Adv Data Anal Classif (2011) 5:231–246
DOI 10.1007/s11634-011-0086-7

REGULAR ARTICLE

Multiple imputation in principal component analysis

Julie Josse · Jérôme Pagès · François Husson

Received: 8 January 2010 / Revised: 17 January 2011 / Accepted: 16 February 2011 /
Published online: 6 March 2011

Abstract The available methods to handle missing values in principal component analysis only provide point estimates of the parameters (axes and components) and estimates of the missing values. To take into account the variability due to missing values a multiple imputation method is proposed. First a method to generate multiple imputed data sets from a principal component analysis model is defined. Then, two ways to visualize the uncertainty due to missing values onto the principal component analysis results are described. The first one consists in projecting the imputed data sets onto a reference configuration as supplementary elements to assess the stability of the individuals (respectively of the variables). The second one consists in performing a principal component analysis on each imputed data set and fitting each obtained configuration onto the reference one with Procrustes rotation. The latter strategy allows to assess the variability of the principal component analysis parameters induced by the missing values. The methodology is then evaluated from a real data set.

Keywords Principal component analysis · Missing values · EM algorithm · Multiple imputation · Bootstrap · Procrustes rotation

Mathematics Subject Classification (2000) 62H25 · 62G09

1 Introduction

Single imputation methods fill in missing values with plausible values. These methods are quite appealing because they lead to a complete data set that can be analysed with

J. Josse (✉) · J. Pagès · F. Husson
Agrocampus Ouest, 65 rue de St-Brieuc, 35042 Rennes, France
e-mail: julie.josse@agrocampus-ouest.fr

any statistical methods. However, the most classical imputation methods such as mean imputation or regression imputation distort the marginal and joint distribution of the variables (Little and Rubin 1987, 2002). It is problematic since many statistical methods rely on the estimation of the vector of means and of the covariance matrix. These distributions can be preserved using more complex imputation methods such as the stochastic regression imputation. This latter consists in imputing with the predicted values from a regression model plus a random noise drawn from a normal distribution with variance equal to the residual variance. However, "the imputed data set still fails to provide an accurate measure of variability; it fails to account for missing data uncertainty" (Schafer and Olsen 1998). Indeed, if the imputed values are considered as observed values in the analysis, we forget that they are not observed values but only predicted ones. Consequently, the uncertainty of the prediction is not taken into account in the subsequent analyses. This implies that standard errors of the parameters calculated from the imputed data set are underestimated (Little and Rubin 1987, 2002, p.65) which lead to confidence intervals and tests that are not valid even if the imputation model is correct.

Multiple imputation (MI) has been proposed by Rubin (1987) to provide both estimation of the parameters of interest and estimation of their variability in the missing data framework. It relies on the principle that a unique value cannot reflect the uncertainty of the prediction of the missing value. Multiple imputation consists first in generating D plausible values for each missing value which leads to D imputed data sets. A way to generate multiple imputed data sets could be to repeat the stochastic regression imputation process: D draws are generated from the predictive distribution of the missing values given the observed values and the estimation of the parameters. However, Little and Rubin (1987, 2002, p.214) qualified this multiple imputation as "improper" since the regression parameters are considered as "true" parameters while there are only sample estimates. A "proper" imputation reflects the uncertainty of the parameters from one imputation to the next. After obtaining D completed data sets, multiple imputation consists in performing the statistical analysis of each imputed data set and estimating the quantity of interest θ. Finally, results are combined to obtain an estimate for θ and for its variability which takes into account the uncertainty due to the missing values.

In this paper we focus on handling missing values in an exploratory multivariate data analysis framework and especially in principal component analysis (PCA). The missing data mechanism is considered as missing at random (MAR) in the sense of Little and Rubin (1987, 2002). In PCA, several algorithms have been proposed to deal with missing values Gabriel and Zamir 1979; Kiers 1997; Josse et al. 2009. These algorithms give both a point estimation of the parameters (principal axes and components) and an estimation of the missing values. However, very few studies (Adams et al. 2002) have focused on computing the standard deviation of the parameters in a missing values framework. This paper proposes a version of multiple imputation adapted to the framework of principal component analysis. Contrary to the methodology proposed in the missing data literature (Little and Rubin 1987, 2002) and in Adams et al. (2002), we do not evaluate the standard errors of the parameters of interest (axes and components) taking into account missing data uncertainty but we focus on assessing the variability due to missing values. It means we focus on assessing the

influence of the different possible predictions of the missing values obtained from a PCA model on the PCA results. Such a method allows the user to know if the results obtained from PCA algorithms which handle missing values are reliable.

Section 2 describes different approaches to deal with missing values in PCA and gives their properties. Sect. 3 describes first a methodology to generate D imputed data sets from a PCA model. Then, two approaches are proposed to visualize the uncertainty due to the missing values onto the PCA results. Finally, Section 4 illustrates the methodology on an example.

2 PCA with missing values

2.1 Weighted least squares

Let $\mathbf{X}$ be an $I \times K$ data matrix and $\|\mathbf{A}\| = \sqrt{\text{tr}(\mathbf{A}\mathbf{A}')}$ the Frobenius norm of matrix $\mathbf{A}$. PCA is a well known method to reduce the dimensionality of a data set. It provides the best low rank $S < K$ approximation of a matrix $\mathbf{X}$ in the least squares sense. Indeed, it finds $\mathbf{F}_{I\times S}$ and $\mathbf{U}_{K\times S}$ that minimize the reconstruction error:

$$\mathcal{C} = \|\mathbf{X} - \mathbf{M} - \mathbf{F}\mathbf{U}'\|^2 = \sum_{i=1}^{I}\sum_{k=1}^{K}\left(x_{ik} - m_k - \sum_{s=1}^{S} f_{is}u_{ks}\right)^2, \quad (1)$$

where $\mathbf{M}$ is an $I \times K$ matrix with each row equal to $(m_1, \ldots m_K)$ the vector of the columns mean. With the additional constraints that the columns of $\mathbf{U}$ are orthogonal and of unit norm, the solution is given by the principal components $\hat{\mathbf{F}}$ (the scores matrix, such that the variance of each column is equal to the corresponding eigenvalue) and the principal axes $\hat{\mathbf{U}}$ (the loadings matrix), eigenvectors of respectively the inner-product matrix and the covariance matrix. These parameters can be obtained, for example, by the singular value decomposition of $(\mathbf{X} - \mathbf{M})$ or by the use of alternating least squares algorithms.

