Document downloaded from:

http://hdl.handle.net/10251/64900

This paper must be cited as:

Folch-Fortuny, A.; Arteaga Moreno, FJ.; Ferrer Riquelme, AJ. (2015). PCA model building with missing data: New proposals and a comparative study. Chemometrics and Intelligent Laboratory Systems. 146:77-88. doi:10.1016/j.chemolab.2015.05.006.

The final publication is available at

https://dx.doi.org/10.1016/j.chemolab.2015.05.006

Copyright Elsevier

Additional Information

PCA model building with missing data: new proposals and a comparative study

A. Folch-Fortuny[1], F. Arteaga[2], A. Ferrer[1]

[1]Dep. de Estadística e Investigación Operativa Aplicadas y Calidad, Universitat Politècnica de Valencia, Camino de Vera s/n, Edificio 7A. 46022 Valencia (Spain), {abfolfor@upv.es, aferrer@eio.upv.es}

[2]Dep. of Biostatistics and Investigation, Universidad Católica de Valencia San Vicente Mártir, C/Quevedo, 3. 46001 Valencia (Spain), francisco.arteaga@ucv.es

Abstract

This paper introduces new methods for building principal component analysis (PCA) models with missing data: projection to the model plane (PMP), known data regression (KDR), KDR with principal component regression (PCR), KDR with partial least squares regression (PLS) and trimmed scores regression (TSR). These methods are adapted from their PCA model exploitation version to deal with the more general problem of PCA model building when the training set has missing values. A comparative study is carried out comparing these new methods with the standard ones, such as the modified nonlinear iterative partial least squares (NIPALS), the iterative algorithm (IA), the data augmentation method (DA) and the nonlinear programming approach (NLP). The performance is assessed using the mean squared prediction error of the reconstructed matrix and the cosines between the actual principal components and the ones extracted by each method. Four data sets, two simulated and two real ones, with several percentages of missing data, are used to perform the comparison.

Keywords: Missing data, PCA model building, PCA model exploitation.

1. Introduction

Multivariate data sets are usually arranged in matrices having non-registered cells, *i.e.* missing values. Missing data (MD) can appear in a wide range of contexts and for a different number of reasons: respondents not answering to some questions in surveys, values outside the instrument range or missing owing to malfunctions of the sensor, failure in the communication between the instrumentation and the digital control system (DCS), sensor with different sampling rates, errors during data acquisition, and so on

[1,2]. In model building stages, in chemometrics environments, practitioners usually deal with 5-20% of missing values. In complex industrial processes, where hundreds of variables are collected per batch, 30-60% of missing data can appear in their historical data sets. Finally, with the paradigm of Big Data, thousands of variables are collected for a huge set of individuals, having sometimes more than 70% of missing values in their data sets.

Principal component analysis (PCA) is one of the most used chemometrics tools for multivariate data analysis. In this context, two problems related to missing data appear: (1) exploiting fitted PCA models when some measurements are missing in new observations, *i.e.* the *model exploitation* problem (PCA-ME); and (2) building PCA models from data sets with missing measurements, *i.e.* the *model building* problem (PCA-MB).

In PCA-ME a lot of methods have been reported in the literature. Wise and Ricker [3] present a method that consists of imputing the values that minimise the squared prediction error (SPE) for the new incomplete observation, based on the known PCA model. Nelson *et al*. [4] study and compare several methods: the single component projection method (SCP), the projection to the model plane method (PMP) and the conditional mean replacement method (CMR). Walczak and Massart [5] study the adaptation of the iterative algorithm (IA) to the prediction of scores for new objects with missing elements. Arteaga and Ferrer [6] also introduce several methods: the trimmed scores method (TRI), the known data regression method (KDR) and the trimmed scores regression method (TSR). Additionally, they show that the regression-based methods (KDR and TSR) are statistically more efficient than the other methods studied. Arteaga and Ferrer [7] propose a framework that allows writing the regression-based methods by a unique expression, function of a key matrix.

Regarding PCA-MB there are two methods that are frequently used by the practitioners. The first one consists of adapting the nonlinear iterative partial least squares algorithm (NIPALS) [8] to deal with incomplete observations by performing the iterative regressions using the present data and ignoring the missing data [9]. The second one is the aforementioned IA [5] that basically consists of filling in the missing data with the predictions obtained from previous PCA models iterated recursively until convergence. Other methods rely on maximum likelihood-based estimations of missing data, like the expectation maximization (EM) algorithm [10-12]. A more complex method is data augmentation (DA) [10-13]. DA is a multiple imputation method, *i.e.* for each missing

value several values are imputed randomly, and it requires the computation of prior distributions of the parameters. Both EM and DA are not so widely used as the previous ones (NIPALS or IA). The reason is that the usual chemometrics data sets have strongly correlated variables with a low number of observations. The use of either DA or EM implies the inversion of the covariance submatrix corresponding to the known variables given an observation, which in these data sets often is not feasible due to submatrices are singular. A recent approach for PCA-MB with MD is the nonlinear programming approach (NLP). In this method the PCA model is obtained solving a nonlinear programming problem, in which the errors between the non-missing values and the model estimations are minimised [14].

In a recent paper [15], Liu and Brown compared the performance of several methods: the singular value decomposition of Krzanowski [16], the general iterative principal component imputation (GIP) [17], the multiple imputation by chained equations (MICE) [18], and two regularized versions of the known EM algorithm: one based on ridge regression (r-EM) [19] and the other one based on a truncated total least squares regression (t-EM) [20]. Some similarities can be drawn from these methods to the aforementioned ones. The GIP algorithm basically consists of filling in the missing data with the predictions obtained from a PCA model on the available data, and then iterated recursively until convergence, so it is equivalent to the IA. MICE is similar to the KDR adaptation for PCA-MB, the former performed variable-wise, the latter observation-wise. r-EM overcomes the problem of the inversion of the covariance matrix of the available data in EM by using ridge regression, while KDR uses the pseudoinverse (see Sections 2-3). The t-EM method is equivalent to the PMP method for PCA-MB. The KDR and PMP adaptations from model exploitation to model building will be presented in the next section. The equivalences between GIP, MICE, t-EM and IA, KDR, PMP, respectively, are shown in Appendix A.

There are also other imputation methods compared in the literature [21] that are strongly not recommended, like the listwise deletion (in which any observation with missing values is removed) and the unconditional mean imputation. The former implies a huge loss of information, leading to loss of precision and bias. The latter distorts the multivariate empirical distribution of the samples, *i.e.* tends to deform nonlinear quantities (*e.g.* variances, covariances) [2]. The nearest neighbour method has been also suggested to fulfil the missing values in incomplete datasets [22], however, its applicability with high percentages of missing values is limited.

The presence of outliers may considerably affect the performance of MD imputation methods. Different approaches have been proposed in the literature dealing simultaneously with missing data and outliers [22-25]. However, this is out of the scope of this paper and would deserve future research.

Based on the good performance of the regression-based methods in the model exploitation context, the main goal of this paper is to verify if this is also true in the model building context. In this paper we propose new methods for building PCA models with missing data by adapting PCA-ME methods to deal with the more general problem of PCA-MB, when the training set has missing values. The new adapted methods proposed here are PMP, KDR, KDR with principal component regression (PCR), KDR with partial least squares regression (PLS) and trimmed scores regression (TSR). They are all introduced in Section 2. In order to evaluate these new methods against the established ones (NIPALS, IA, DA and NLP), four data sets (two simulated and two actual ones) are employed to first generate missing values in an incremental manner. The criteria to assess the different methods are the mean squared prediction error (*MSPE*) of the reconstructed data set and the cosines between the actual principal components and the ones extracted by each method. This is illustrated in Section 3. Finally, Section 4 presents the conclusions of the study.

2. Methodology

Let $\mathbf{X}$ be an N by K matrix with missing values where $\mathbf{x}_i^{\mathrm{T}} = [x_{i1} \ \ldots \ x_{iK}]$ is the i-th row containing the information for observation i, and x_{ij} is the value of the j-th variable for observation i. We define the missing data indicator matrix $\mathbf{M}$ as the binary N by K matrix such that $m_{ij} = 1$ if x_{ij} is missing, and $m_{ij} = 0$ if x_{ij} is known. The matrix $\overline{\mathbf{M}}$ is the complement of $\mathbf{M}$, that is, $\overline{m}_{ij} = 1 - m_{ij}$. And finally, let $\mathbf{Z}$ denote the resulting $\mathbf{X}$ matrix after filling in the unknown values with zeroes, that is, $\mathbf{Z} = \overline{\mathbf{M}} \circ \mathbf{X}$, where $\circ$ is the Hadamard element wise-product.

