Journal de la Société Française de Statistique
Vol. 153 No. 2 (2012)

Handling missing values in exploratory multivariate data analysis methods

Titre : Gestion des données manquantes en analyse factorielle

Julie Josse and François Husson

Abstract: This paper is a written version of the talk Julie Josse delivered at the 44 Journées de Statistique (Bruxelles, 2012), when being awarded the Marie-Jeanne Laurent-Duhamel prize for her Ph.D. dissertation by the French Statistical Society. It proposes an overview of some results, proposed in Julie Josse and François Husson's papers, as well as new challenges in the field of handling missing values in exploratory multivariate data analysis methods and especially in principal component analysis (PCA). First we describe a regularized iterative PCA algorithm to provide point estimates of the principal axes and components and to overcome the major issue of overfitting. Then, we give insight in the parameters variance using a non parametric multiple imputation procedure. Finally, we discuss the problem of the choice of the number of dimensions and we detail cross-validation approximation criteria. The proposed methodology is implemented in the R package missMDA.

Résumé : Cet article fait suite à la conférence de Julie Josse sur ses travaux de thèse lors de la réception du prix Marie-Jeanne Laurent-Duhamel, dans le cadre des 44e Journées de Statistique (Bruxelles, 2012). Il reprend les principaux résultats des papiers de Julie Josse et François Husson sur la gestion des données manquantes en analyse factorielle et décrit de nouvelles avancées sur le sujet. Dans un premier temps, nous détaillons un algorithme d'ACP itérative régularisée qui permet d'estimer les axes et composantes principales en présence de données manquantes et qui pallie le problème majeur du surajustement. L'estimation ponctuelle est enrichie par la construction de zone de confiance. Une méthode d'imputation multiple non-paramétrique est alors développée pour prendre en compte l'incertitude due aux données manquantes. Enfin, nous abordons le problème récurrent du choix du nombre de dimensions et définissons des approximations de la validation croisée de type validation croisée généralisée. Tous ces travaux sont mis à disposition de l'utilisateur grâce au package missMDA du logiciel libre R.

Keywords: Missing values, PCA, Multiple imputation, MCA, EM algorithm, Regularization, Residual bootstrap, Number of dimensions, Generalized cross-validation
Mots-clés : Données manquantes, ACP, Imputation multiple, ACM, Algorithme EM, Regularization, Bootstrap des résidus, Nombre de dimensions, Validation croisée généralisée

1. Introduction

Multivariate exploratory data analysis methods also known as principal component methods are dimensionality reduction techniques often used to sum-up data where individuals are described by continuous and or categorical variables. These methods allow one to study the similarities between individuals from a multidimensional point of view, to study the relationship between variables and to characterize the individuals using the variables. Traditionally, in "the French school of data analysis", a current initiated by Jean-Paul Benzécri [2] in the 1960s, the graphical

[1] Agrocampus Ouest Rennes.
E-mail: julie.josse@agrocampus-ouest.fr

 ISSN: 2102-6238

representations of the results are at the core of the interpretation and quite often a same emphasis is given to both the representation of the individuals and of the variables.

Whatever the structure and the nature of the data, missing values are ubiquitous and occur for a number of reasons: individuals who do not answer to items from a questionnaire, machines that fail, etc. Missing values are problematic since most statistical methods cannot be applied directly to an incomplete dataset. One of the most popular approach to deal with missing values consists in using "single imputation" methods. These methods fill in missing values with plausible values which leads to a completed dataset that can be analysed by any statistical method. However, the most classical imputation methods have drawbacks that have been well documented in the literature. The famous mean imputation preserves the mean of the imputed variable but reduces its variance and distorts the correlation with the other variables. Imputing by regression for example improves the imputation taking into account the relationship between variables. However, the marginal and joint distribution of the variables are still distorted. These distributions can be preserved using more complex imputation methods such as the stochastic regression imputation. The 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, even with such a strategy, "the imputed dataset ... fails to account for missing data uncertainty" [42]. Indeed, imputed values are considered as observed values and 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 dataset are underestimated [p.65][30] which leads to confidence intervals and tests that are not valid even if the imputation model is correct.

Highly recommended alternative methods [41, 30] are the use of two families of methods: multiple imputation (MI) [40] and the maximum likelihood approach. The former consists first of generating different plausible values for each missing value leading to different imputed datasets. The variation among the different imputations reflects the variance of the prediction of the missing values from the observed ones. Then, MI performs the same statistical analysis on each imputed dataset and it combines the results to obtain point estimates and standard errors of the parameters in a missing data framework. For the latter, the practice is to use an expectation-maximization (EM) algorithm [9] to obtain point estimates of the parameters in spite of the missing values and other algorithms to obtain an estimation of their variability that incorporates the variability due to missing values. Note that these two approaches are recommended when the mechanism that leads to missing values is either missing completely at random (MCAR) or missing at random (MAR) as defined by [39]. In this paper, we assume such a framework. However, in practice it is impossible to test the nature of the mechanism. Maximum likelihood and multiple imputation approaches have pros and cons but what is important to highlight is that the aim of these methods is to provide the best estimation of the parameters and of their variance despite the missing values and not to provide the best prediction of the missing entries.

We also pursue this aim when dealing with missing values in principal component methods such as principal component analysis (PCA) for continuous data and multiple correspondence analysis (MCA) for categorical data. Since all the principal component methods can be presented as PCA on matrices with specific row weights and column weights, we discuss here to a greater extent how to deal with missing values in PCA. In section 2 we will thoughtfully discuss PCA in the complete case to define some notions that will be used in section 3 to perform PCA with missing values. In both section we start by the point estimates of the parameters followed by

 ISSN: 2102-6238

notions of variability. We also propose a solution to the recurrent problem of the choice of the number of underlying dimensions. At the end of section 3, we illustrate the methodology on a small dataset and also give insights on how to perform MCA with missing values.

2. Complete case

2.1. Estimation of the principal axes and components

PCA can be presented from geometrical or model points of view.

2.1.1. Geometrical point of view

PCA provides a subspace that minimizes the distances between individuals and their projections onto the subspace. It amounts to finding a matrix of low rank S that provides the best approximation of the matrix $\mathbf{X}_{I\times K}$ in the least squares sense which is equivalent to find two matrices $\mathbf{F}_{I\times S}$ and $\mathbf{U}_{K\times S}$ that minimize the following criterion:

$$\mathscr{C} \quad = \quad \|\mathbf{X}-\mathbf{M}-\mathbf{F}\mathbf{U}'\|^2 = \sum_{i=1}^{I}\sum_{k=1}^{K}(X_{ik}-m_k-\sum_{s=1}^{S}F_{is}U_{ks})^2, \tag{1}$$

where $\mathbf{M}$ is an $I\times K$ matrix with each row equal to $(m_1,...,m_K)$, *i.e.* the vector with the mean of each variable. With the additional constraints that the columns of $\mathbf{U}$ are orthogonal and of unit norm, the solution is given by the eigenvectors of the inner-product matrix, namely the principal components $\hat{\mathbf{F}}$ (the scores matrix such that the variance of each column is equal to the corresponding eigenvalue) and the eigenvectors of the covariance matrix, namely the principal axes $\hat{\mathbf{U}}$ (the loadings matrix or the coefficient matrix). The solution can be obtained either by the singular value decomposition of $(\mathbf{X}-\hat{\mathbf{M}})$ or by the use of an alternating least squares algorithm (ALS). The latter consists in minimizing (1) by alternating two multiple regressions until convergence, one for estimating the loadings and one for the scores. The solution satisfies the following two equations:

$$\hat{\mathbf{U}}' \quad = \quad (\hat{\mathbf{F}}'\hat{\mathbf{F}})^{-1}\hat{\mathbf{F}}'(\mathbf{X}-\hat{\mathbf{M}}), \tag{2}$$

$$\hat{\mathbf{F}} \quad = \quad (\mathbf{X}-\hat{\mathbf{M}})\hat{\mathbf{U}}(\hat{\mathbf{U}}'\hat{\mathbf{U}})^{-1}. \tag{3}$$

2.1.2. Model point of view

Two main models can be distinguished.

Fixed effect model A classical PCA model is a bilinear model where the data are generated as a *fixed structure* corrupted by noise as follows:

$$X_{ik} \quad = m_k + (\mathbf{F}\mathbf{U}')_{ik} + \varepsilon_{ik}\ ,\ \varepsilon_{ik} \sim \mathscr{N}(0,\sigma^2). \tag{4}$$

It is known either as the "fixed effect model" [4] or the "fixed factor score model" [8]. In this model the individuals have different expectations and the randomness is only due to the error term.