An approach commonly used to deal with missing values consists in ignoring the missing values by minimizing the reconstruction error over all non-missing elements. Let $\mathbf{W}$ be a weight matrix ($w_{ik} = 0$ if x_{ik} is missing and $w_{ik} = 1$ otherwise), the criterion becomes:

$$\mathcal{C} = \|\mathbf{W} * (\mathbf{X} - \mathbf{M} - \mathbf{F}\mathbf{U}')\|^2 = \sum_{i=1}^{I}\sum_{k=1}^{K} w_{ik}\left(x_{ik} - m_k - \sum_{s=1}^{S} f_{is}u_{ks}\right)^2, \quad (2)$$

with $*$ the Hadamard product. In contrast to the complete case, there is no explicit solution to minimize the criterion (2) and it is necessary to resort to iterative algorithms. The mean parameter $\mathbf{M}$ has to be included in the criterion (2) because, as $\mathbf{F}$ and $\mathbf{U}$, this parameter has to be updated during the estimation process.

NIPALS (Wold 1966) is a well known algorithm which proceeds dimension by dimension and estimates the first axis and the first component by alternating two weighted simple regressions: for a fixed $\mathbf{f}_1$, the axis $\mathbf{u}_1$ is estimated by weighted least

squares and for a fixed $\mathbf{u}_1$, the component $\mathbf{f}_1$ is estimated by weighted least squares. The second dimension is obtained by applying the same procedure to the residual matrix $\mathbf{X} - \hat{\mathbf{f}}_1 \hat{\mathbf{u}}_1'$ and so on for the next dimensions. NIPALS can be seen as an extension of the power method, algorithm which computes the first eigenvector of a matrix, for the incomplete case (Golub and Van Loan 1996). NIPALS provides quite reasonable results when there are few missing values but encounters many difficulties when the number of missing values increases. Moreover, the dimensions are not orthogonal, centring and scaling are not updated during the estimation process (Josse et al. 2009), the criterion (2) is not minimized, etc. That is why many authors (Gabriel and Zamir 1979; Grung and Manne 1998; Kroonenberg 2008) have suggested to resort to criss-cross multiple regression (Gabriel and Zamir 1979). This method is an extension of NIPALS which seeks the entire subspace ($S > 1$) directly rather than sequentially. Two weighted multiple regressions, instead of two simple regressions, are repeated until convergence. After each iteration the criterion (2) decreases until the algorithm converges to a (possibly local) minimum.

2.2 Iterative PCA

In exploratory multivariate data analysis methods and first in correspondence analysis (CA), methods of iterative imputation have been proposed by Nora-Chouteau (1974) and Greenacre (1984, p.238) to deal with missing values. These methods consist in setting the missing elements at initial values, performing the analysis (such as CA) on the completed data set, filling-in the missing values with the reconstruction formulae (the CA model) using a predefined number of dimensions and repeating the procedure on the newly obtained matrix until the total change in the matrix falls below an empirically determined threshold. The missing values and the parameters are estimated simultaneously. Such an algorithm is available in PCA and is named iterative PCA. Kiers (1997) has shown that this latter algorithm also minimizes the criterion (2) and can be considered as an alternative to the criss-cross multiple regression algorithm. The iterative PCA algorithm is:

1. initialization $\ell = 0 : \mathbf{X}^0$ is obtained by substituting missing values with initial values (for example with column means on the non-missing entries); $\hat{\mathbf{M}}^0$ is computed;
2. step ℓ :
   (a) find $(\hat{\mathbf{F}}^\ell_{I\times S}, \hat{\mathbf{U}}^\ell_{K\times S})$ such as :
   $(\hat{\mathbf{F}}^\ell, \hat{\mathbf{U}}^\ell) = \text{argmin}_{(\mathbf{F},\mathbf{U})} \|\mathbf{X}^{\ell-1} - \mathbf{F}\mathbf{U}' - \hat{\mathbf{M}}^{l-1}\|^2$;
   (b) missing values in $\mathbf{X}$ are replaced by the fitted values $\hat{\mathbf{X}}^\ell = \hat{\mathbf{F}}^\ell \hat{\mathbf{U}}^{\ell\prime} + \hat{\mathbf{M}}^{\ell-1}$. The new imputed data set is $\mathbf{X}^\ell = \mathbf{W} * \mathbf{X} + (1 - \mathbf{W}) * \hat{\mathbf{X}}^\ell$;
   (c) $\hat{\mathbf{M}}^\ell$ is computed on $\mathbf{X}^\ell$.
3. steps (2.a) (2.b) and (2.c) are repeated until convergence.

The PCA step (2.a) can be computed by the singular value decomposition of $(\mathbf{X}^{\ell-1} - \hat{\mathbf{M}}^{l-1})$ or by an alternating least squares algorithm since both minimize the criterion. To reduce the computational cost, the step (2.a) can be replaced by only

one step of the alternating least squares algorithm:

$$\hat{\mathbf{U}}^{\ell} = (\mathbf{X}^{\ell-1} - \hat{\mathbf{M}}^{l-1})'\hat{\mathbf{F}}^{\ell-1}(\hat{\mathbf{F}}^{\ell-1\prime}\hat{\mathbf{F}}^{\ell-1})^{-1},$$
$$\hat{\mathbf{F}}^{\ell} = (\mathbf{X}^{\ell-1} - \hat{\mathbf{M}}^{l-1})\hat{\mathbf{U}}^{\ell}(\hat{\mathbf{U}}^{\ell\prime}\hat{\mathbf{U}}^{\ell})^{-1}.$$

In that case, the criterion in step (2.a) is decreased instead of minimized but the general criterion (2) is still minimized.