A common procedure for building a PCA model from $\mathbf{X}$ is the known iterative algorithm (IA) [5] that consists of filling in the missing data with initial values (usually zeros, although other imputations such as the mean of the known values of the corresponding column or the mean of the corresponding rows and columns are also used), yielding a reconstructed data set from which a PCA model is fitted. By replacing the original missing data by their predictions from this PCA model, a new reconstructed

data set is obtained, and a new PCA model is fitted. This process is iterated until convergence of the predicted values for the missing data, as shown in Figure 1.

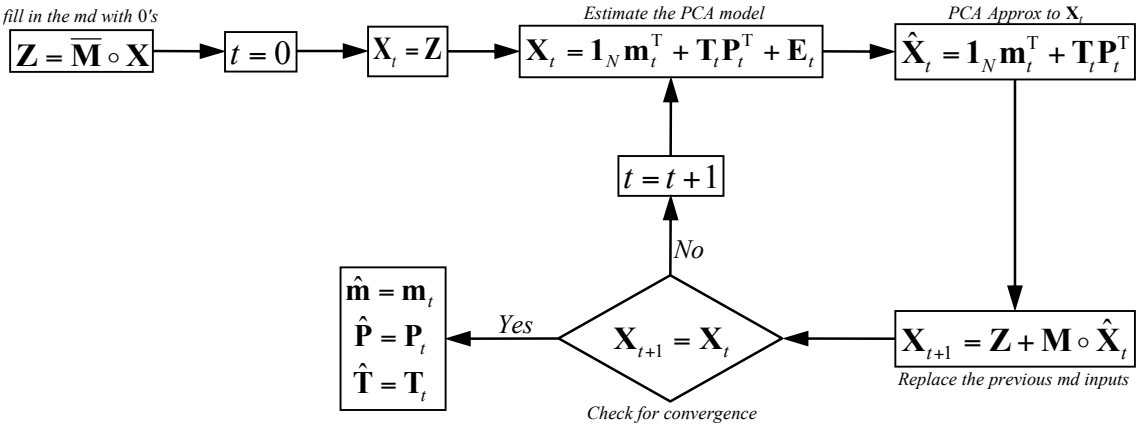

Figure 1. IA method in PCA-MB with missing data.

Walczak and Massart [5] introduce an adaptation of the IA for estimating the scores of new incomplete observation from a pre-built PCA model, fixed and known (*i.e.* model exploitation). This method has the following structure: i) fill in the missing positions of the new observation with an initial estimate; ii) predict the scores for the filled-in observation, using the loadings matrix $\mathbf{P}$ from the fitted PCA model; iii) re-estimate the missing values by employing the predicted scores and the loadings of the known PCA model; and iv) iterate until convergence.

Consider that the new observation $\mathbf{x}^{\mathrm{T}}$ has some unmeasured variables and these can be taken to be the first R elements of the row vector, without loss of generality. Thus, the vector can be partitioned as $\mathbf{x}^{\mathrm{T}} = [\mathbf{x}^{\#\mathrm{T}}\ \mathbf{x}^{*\mathrm{T}}]$, where $\mathbf{x}^{\#\mathrm{T}}$ denotes the missing measurements and $\mathbf{x}^{*\mathrm{T}}$ the observed variables. This induces the following partition in $\mathbf{X}$: $\mathbf{X} = [\mathbf{X}^{\#}\ \mathbf{X}^{*}]$ where $\mathbf{X}^{\#}$ is the submatrix containing the first R columns of $\mathbf{X}$ (corresponding to the variables that are missing in $\mathbf{x}^{\mathrm{T}}$), and $\mathbf{X}^{*}$ accommodates the remaining $K - R$ columns (corresponding to the observed variables in $\mathbf{x}^{\mathrm{T}}$).

Likewise, the loadings matrix $\mathbf{P}$ can be partitioned as $\mathbf{P} = \begin{bmatrix} \mathbf{P}^{\#} \\ \mathbf{P}^{*} \end{bmatrix}$, where $\mathbf{P}^{\#}$ is the submatrix made up of the first R rows of $\mathbf{P}$, and matrix $\mathbf{P}^{*}$ contains the remaining $K - R$ rows. These induced partitions are illustrated in Figure 2.

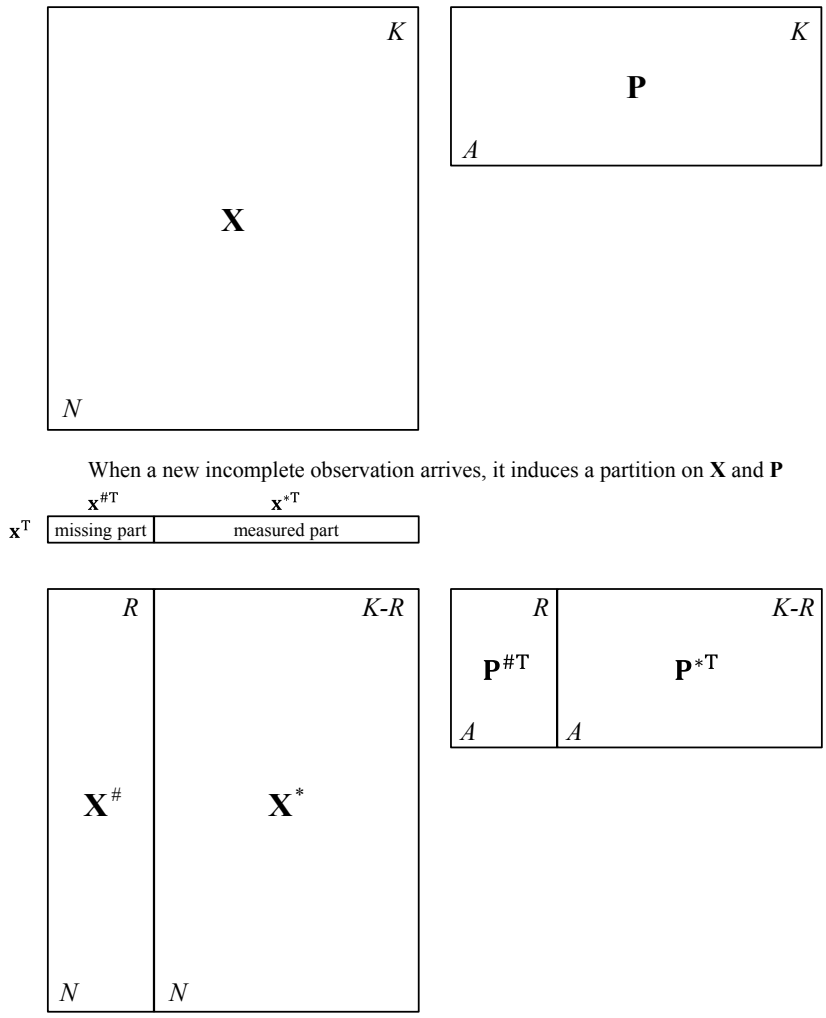

Figure 2. Data set partition induced by observation $\mathbf{x}^{\mathrm{T}}$.

Arteaga and Ferrer [6] show that, under general conditions, the IA adaptation from Walczack and Massart [5] is equivalent to the projection to the model plane (PMP) estimator, studied by Nelson *et al*. [4], that is, the least square estimator based on the observed variables: $\hat{\boldsymbol{\tau}} = (\mathbf{P}^{*\mathrm{T}}\mathbf{P}^{*})^{-1}\mathbf{P}^{*\mathrm{T}}\mathbf{x}^{*}$, where $\hat{\boldsymbol{\tau}}$ is the estimated vector of scores for observation $\mathbf{x}^{\mathrm{T}}$. Figure 3 shows a flow diagram of this IA adaptation. Note that the known part of the new individual, $\mathbf{x}^{*\mathrm{T}}$, is assumed to be centred with the mean of the corresponding columns of the **X** matrix, $\mathbf{X}^{*}$. Note also that in this adaptation of the IA the PCA model does not change, and thus loadings in matrix **P** are fixed.