 ISSN: 2102-6238

The maximum likelihood estimates correspond to the least squares solution, *i.e.* to the principal axes and components. Remark that the inferential framework associated to this model is not usual. Indeed, when the number of individuals increases, the parameters' space increases also. Consequently, in this model, the asymptotic comes down to considering that the variance of the noise tends to 0.

This model is in agreement with cases where PCA is performed on data where each individual is an object of interest, *i.e.* the set of individuals is not a sample drawn from a population of individuals. Such situations arise quite frequently in practice. For example in sensory analysis, individuals can be food products (such as beers or chocolates) and variables sensory descriptors (such as bitterness, sweetness, etc.); the aim of the analysis is to describe these specific products. In such studies, it makes sense to estimate the individuals' parameters ($\hat{\mathbf{F}}$) and to examine the graphical representation of the cloud of individuals in a lower dimensional space. Throughout this work, we will favor this model because it fits many applications we encounter in practice.

Remark:

Model (4) is related to the biadditive models defined by [10] also known as additive main effects and multiplicative interaction models (AMMI). Such models are defined in an analysis of variance framework and often used in the field of plant breeding [33].

Random effect model In the beginning of the 2000, [38] and [45] defined what they called a probabilistic PCA (PPCA) model. One of their motivation was the absence of a probabilistic formulation of PCA. In fact, they didn't pay much attention to the works mentioned previously (fixed effect model). The model presented is a particular case of a factor analysis (FA) model [1] with isotropic noise (instead of leaving the variance of the noise free):

$$X_{ik} = \mathbf{m}_k + (\mathbf{Z}\mathbf{B}')_{ik} + \varepsilon_{ik}\ ,\ \varepsilon_{ik} \sim \mathcal{N}(0, \sigma^2) \tag{5}$$

with $\mathbf{B}_{K\times S}$ the matrix of unknown coefficients and $\mathbf{Z}_{I\times S}$ the matrix of latent variables. Each row of $\mathbf{Z}$ and of ε are mutually independent standard normal. In this model, the data are generated as *random structure* corrupted by noise. It induces a Gaussian distribution for the individuals with a specific structure of variance-covariance:

$$X_{i.} \sim \mathcal{N}(0, \Sigma) \text{ with } \Sigma = \mathbf{B}\mathbf{B}' + \sigma^2 \mathbb{I}_K, \text{ for } i = 1, ..., I,$$

where $\mathbb{I}_K$ is the identity matrix of size K. One of the most important property of factor analysis is the one of conditional independence: $X_{i.}|Z_{i.} \sim \mathcal{N}(\mathbf{B}Z_{i.}, \sigma^2)$. In PPCA, there is an explicit solution for the maximum likelihood estimates given by:

$$\hat{\mathbf{B}} = \hat{\mathbf{U}}(\hat{\Lambda} - \hat{\sigma}^2 \mathbb{I}_S)^{1/2}\mathbf{R} \text{ and } \hat{\sigma}^2 = \frac{1}{K-S}\sum_{s=S+1}^{K} \hat{\lambda}_s, \tag{6}$$

where $\hat{\mathbf{U}}$ is defined as previously, *i.e.* as the matrix of the S first eigenvectors of the covariance matrix, $\hat{\Lambda}$ is a diagonal matrix with the associated eigenvalues and $\mathbf{R}_{S\times S}$ an orthogonal matrix (usually $\mathbf{R} = \mathbb{I}_S$).

This model seems more in agreement with cases where PCA is performed on sample data such as survey data for example. Indeed, contrary to the model (4), the individuals are independent

 ISSN: 2102-6238

and identically distributed. It means that individuals are just considered for the information they bring on the relations between variables. In such studies, at first sight, it does not make sense to consider an "estimation" of the individuals' parameters since no parameters are associated to the individuals, only random variables are ($\mathbf{Z}$). However, it is customary in factor analysis to take as an estimation for the scores the expectation of the latent variables given the observed variables $\mathbb{E}(Z_{i.}|X_{i.})$, which gives:

$$\hat{\mathbf{Z}} = \mathbf{XB}(\mathbf{B}'\mathbf{B} + \sigma^2 \mathbb{I}_S)^{-1}. \tag{7}$$

Remark:
In the framework of mixed effect models, "estimating" random effects is also common [37]. An equation as (7) is known as BLUP (best linear unbiased prediction).

A Bayesian interpretation of the fixed effect model Another way of looking at the *random effect* model (5) is to consider it as a Bayesian treatment of the fixed effect model (4). It means that one assumes a Gaussian *a priori* distribution for the individuals' parameters (the scores). In such a case, the maximum *a posteriori* estimation for the scores corresponds to the equation (7). Assuming a distribution on the scores is a way to impose a constraint on the model. That is why, equation (7) corresponds to a ridge regression and not to an ordinary regression as in equation (3). Consequently, the *random effect* model can be regarded as a shrinkage version of the *fixed effect* model.
Remark:
This phenomenon is known in the framework of linear regression analysis where a Bayesian treatment of the model (putting a Gaussian prior on the parameters β) provides a probabilistic interpretation of the ridge estimator [p.64][19].

2.2. Confidence areas

Even if traditionally PCA is used as a descriptive technique, there is a growing interest in inferential issues. One of the major approaches to assess the variability of the results is to resort to bootstrap procedures. The most common practice is to bootstrap the individuals [5, 44]. It implicitly implies that individuals are a random sample from a larger population which is in agreement with the *random effect* model (5). We answer the question what would be the estimation of the parameters with another sample? The parameters are restricted to the variables coordinates since the individuals are bootstrapped. Consequently, this procedure provides confidence areas around the position of the variables.

When the individuals are not exchangeable, it may be interesting to obtain confidence areas around both the position of the individuals and of the variables. The model (4) suggests studying the variability of the parameters using a residuals bootstrap procedure which can be carried out using the following steps:

1. Perform PCA on $\mathbf{X}$ to obtain an estimation of the parameters $\hat{\mathbf{M}}$, $\hat{\mathbf{F}}$ and $\hat{\mathbf{U}}$
2. Reconstruct the data with the first S dimensions ($\hat{\mathbf{X}}^{(S)} = \hat{\mathbf{M}} + \hat{\mathbf{F}}\hat{\mathbf{U}}'$) and calculate the matrix of residuals $\hat{\boldsymbol{\varepsilon}} = \mathbf{X} - \hat{\mathbf{X}}^{(S)}$

 ISSN: 2102-6238

3. For $b = 1, .., B$
   (a) bootstrap the residuals $\hat{\varepsilon}$ by cells to obtain a new matrix of residuals ε^* or draw ε^*_{ik} from $\mathcal{N}(0, \hat{\sigma}^2)$;
   (b) generate a new data table: $\mathbf{X}^* = \hat{\mathbf{X}}^{(S)} + \varepsilon^*$;
   (c) perform PCA on table $\mathbf{X}^*$ to obtain new estimates of the parameters $(\hat{\mathbf{M}}^*, \hat{\mathbf{F}}^*, \hat{\mathbf{U}}^*)$

Many comments can be done regarding this procedure. First, the method requires the validity of the model (4) and consequently the choice of the number of dimensions is important. Then, like in ordinary regression analysis, the residuals underestimate the true errors and the maximum likelihood estimate of σ is biased. It is thus required to correct the residuals or their variance. These two points will be discussed in section 2.3. Finally, the algorithm provides B estimations of the parameters $(\hat{\mathbf{M}}^1, \hat{\mathbf{F}}^1, \hat{\mathbf{U}}^1), ..., (\hat{\mathbf{M}}^B, \hat{\mathbf{F}}^B, \hat{\mathbf{U}}^B)$ and it is not trivial to know how to combine and how to visualize the results. Indeed, performing bootstrap methods with singular value decomposition methods based is a difficult task since from one replication bootstrap to another, problems of rotation of the configurations may occur. One solution can be to resort to Procrustes rotations [15]. Another solution can be to compute the model matrix for each bootstrap replication, that is to say $\hat{\mathbf{X}}^{(S)b} = \hat{\mathbf{M}}^b + \hat{\mathbf{F}}^b \hat{\mathbf{U}}^{b'}$.