Iterative PCA can also be seen as a particular expectation-maximization (EM) algorithm of the "fixed effect" model (Caussinus 1986) where data are generated as a structure corrupted by noise:

$$x_{ik} = m_k + \sum_{s=1}^{S} f_{is}u_{ks} + \varepsilon_{ik}, \quad \text{with } \varepsilon_{ik} \sim \mathcal{N}(0, \sigma^2). \tag{3}$$

The EM algorithm, developed by Dempster et al. (1977), is an iterative algorithm which provides a maximum likelihood estimate from an incomplete data set when an explicit solution is not available. The expectation step corresponds here to the imputation by the expectation of the missing values given the observed values and the value of the parameters at the iteration ℓ : $\hat{x}_{ik}^{\ell} = \sum_{s=1}^{S} \hat{f}_{is}^{\ell}\hat{u}_{ks}^{\ell} + \hat{m}_k^{\ell-1}$. The maximization step corresponds to the maximization of the "complete likelihood" which is equivalent to carrying out a PCA on the imputed data set. The iterative PCA algorithm is then often named EM-PCA. An iterative PCA algorithm with one step of the alternating least squares algorithm corresponds to a generalized EM (GEM-PCA) algorithm (Josse et al. 2009) where the likelihood is increased rather than maximized at each step.

2.3 Properties

Imputation. Even if the aim of the EM-PCA algorithm is to estimate the principal axes and components in spite of missing values (missing values are skipped), it can also be seen as an imputation method (Kiers 1997; Bro 1998). The imputation is carried out by simultaneously taking into account the similarities between individuals and the relationships between variables. That is why imputing with the PCA model has known a great success in the data matrix completion framework such as in the Netflix problem (Netflix 2009).

Number of components. Solutions provided by the algorithms are not nested (the solution with S dimensions is not included in the solution with $S + 1$ dimensions). Hence, the choice of the number of dimensions, which is done a priori, is a crucial step. Under the hypothesis that the model (3) is correct, if a too low number of components is kept then relevant information is not taken into account in the analysis. On the other hand, if the information is contained in few dimensions and if more dimensions are taken into account, then noise is fitted by the model which leads to unstable results. This situation leads to overfitting (*cf.* §*Overfitting*) because too many parameters are estimated from the observed values (Raiko et al. 2007; Ilin and Raiko 2010). Several

strategies are available to choose the number of components in the complete case (Dray 2008; Peres-Neto et al. 2005) and recently Bro et al. (2008) have described a cross-validation procedure. This procedure can be easily extended to the incomplete case. For a fixed S, it consists in removing each observed value alternatively (leave-one-out) and predicting it using the EM-PCA algorithm. Then, the prediction error $(x_{ik} - \hat{x}_{ik}^{-ik})$ is computed for all elements $\{ik\}$ leading to a matrix of prediction errors; with $\hat{x}_{ik}^{-ik}$ the predicted value for x_{ik} calculated without the element $\{ik\}$. The mean squared error of prediction (MSEP) is then calculated $\frac{1}{IK}\sum_i \sum_j (x_{ik} - \hat{x}_{ik}^{-ik})^2$. The number S that leads to the smallest MSEP is retained. This criterion often provides satisfactory results but its main drawback is its computational cost.

Overfitting. In the EM-PCA algorithm overfitting is a serious problem. Indeed, the criterion (2) frequently happens to be very low while the estimation of the parameters (axes and components) and the predictions of the missing values are not reliable. To improve the prediction accuracy, "regularized" algorithms can be used. Josse et al. (2009) recently described an algorithm which is quite similar to the EM-PCA one but the reconstruction step is substituted by a "shrunk" reconstruction step:

1. initialization $\ell = 0$: $\mathbf{X}^0$ is obtained by setting missing elements to an initial value; The mean matrix $\hat{\mathbf{M}}^0$ is computed;
2. iteration ℓ :
   (a) calculate $\hat{\mathbf{F}}^\ell$, λ_s and $\hat{\mathbf{U}}^\ell$ by the singular value decomposition of $(\mathbf{X}^{\ell-1} - \hat{\mathbf{M}}^{\ell-1})$:
   $\mathbf{X}^{\ell-1} - \hat{\mathbf{M}}^{\ell-1} = \frac{\hat{\mathbf{F}}^\ell}{\|\hat{\mathbf{F}}^\ell\|} \operatorname{diag}\left(\sqrt{\lambda_s}\right)_{s=1,\ldots,K} \hat{\mathbf{U}}^{\ell\prime}$; S dimensions are kept.
   σ^2 is estimated as $\hat{\sigma}^2 = \frac{1}{K-S}\sum_{s=S+1}^{K} \lambda_s$ which is the mean of the last eigenvalues;
   (b) missing values in $\mathbf{X}$ are replaced by the fitted values:
   $\hat{\mathbf{X}}^\ell = \frac{\hat{\mathbf{F}}^\ell}{\|\hat{\mathbf{F}}^\ell\|} \operatorname{diag}\left(\sqrt{\lambda_s} - \frac{\hat{\sigma}^2}{\sqrt{\lambda_s}}\right)_{s=1,\ldots,S} \hat{\mathbf{U}}^{\ell\prime} + \hat{\mathbf{M}}^{\ell-1}$. The new imputed data set is $\mathbf{X}^\ell = \mathbf{W} * \mathbf{X} + (1 - \mathbf{W}) * \hat{\mathbf{X}}^\ell$;
   (c) $\hat{\mathbf{M}}^\ell$ is computed;
3. steps (2.a) (2.b) and (2.c) are repeated until convergence. A postprocess step consists in performing a PCA on the completed data set to obtain $\hat{\mathbf{M}}$, $\hat{\mathbf{F}}$ and $\hat{\mathbf{U}}$.