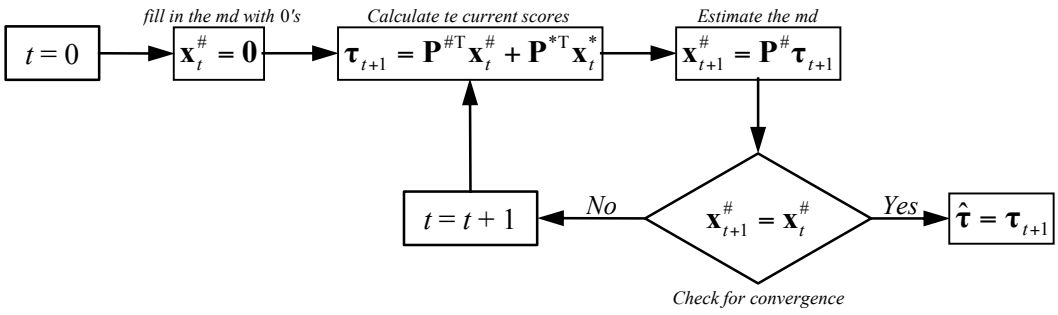

Figure 3. IA method in PCA-ME for a new incomplete observation.

From this equivalence we can state that, in model exploitation, the PMP estimator summarises the iterations of the IA method in one step.

Based on these results, in this paper we propose to adapt the IA method in PCA-MB from Figure 1, by replacing the prediction of missing values from the PCA model to that resulting when we treat each incomplete row in the data set as a new observation with missing values, and apply the PMP method for PCA-ME. This is illustrated in Figure 4. As was mentioned in the Introduction section, the PMP adaptation to PCA-MB uses the same regression coefficients for each incomplete observation as the t-EM algorithm [20]. Note that in this case the scores and loadings matrices **T** and **P**, respectively, change at each iteration.

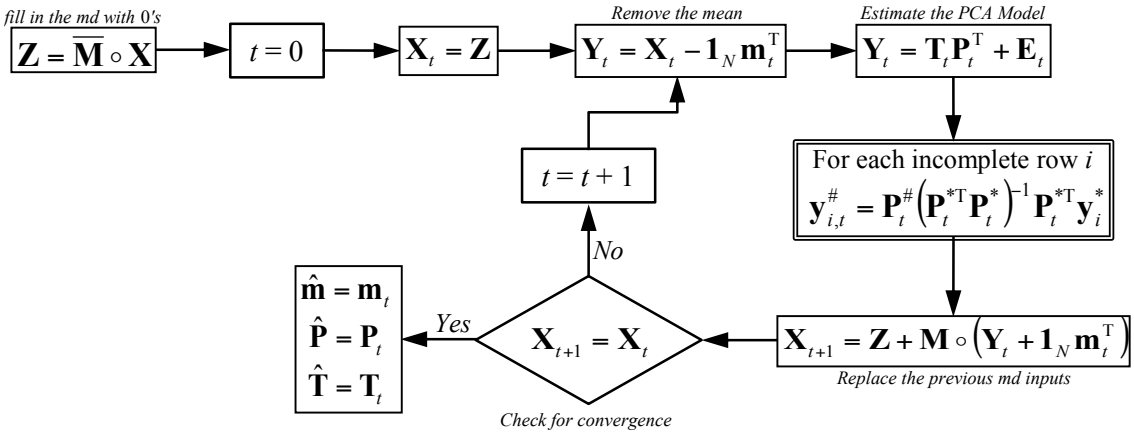

Figure 4. PMP method adapted for PCA-MB with missing data.

Arteaga and Ferrer [6] also present two new regression-based methods for estimating the scores of a new incomplete observation: the known data regression (KDR) method and the trimmed scores regression (TSR) method.

The KDR method in PCA-ME, when a new incomplete observation **x** is registered, consists of the following steps:

1. Fit the regression model $\mathbf{X}^{\#} = \mathbf{X}^{*}\mathbf{B} + \mathbf{U}$, yielding $\widehat{\mathbf{B}} = (\mathbf{X}^{*\mathrm{T}}\mathbf{X}^{*})^{-1}\mathbf{X}^{*\mathrm{T}}\mathbf{X}^{\#}$.

2. Estimate the missing part $\mathbf{x}^{\#\mathrm{T}}$ as $\hat{\mathbf{x}}^{\#} = \mathbf{X}^{\#\mathrm{T}}\mathbf{X}^{*}(\mathbf{X}^{*\mathrm{T}}\mathbf{X}^{*})^{-1}\mathbf{x}^{*}$, that is, $\hat{\mathbf{x}}^{\#} = \mathbf{S}^{\#*}(\mathbf{S}^{**})^{-1}\mathbf{x}^{*}$ [7], where $\mathbf{S}^{**}$ is the estimated covariance matrix of $\mathbf{X}^{*}$, $\mathbf{S}^{**} = \mathbf{X}^{*\mathrm{T}}\mathbf{X}^{*}/(N-1)$, and $\mathbf{S}^{\#*}$ is a R by $K-R$ matrix containing the estimated covariances of the combinations of columns of $\mathbf{X}^{\#}$ and columns of $\mathbf{X}^{*}$, $\mathbf{S}^{\#*} = \mathbf{X}^{\#\mathrm{T}}\mathbf{X}^{*}/(N-1)$.

In the same conditions, the TSR method can be summarized as:

1. Fit the regression model $\mathbf{X}^{\#} = (\mathbf{X}^{*}\mathbf{P}^{*})\mathbf{B} + \mathbf{U}$, where $\mathbf{X}^{*}\mathbf{P}^{*}$ is the trimmed scores matrix, *i.e.* the score matrix that corresponds only to the known variables and their associated loadings (note that $\mathbf{T} = \mathbf{XP} = \mathbf{X}^{*}\mathbf{P}^{*} + \mathbf{X}^{\#}\mathbf{P}^{\#}$), yielding $\widehat{\mathbf{B}} = (\mathbf{P}^{*\mathrm{T}}\mathbf{X}^{*\mathrm{T}}\mathbf{X}^{*}\mathbf{P}^{*})^{-1}\mathbf{P}^{*\mathrm{T}}\mathbf{X}^{*\mathrm{T}}\mathbf{X}^{\#}$.

2. Estimate the missing part $\mathbf{x}^{\#\mathrm{T}}$ as $\hat{\mathbf{x}}^{\#} = \mathbf{X}^{\#\mathrm{T}}\mathbf{X}^{*}\mathbf{P}^{*}(\mathbf{P}^{*\mathrm{T}}\mathbf{X}^{*\mathrm{T}}\mathbf{X}^{*}\mathbf{P}^{*})^{-1}\mathbf{P}^{*\mathrm{T}}\mathbf{x}^{*}$, that is, $\hat{\mathbf{x}}^{\#} = \mathbf{S}^{\#*}\mathbf{P}^{*}(\mathbf{P}^{*\mathrm{T}}\mathbf{S}^{**}\mathbf{P}^{*})^{-1}\mathbf{P}^{*\mathrm{T}}\mathbf{x}^{*}$.

Arteaga and Ferrer [7] show that TSR and KDR methods are particular cases of a general framework of methods derived from the generalized regression model $\mathbf{X}^{\#} = (\mathbf{X}^{*}\mathbf{L})\mathbf{B} + \mathbf{U}$, where the key matrix $\mathbf{L}$ takes different expressions depending on which method is applied. The key matrix for the KDR method is the identity matrix, $\mathbf{L} = \mathbf{I}_{K-R}$; in KDR with principal component regression (PCR), $\mathbf{L} = \mathbf{V}_{1:\rho}$, where $\mathbf{V}$ is the eigenvector matrix of $\mathbf{S}^{**}$ and $\rho \leq rank(\mathbf{S}^{**})$; in KDR with partial least squares (PLS), $\mathbf{L} = \mathbf{W}^{*}$, where $\mathbf{W}^{*}$ is the loadings matrix that allows writing the PLS scores $\mathbf{T}_{\mathrm{PLS}}$ as $\mathbf{T}_{\mathrm{PLS}} = \mathbf{X}^{*}\mathbf{W}^{*}$ in the PLS model for estimating $\mathbf{X}^{\#}$ from $\mathbf{X}^{*}$; finally, for the TSR method, $\mathbf{L} = \mathbf{P}^{*}$.