2.3. *Choosing the number of dimensions*

The choice of the number of dimensions in PCA is a core issue and a very difficult task. A huge amount of criteria have been proposed in the literature and none of them has really proved its superiority. It may be explained because PCA can be used for different purposes (just to describe the data, to reduce the dimensionality, to obtain new variables, etc.) and on very different datasets with very different underlying structures. Is a unique variable orthogonal to the other variables represents a dimension or noise? Nevertheless, [23] distinguished three kinds of methods: 1) ad-hoc rules (such as the scree test), 2) tests based on distributional assumptions and 3) computational methods such as bootstrap, permutation and cross-validation. Leave-one-out cross-validation, as described in [3], first consists in removing one cell (i,k) of the data matrix $\mathbf{X}$. Then, for a fixed number of dimensions S, it consists in predicting its value using the PCA model obtained from the dataset that excludes this cell. The value of the predicted cell is denoted $(\hat{X}_{ik}^{-ik})^{(S)}$. Finally, the prediction error is computed and the operation is repeated for all the cells in $\mathbf{X}$. The number S that minimizes the mean square error of prediction

$$MSEP(S) = \frac{1}{IK} \sum_{i=1}^{I} \sum_{k=1}^{K} \left(X_{ik} - (\hat{X}_{ik}^{-ik})^{(S)} \right)^2 \tag{8}$$

is retained. Such a procedure requires a method which is able to estimate the PCA parameters from an incomplete dataset (more details will be given in section 3.1). The main drawback of the cross-validation approach is its computational cost. To avoid resorting to intensive computational methods, we proposed in [26] two cross-validation approximation criteria. The proposed methodology is inspired by works in linear methods such as linear regression.

Let us denote a linear model as $\hat{\mathbf{y}} = \mathbf{P}\mathbf{y}$ with $\mathbf{y}$ a response vector and $\mathbf{P}$ a smoothing matrix. The computational cost problems of cross-validation are avoided thanks to a formula linking

 ISSN: 2102-6238

the prediction error and the fitting error. More precisely, using the "leaving-out-one" lemma, [7] showed that the prediction error $y_i - \hat{y}_i^{-i}$ for individual i can be written as:

$$y_i - \hat{y}_i^{-i} = \frac{y_i - \hat{y}_i}{1 - P_{i,i}}, \tag{9}$$

where $P_{i,i}$ is the ith diagonal element of $\mathbf{P}$. Consequently, there is no need to perform cross-validation to compute a mean square error of prediction. It is sufficient to compute the fitted values and the smoothing matrix.

The rationale of the approach that we proposed in [26] is to write PCA as "$\hat{\mathbf{X}} = \mathbf{PX}$" in order to use such a formula (9) to propose criteria so as to avoid performing cross-validation explicitly. We showed that PCA can be written as:

$$\begin{aligned} \text{vec}(\hat{\mathbf{X}}^{(S)}) &= \mathbf{P}^{(S)}\,\text{vec}(\mathbf{X}) \text{ with,} \\ \mathbf{P}^{(S)} &= \left(\mathbb{I}_K \otimes \frac{1}{I}\mathbb{1}\mathbb{1}'\right) + \left(\mathbf{P}_\mathbf{U}' \otimes \left(\mathbb{I}_I - \frac{1}{I}\mathbb{1}\mathbb{1}'\right)\right) \\ &\quad + (\mathbb{I}_K \otimes \mathbf{P}_\mathbf{F}) - \left(\mathbf{P}_\mathbf{U}' \otimes \mathbf{P}_\mathbf{F}\right), \end{aligned} \tag{10}$$

where vec is the operator of vectorization: $\text{vec}(\hat{\mathbf{X}})$ is a vector of size IK with the columns of $\hat{\mathbf{X}}$ stacked below each other; $\mathbb{1}$ is the I-vector of 1's; $\otimes$ is the Kronecker product; $\mathbf{P}_\mathbf{F} = \hat{\mathbf{F}}(\hat{\mathbf{F}}'\hat{\mathbf{F}})^{-1}\hat{\mathbf{F}}'$ and $\mathbf{P}_\mathbf{U} = \hat{\mathbf{U}}(\hat{\mathbf{U}}'\hat{\mathbf{U}})^{-1}\hat{\mathbf{U}}'$. $\mathbf{P}_\mathbf{F}$ and $\mathbf{P}_\mathbf{U}$ are the two orthogonal projection matrices involved in PCA, projection respectively onto the space spanned by $\hat{\mathbf{F}}$ (projection onto the column space of $(\mathbf{X} - \hat{\mathbf{M}})$, equation 3) and onto the space spanned by $\hat{\mathbf{U}}$ (projection onto the row space of $(\mathbf{X} - \hat{\mathbf{M}})$, equation 2). The matrix $\mathbf{P}$ of size $IK \times IK$ is an orthogonal projection matrix.

The MSEP for S dimensions can be approximated (it is only an approximation since the matrix $\mathbf{P}$ defined in equation 10 depends on the data) by the quantity denoted SACV for smoothing approximation of cross-validation:

$$\text{SACV}(S) = \frac{1}{IK}\sum_{i=1}^{I}\sum_{k=1}^{K}\left(\frac{\hat{X}_{ik}^{(S)} - X_{ik}}{1 - P_{ik,ik}}\right)^2, \tag{11}$$

with $P_{ik,ik}$, the ik-th diagonal element of $\mathbf{P}$. It is also possible to define a generalized cross-validation (GCV) criterion [7] which consists in substituting each element $P_{ik,ik}$ by its average value, *i.e.* $\text{tr}(\mathbf{P})/(IK)$. Using $\text{tr}(\mathbf{P}_\mathbf{U}') = \text{tr}(\mathbf{P}_\mathbf{F}) = S$, the trace of the projection matrix (10) is:

$$\text{tr}\left(\mathbf{P}^{(S)}\right) = K \times 1 + S \times (I-1) + K \times S - S^2. \tag{12}$$

This trace corresponds to the number of independent estimated parameters, *i.e.* K for the centering of $\mathbf{X}$ plus $(I-1)S$ for the (centred) scores plus KS for the loadings minus S^2 that represents the orthogonality constraints. The degrees of freedom of the residuals are defined as $\text{tr}\left(\mathbb{I}_{IK} - \mathbf{P}^{(S)}\right) = IK - (K + IS - S + KS - S^2) = IK - K - IS - KS + S^2 + S$. The GCV criterion is then defined as follows:

$$\begin{aligned} \text{GCV}(S) &= \frac{1}{IK} \times \frac{\sum_{i=1}^{I}\sum_{k=1}^{K}(\hat{X}_{ik}^{(S)} - X_{ik})^2}{(1 - \text{tr}(\mathbf{P}^{(S)})/(IK))^2}, \\ &= \frac{IK \times (\sum_{i=1}^{I}\sum_{k=1}^{K}(\hat{X}_{ik}^{(S)} - X_{ik})^2)}{(IK - K - IS - KS + S^2 + S)^2}. \end{aligned} \tag{13}$$

 ISSN: 2102-6238

The GCV criterion is a residual sum of squares criterion times a correction term, which penalizes for model complexity. Indeed, when the number of components increases, the residual sum of squares decreases but this decreasing is balanced by the squares of the degrees of freedom of the residuals. Simulations were performed to compare these two approximations to others well-known methods to choose the number of components in PCA and the results obtained were very competitive.

We can remark that with the degrees of freedom of the residuals, it is possible to correct the maximum likelihood estimate of σ^2 from model (4) as followed:

$$\hat{\sigma}^2_{\text{corrected}} = \frac{I \sum_{s=S+1}^{K} \hat{\lambda}_s}{IK - K - IS - KS + S^2 + S}.$$

This new estimation can thus be used in the bootstrap residuals procedure (section 2.2). It is also possible to modify residuals by correcting the residuals with the leverages of the observations as $\tilde{\varepsilon}_{ik} = \frac{\hat{X}_{ik}^{(S)} - X_{ik}}{\sqrt{1 - P_{ik,ik}}}$. Work has to be done to assess the latter proposal.

3. Incomplete case

3.1. Estimation of the principal axes and components with missing values

3.1.1. The iterative PCA algorithm

A common approach dealing with missing values in PCA consists in ignoring the missing values by minimizing the least squares criterion (1) over all non missing entries. This can be achieved by the introduction of a weighted matrix $\mathbf{W}$ in the criterion, with $W_{ik} = 0$ if X_{ik} is missing and $W_{ik} = 1$, otherwise:

$$\mathscr{C} = \sum_{i=1}^{I} \sum_{k=1}^{K} W_{ik} (X_{ik} - m_k - \sum_{s=1}^{S} F_{is} U_{ks})^2. \tag{14}$$

In contrast to the complete case, there is no explicit solution to minimize the criterion (14) and it is necessary to resort to iterative algorithms. Many algorithms are available in the literature such as criss-cross multiple regression [13] which consists in alternating two weighted multiple regressions until convergence. It is a direct extension of the ALS algorithm (equation 2 and 3) to the incomplete case. The geometric interpretation of this algorithm is the following: the matrix $\mathbf{W}$ implies that the weights of the variables are different from one individual to another and reciprocally the weights of the individuals are different from one variable to another [27].

