental conditions:	
– Number of variables in $\mathbf{X}$	$p = 3; 5; 10$
– Number of units in $\mathbf{X}$	$n = 500; 1{,}000$
– Percentage of MCAR missing values	5 %; 10 %; 20 %
Data generation from $N_p(\mathbf{0}, \mathbf{R})$:	
– Correlation coefficient	$\rho = 0; 0.3; 0.7$
Data generation from $\text{MSN}_p(\boldsymbol{\Omega}, \boldsymbol{\alpha})$:	
– Skewness parameter	$\alpha = 1; 4; 10; 30$
– Correlation parameter in $\boldsymbol{\Omega}$	$\omega = 0; 0.5; 0.8$

skewness could affect the imputation capability of the three methods. Accordingly, complete data matrices were randomly generated from both the multivariate normal (*MVN*) distribution and the multivariate skew normal (*MSN*) family of distributions, the latter being an extension of the multivariate normal distribution allowing for the presence of skewness (Azzalini and Capitanio 1999; Azzalini and Dalla Valle 1996). To better understand the role of *MSN* parameters involved in the simulation study, it is worth recalling that a p-dimensional random vector $\boldsymbol{X}$ is $\text{MSN}_p(\boldsymbol{\Omega}, \boldsymbol{\alpha})$ distributed if its density function (d.f.) can be expressed as:

$$f(\mathbf{x}; \boldsymbol{\Omega}, \boldsymbol{\alpha}) = 2\phi_p(\mathbf{x}; \boldsymbol{\Omega})\Phi(\boldsymbol{\alpha}^t \mathbf{x}), \tag{2}$$

where: $\phi_p(\mathbf{x}; \boldsymbol{\Omega})$ is the $N_p(\mathbf{0}, \boldsymbol{\Omega})$ d.f., with $\boldsymbol{\Omega}$ a correlation matrix of full rank; $\Phi(\cdot)$ is the $N(0, 1)$ distribution function, and $\boldsymbol{\alpha}$ is a p-dimensional parameter vector regulating the skewness. In particular, if: $\boldsymbol{\alpha} = \mathbf{0}$, then the d.f. (2) reduces to a multivariate normal: $\boldsymbol{X} \sim N_p(\mathbf{0}, \boldsymbol{\Omega})$.

We generated data from both *MVN* and *MSN* distributions according to the simulation settings reported in Table 1, for a total number of, respectively, 54 scenarios in the case of *MVN*, and 216 in the case of *MSN*. Specifically, in each scenario a complete data matrix $\mathbf{X}^*$ was generated from an *MVN* or an *MSN* distribution, and then 1,000 matrices $\mathbf{X}_t$ were formed from it with a given percentage of MCAR missing data, $t = 1, \ldots, 1{,}000$ (Table 1). Then, *missForest*, *IPCA* and *ForImpPCA* were applied with the following options. For *missForest*, the maximum number of iterations was increased from 10 (the default in the R library *missForest*, Stekhoven and Bühlmann 2012) to 50. For *IPCA*, the number of extracted principal components was fixed to the maximum possible, i.e. $p - 2$, with $p \geq 3$ (R library *missMDA*, Josse et al. 2011). For *ForImpPCA*, we considered the Euclidean distance ($r = 2$ in formula (1)), and the first q-th quantile of such distances with $q = 0.05; 0.1; 0.15; 0.2$ in order to detect donors.