To adapt the PCA-ME framework methods for PCA-MB we only need to substitute, in the IA adaptation (Figure 4), the estimation of the missing part of each incomplete observation $\mathbf{y}_{i,t}^{\#} = \mathbf{P}_{t}^{\#}(\mathbf{P}_{t}^{*\mathrm{T}}\mathbf{P}_{t}^{*})^{-1}\mathbf{P}_{t}^{*\mathrm{T}}\mathbf{y}_{i}^{*}$ by the expression $\mathbf{y}_{i,t}^{\#} = \mathbf{S}_{t}^{\#*}\mathbf{L}_{t}(\mathbf{L}_{t}^{\mathrm{T}}\mathbf{S}_{t}^{**}\mathbf{L}_{t})^{-1}\mathbf{L}_{t}^{\mathrm{T}}\mathbf{y}_{i}^{*}$. This is illustrated in Figure 5.

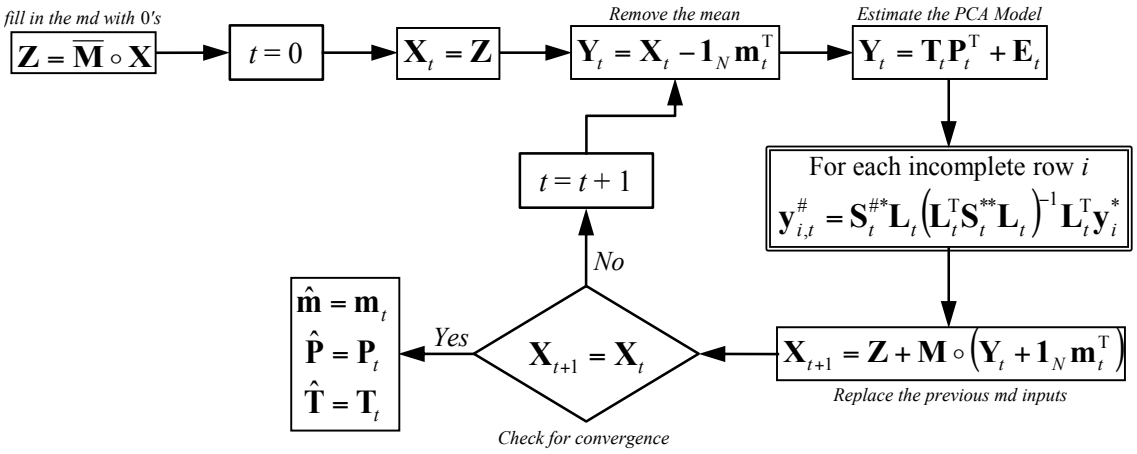

Figure 5. Regression-based framework adapted for PCA-MB with missing data.

Assuming data follows a multivariate normal distribution the adaptation of the KDR method results in the known EM algorithm [10-12]. The other adapted framework members (*i.e.* KDR with PCR, KDR with PLS, and TSR) are approximations to KDR

method that are useful when the covariance matrix $\mathbf{S}^{**}$ is ill-conditioned or singular, because the matrix $\mathbf{L}$ makes $\mathbf{L}^{\mathrm{T}}\mathbf{S}^{**}\mathbf{L}$ to have best conditioning properties than $\mathbf{S}^{**}$.

Note that in all the cited approximations, the key matrix $\mathbf{L}_t$ depends on the missing data combination. This implies that, at each iteration, two incomplete observations with different missing data combinations require two different $\mathbf{L}_t$ matrices. In KDR and TSR methods this is not a problem because in KDR $\mathbf{L}_t$ is the identity matrix, and in TSR $\mathbf{L}_t$ is $\mathbf{P}^{*}$, that is, a submatrix of $\mathbf{P}$. Nevertheless, in KDR with PCR a singular value decomposition for each missing data combination at each iteration is needed, and in KDR with PLS, a PLS regression for each missing data combination at each iteration has to be fitted. This causes PCR and PLS adaptations to be more computing time demanding than KDR and TSR adaptations. Nevertheless, as commented before, KDR may not be useful in practice due to ill-conditioning or singular problems of the covariance matrix $\mathbf{S}^{**}$.

The previously studied imputation methods impute a unique number for each missing value. These single imputation methods permit us to estimate the parameter's values, but ignore the variability of the estimates, leading to underestimation of standard errors and confidence intervals for the estimated parameters. That is, the single value being imputed cannot reflect the sampling variability around the actual value. Multiple imputation [12,27] overcomes this disadvantage. Multiple imputation, basically, creates several (M) values for each missing value representing a distribution capable of reflecting the sampling variability. Then, we have M complete data sets that we can analyse with the standard statistical techniques to estimate their parameters of interest. This allows calculation of variances of the parameters by combining the variability of estimates from within each imputed data set with the variability of the estimates across M imputed data sets. Rubin [26] shows how to combine both sources of variability in order to obtain confidence intervals for the estimated parameters.

Multiple imputation is based on three main assumptions: a probability model on complete data (observed and missing values), a prior distribution reflecting the uncertainty of the parameters for the imputation model, and that the missing data mechanism is ignorable (*i.e.* Missing at Random: MAR or Missing Completely at Random: MCAR).

Multiple imputation can be made in several manners, but the most popular is the data augmentation (DA) algorithm [13]. This is an iterative process that alternatively fills in the missing data and makes inferences about the unknown parameters, but unlike the

EM algorithm, this is made in a stochastic or random fashion. DA first performs a random imputation of missing data under assumed values of the parameters, and then draws new parameters from a Bayesian posterior distribution based on the observed and imputed data. DA starts with some value of the set of parameters $\boldsymbol{\Theta}$, usually these initial estimates are obtained with the EM algorithm, and in iteration *t* alternates between two steps:

I step (Imputation), draws $\mathbf{X}_t^{\#}$ from their conditional distribution given $\mathbf{X}^*$ and $\boldsymbol{\Theta}_{t-1}$.

P step (Posterior), draws $\boldsymbol{\Theta}_t$ from their posterior distribution given $\mathbf{X}^*$ and $\mathbf{X}_t^{\#}$.

where $\mathbf{X}_t^{\#}$ is the missing part of $\mathbf{X}$ in iteration *t*, and $\mathbf{X}^*$ is the known part of $\mathbf{X}$ that remains the same along iterations. The procedure of alternately simulating missing data and parameters creates a Markov chain that eventually stabilizes or converges in distribution [11]

In practice, the length of the Markov chain should be long enough to assure stability (convergence) and thus no dependency on initial values. The EM algorithm is recommended in practice to provide initial estimates of model parameters $\boldsymbol{\Theta}$ and number of iterations that DA algorithm needs to converge.

The EM algorithm converges to a single set of values and then the convergence can be easily assessed by checking the change in the parameter estimations from one iteration to the next. For DA, the algorithm converges to a probability distribution, not a single set of values. This makes it rather difficult to determine whether convergence has, in fact, been achieved [12]. The rate of convergence of the EM algorithm is a useful indication of the rate of convergence for DA. A good rule of thumb is that the number of iterations for DA should be at least as large as the number of iterations required for EM and then it is useful to run the EM algorithm before DA.

Finally, López-Negrete de la Fuente *et al.* [14] present a new approach for PCA-MB with incomplete data sets. This methodology solves a nonlinear programming problem (NLP) (see Equation 1) using the IPOPT solver [27].

$$\begin{aligned} Min \quad & \|\bar{\mathbf{M}} \circ (\mathbf{X} - \mathbf{T}\mathbf{P}^{\mathrm{T}})\|_F^2 \\ s.t. \quad & \begin{cases} \mathbf{p}_a^{\mathrm{T}}\mathbf{p}_b = \delta_{a,b} & a,b = 1,\ldots,A \\ \mathbf{t}_a^{\mathrm{T}}\mathbf{t}_b = 0 & a \neq b \; and \; a,b = 1,\ldots,A \\ \mathbf{1}^{\mathrm{T}}\mathbf{t}_a = 0 & a = 1,\ldots,A \end{cases} \end{aligned} \tag{1}$$

where $\delta_{a,b}$ is the Kronecker delta and *A* is the number of PCs extracted by the PCA model. The objective function minimises the squared error between the known values

and their estimations from the PCA model, subject to the constraints defined by the PCA assumptions: the loadings have to be orthonormal, and the scores have to be orthogonal with zero mean.