The *iterative PCA* algorithm proposed by [28] also minimizes the criterion (14). In this algorithm, an imputation of the missing values is achieved during the estimation process. More precisely, the algorithm is the following:

1. Initialization $\ell = 0$: $\mathbf{X}^0$. Missing elements are replaced by initial values such as for example the mean of each variable
2. Step ℓ:

(a) PCA is performed on the completed dataset to estimate the parameters $\rightarrow (\hat{\mathbf{M}}^{\ell}, \hat{\mathbf{F}}^{\ell}, \hat{\mathbf{U}}^{\ell})$; S dimensions are kept

(b) missing values are imputed with the fitted values $\hat{\mathbf{X}}^{\ell} = \hat{\mathbf{M}}^{\ell} + \hat{\mathbf{F}}^{\ell}\hat{\mathbf{U}}^{\ell\prime}$; the new imputed dataset is $\mathbf{X}^{\ell} = \mathbf{W} * \mathbf{X} + (1 - \mathbf{W}) * \hat{\mathbf{X}}^{\ell}$: observed values are the same and missing values are replaced by the fitted one

3. Steps 2.(a) of estimation of parameters via PCA and 2.(b) of imputation are repeated until convergence

The *iterative PCA* corresponds to an expectation maximization algorithm [9] associated to the model (4) and is thus often named EM-PCA algorithm. This property is important and allows us to position the methodology in the missing data theory. When doing maximum likelihood based inference, the mechanism that leads to missing values can be ignored if it is a MAR mechanism. EM-algorithms are special cases of MM (majorize/minimize) algorithms [29] where a complex criterion (the weighted least squares criterion, equation 14) is minimized by successive minimizations of a simple criterion (the least squares criterion, equation 1, minimized via the PCA step 2.(a)).

Such algorithms of iterative imputation have first been proposed in the framework of exploratory data analysis methods in correspondence analysis (CA) by [35] and [16, p.238] and these kind of algorithms can be traced back to the work of [20].

3.1.2. A single imputation method

Even if the aim of the algorithm is to estimate the parameters in spite of the missing values, an imputation is achieved. Consequently, the *iterative PCA* algorithm can also be viewed as a single imputation method. This is a strong and highly appealing point. We thus have both parameters estimate with welcome properties (since they are maximum likelihood estimates), and a completed dataset that can be analyzed (cautiously) by any statistical method. Moreover, since the imputation is based on the PCA model (on the scores and loadings), it takes into account the similarities between individuals and the relationships between variables. The *iterative PCA* method improves the prediction of the missing values compared to the mean imputation (which is the first step in the algorithm).

That is why for a few years there has been a surge of interest in imputing with the PCA model, especially in the machine learning community with data matrix completion problems. One of the most famous illustration is the Netflix Prize [34, 22]. However, it is very important to note that their aim is to provide the best prediction of the missing entries and not to best estimate parameters despite missing values. The conceptual basis is thus very different from what we are interested in this paper, even though the techniques used are very similar.

3.1.3. Scaling

After each imputation step, the means of the variables change. Consequently the vector $\mathbf{M}$ has to be updated after each imputation step. This updating seems self-evident since the offset term is part of the model (4). It is important to adopt the same point of view regarding scaling. If one wishes to perform a standardized PCA (to give the same weight to each variable in the

 ISSN: 2102-6238

analysis) with missing values, a rescaling step is then incorporated after each imputation step. In the complete case, the scaling is often carried out prior to the analysis and consequently it is often regarded as a pre-processing step. In the incomplete case, it is necessary to consider the scaling process as a part of the analysis. Surprisingly, including the scaling terms in the model (4) is not so trivial and there is still work to be done in this direction.

3.1.4. Overfitting

The major problem of the algorithms minimizing the criterion (14) is the overfitting problem. To illustrate this problem, let us consider an example with a simulated data generated as a fixed structure in two dimensions corrupted by noise: $\mathbf{X}_{41\times 6} = \mathbf{F}_{41\times 2}\mathbf{U}'_{2\times 6} + \varepsilon$ with $\varepsilon_{ik} \sim \mathcal{N}(0, 0.5)$. The two-dimensional configuration of the individuals associated to this dataset is given figure 1 on the left. Then, 50% of values are removed completely at random from $\mathbf{X}$ and the *iterative PCA* algorithm is performed on this incomplete dataset. The configuration obtained (figure 1 in the middle) is not satisfactory: individuals are going far from the center of gravity. The fitting error (corresponding to the criterion 14) is low (0.48), whereas the error of prediction (using $(1 - W_{ik})$ rather than W_{ik} in the criterion 14) is very high (5.58), which is characteristic of the overfitting problem. It means that the quality of prediction of the missing values and the quality of estimation of the parameters are very poor. It is also important to note that the percentage of variability explained by the first two dimensions is higher (94%) than the one obtained from the completed dataset (82%).

Problems of overfitting are exacerbated with increasing number of missing values, with increasing number of dimensions, since many parameters are estimated with respect to the number of observed values, and with a decreasing signal-to-noise ratio. The reason is that with limited and/or noisy data, the model fits the available data too well and places too much "trusts" in the relationships between observed variables even if such relations are far from the true ones. We can also remark that the problem of missing values can be regarded either as a special case of small sample data, or in our particular situation where an imputation is associated to the estimation of the parameters, as a generalization of a prediction problem. Both correspond to the usual situations where overfitting problems may occur. Moreover, the algorithm used to deal with missing values is an EM-type algorithm and it can be expected that such an algorithm is prone to overfitting problems.

3.1.5. A regularized iterative PCA algorithm

A first way to reduce overfitting can be to reduce the number of dimensions S used in the algorithm in order to estimate less parameters; however it is important not to remove too many components since information can be lost. Another solution can be to resort to early stopping but it is not really satisfactory. One of the other major approaches is to resort to shrinkage methods. We detailed in [27] a *regularized iterative PCA* algorithm to overcome the overfitting problem. The algorithm is very similar to the *iterative PCA* one. There is an initialisation step of the missing values, an estimation step of the parameters and an imputation step of the missing values. However, the imputation step 2.(b) which was written as:

 ISSN: 2102-6238

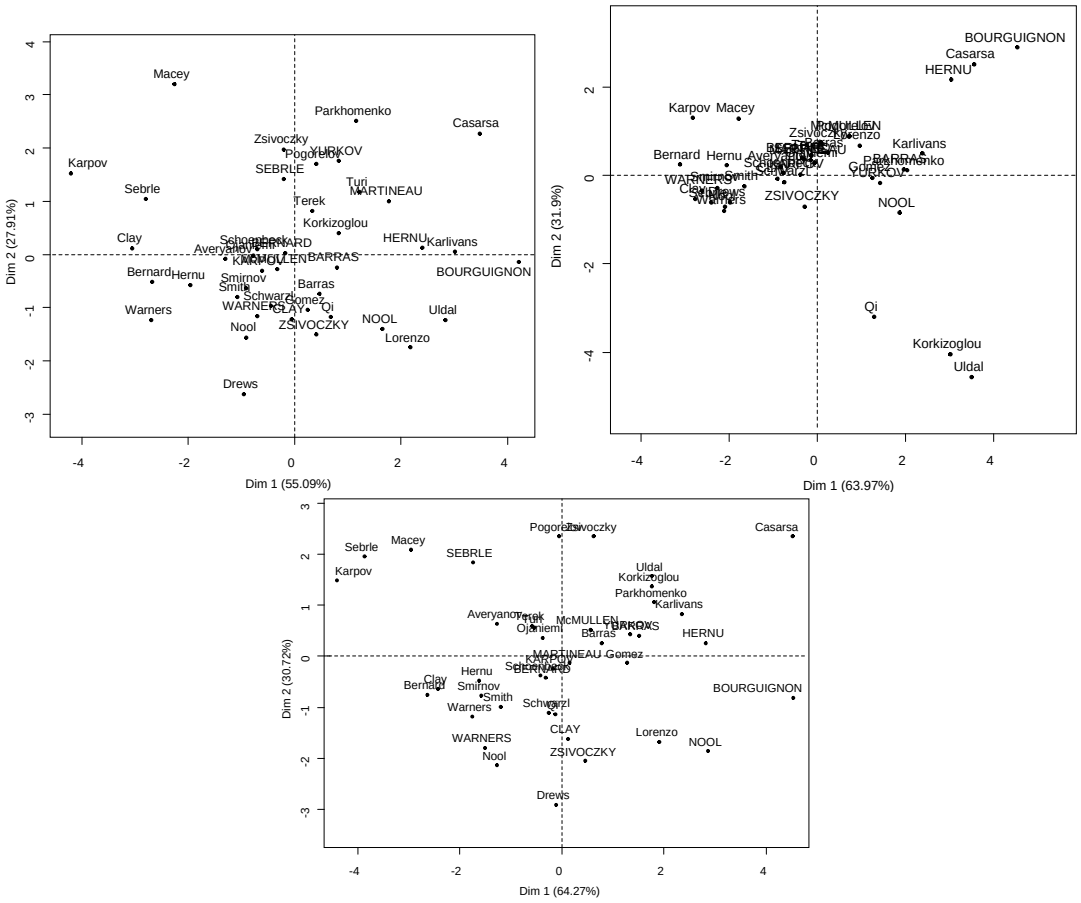

FIGURE 1. *Illustration of the overfitting problem on the individuals PCA map obtained on a dataset with a strong structure and 50% of missing values. The true configuration is on the left, the configuration obtained with the* iterative PCA *algorithm is in the middle, the configuration obtained with the* regularized iterative PCA *algorithm is on the right.*