Simulation results were synthesized, and comparisons among the three methods performed, through the Relative Mean Square Error (*RMSE*) computed as a function of the difference between the complete data matrix $\mathbf{X}^*$ and the imputed data matrix $\tilde{\mathbf{X}}_t$ at the t-th simulation run: $RMSE_t = \sum_{j=1}^{p} \frac{1}{n\sigma_j^2}(\mathbf{x}_j^* - \tilde{\mathbf{x}}_{j,t})^t(\mathbf{x}_j^* - \tilde{\mathbf{x}}_{j,t})$, where $\mathbf{x}_j^*$ is the j-th column vector of $\mathbf{X}^*$, $\tilde{\mathbf{x}}_{j,t}$ is the j-th column vector of $\tilde{\mathbf{X}}_t$, and σ_j^2 is the

variance of the j-th variable in $\mathbf{X}^*$, ($t = 1, \ldots, 1{,}000$). Codes of *ForImpPCA* were implemented and simulations performed in the R environment (R Development Core Team 2012).

3.1 Simulation Results

Figure 1 shows line plots of *RMSE* median values, plotted against the percentages of MCAR missing values (5 %; 10 %; 20 %), obtained for the three methods (*ForImpPCA* with $q = 0.1$) under a subset of the scenarios considered, with the number of variables varying ($p = 3; 5; 10$), number of units fixed to $n = 1{,}000$, and data generated from *MVN* (with $\rho = 0; 0.3; 0.7$) and *MSN* (with $\omega = 0; 0.5; 0.8$ and $\alpha = 4; 30$). The other omitted results exhibit the same trend. Two remarks are worth making. First, as expected, *RMSE* increases as the complexity of the data increases, that is, the number of variables and the proportion of missing values. Moreover, *ceteris paribus*, *RMSE* tends to decrease as the correlation between variables increases, thus indicating that the imputation process is more effective if variables are closely related. Second, the three methods produce very similar *RMSE* values with a low percentage of missing values, whereas they display a noticeably different performance in the presence of higher proportions of missing data. In particular, *IPCA* turns out to be the best imputation method in the case of normally distributed data (1st row of panels, Fig. 1), and highly correlated variables (2nd and 3rd rows, last column, Fig. 1), while *ForImpPCA* tends to perform best with skew distributions and variables with small/medium correlations (2nd and 3rd rows, first two columns, Fig. 1). Finally, *missForest* tends to produce the highest *RMSE* values in most scenarios considered, although it must be remembered that it is designed especially for imputation in the case of mixed-type data.

Figure 2 displays a more detailed picture of the results achieved in the specific scenarios with $p = 5$ variables, $n = 1{,}000$ units, and 20 % of missing data. In addition to *missForest* and *IPCA*, boxplots of *RMSE* distributions are shown also for *ForImpPCA* with different donors' quantiles ($q = 0.05; 0.1; 0.15; 0.2$), in order to check their effect on the imputation task. The above remarks concerning *IPCA* and *ForImpPCA* can now be understood more clearly. The best performance of *IPCA* can be observed in the first row of panels, while 2nd to 4th rows in the first two columns highlight the best performance of *ForImpPCA*. Moreover, a comparison among boxplots of *ForImpPCA* pertaining to different donors' quantiles suggests that, overall, having a high percentage of donors is not a convenient choice if variables are highly correlated (last column of panels, Fig. 2), while having few donors is not suitable if variables are uncorrelated or little correlated (1st column, Fig. 2). This would seem to indicate that a good choice is to select donors that correspond to the first $q = 0.1$ or $q = 0.15$ quantile of Euclidean distances.