3. Comparative study

In this section we compare the performance of different methods for PCA-MB with missing data (MD). The methods under study are the standard methods for PCA-MB: NIPALS and IA, and the proposed PCA-ME adapted methods: PMP and the regression-based framework adapted methods (KDR, PCR, PLS and TSR).

In order to improve the convergence properties, in the implementation of the above methods we have used pseudoinverse to calculate $(\mathbf{P}_t^{*\mathrm{T}}\mathbf{P}_t^*)^{-1}$ in PMP, and $(\mathbf{L}_t^{\mathrm{T}}\mathbf{S}_t^{**}\mathbf{L}_t)^{-1}$ in the framework methods.

The MATLAB implementations for PMP, KDR, PCR, PLS and TSR have been developed for the present paper (see Appendix C). The codes for IA and DA have been reproduced from the original papers (see [5] for IA and [11,13] and Appendix C for DA). For the NLP method we used its implementation in the Phi toolbox (version 1.7) [14].

On the following subsections four data sets with missing values are used. Two of them are simulated and the other two are taken from the literature. The strategy to generate the MD is the same in all of them. Nine incremental levels of MD are considered in each data matrix (5%, 10%, 15%, 20%, 30%, 40%, 50%, 60% and 70%). 80% and 90% of missing values are also included in the last example. And for each data set and percentage, 50 possible data sets are simulated, following missing completely at random (MCAR) mechanism [2]. Therefore, 1900 data sets are analysed in this study: 4 data sets times 9 MD levels (11 for the last example) times 50 possible data sets. The original data sets can be found in Appendix B. Supplementary Material.

The principal performance criterion for each method is the mean squared prediction error *MSPE* (equation 2).

$$MSPE(Method) = \frac{\sum_{i=1}^{N}\sum_{j=1}^{K}\left(\hat{x}_{ij} - \hat{x}_{ij}^{Method}\right)^2}{NK} \tag{2}$$

where $\hat{x}_{ij}$ is the predicted value for the *j*-th variable of the *i*-th observation in the prediction matrix $\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}}$ obtained from the complete data set; and $\hat{x}_{ij}^{Method}$ the

analogous prediction obtained after applying the corresponding method on the incomplete data set.

In order to assess whether the differences among methods, in terms of *MSPE*, are statistically significant, a mixed-effects ANOVA model is fitted per each case study. The factors considered are the method, the percentage of missing values, and the simulated data set, being the latter nested to the percentage factor. Method and percentage are fixed-effects factors; the data set is a random-effect factor. Given the positive skewness of *MSPE*, a logarithmic transformation is used. This transformation also expands the differences for low percentages of MD, easing the visualization of the plots. In case any effect or interaction is statistically significant, the 95% LSD (least significance difference) intervals are calculated to assess which groups are different from others.

In order to understand the degradation in the PCA model due to missing values the cosine between each loading vector obtained using the full data matrix and its corresponding from the incomplete data set is calculated. The closer to one it is, the more similar are both loadings for a particular component. However, if more than one PC is extracted from data, the cosines of further PCs are being strongly influenced by the previous ones, since they have to be orthogonal to the estimated first PC.

3.1 Olive Oil data set

This data set consists of the percentage composition of eight fatty acids: palmitic, palmitoleic, stearic, oleic, linoleic, linolenic, arachidic and eicosenoic, found in the lipid fraction of 572 Italian olive oils. In the original data set [28] there are nine collection areas from three different regions of Italy. In order to reduce the computation time, our data set is built with 75 randomly chosen wines from South Apulia. One PC is extracted from these data, explaining 59% of variance.

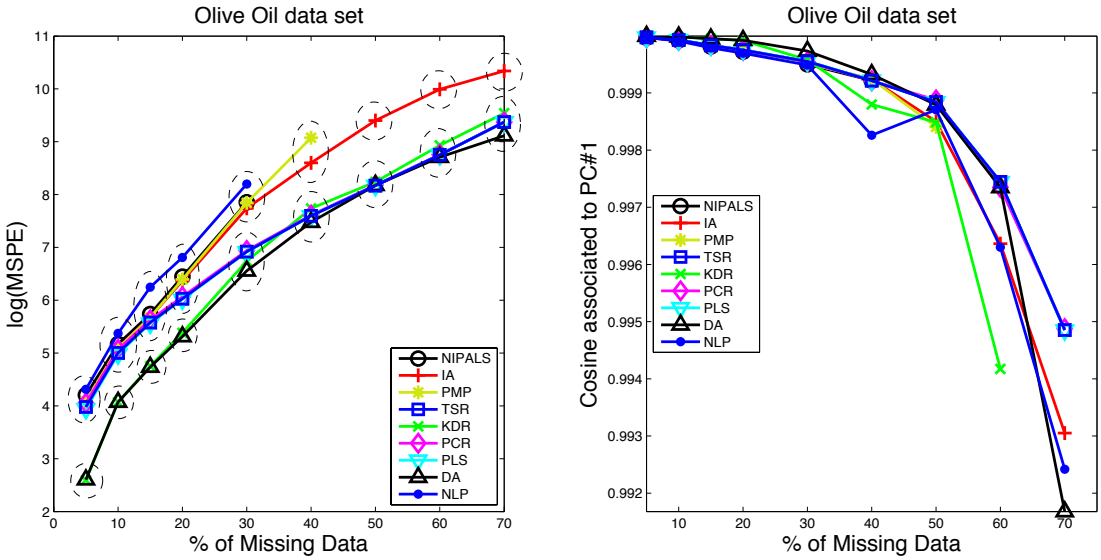

Figure 6. log(*MSPE)* (left) and cosines associated to the first PC (right) for the reconstructed Olive Oil data set from a PCA model with different methods: NIPALS, IA, PMP, TSR, KDR, PCR, PLS, DA and NLP. The missing values in the figure correspond to NaNs (which implies convergence problems for the method). Some other higher values are not shown (especially for the highest %) in order to appreciate the differences among the most accurate methods. Dashed ellipses mark the statistically significant differences between groups of methods, *i.e.* the differences exist between-groups, not within-groups.

As shown in Figure 6, DA and KDR are statistically superior to all other methods from 5-20% of missing data. There exist no significant differences between DA, KDR, PCR, PLS and TSR for further percentages. From 30% of MD onwards, NLP, PMP, IA and NIPALS perform statistically worse than the other methods. From 40% upwards NLP is unstable for some combinations of missing values, *i.e.* some missing values are poorly imputed and therefore the *MSPE* value and the cosine are strongly affected by them. NIPALS and PMP are unable to converge for high percentages of missing values (60-70%). Some *MSPE*s are not shown on Figure 6 up to some percentage, *e.g.* NLP, in order to appreciate the differences among the most accurate methods. Figure 6 also shows that the degradation in cosine associated to the first PC matches the increment in *MSPE*.

3.2 Diesel data set

This data set is obtained from the Eigenvector Research Inc. data library (http://www.eigenvector.com/data/index.html). The data set contains the NIR spectra of several diesel fuels (N = 40) obtained at the Southwest Research Institute (SWRI) on a project sponsored by the U.S. Army [29]. The fuels were originally scanned from the

wavelength 750 nm to 1550 nm in 2 nm increments (K = 401 variables). Two PCs are extracted explaining 60% and 25% of variance in data, respectively.