Compared to the EM-PCA algorithm, step (2.a) is the same but step (2.b) has changed because the singular value $\sqrt{\lambda_s}$ has been substituted by $\left(\sqrt{\lambda_s} - \frac{\hat{\sigma}^2}{\sqrt{\lambda_s}}\right)$. The regularization term comes from the probabilistic formulation of PCA proposed by Tipping and Bishop (1999) and appears to be a well-adapted regularization term to avoid overfitting in the missing data framework. Indeed, the regularization term increases with s: individual coordinates are more shrunk on the last dimensions and consequently more importance is given to the first dimensions. Moreover, when there is a lot of missing values and/or a lot of noise (a low structure of correlation between the variables), the value of $\hat{\sigma}^2$ is high. It implies that individual coordinates that are shrunk on axis s by $\frac{\lambda_s - \hat{\sigma}^2}{\lambda_s}$ get closer to the centre of gravity. Consequently, the algorithm tends to impute missing values with the mean of each variable. This behaviour is quite good since mean imputation is reasonable when there is "nothing" in the data.

This algorithm, denoted REM-PCA for regularized EM-PCA, improves the estimation of the axes and components and the prediction of missing values. Moreover, as singular values are shrunk, it reduces the impact of a bad choice of the number of dimensions (when too many dimensions are kept).

Reduction of variability. The variability of the imputed data set is underestimated. Indeed, the observed values consists of signal plus noise (under the hypothesis that the model (3) is true) whereas the residual part is omitted for the imputed values (which are imputed with the conditional mean). Moreover, the variability is reduced in a second way (Kroonenberg 2008): while the true values corresponding to the missing values are unknown, it is not possible to know how they would have influenced the solution. Multiple imputation is a solution to overcome this reduction of variability. The aim of the next section is then to adapt multiple imputation in the framework of PCA so that the uncertainty due to missing values is taken into account.

3 Multiple imputation

3.1 Multiple imputation with a PCA model

There are different steps in multiple imputation (van Buuren 2007) and the first one consists in generating D completed data sets. "The variation among the D imputations reflects the uncertainty with which the missing values can be predicted from the observed ones" (Schafer and Olsen 1998). As the variance of a prediction is composed of two parts, the first one reflecting the uncertainty in the estimation of the parameters and the second one reflecting the variability of the noise, the algorithm to generate multiple imputed data sets with a PCA model is composed of two parts:

- obtaining D plausible sets of parameters, $(\hat{\mathbf{M}}, \hat{\mathbf{F}}, \hat{\mathbf{U}}')^1$, ..., $(\hat{\mathbf{M}}, \hat{\mathbf{F}}, \hat{\mathbf{U}}')^D$
- for $d = 1, ..., D$, imputing all missing values x_{ik}^d by drawing from the predictive distribution of the missing values given the observed values and the parameters: $\hat{m}_k^d + \sum_{s=1}^{S} \hat{f}_{is}^d \hat{u}_{ks}^d + \tilde{\varepsilon}$, with $\tilde{\varepsilon}$ a residual drawn from the observed distribution of the residuals.

The first step consists in reflecting the variability of the parameters which is called "external stability" by Greenacre (1984). The second step represents the additional variance due to the unknown value of ε_{ik} associated to the missing value x_{ik}. Then we detail what is the "variability of the parameters" and after we describe the algorithm.

Variability of the parameters. Timmerman et al. (2007) compared different approaches to assess variability in PCA. They showed that bootstrap methods are more flexible and give better confidence areas (in the sense of the coverage) than asymptotic ones. A bootstrap sample can be obtained in a nonparametric or a semiparametric way. Nonparametric bootstrap consists in drawing randomly with replacement I individuals from the observed data. Semiparametric bootstrap requires to define an explicit model and consists in resampling the residuals. The resampling procedure is then repeated several times and the parameters of interest are estimated for each new data set. When individuals can be considered as a random sample from a population, as for

example in survey analyses, bootstrapping the individuals is a good option since they can be considered as independent and identically distributed (iid). When data represent a full population of individuals, which is often the case in PCA, bootstrapping the residuals seems more appropriate. This latter strategy is in agreement with the bilinear model (3) where individuals have different means and are not iid. In the model (3), the randomness is only due to the error term. The two approaches involve different kind of variability (the sampling variability and the variability due to the error term) and so provide different confidence areas. The individuals bootstrap procedure provides larger confidence areas since all the dimensions are bootstrapped whereas in the residual bootstrap procedure only the last dimensions are bootstrapped. However, even if only the last dimensions are bootstrapped, all the parameters (even those associated with the first dimensions) change since "the noise goes everywhere". The approach which is described below is to bootstrap the residuals since we assume model (3). This method requires the validity of the model and consequently the choice of the number of dimensions is important.

MI-PCA algorithm. The algorithm proposed to generate D imputations with a PCA model is:

1. initialization
   (a) perform the EM-PCA or REM-PCA algorithm on $\mathbf{X}$ to obtain an estimation of the parameters $\hat{\mathbf{M}}$, $\hat{\mathbf{F}}$ and $\hat{\mathbf{U}}$;
   (b) reconstruct the data with the first S dimensions ($\hat{\mathbf{X}} = \hat{\mathbf{M}} + \hat{\mathbf{F}}\hat{\mathbf{U}}'$) and calculate the matrix of residuals $\hat{\varepsilon} = \mathbf{X} - \hat{\mathbf{X}}$. The matrix $\hat{\varepsilon}$ is incomplete since $\mathbf{X}$ is incomplete;

For $d = 1, .., D$ **perform steps 2 and 3**

2. variability of the parameters: taking into account the uncertainty on the axes and on the components
   (a) bootstrap the residuals $\hat{\varepsilon}$ to obtain a new matrix of residuals ε^* (only the observed residuals are bootstrapped to keep the same pattern of missing values);
   (b) generate a new data table: $\mathbf{X}^* = \hat{\mathbf{X}} + \varepsilon^*$;
   (c) perform the missing data algorithm on table $\mathbf{X}^*$ to obtain new estimates of the parameters ($\hat{\mathbf{M}}^*$, $\hat{\mathbf{F}}^*$, $\hat{\mathbf{U}}^*$);

3. drawing from the predictive distribution: taking into account the uncertainty on the cell
   (a) impute each missing value x_{ik}^d from its conditional mean $(\hat{\mathbf{M}}^* + \hat{\mathbf{F}}^*\hat{\mathbf{U}}^{*\prime})_{ik}^d$;
   (b) for all the imputed values, add a residual $\tilde{\varepsilon}$ drawn from the observed distribution of the residuals $\hat{\varepsilon}$;
   the imputed data table is $\mathbf{X}^d = \mathbf{W} * \mathbf{X} + (1 - \mathbf{W}) * ((\hat{\mathbf{M}}^* + \hat{\mathbf{F}}^*\hat{\mathbf{U}}^{*\prime}) + \tilde{\varepsilon})$;

This MI-PCA algorithm leads to D imputed data sets. The residuals used in the steps 2.b and 3.b can be drawn from a Gaussian distribution with mean 0 and standard deviation equal to the standard deviation of the residuals.