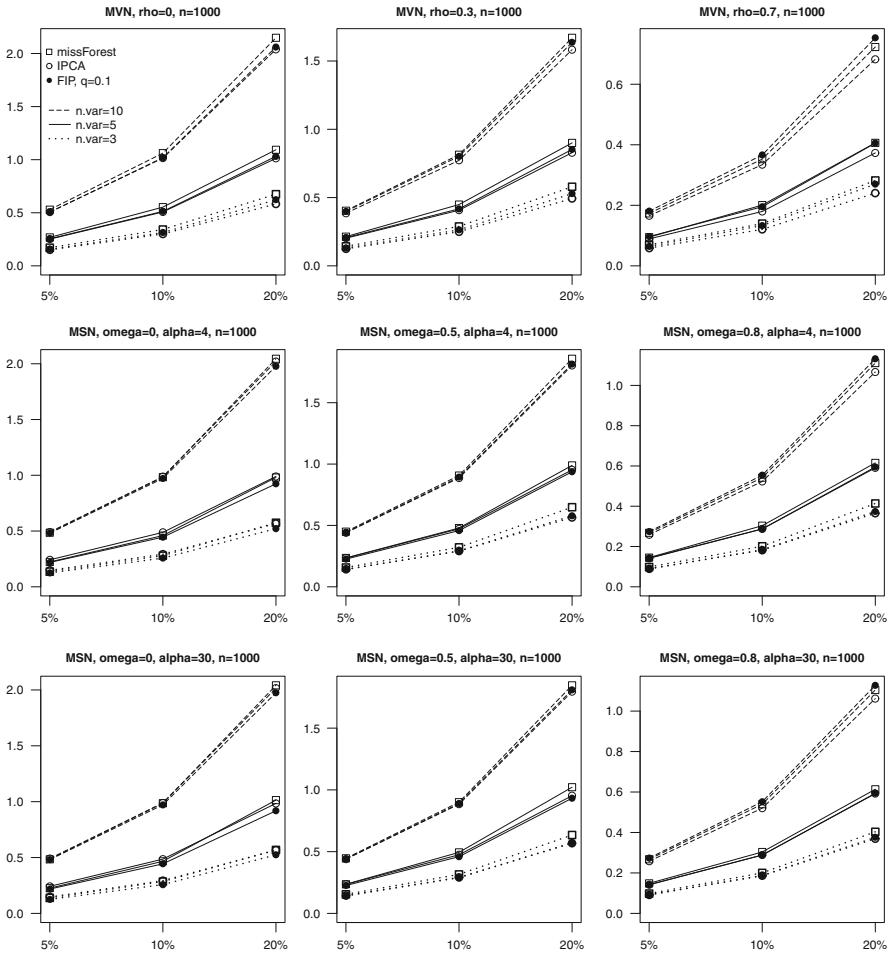

Fig. 1 Line plots of *RMSE* median values of *missForest*, *IPCA*, and *ForImpPCA* (FIP), plotted against percentages of MCAR missing data with $p = 3; 5; 10$ variables and $n = 1{,}000$ units

4 Discussion and Future Work

In the light of our current results, *ForImpPCA* seems to be promising as a single imputation method. It performs best with skew distributions and variables which are not highly correlated, characteristics typically encountered in real data. Nonetheless, further studies would help investigate the performance of *ForImpPCA* more thoroughly. For example, the results obtained indicate that it would be useful to examine *ForImpPCA*, and to then compare it with other methods, in the presence of data contaminations such as multivariate outliers, or a different generating mechanism of missing data, such as MAR (Little and Rubin 2002). From a methodological point of