$$\hat{X}_{ik}^{\ell} = \hat{m}_k + \sum_{s=1}^{S} \hat{F}_{is}^{\ell} \hat{U}_{ks}^{\ell} = \hat{m}_k + \sum_{s=1}^{S} \frac{\hat{F}_{is}^{\ell}}{\|\hat{\mathbf{F}}_s^{\ell}\|} (\sqrt{\hat{\lambda}_s}) \hat{U}_{ks}^{\ell}$$

is replaced by a "shrunk" imputation step:

$$\hat{X}_{ik}^{\ell} = \hat{m}_k + \sum_{s=1}^{S} \frac{\hat{F}_{is}^{\ell}}{\|\hat{\mathbf{F}}_s^{\ell}\|} \left(\sqrt{\hat{\lambda}_s} - \frac{\hat{\sigma}2}{\sqrt{\hat{\lambda}_s}} \right) \hat{U}_{ks}^{\ell} \tag{15}$$

$$= \hat{m}_k + \sum_{s=1}^{S} \left(\frac{\hat{\lambda}_s - \hat{\sigma}2}{\hat{\lambda}_s} \right) \hat{F}_{is}^{\ell} \hat{U}_{ks}^{\ell}, \tag{16}$$

with $\hat{\sigma}^2 = \frac{1}{K-S} \sum_{s=S+1}^{K} \hat{\lambda}_s$. Implicitly, it is assumed that the last dimensions are restricted to noise, that is why the noise variance is estimated by the mean of the last eigenvalues. The rationale of the algorithm is thus to remove the noise to avoid instabilities in the predictions. In the extreme case

 ISSN: 2102-6238

when there is no noise ($\hat{\sigma}$ is equal to zero), the *regularized iterative PCA* algorithm is equivalent to the *iterative PCA* algorithm. At the opposite, when the noise is very important which implies that the quantity $\left(\sqrt{\hat{\lambda}_s} - \frac{\hat{\sigma}^2}{\sqrt{\hat{\lambda}_s}}\right)$ is close to 0, the algorithm tends to impute with the mean of each variable. It is thus no longer using the structure of the data to impute missing values. This is acceptable since the structure is either too weak or can not be trusted since it is learned on the basis of too few or too noisy observations. Between these two extreme cases, each singular value $\sqrt{\hat{\lambda}_s}$ is thresholded by the term $\left(\frac{\hat{\sigma}^2}{\sqrt{\hat{\lambda}_s}}\right)$ which implies that the amount of thresholding is greater for the smallest singular values (among the first S ones). This is also acceptable since these singular values can be seen as responsible for the instabilities in the predictions.

We explained in [27] that the thresholded terms comes from the probabilistic formulation of PCA (section 2.1.2). The regularized model matrix (16) is defined using the "estimation" of the scores (equation 7) and the estimation of the loadings (equation 6):

$$\hat{X}_{regularized} = \hat{\mathbf{M}} + \hat{\mathbf{Z}}\hat{\mathbf{B}}' = \hat{\mathbf{M}} + \sum_{s=1}^{S} \frac{\hat{\lambda}_s - \hat{\sigma}^2}{\hat{\lambda}_s} \mathbf{F}_s \mathbf{U}_s'$$

However, since the *regularized iterative PCA* algorithm is also used in cases where individuals are not coming from a larger population of individuals, it is better from a conceptual point of view to consider the regularization as coming from a Bayesian treatment of the *fixed effect* model (section 2.1.2). It is also possible to note that under model (4), the variance of the signal is equal to $(\lambda_s - \sigma^2)$ for each dimension. Consequently, the quantity $\frac{\hat{\lambda}_s - \hat{\sigma}^2}{\sqrt{\hat{\lambda}_s}}$ which can also be written as $\sqrt{\hat{\lambda}_s} \times \frac{\hat{\lambda}_s - \hat{\sigma}^2}{\hat{\lambda}_s}$ shows that the singular values are shrunk by a term which can be considered as the ratio of the signal over the signal plus noise, which is a common term for shrinkage methods.
Remark:
Such algorithms, which can be called "spectral thresholding algorithms", can also be found in the matrix completion literature. For example, [31] used a soft thresholding algorithm which means that the singular values are thresholded by a same quantity. Here, the last singular values are killed and the others are thresholded by a specific amount for each singular value. Our proposition is more related to a recent proposal from [14] who defined a weighted version of the soft thresholding algorithm.

The *regularized iterative PCA* algorithm is applied on the example (figure 1). The fitting error and the prediction error are approximately of the same magnitude (0.52 and 0.67) and there are no points distant from the centre of gravity in the PCA map (figure 1 on the bottom). The regularization comes down to shrinking the coordinates of the individuals towards the origin. The PCA map is very close to the observed one and the percentage of variability as explained by the first map is also closest to the one obtained from the complete dataset.

We conducted in [27] a simulation study to assess the performances of the *regularized iterative PCA* algorithm. We compared the proposed method to the non regularized algorithm as well as to other well-known methods to deal with missing values in PCA such as the NIPALS algorithm [6, 47]. We varied many parameters such as the percentage of missing values, the relationship

 ISSN: 2102-6238

between variables, etc. The main points are the following. The regularized algorithm clearly outperforms the non regularized one especially in noisy schemes with many missing values (it corresponds to situations where overfitting problems are very important). When the noise is very important, the regularized algorithm tends to impute the missing values with the mean of the variables(such behavior is expected). If the algorithms are used with too many dimensions (compared to the true underlying ones known in a simulation study), the performances of the EM algorithm are catastrophic whereas the regularized algorithm still provides satisfactory results. It is because the regularization limits the influence of taking too many dimensions which are made only by noise. The regularized algorithm also clearly outperforms the NIPALS algorithm. The latter is very unstable (it doesn't converge all the time) especially when there are many missing values. Moreover, we showed that NIPALS is not so well suited to deal with missing values. Indeed, NIPALS proceeds by deflation, it means that the first dimension is estimated, then the second dimension is found from the residual matrix, and so on for the other dimensions. Contrary to the complete case, this procedure doesn't ensure the minimization of the global criterion (14). In addition, due to the deflation procedure, it is not possible to perform a standardized PCA with missing values (it is not possible to update the standard deviation of the variables).

3.2. Multiple imputation in PCA

After the point estimate of the parameters from an incomplete data, it is natural to focus on notions of variability. In a missing data framework, a new source of variability should be taken into account: the "variability due to missing values". A solution can be to resort to multiple imputation. A first way to generate multiple imputed datasets could be to generate multiple draws from the predictive distribution of the missing values given the observed values and the estimation of the parameters obtained from the *(regularized) iterative PCA* algorithm: for $d = 1,...,D$, $X_{ik}^d \sim \mathcal{N}\left(\hat{m}_k + (\hat{\mathbf{F}}\hat{\mathbf{U}}')_{ik}, \hat{\sigma}^2\right)$. However, [30, p.214] qualified this kind of multiple imputation as "improper" since the parameters are considered as "true" parameters when there are in fact only an estimation. A "proper" imputation requires to reflect the uncertainty of the estimation of the parameters from one imputation to the next. In order to do so and to obtain D plausible sets of parameters $(\hat{\mathbf{M}}, \hat{\mathbf{F}}, \hat{\mathbf{U}}')^1, ..., (\hat{\mathbf{M}}, \hat{\mathbf{F}}, \hat{\mathbf{U}}')^D$, we used in [25] a residual bootstrap procedure applied to the incomplete dataset. Then, we generated D imputed datasets by drawing missing values X_{ik}^d from the predictive distribution $\mathcal{N}\left(\hat{m}_k^d + (\hat{\mathbf{F}}\hat{\mathbf{U}}')_{ik}^d, \hat{\sigma}^2\right)$.