A better approach is to use corrected residuals in the steps 2.b and 3.b in the same vein as in regression or to draw the residuals from a Gaussian distribution with mean 0 and standard deviation equal to the corrected standard deviation of the residuals:

$$\hat{\sigma}^2 = \frac{\|\mathbf{W} * (\mathbf{X} - \hat{\mathbf{X}})\|^2}{IK - \text{nb}_{\text{miss}} - IS - KS - K + S + S^2}$$

with $\text{nb}_{\text{miss}} = \sum_{i,k}(1 - W_{ik})$ the number of missing values in $\mathbf{X}$. As usual, the denominator corresponds to the number of the degree of freedom of the residuals (number of observed values minus the number of independent parameters estimated). This formula is detailed without missing values in Denis (1991) and Josse and Husson (2011).

Remark: benefits from imputing with PCA. The proposed algorithm MI-PCA is an alternative to the methods available to generate multiple imputed data sets. On of the most popular algorithm is NORM (Schafer 1997) which uses the data augmentation algorithm (Tanner and Wong 1987). NORM is based on the multivariate normal model and requires the estimation and the inversion of the variance-covariance matrix. Consequently, when variables are highly correlated or when $I > K$, the estimation encounters difficulties or is not possible. Schafer (1997) proposed then to use a ridge prior to stabilize the estimation. The MI-PCA algorithm also overcomes this problem since the number of parameters is reduced compared to the estimation of the full covariance matrix. Moreover, the method is easy to implement and computationally fast. Song (1999) has proposed in a Bayesian framework an algorithm in the same vein as the MI-PCA one which simplifies the covariance structure with a factor analysis model.

3.2 Multiple imputation in PCA

This section presents how to deal with the D multiple imputed data sets in order to take into account the variability due to the missing values in the PCA results. The imputed data sets are juxtaposed as shown in Fig. 1: all the values except the missing ones (boxes) are the same for all the data sets. The first imputed data set on the left (with black boxes) is the one obtained from the REM-PCA or EM-PCA algorithm (missing values are imputed from their conditional mean without adding any variability). The other tables are obtained from the multiple imputation procedure using the MI-PCA algorithm.

The two subspaces obtained from the PCA on the data table on the left are considered as the reference configurations (one for the individuals and one for the variables). Two approaches are described to visualize the uncertainty induced by the missing values on these reference configurations.

Projection of the completed data sets as supplementary elements. Each imputed data set is projected as supplementary information onto the reference configurations as illustrated in Fig. 2 for the individuals. Individuals (respectively variables) without missing values are projected exactly on their corresponding point on the reference

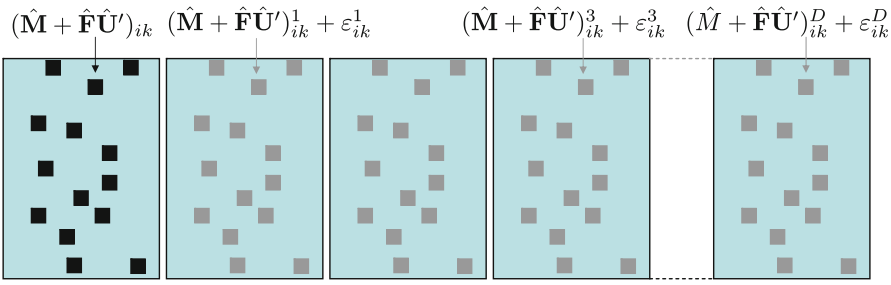

Fig. 1 Multiple imputed data sets. A *black box* corresponds to the imputation of one missing value obtained with the REM-PCA algorithm. A *grey box* corresponds to one imputed values obtained with the MI-PCA algorithm. *D* imputed data sets are generated

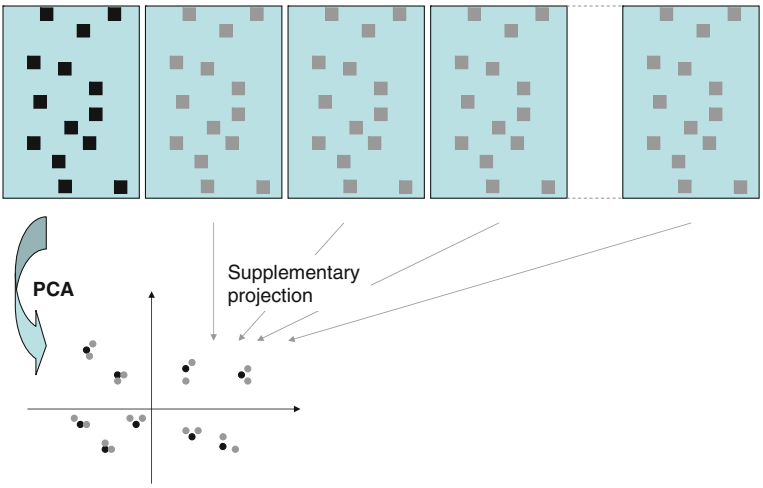

Fig. 2 Supplementary projection of the multiple imputed data sets onto the reference configuration (*in black*)

configuration, and individuals (respectively variables) with several missing values are projected around their corresponding point. These projections materialize the confidence areas around the points and can be summarized with convex hulls or ellipses for the individuals. This strategy is in the same vein as the "partial bootstrap" proposed by Chateau and Lebart (1996) and Greenacre (1984) in the context of sampling variability in exploratory multivariate data analyses. Indeed, the components and the axes are not recalculated for each table. This approach allows one to visualize the stability of the individuals (respectively of the variables) due to missing values.