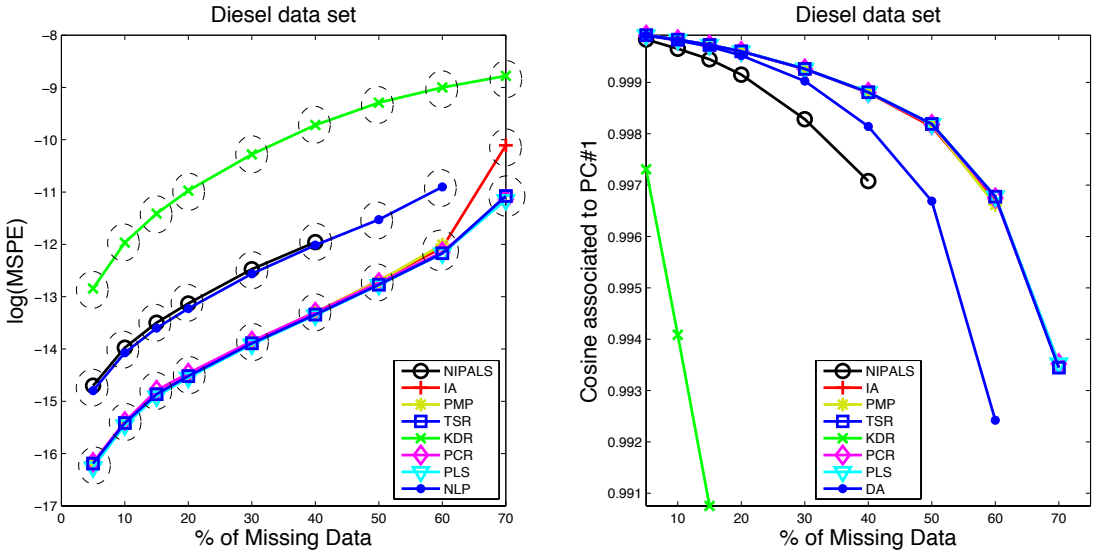

Figure 7. log(*MSPE)* (left) and cosines associated to the first PC (right) for each method in the Diesel data set. DA is not applicable in this data set. See Figure 6 caption for more details.

DA is the only method that is unable to analyse the present data set, regardless the percentage of missing values. The main reason is the singularity of the $\mathbf{S}^{**}$ matrix for the different combinations of missing values. This also affects the KDR method, which is statistically the worst in terms of *MSPE* (see Figure 7).

NLP and NIPALS offer better results than KDR, but are statistically worse than the other methods for all percentages of missing values. NIPALS does not converge with 70% of MD, and neither does NLP. TSR, PCR, PLS, PMP and IA show the best performances for 10-60% of missing data. For 70% of MD IA and PMP are statistically worse than the previous regression-based methods. These results are coherent with the degradation of the cosine of the first and second loadings (see Figure 7). The cosines of the second PC are included in Appendix B.

3.3 Simulated data set

A three-component multivariate data set is generated [30,31] to compare the performance of the different methods. This data set has ten variables (K = 10) and a hundred samples (N = 100), and follows a multivariate normal distribution with zero means and unit variances. We forced the highest eigenvalue of the correlation matrix to be 4.0, the second one 3.0 and the last one 2.0, explaining 40%, 30% and 20% of the variance from the ten variables, respectively.

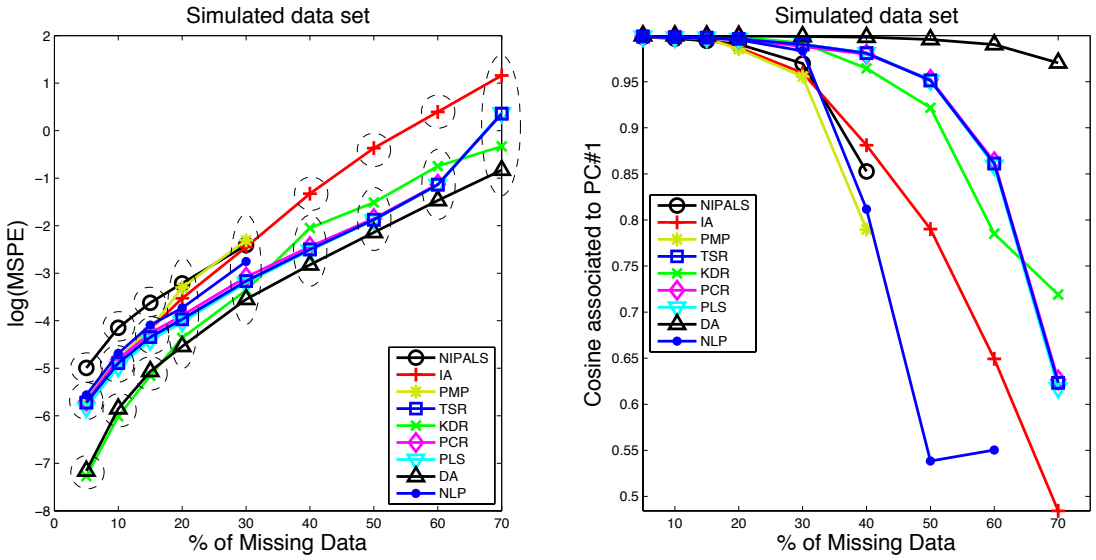

Figure 8. log(*MSPE)* (left) and cosines associated to the first PC (right) for each method in the simulated data set. See Figure 6 caption for more details.

In the simulated data set, PMP, NLP and NIPALS have again problems with the convergence (see Figure 8), the first one does not converge with 40% of MD, the second one with 50%, and the last one with 60%. Now, at early stages, there exist statistically significant differences between methods. KDR, jointly with DA, is statistically superior to all other methods with 5%-15% of missing values, being NIPALS the worst method. For medium and high percentages of MD, DA and KDR are not significantly superior to PCR, PLS and TSR. The performance of IA in this data set is coherent with the previous data sets, *i.e.* with low percentages of missing data its performance is similar to the regression-based methods, for higher percentages IA performs significantly worse. Again, the cosine degradations (Figure 8) agree with the results observed in *MSPE*. The cosines of the second-third PCs are included in Appendix B.

A comment regarding the cosine values is here in due. A value of 0.9 implies a deviation of 25 degrees between the imputed and the actual PC, which is, in fact, a huge rotation of the basis, of the PCs. However, even when the cosines are below 0.9, the imputations of the best methods are still useful, based on their *MSPE* values.

3.4 Big data set

An additional data set, from the big data perspective [32], is analysed here. Using again [30,31] a simulated data set is generated with 10 million data entries (100000 observations x 100 variables). 4 PCs are used to generate the data, explaining 35%, 25%, 15% and 15% of the variance, respectively. In this dataset we decided to force the

methods to deal with even more missing values than in the previous datasets. Here, 80% and 90% of missing values are also simulated.

Based on the results of the previous subsections we decided not to apply in this dataset the methods that have problems with convergence and/or instability, such as NIPALS, PMP and NLP. IA is applied here, since it is a fast imputation method and showed no problems with convergence in the previous data sets. DA was initially applied but since this method, jointly with KDR, PCR and PLS, are more time consuming than the rest of methods the imputation was not obtained in a reasonable time period. Therefore, only IA and TSR are applied to this data set.

The *MSPE* results and the 1st PC cosines are depicted in Figure 9. The cosines of the 2nd-4th PCs are included in Appendix B. We found statistically significant differences between both methods for all percentages. TSR offers a lower *MSPE* than IA for all simulations; however, this cannot be appreciated in Figure 9. It is interesting that the results with this huge data set are better than in the previous simulated one. This is due to the more individuals has the dataset, the more accurate are the estimations of the covariance matrices that TSR and IA perform internally, which implies that the estimation is more coherent that with less observations. It is worth noting that the imputation in the most extreme case, 90% of missing data, is indeed difficult, taking around an hour.

As stated at the beginning of Section 3, the deviations of the last PCs, with respect to the true ones, are strongly affected by the deviations of the first ones. This is clearly seen in the cosines of 3rd-4th PCs (see Appendix B). However, the imputations performed by both IA and TSR show low *MSPEs* since the deviations of 1st-2nd PCs, the ones explaining more variance, are small.