The imputed datasets are juxtaposed as shown in figure 2: all the values in blue except the missing ones (boxes) are the same for all the datasets. The first imputed dataset on the left (with black boxes) is the one obtained from the *(regularized) iterative PCA* algorithm (missing values are imputed from their conditional mean without adding any variability). The other tables are obtained from the multiple imputation procedure. The representation of the individuals and variables obtained from the PCA on the data table on the left are considered as the reference configurations. We described two approaches to visualize the different imputed datasets. The first one consists in projecting each imputed data as supplementary information onto the reference configurations as illustrated in figure 3 for the individuals. Individuals (and likewise the variables) without missing values are projected exactly on their corresponding point on the reference configuration, and individuals (and likewise the variables) with several missing values are projected around their corresponding point. Convex hulls or ellipses can be constructed for the individuals. This approach

 ISSN: 2102-6238

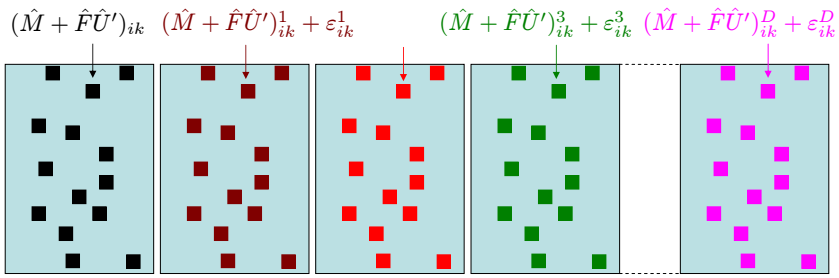

FIGURE 2. *Multiple imputed datasets. The table on the left is the one obtained with the (regularized) iterative PCA algorithm. The others are obtained with the multiple PCA imputation algorithm.*

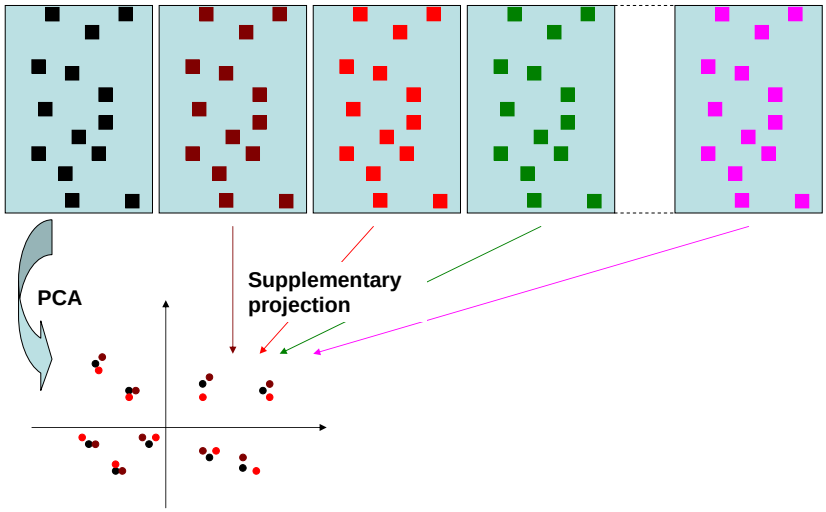

FIGURE 3. *Supplementary projection of the multiple imputed datasets onto the reference configuration (in black).*

allows one to visualize the position of the individuals (and likewise the variables) with different predictions for the missing values. The second approach consists of performing a PCA on each imputed dataset leading to different values for the parameters (axes and components). In order to compare the results provided by each PCA, it is possible to resort to Procrustes rotations [15] to fit the PCA configurations obtained from the imputed data tables toward the fixed reference configuration as illustrated in figure 4 for the individuals. In this approach, all individuals have confidence areas, even those who do not have missing values. Indeed, even if only the prediction of the missing values change, it impacts all the parameters. This approach allows one to visualize the influence of the different predictions on the estimation of the parameters, it means the between-imputation variability. It does not represent the total variance of the parameters with missing values which would be a mixed of the within and between-imputation variability (a combination of the variance due to the noise and to the missing values). Nevertheless, the confidence areas are precious for the user in order to know which credit can be given to a solution obtained from incomplete data.

 ISSN: 2102-6238

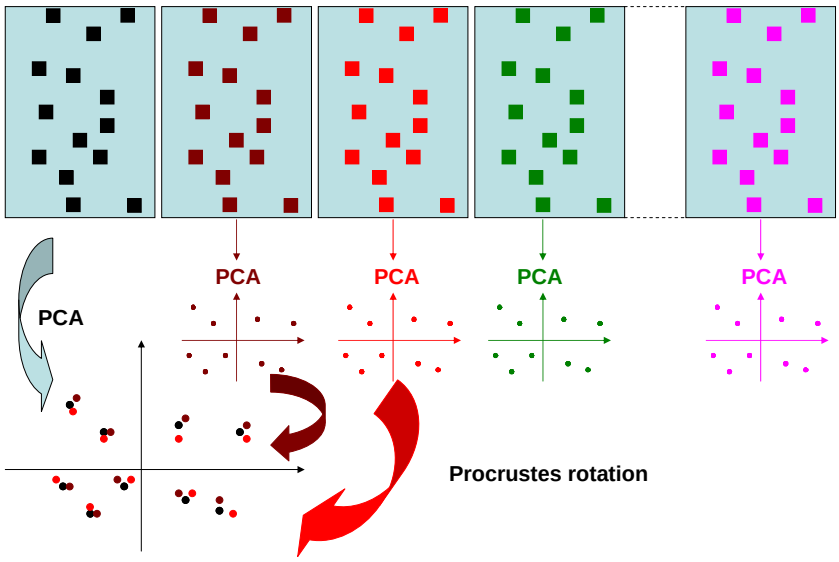

FIGURE 4. *Procrustes rotations of the PCA configuration obtained from the multiple imputed datasets onto the reference configuration.*

3.3. Choosing the number of dimensions with missing values

In the incomplete case, the choice of the number of dimensions is also an important issue. Contrary to the complete case, solutions are not nested in the incomplete case: 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* and used in the *(regularized) iterative PCA* algorithm, is all the more important. If too few dimensions are kept, information can be lost and if too many dimensions are kept, overfitting problems may occur. The number of dimensions has also to be specified for the multiple imputation procedure (when using the residual bootstrap procedure). A method to select this number from an incomplete data is thus required. No method is available in the literature. The cross-validation procedure described section 2.3 can be extended to the incomplete case. However this procedure is all the more unstable as the dataset contains already many missing values and is time-consuming. We propose to use an extended version of the GCV criterion:

$$\text{GCV}(S) = \frac{(IK - nb\ miss) \times (\sum_{i=1}^{I} \sum_{k=1}^{K} (W_{ik}(\hat{X}_{ik}^{(S)} - X_{ik}))^2)}{(IK - nb\ miss - K - IS - KS + S^2 + S)^2},$$

with *nb miss* the number of missing cells. The least squares criterion is replaced by a weighted least squares criterion and the number of observations IK by the number of observed values $(IK - nb\ miss)$.

 ISSN: 2102-6238

3.4. *Implementation with the missMDA package*

In this section, we illustrate the methodology on an incomplete dataset. To carry out the PCA with missing values, we use the `missMDA` package (from the R statistical software [36]) which is dedicated to handle missing values in principal component methods. We illustrate the method on the dataset orange juice where 12 orange juices are described by sensory descriptors (left table in figure 6). Each cell corresponds to the average of the scores given by a set of experts for one orange juice and one sensory descriptor. To perform PCA on an incomplete dataset, the first step consists in estimating the number of dimensions that will be used in the *regularized iterative PCA* algorithm. Figure 5 represents the prediction error for different number of dimensions calculated by cross-validation (CV) and GCV. The error for the model without components corresponds to

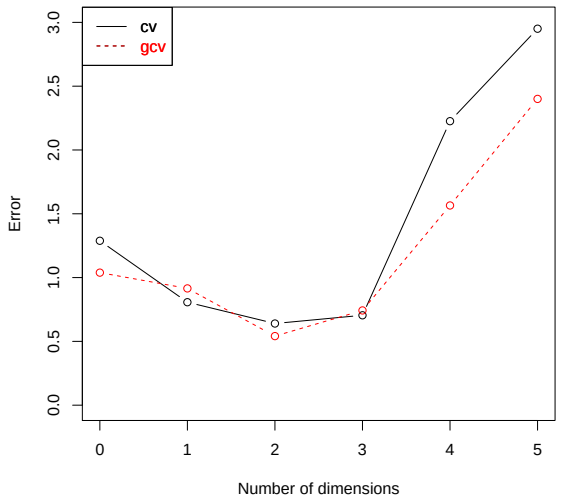

FIGURE 5. *Cross-validation (CV) and its GCV approximation.*

a reconstruction of order 0, *i.e.* $\hat{\mathbf{X}}_{ik}^{(S)} = \hat{\mathbf{M}}$. Cross-validation and its GCV approximation have a well-marked minimum for two components. The shape of the curves is very appealing since it decreases until it reaches a minimum and increases thereafter.