Representation of the uncertainty on the PCA parameters generated by the imputed values. A PCA is performed on each imputed data set leading to different values for the parameters (axes and components). In order to compare the results provided by each PCA, it is necessary to minimize the "nuisance variation" as called by Milan (1995), it means the possible translation, reflection, dilatation or rotation of

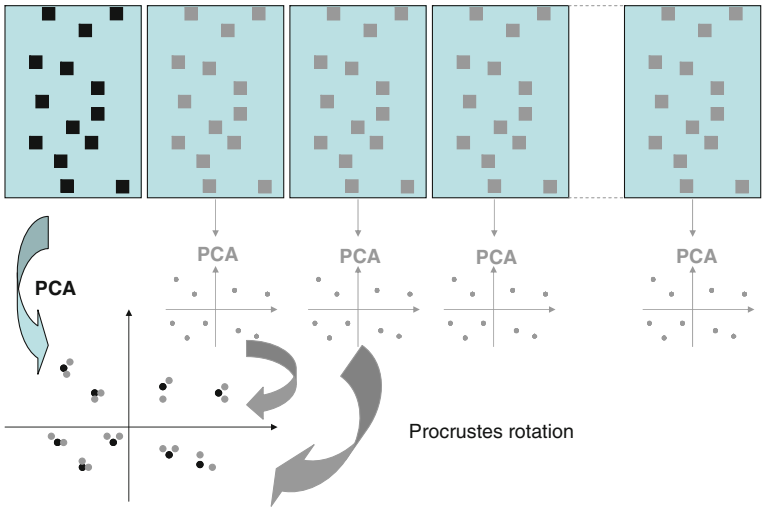

Fig. 3 Procrustes rotations of the PCA configuration obtained from the multiple imputed data sets onto the reference configuration

the different configurations. To this end, it is possible to resort to Procrustes rotations (Gower and Dijksterhuis 2004) to fit the PCA configurations obtained from the imputed data tables toward the fixed reference configuration as illustrated in Fig. 3 for the individuals. This strategy is quite similar to the "total bootstrap" (type 3) proposed by Chateau and Lebart (1996) and Greenacre (1984) in the framework of the stability in principal component methods. This approach allows one to visualize the uncertainty on the PCA dimensions (axes and components) induced by the missing values, it means to assess the influence of the different predictions of the missing values on the estimation of the parameters.

4 Results

A simulation study has been conducted to evaluate the procedure for different number of individuals and variables, varying the dimensionality of the data sets (the rank S) and the percentage of missing values. Results presented in this section on the data set "Wine" (Escofier and Pagès 2008) illustrate all the results obtained. The data table is composed of 21 wines described by 10 odour sensory attributes. This data set has no missing values and has a strong two-dimensional structure (more than 70% of variability is explained by the first two dimensions). First, 10% of the values are removed completely at random (first data set named 10%) and then additional values are removed leading to 30% of missing values (second data set named 30%).

First, the REM-PCA algorithm using two dimensions is performed on the 10% (respectively 30%) incomplete data set leading to the reference configurations. The reference configurations are presented in Fig. 4 for the individuals (represented by points, graphs on the left) and for the variables (represented by arrows, graphs on the

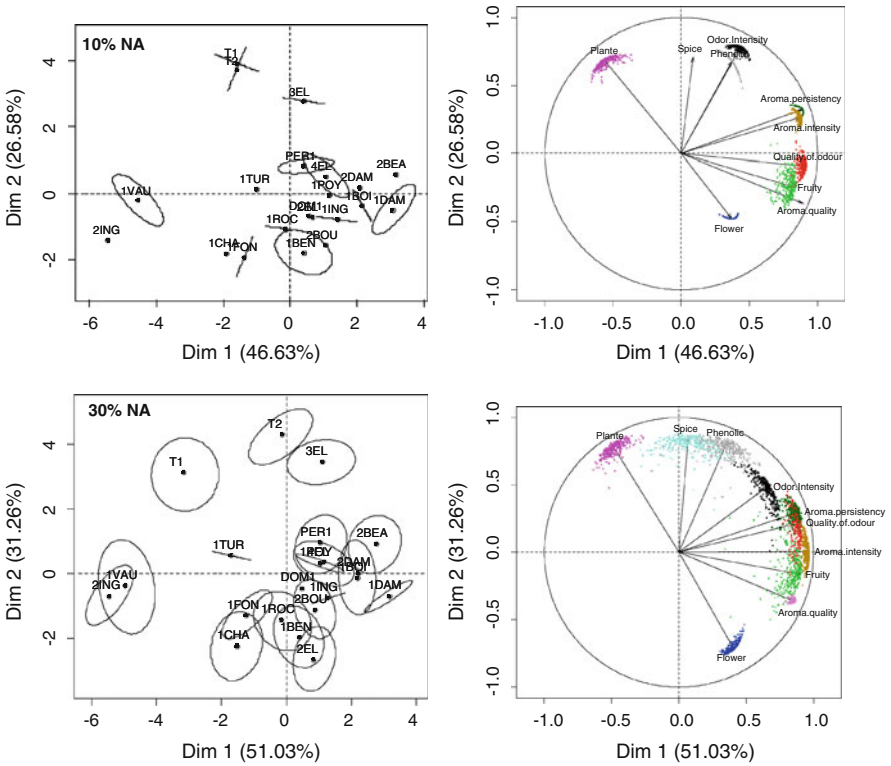

Fig. 4 Visualization of the uncertainty on the map of individuals and on the map of variables due to missing values

right). Note that it is always possible to obtain a configuration of individuals and variables whatever the dataset and the pattern of missing values. Unfortunately, with only these configurations, there is no way to know if the results obtained are plausible and if the user can interpret the results. Consequently, it is crucial to have a tool, such as confidence areas provided from the MI-PCA algorithm, which allows the researcher to decide if he can analyse or not its results.