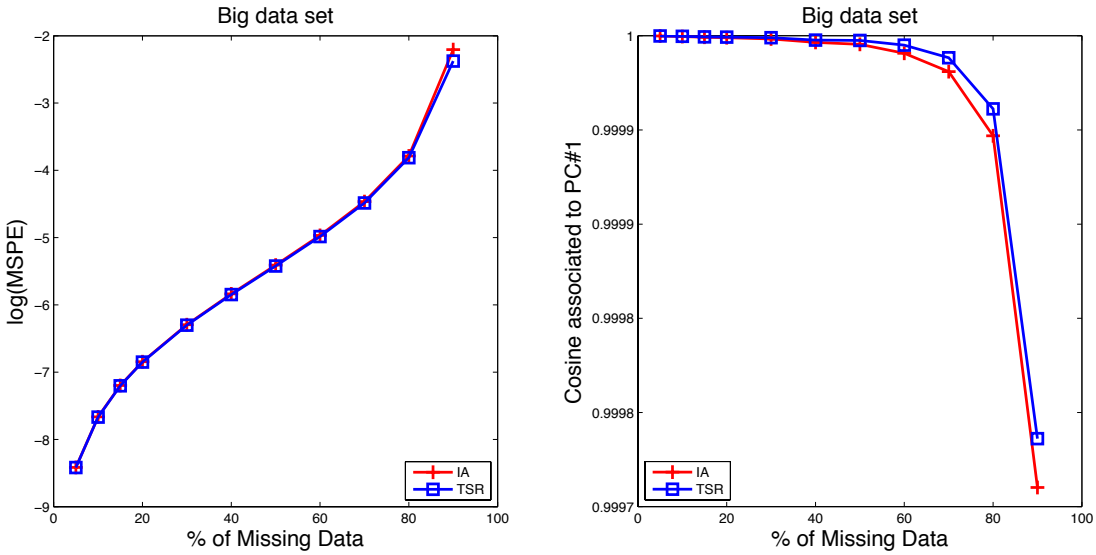

Figure 9. log(*MSPE)* (left) and cosines associated to the first PC (right) for each method in the Big simulated data set. See Figure 6 caption for more details.

4. Discussion and conclusions

Trimmed Scores Regression (TSR) method performed extraordinarily well in all the data structures and missing data percentages analysed throughout this paper. This missing data imputation method, adapted here from the PCA model exploitation context to model building, represents the best compromise solution among prediction quality, robustness against data structure and computation time. From the other regression-based methods adapted here, the known data regression methods with PCR and PLS offer also good solutions, however, they are more time-consuming, since they fit additional PCR and PLS models.

From the rest of the methods analysed here, Data Augmentation (DA) and KDR have excellent performances with thin data sets (*i.e.* more observations than variables). Nevertheless, they have two important drawbacks. The first one is that both methods, especially DA, since it is a multiple imputation method, are strongly more time consuming than, *e.g.* TSR. The second drawback is that with fat data sets (*i.e.* more variables than observations, typical in batch processes or with spectral data) DA is unfeasible, and KDR has the worst performance among the rest of the methods.

The NIPALS method for missing data imputation, a procedure implemented in many commercial statistical packages such as ProMV [33], SIMCA-P [34] and PLS Toolbox [35], is unable to deal with most of the missing data scenarios analysed in this study. Regarding the rest of the methods, PMP and NLP have also convergence problems

when high percentages of missing data are generated in all datasets. IA is the only method, jointly with TSR, applied to the big data set, due to its fast performance in the previous data sets and its robustness against high percentages of missing data. However, its performance level in all four data sets is statistically worse than TSR's.

Two key points in PCA model building with missing data are not addressed here. The first one is how the possible outliers are affecting the imputation performed by the missing data methods analysed here. The second one is the subtle issue of choosing of the appropriate number of principal components when the data set has missing values. These two points deserve future research.

Acknowledgements

Research in this study was partially supported by the Spanish Ministry of Science and Innovation and FEDER funds from the European Union through grant DPI2011- 28112-C04-0, and the Spanish Ministry of Economy and Competitiveness through grant ECO2013-43353-R. The authors gratefully acknowledge Salvador García-Muñoz for providing the Phi toolbox (version 1.7) to perform the nonlinear programming approach (NLP) method.

Appendix A. Methods equivalences.

In this section, the equivalence between some methods described in [15] and the ones presented here are proven. The general iterative principal components imputation (GIP) [15,17] is equivalent to the Iterative Algorithm (IA) [5] extracting one principal component. The GIP steps are the following:

1. A missing indicator matrix **R** is defined as $r_{ij} = 1$ if x_{ij} is observed, and 0 otherwise.
2. An initial guess for the missing data is imputed.
3. The correlation matrix **C** is obtained using the available data, and then the largest eigenvalue λ, and its associated eigenvector **v**, is obtained. In this way, the first principal component score for sample i is $\tau_i = \sum_{j=1}^{K} v_j x_{ij}$, where K is the number of variables in the data set.
4. The missing elements are replaced by their projection using the score, *i.e.* if $r_{ij} = 0$, then $\hat{x}_{ij} = \tau_i v_j$.

5. Steps 3-4 are repeated until the consecutive imputed values are within the specified tolerance.

The indicator matrix **R** is exactly the matrix $\overline{\mathbf{M}}$, used in IA (see Figure 1). The principal component scores are calculated in the same way as the estimation of the PCA model in IA (box 4 in Figure 1). And the projection for the missing values is basically the PCA approximation to $\mathbf{X}_t$ matrix (box 5 in Figure 1). Finally, the process is iterated using the same condition as in IA. Therefore, GIP method is mathematically equivalent to IA when one principal component is extracted.

The main similarity between the so-called multiple imputation by chained equation method (MICE) defined in [15,18] and KDR is that the former works variable-wise and the latter observation-wise. The steps of MICE algorithm are detailed here:

1. Initial guesses for all missing elements are provided.
2. For each variable with missing elements, $\mathbf{x}_j$, the data are split into two sub-vectors: $\mathbf{x}_{ja}$ a sub-vector that contains all available data, and $\mathbf{x}_{jm}$ a sub-vector that contains all missing data. The available sub-vector $\mathbf{x}_{ja}$ is regressed on all other variables, which are restricted to the samples in $\mathbf{x}_{ja}$; that is $\mathbf{x}_{ja} = f(\mathbf{x}_1, \ldots, \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, \ldots, \mathbf{x}_K)$.
3. The missing sub-vector $\mathbf{x}_{jm}$ is then predicted from the regression and its missing entries are replaced with the predictions from the regression. The regression procedure is repeated for all variables with missing elements.
4. After all missing elements are imputed, the regressions and predictions are repeated until consecutive iterates are within the specified tolerance for each of the imputed values.

In MICE the regression model is performed within each column, predicting the observations with the missing values from the observations with available data. KDR follows the same algorithm as MICE, but the data is split based on the missing and available values of an observation (see Figure 5, taking **L** as the identity matrix). Then, the calibration is performed between the submatrix of **X** corresponding to the missing elements in row *i*, $\mathbf{X}^{\#}$, and the submatrix of available measurements, $\mathbf{X}^{*}$, following the expression $\mathbf{X}^{\#} = \mathbf{X}^{\#}\mathbf{B} + \mathbf{U}$, where **B** is the regression coefficients matrix and **U** is the residual matrix. The prediction step is also performed observation-wise: using the model between the missing and the available submatrices, the missing elements in row *i* are predicted based on its available measurements (see Figure 5).

Finally, the equivalence between the regularised t-EM method [15,20] and PMP [4] is drawn. t-EM algorithm is defined as follows:

1. Estimate the covariance matrix, $\widehat{\mathbf{\Sigma}}$.
2. Calculate the singular value decomposition of the covariance matrix $\mathbf{\Sigma} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\mathrm{T}}$.
3. Build the regression model $\hat{\mathbf{x}}_m = \widehat{\boldsymbol{\mu}}_m + (\mathbf{x}_a - \widehat{\boldsymbol{\mu}}_a)\mathbf{B}$, being $\mathbf{B} = \mathbf{V}_{aq}(\mathbf{V}_{aq}^T\mathbf{V}_{aq})^{-1}\mathbf{V}_{mq}$, where the row vectors $\hat{\mathbf{x}}_m$, $\widehat{\boldsymbol{\mu}}_m$, $\mathbf{x}_a$ and $\widehat{\boldsymbol{\mu}}_{\mathrm{a}}$ are the estimated missing part of row **x**, the estimated mean of the missing part, the available measurements of **x** and its estimated mean vector, respectively.
4. Iterate the process until convergence.

Since in Figure 4, the rows of matrix **X** are represented by columns, the row-wise representation of the box 6 is $\mathbf{y}_i^{\#\,\mathrm{T}} = \mathbf{y}_i^{*\,\mathrm{T}}\mathbf{P}^*(\mathbf{P}^{*\,\mathrm{T}}\mathbf{P}^*)^{-1}\mathbf{P}^{\#\,\mathrm{T}}$. This equation, bearing in mind that the data in Figure 4 is previously mean-centred, is exactly the same as Step 3 in t-EM, being $\mathbf{P}^*$ the first q loadings (significant ones) of the available values of the covariance matrix (which are the same as in **X** matrix) and $\mathbf{P}^{\#}$ the first q loadings corresponding to the variables of the missing part.