Then we perform the *regularized iterative PCA* algorithm with 2 dimensions and obtain a completed dataset (right table in figure 6).

```
 Sweet Acid Bitter Pulp Typicity
   NA   NA  2.83   NA     5.21
5.46 4.13  3.54 4.62     4.46
   NA 4.29  3.17 6.25     5.17
4.17 6.75    NA 1.42     3.42
...
   NA   NA    NA 7.33     5.25
4.88 5.29  4.17 1.50     3.50
```

```
Sweet Acid Bitter Pulp Typicity
 5.54 4.13   2.83 5.89     5.21
 5.46 4.13   3.54 4.62     4.46
 5.45 4.29   3.17 6.25     5.17
 4.17 6.75   4.73 1.42     3.42
...
 5.71 3.87   2.80 7.33     5.25
 4.88 5.29   4.17 1.50     3.50
```

FIGURE 6. *Incomplete dataset and the completed dataset obtained after the* regularized iterative PCA *algorithm.*

The figure 7 shows the first two dimensions of the PCA performed on this completed dataset.

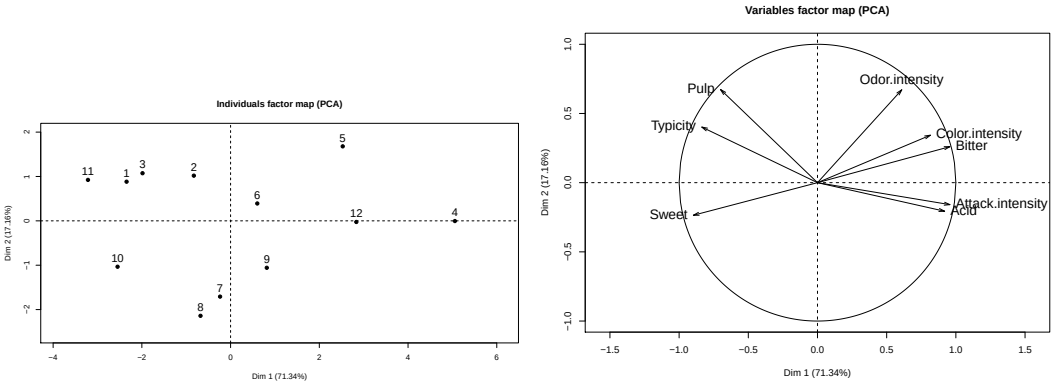

FIGURE 7. *PCA results from the incomplete dataset: individuals and variables representation on the two first dimensions.*

We then carry out the multiple imputation procedure which gives as an output different imputed datasets. The figure 8 represents the projection of the different imputed datasets onto the reference configuration (figure 7) as supplementary elements (supplementary individuals and supplementary columns).

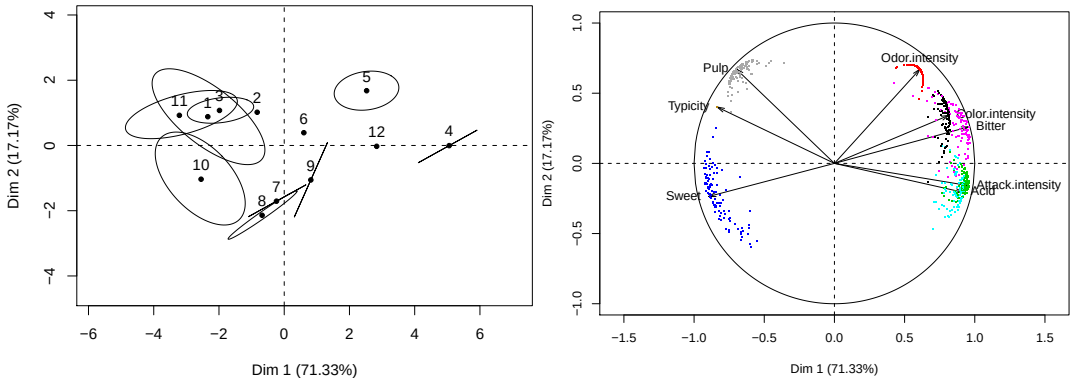

FIGURE 8. *Representation of the uncertainty due to missing values on the position of the individuals and of the variables.*

Individual 12 has no missing value and consequently no ellipse. Individuals 4 has only one missing value for the variable "Bitter". Similarly, the "Odor intensity" variable has one missing value while the "Sweet" variable has several missing values. The sizes of the ellipses are not too large and the general interpretation of the PCA is not affected by the missing values. These ellipses are the only way to assess the relevance of performing a PCA on such an incomplete dataset.

 ISSN: 2102-6238

	V1	V2	V3	...	V14
ind 1	a	NA	g	...	u
ind 2	NA	f	g		u
ind 3	a	e	h		v
ind 4	a	e	h		v
ind 5	b	f	h		u
ind 6	c	f	h		u
ind 7	c	f	NA		v
...	...	...	...		...
ind 1232	c	f	h		v

	V1_a	V1_b	V1_c	V2_e	V2_f	V3_g	V3_h	...
ind 1	1	0	0	0.71	0.29	1	0	...
ind 2	0.12	0.29	0.59	0	1	1	0	...
ind 3	1	0	0	1	0	0	1	...
ind 4	1	0	0	1	0	0	1	...
ind 5	0	1	0	0	1	0	1	...
ind 6	0	0	1	0	1	0	1	...
ind 7	0	0	1	0	1	0.37	0.63	...
...	...	...	...	...	...	...	...	...
ind 1232	0	0	1	0	1	0	1	...

FIGURE 9. *On the left: an incomplete data with categorical variables. On the right: the imputed indicator matrix obtained after performing the regularized iterative MCA algorithm.*

3.5. Multiple Correspondence Analysis

MCA is an exploratory method to visualise categorical data and which is often used to analyse questionnaire. MCA allows one to study the similarities between individuals (the distance used is the chi-square distance), to study the associations between categories and between variables. Even if MCA is a method in itself with specific properties and specific graphical representations, the core of MCA is a specific weighted PCA. Consequently, the methodology developed to handle missing values in PCA can be extended to MCA.

Let us consider a dataset with I individuals and J categorical variables $\mathbf{v}_j,\ j = 1,...,J$ with k_j categories. The data are coded using the indicator matrix of dummy variables, denoted $\mathbf{X}$ of size $I \times K$ with $K = \sum_{j=1}^{J} k_j$. MCA [17] can be presented as the PCA of the following $(\mathbf{Z}, \mathbf{M}, \mathbf{D})$ triplet:

$$(I\mathbf{X}\mathbf{D}_\Sigma^{-1}, \frac{1}{IJ}\mathbf{D}_\Sigma, \frac{1}{I}\mathbb{I}_I),$$

with $\mathbf{D}_\Sigma = \text{diag}\left((I_k)_{k=1,...,K}\right)$ the diagonal matrix of the column margins of the matrix $\mathbf{X}$. The matrix $\mathbf{D} = \frac{1}{I}\mathbb{I}_I$ corresponds to the row masses and the matrix $\mathbf{M} = \frac{1}{IJ}\mathbf{D}_\Sigma$ is the metric (used to compute distances between rows). We defined a *regularized iterative MCA* algorithm in [24] to perform MCA with missing values. Roughly speaking, the algorithm consists of a step of initialization where the missing values in the indicator matrix are imputed by initial values such as the proportion of the category. This initialisation for the categorical variables is the equivalent to the mean imputation for continuous variables and corresponds to the method named *missing fuzzy average*. Note that the entries can be non-integer but the sum of the entries corresponding to one individual and one variable has to be equal to one. Then MCA is performed on the imputed indicator matrix to obtain an estimation of the parameters and the missing values are imputed using the model matrix. After the imputation step, the margins $\mathbf{D}_\Sigma$ change, and consequently it is necessary to incorporate a step in order to update the margins in a similar way as for the scaling in section 3.1.3. When the algorithm converges, it provides both the classical outputs of MCA (scores and loadings) obtained from the incomplete data, as well as an imputed indicator matrix as illustrated figure 9. The margins of the imputed fuzzy indicator matrix are still equal to one per variable which ensures to preserve many MCA properties. Moreover, each imputed value can be seen as a degree of membership of the corresponding category and the original data may be imputed with the most plausible category (for example with the category c for the individual 2 for the first variable). We compared the *regularized iterative MCA* method to other proposals to

deal with missing values in MCA such as the *missing passive* [32], the *missing passive modified margin* [11], and *subset MCA* [18]. We discussed which method is well suited to which kind of missing values and we presented a real data analysis. Note that the most usual approach to deal with missing values consists in coding the missing values as a new category (for example NA for not available) and then performing the MCA on the new dataset. This allows one to see the possible associations between missing entries, *i.e.* the pattern of missing values (indicating that individuals who do not have answered a question do not have answered to other questions). This analysis is also useful when the missing values correspond to a new category or have a specific meaning. On the contrary, the *regularized iterative MCA* algorithm is dedicated to missing entries that mask an underlying category among the available categories.