The MI-PCA algorithm is then performed to generate 500 imputed data sets. The different imputed values reflect the prediction uncertainty of the missing values obtained with a PCA model.

First we visualize the variability of the different imputed data sets: the 500 imputed data sets are projected onto the reference configuration (Fig. 4). The 95% confidence ellipses show the uncertainty around the individuals and the clouds of points represent the uncertainty for the variables. In the example with 10% of missing values, individual 2ING has no missing value and consequently no ellipse. Individuals 1FON and T1 have only one missing value for the variable "Odor intensity" and 1DAM has several missing values. Similarly, the "Flower" variable has one missing value while the "Plante" variable has several missing values. As expected, the size of the ellipses and the variability of the clouds of points for the variables increase with the

number of missing values. However, the ellipses for the individuals and the clouds of points for the variables are not too large and the general interpretation of the PCA results is not really affected by the missing values even with 30% of missing values. The first dimension can then be named "Quality of the aroma" and opposes wines such as 2BEA which smell good to wines such as 2ING which smell bad. The second dimension corresponds to wines such as T1 and T2 with very particular aroma such as "spice". The small sizes of the confidence areas is due to the strong structure of the data set (the correlation between variables). When the structure is lower, the size of the ellipses becomes larger which leads to be careful in the interpretation of the results or even not to interpret the results.

Now, we visualize the influence of the different imputations on the PCA dimensions to take into account the uncertainty due to the missing values on the estimation of the parameters. A PCA is performed on each of the 500 imputed data sets and the 500 obtained configurations are compared to the reference one. To this end, each two-dimensional configuration is fitted onto the reference configuration with a Procrustes rotation (Fig. 5). As previously done, the graphs on the left represent the uncertainty in the individuals. The different imputed values in each data set imply slightly different axes and components. Consequently, since the dimensions change, the position of all the individuals changes even if an individual, as in the case of wine 2ING, is not corrupted by any missing value. However, the uncertainty around the individuals is all the more important that they have a lot of missing values. The graphs on the right represent the projections of the two first principal components obtained from each imputed data set onto the reference configuration. It is thus a synthetic visualization of the impact of the missing values onto the PCA dimensions. Results appear to be stable even for 30% of missing values and consequently the interpretation of the dimensions of the reference configuration is quite accurate.

5 Discussion

This paper proposes a method to handle missing values in principal component analysis providing point estimations of the parameters and of the missing values with a notion of uncertainty. A multiple imputation method is proposed to assess the supplement variability due to missing values. Two approaches have been proposed to visualize the impact of the different predictions of the missing values onto the PCA results without reconsidering the reference configurations which are held as fixed. The graphical representations provide a useful tool which can help in the interpretation of the PCA results obtained in a missing data framework. Indeed, it allows the user to comfort the PCA results if the confidence areas are small or to be careful otherwise. The MI-PCA algorithm is implemented in the R (R Development Core Team 2009) package named missMDA (Husson and Josse 2010).

Another approach to deal with the D imputed data sets could be to use three-way methods such as multiple factor analysis in the sense of Escofier and Pagès (2008). Indeed, these methods are dedicated to analyse data tables in which individuals are described by several groups of variables. They can be used on the whole multiple imputed data set to compare the results provided by several PCA. They look for a

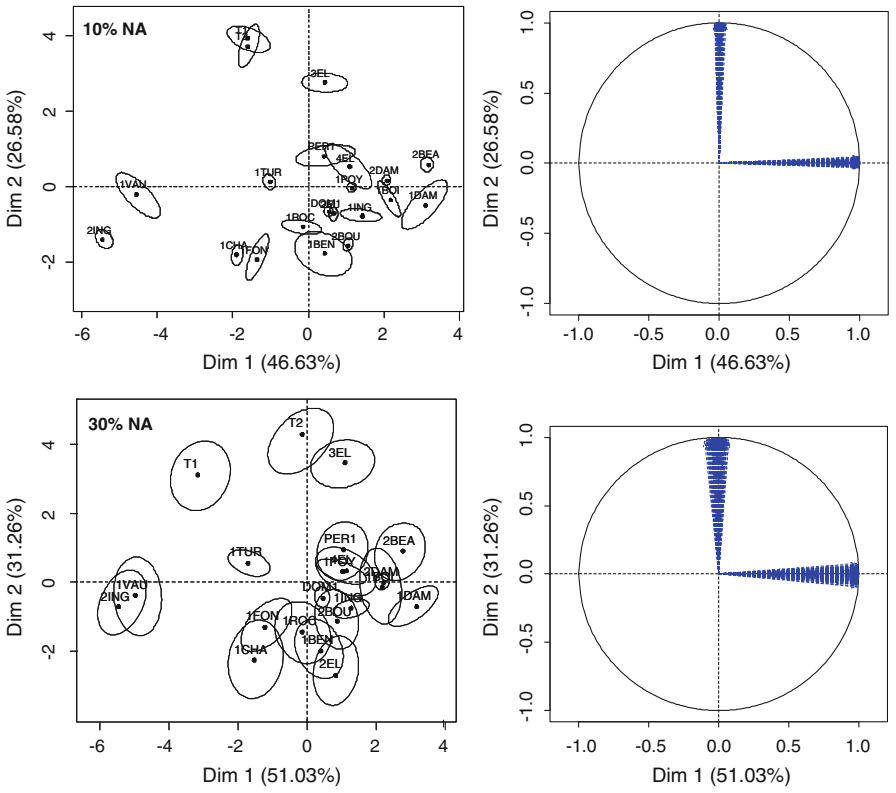

Fig. 5 Visualization of the uncertainty on the PCA dimensions induced by the missing values. On the left, uncertainty of the individual map; on the right, representation of the two first dimensions obtained on the 500 data sets

common structure and highlight the commonalities and the discrepancies between data tables. The main difference with our approach concern the reference configuration. Indeed, in this kind of method, the common configuration is not the reference configuration but a weighted mean of all the configurations on which the different data tables are represented.