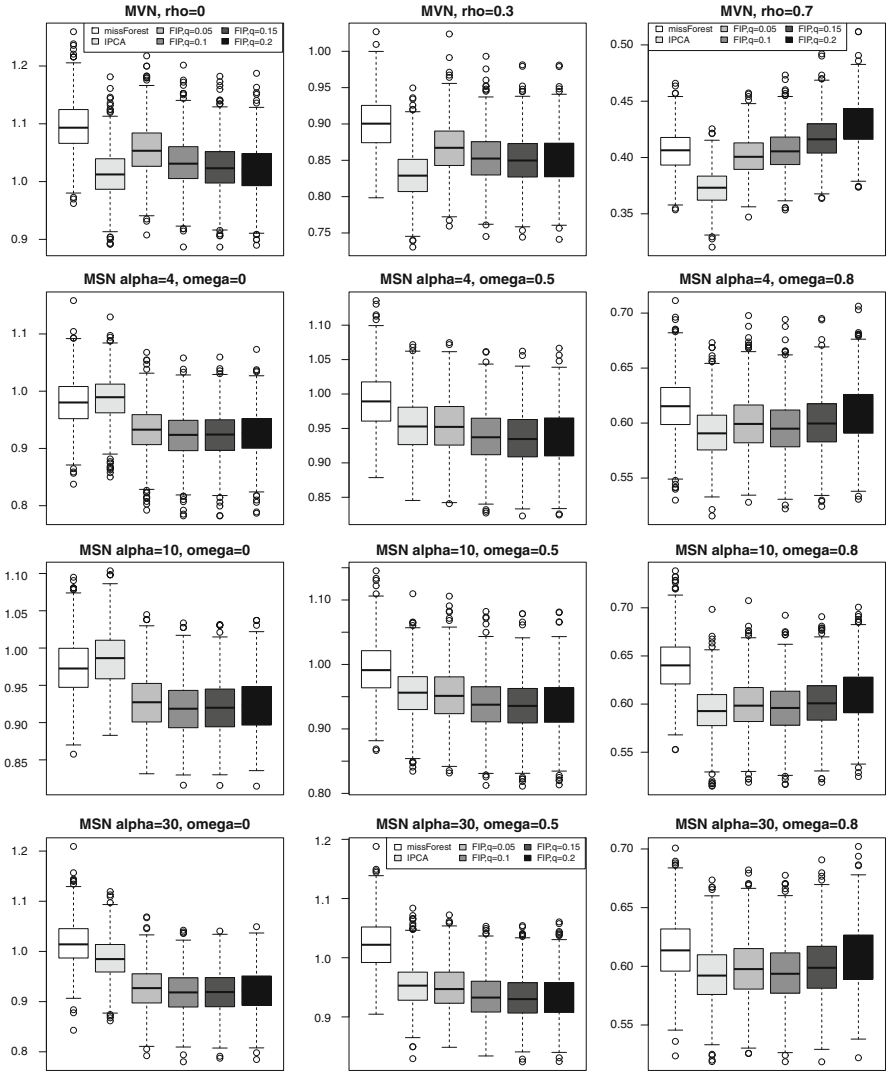

Fig. 2 Boxplots of *RMSE* distributions of *missForest*, *IPCA*, and *ForImpPCA* (FIP) with $q = 0.05, 0.1, 0.15, 0.2$ donors' quantile, under the scenarios with $p = 5$ variables, $n = 1{,}000$ units, and 20 % of MCAR missing data

view, the potentially optimal properties of *ForImpPCA* along with its performance in cases of more complex data structures need to be further investigated in order to highlight the capacity of *ForImpPCA* to manage different skew distributions better.

References

Aluja-Banet, T., Daunis-i-Estadella, J., & Pellicer, D. (2007). GRAFT, a complete system for data fusion. *Computational Statistics & Data Analysis, 52*, 635–649.

Azzalini, A., & Capitanio, A. (1999). Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society: Series 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. *Machine Learning, 45*(1), 5–32.

Ferrari, P. A., Annoni, P., Barbiero, A., & Manzi, G. (2011). An imputation method for categorical variables with application to nonlinear principal component analysis. *Computational Statistics & Data Analysis, 55*, 2410–2420.

Greenacre, M. (1984). *Theory and applications of correspondance analysis*. London: Academic.

Josse, J., Pagès, J., & Husson, F. (2011). Multiple imputation in principal component analysis. *Advances in Data Analysis and Classification, 5*, 231–246.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley.

Nora-Chouteau, C. (1974). Une méthode de reconstitution et d'analyse de données incomplètes. Ph.D. thesis, Université Pierre et Marie Curie.

R Development Core Team (2012). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Saporta, G. (2002). Data fusion and data grafting. *Computational Statistics & Data Analysis, 38*, 465–473.

Stekhoven, D. J., & Bühlmann, P. (2012). MissForest - non-parametric missing value imputation for mixed-type data. *Bioinformatics, 28*(1), 112–118.