4. Conclusion

This work sums up our contributions on the topic of handling missing values in principal component methods and especially in PCA. The *regularized iterative PCA* algorithm allows one to obtain a point estimate of the parameters and to overcome the major problem of overfitting. Then, the multiple imputation procedure gives indications on the variability of the parameters. Finally, a solution is proposed to the recurrent problem of the choice of the number of dimensions. The methodology proposed in PCA serves as a base for handling missing data for other methods. The case of MCA has been discussed quickly. We are studying handling missing values procedures for other methods dedicated to the description of multi-table data such as multiple factor analysis [12] and multi-level simultaneous component analysis [43]. The methodology is implemented in the package *missMDA* of the free R software [36].

We can mention one main perspective for this work. It may be interesting to assess the performances of the *regularized iterative PCA* and of the *regularized iterative MCA* algorithms as single imputation methods and to compare it to several approaches dealing with missing values for continuous and categorical data. The first attempts show that the approaches are competitive. In the same way, the multiple imputation method based on PCA can be seen as an alternative to other multiple imputation methods [46, 21].

References

[1] D. J. Bartholomew. *Latent Variable Models and Factor Analysis*. Griffin, 1987.

[2] J.-P. Benzécri. *L'analyse des données. Tome II: L'analyse des correspondances*. Dunod, 1973.

[3] R. Bro, K. Kjeldahl, A. K. Smilde, and H. A. L. Kiers. Cross-validation of component model: a critical look at current methods. *Anal Bioanal Chem*, 390:1241–1251, 2008.

[4] H. Caussinus. Models and uses of principal component analysis (with discussion). In J. de Leeuw, W. Heiser, J. Meulman, and F. Critchley, editors, *Multidimensional Data Analysis*, pages 149–178. DSWO Press, 1986.

[5] F. Chateau and L. Lebart. Assessing sample variability in the visualization techniques related to principal component analysis: bootstrap and alternative simulation methods. In A. Prats, editor, *COMPSTAT, Physica-Verlag*, pages 205–210, 1996.

[6] A. Christofferson. *The one-component model with incomplete data*. PhD thesis, Uppsala University, Institute of statistics, 1969.

[7] P. Craven and G. Wahba. Smoothing noisy data with spline functions. *Numer. Math.*, 31(4):377–403, 1979.

[8] J. de Leeuw, A. Mooijaart, and R. van der Leeden. Fixed factor score models with linear restrictions. Technical report, Leiden: Department of Datatheory, 1985.

 ISSN: 2102-6238

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society B*, 39(1):1–38, 1977.

[10] J.-B. Denis and J. C. Gower. Asymptotic confidence regions for biadditive models: interpreting genotype-environment interactions. *Applied Statistics*, 45(4):479–493, 1996.

[11] B. Escofier. Traitement des questionnaires avec non réponse, analyse des correspondances avec marge modifiée et analyse multicanonique avec contrainte. *Pub. Inst. Stat. Univ.*, 32(3):33–69, 1987.

[12] B. Escofier and J. Pagès. *Analyses Factorielles simples et multiples*. Dunod, 2008.

[13] K. R. Gabriel and S. Zamir. Lower rank approximation of matrices by least squares with any choice of weights. *Technometrics*, 21(4):236–246, 1979.

[14] S. Gaiffas and G. Lecué. Weighted algorithms for compressed sensing and matrix completion. *Submitted*, 2011.

[15] J. C. Gower and G. B. Dijksterhuis. *Procrustes Problems*. New York: Oxford University Press, 2004.

[16] M. Greenacre. *Theory and Applications of Correspondence Analysis*. Acadamic Press, 1984.

[17] M. Greenacre and J. Blasius. *Multiple Correspondence Analysis and Related Methods*. Chapman & Hall/CRC, 2006.

[18] M. Greenacre and R. Pardo. Subset correspondence analysis: visualizing relationships among a selected set of response categories from a questionnaire survey. *Sociological methods and research*, 35(2):193–218, 2006.

[19] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning. Data Mining, Inference and Prediction. Second Edition*. Springer series in statistics, 2009.

[20] M. J. R. Healy and M. Wesmacott. Missing values in experiments analyzed on automatic computers. *Applied statistics*, 5(3):203–206, 1956.

[21] J. Honaker, G. King, and M. Blackwell. *Amelia: Amelia II: A Program for Missing Data*, 2010. R package version 1.2-16.

[22] A. Ilin and T. Raiko. Practical approaches to principal component analysis in the presence of missing values. *Journal of Machine Learning Research*, 11:1957–2000, 2010.

[23] I. T. Jolliffe. *Principal Component Analysis*. Springer, 2002.

[24] J. Josse, M. Chavent, B. Liquet, and F. Husson. Handling missing values with regularized iterative multiple correspondence analysis. *Journal of classification*, 29(1):91–116, 2012.

[25] J. Josse and F. Husson. Multiple imputation in pca. *Advances in data analysis and classification*, 5(3):231–246, 2011.

[26] J. Josse and F. Husson. Selecting the number of components in pca using cross-validation approximations. *Computational Statististics and Data Analysis*, 56(6):1869–1879, 2011.

[27] J. Josse, J. Pagès, and F. Husson. Gestion des données manquantes en analyse en composantes principales. *Journal de la Société Française de Statistique*, 150(2):28–51, 2009.

[28] H. A. L. Kiers. Weighted least squares fitting using ordinary least squares algorithms. *Psychometrika*, 62(2):251–266, 1997.

[29] K. Lange. *Optimization*. Springer-Verlag, New-York, 2004.

[30] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. Wiley series in probability and statistics, New-York, 1987, 2002.

[31] R. Mazumder, T. Hastie, and R. Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal machine learning research*, 11:2287–2322, 2009.

[32] J Meulman. *Homgeneity Analysis of Incomplete Data*. D.S.W.O.-Press, Leiden, 1982.

[33] J. Moreno-Gonzalez, J. Crossa, and P. L. Cornelius. Additive main effects and multiplicative interaction model: I. theory on variance components for predicting cell means. *Crop Science*, 43:1967–1975, 2003.

[34] Netflix. Netflix challenge, 2009.

[35] C. Nora-Chouteau. *Une méthode de reconstitution et d'analyse de données incomplètes*. PhD thesis, Université Pierre et Marie Curie, 1974.

[36] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

[37] G. K. Robinson. Blup is a good thing: The estimation of random effects. *Statistical Science*, 6(1):15–51, 1991.

[38] S. Roweis. Em algorithms for pca and sensible pca. *Advances in Neural Information Processing Systems*, 10:626–632, 2008.

 ISSN: 2102-6238

[39] D B Rubin. Inference and missing data. *Biometrika*, 63:581–592, 1976.

[40] D. B. Rubin. *Multiple Imputation for Non-Response in Survey*. Wiley, 1987.

[41] J. Schafer. *Analysis of Incomplete Multivariate Data*. Chapman & Hall/CRC, 1997.

[42] J. L. Schafer and M. K. Olsen. Multiple imputation for missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research*, 33(4):545–571, 1998.

[43] M. E. Timmerman. Multilevel component analysis. *British Journal of Mathematical and Statistical Psychology*, 59(2):301–320, 2006.

[44] M. E. Timmerman, H. A. L. Kiers, and A. K. Smilde. Estimating confidence intervals for principal component loadings: a comparison between the bootstrap and asymptotic results. *British Journal of Mathematica and Statistical Psychology*, 60(2):295–314, 2007.

[45] M. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society B*, 61(3):611–622, 1999.

[46] S. van Buuren. *Flexible Imputation of Missing Data*. Chapman & Hall/CRC, Boca Raton, 2012.

[47] H Wold and E Lyttkens. Nonlinear iterative partial least squares (nipals) estimation procedures. *Bulletin. Int. Stat. Institut*, 43:29–51, 1969.

 ISSN: 2102-6238