The MI-PCA algorithm allows us to generate D imputed data sets from a PCA model and consequently can be seen as an alternative to the other methods of multiple imputation. This algorithm reflects the uncertainty about the unknown model parameters using a bootstrap procedure and without resorting to Bayesian considerations. It is in the same vein as the nonparametric multiple imputation suggested by Little and Rubin (1987, 2002, p.216). Further work needs to be done in order to assess the performances of the MI-PCA algorithm as a multiple imputation method in a general framework. It can be done by comparing the coverage computed with different methods which is out of the scope of this paper. The main objective of this paper is to handle missing values in PCA so we do not focus on the quality of the multiple imputation of the dataset.

The method proposed for continuous variables in the framework of PCA may be extended to categorical variables and mixed variables. Indeed, methods such as multiple correspondence analysis (MCA) can be seen as a weighted PCA on the indicator matrix and the method proposed in this paper can take into account weights. A (multiple) imputation is performed on the indicator matrix which may lead to non-discrete values. The imputed values can be seen as a degree of membership to the corresponding categories. From this imputed (or multiple imputed) indicator matrix, MCA scores and MCA dimensions can be derived and the construction of confidence areas is thus possible in the framework of MCA.

References

Adams E, Walczak B, Vervaet C, Risha PG, Massart D (2002) Principal component analysis of dissolution data with missing elements. Int J Pharm 234:169–178

Bro R (1998) Multi-way analysis in the food industry—models, algorithms, and applications. Tech. rep., MRI, EPG and EMA, Proc ICSLP 2000

Bro R, Kjeldahl K, Smilde AK, Kiers HAL (2008) Cross-validation of component models: A critical look at current methods. Anal Bioanal Chem 5:1241–1251

Caussinus H (1986) Models and uses of principal component analysis. In: de Leeuw J, Heiser W, Meulman J, Critchley F (eds) Multidimensional data analysis. DSWO Press, pp 149–178

Chateau F, Lebart L (1996) Assessing sample variability in the visualization techniques related to principal componenet analysis: Bootstrap and alternative simulation methods. In: COMPSTAT, pp 205–210

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J Royal Stat Soc B 39:1–38

Denis JB (1991) Ajustements de modèles linéaires et bilinéaires sous contraintes linéaires avec données manquantes. Revue de Statistique Appliquée 39:5–24

Dray S (2008) On the number of principal components: A test of dimensionality based on measurements of similarity between matrices. Comput Stat Data Anal 52:2228–2237

Escofier B, Pagès J (2008) Analyses factorielles simples et multiples, 4th edn. Economica, Paris

Gabriel KR, Zamir S (1979) Lower rank approximation of matrices by least squares with any choice of weights. Technometrics 21:236–246

Golub GH, Van Loan CF (1996) Matrix computations, 3rd edn. Johns Hopkins University Press, Baltimore

Gower JC, Dijksterhuis GB (2004) Procrustes problems. Oxford University Press, New York

Greenacre M (1984) Theory and applications of correspondence analysis. Acadamic Press, London

Grung B, Manne R (1998) Missing values in principal component analysis. Chemiometr Intell Lab Syst 42:125–139

Husson F, Josse J (2010) missMDA: Handling missing values with/in multivariate data analysis (principal component methods). http://www.agrocampus-ouest.fr/math/husson, http://www.agrocampus-ouest.fr/math/josse, R package version 1.2

Ilin A, Raiko T (2010) Practical approaches to principal component analysis in the presence of missing values. J Mach Learn Res 11:1957–2000

Josse J, Pagès J, Husson F (2009) Gestion des données manquantes en analyse en composantes principales. J de la Société Française de Statistique 150:28–51

Josse J, Pagès J, Husson F (2011) Selecting the number of components in principal component analysis using cross-validation approximations (submitted)

Kiers HAL (1997) Weighted least squares fitting using ordinary least squares algorithms. Psychometrica 62:251–266

Kroonenberg PM (2008) Applied Multiway data analysis (chap.7). Wiley series in probability and statistics, New York

Little RJA, Rubin DB (1987) 2002) Statistical analysis with missing data. Wiley series in probability and statistics, New York

Milan M (1995) Application of the parametric bootstrap to models that incorporate a singular value decomposition. J Royal Stat Soc Ser C 44:31–49

Netflix (2009) Netflix challenge. http://www.netflixprize.com

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

Peres-Neto PR, Jackson DA, Somers KM (2005) How many principal components? stopping rules for determining the number of non-trivial axes revisited. Comput Stat Data Anal 49:974–997

R Development Core Team (2009) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org, ISBN 3-900051-07-0

Raiko T, Ilin A, Karhunen J (2007) Principal component analysis for sparse high-dimensional data. In: Neural Information Processing, pp 566–575

Rubin DB (1987) Multiple imputation for non-response in survey. Wiley, New York

Schafer JL (1997) Analysis of incomplete multivariate data. Chapman & Hall/CRC, London

Schafer JL, Olsen MK (1998) Multiple imputation for missing-data problems: A data analyst's perspective. Multivar Behav Res 33:545–571

Song J (1999) Analysis of incomplete high-dimensional multivariate normal data using a common factor model. PhD thesis, Dept. of Biostatistics, UCLA, Los Angeles

Tanner MA, Wong WH (1987) The calculation of posterior distributions by data augmentation. J Am Stat Assoc 82:805–811

Timmerman ME, Kiers HAL, Smilde AK (2007) Estimating confidence intervals for principal component loadings: a comparaison between the bootstrap and asymptotic results. Br J Math Stat Psychol 60:295–314

Tipping M, Bishop CM (1999) Probabilistic principal component analysis. J Royal Stat Soc B 61:611–622

van Buuren S (2007) Multiple imputation of discrete and continuous data by fully conditional specification. Stat Methods Med Res 16:219–242

Wold H (1966) Nonlinear estimation by iterative least squares procedures. In: David FN (ed) Research Papers in Statistics: Festschrift for Jerzy Neyman. Wiley, New York pp 411–444