Appendix B. Supplementary material.

The cosines of the 2nd PC of the diesel data set (Additional Figure 1), the cosines of the 2nd-3rd PCs of the small simulated data set (Additional Figure 2) and the cosines of the 2nd-4th PCs of the Big data set (Additional Figure 3) are available online. An Excel file (Additional File 1) with the original data sets is also included as additional material. The Big data set can be downloaded from http://mseg.webs.upv.es/Software_e.html.

Appendix C. List of MATLAB codes for TSR and DA.

TSR and DA source codes are provided here. The first method represents the best compromise solution between prediction quality, robustness against data structure and computation time. KDR, KDR with PCR and KDR with PLS can be easily computed using TSR source code by modifying the key matrix L (as shown in Methods section). DA source code is also included.

```
% TSR SOURCE CODE:
function [X,m,S,It,Xrec]=pcambtsr(X,A)
%
% Inputs:
```

```
% X: data matrix with NaNs for the missing data.
% A: number of principal components.
%
% Outputs:
% X: original data set with the imputed values.
% m: estimated mean vector of X.
% S: estimated covariance matrix of X.
% It: number of iterations.
% Xrec: PCA reconstruction of X with A components.
%
[n,p]=size(X);
for i=n:-1:1,
  r=~isnan(X(i,:));
  pat(i).O=find(r==1); % observed variables
  pat(i).M=find(r==0); % missing variables
  pat(i).nO=size(pat(i).O,2); % number of observed variables
  pat(i).nM=size(pat(i).M,2); % number of missing variables
end
mis=isnan(X);
[r c]=find(isnan(X));
X(mis)=0;
meanc=sum(X)./(n-sum(mis));
for k=1:length(r),
  X(r(k),c(k))=meanc(c(k));
end
maxiter=5000;
conv=1.0e-10;
diff=100;
It=0;
while It<maxiter & diff>conv,
  It=It+1;
  Xmis=X(mis);
  mX=mean(X);
  S=cov(X);
  Xc=X-ones(n,1)*mX;
  if n>p, [U D V]=svd(Xc,0); else [V D U]=svd(Xc',0); end
  V=V(:,1:A);
  for i=1:n,            % for each row
    if pat(i).nM>0,     % if there are missing values
      L=V(pat(i).O,1:min(A,pat(i).nO)); % L is the key matrix
      S11=S(pat(i).O,pat(i).O);
      S21=S(pat(i).M,pat(i).O);
      z1=Xc(i,pat(i).O)';
```

```
      z2=S21*L*pinv(L'*S11*L)*L'*z1;
      Xc(i,pat(i).M)=z2';
    end
  end
  X=Xc+ones(n,1)*mX;
  d=(X(mis)-Xmis).^2;
  diff=mean(d);
end
S=cov(X);
m=mean(X);
[u d v]=svd(S,0);
P=v(:,1:A);
T=(X-ones(n,1)*m)*P;
Xrec=ones(n,1)*m+T*P'


% DA SOURCE CODE
function [DAm,DAS,mest,Sest,Y]=DataAugmentation(X,M,CL)
%
% Inputs:
% X: data matrix with NaN for the missing data.
% M: number of independent chains (we use M=10).
% CL: length of each chain (we use CL=100).
%
% Outputs:
% DAm:  estimated means for the M chains in M rows.
% DAS:  estimated covariance matrices for the M chains DAS(1).co,...,
DAS(M).co.
% mest: estimated means (mest=mean(DAm)).
%  Sest:  estimated  covariance  matrix  (averaging  DAS(1).co,...,
DAS(M).co).
% Y: original data set with the imputed values.
%
[n,p]=size(X);
for i=n:-1:1,
  r=~isnan(X(i,:));
  pat(i).O=find(r==1); % observed variables
  pat(i).M=find(r==0); % missing variables
  pat(i).nO=size(pat(i).O,2); % number of observed variables
  pat(i).nM=size(pat(i).M,2); % number of missing variables
end
mis=isnan(X);           % mis are the positions of the md in X
[r c]=find(isnan(X));   % r and c store the row and the column for all
the md
```

```
X(mis)=0;               % fill in the md with 0's
meanc=sum(X)./(n-sum(mis)); % average of the known values for each
column
for k=1:length(r),
  X(r(k),c(k))=meanc(c(k)); % fill in each md with the mean of its
column
end
mini=mean(X); Sini=cov(X);   % initial mean vector an covariance
matrix
DAm=zeros(M,p);
for run=1:M,
  S=Sini;
  m=mini;
  for It=1:CL,
    for i=1:n,              % for each row
      if pat(i).nM>0,           % if there are missing values
        m1=m(1,pat(i).O)';          % nOx1
        m2=m(1,pat(i).M)';          % nMx1
        S11=S(pat(i).O,pat(i).O);   % nOxnO
        S12=S(pat(i).O,pat(i).M);   % nOx1
        z1=X(i,pat(i).O)';          % nOx1
        z2=m2+S12'*pinv(S11)*(z1-m1);% nMx1
        X(i,pat(i).M)=z2';          % 1xp
      end
    end
    m=mean(X);
    S=cov(X);
    [m,S]=DrawPost(m,S,10*n*n);
  end
  DAm(run,:)=m;
  DAS(run).co=S;
end
mest=mean(DAm);
Sest=zeros(p,p);
for k=1:M,
  Sest=Sest+DAS(k).co;
end
Sest=Sest/M;
% Applies stochastic regression with the posterior of the mean (mest)
% and the covariance matrix (Sest)
for i=1:n,            % for each row
  if pat(i).nM>0,     % if there are missing values
    m1=mest(1,pat(i).O)';          % nOx1
    m2=mest(1,pat(i).M)';          % nMx1
```

```
    S11=Sest(pat(i).O,pat(i).O);   % nOxnO
    S12=Sest(pat(i).O,pat(i).M);   % nOxnM
    z1=X(i,pat(i).O)';             % nOx1
    z2=m2+S12'*pinv(S11)*(z1-m1);  %nMx1
    X(i,pat(i).M)=z2';  % fill in the md positions of row i
  end
end
Y=X;

function [mpost,Spost]=DrawPost(m,S,n)
d=chol(S/n);
p=size(S,1);
if n<=81+p,
  x=randn(n-1,p)*d;
else
  a=diag(sqrt(chi2rnd(n-(0:p-1))));
  for i=1:p-1,
    for j=i+1:p,
      a(i,j)=randn(1,1);
    end
  end
  x=a*d;
end
Spost=x'*x;
mpost=(m'+chol(Spost/(n-1))*randn(size(S,1),1))';
```

References

[1] B. Grung, R. Manne, Missing Values in Principal Component Analysis. Chemometrics and Intelligent Laboratory Systems 42 (1998) 125–139.

[2] F. Arteaga, A. Ferrer, Missing Data. In: S. Brown, R. Tauler, B. Walczak (eds.) Comprehensive Chemometrics volume 3, Oxford: Elsevier, 2009, pp. 285-314

[3] B.M. Wise, N.L. Ricker. Recent Advances in Multivariate Statistical Process Control: Improving Robustness and Sensitivity, Proceedings of IFAC International Symposium, ADCHEM'91, Toulouse, France, 1991, 125-130.

[4] P.R.C. Nelson, P.A. Taylor, J.F. MacGregor, Missing data methods in PCA and PLS: Score calculations with incomplete observations, Chemometrics and Intelligent Laboratory Systems 35 (1996) 45-65.

[5] B. Walczak, D.L. Massart, Dealing with missing data Part I, Chemometrics and Intelligent Laboratory Systems 58 (2001) 15-27.

[6] F. Arteaga, A. Ferrer, Dealing with missing data in MSPC: several methods, different interpretations, some examples, Journal of Chemometrics 16 (2002) 408-418.

[7] F. Arteaga, A. Ferrer, Framework for regression-based missing data imputation methods in on-line MSPC, Journal of Chemometrics 19 (2005) 439-447.

[8] S. Wold, C. Albano, W.J. Dunn, K. Esbensen, S. Hellberg, E. Johansson, M. Sjöström. In H. Martens and H. Russwurm, Jr. (Editors), Food Research and Data Analysis, Applied Science Publishers, London, 1983, pp. 183-185.

[9] P.R.C. Nelson: Treatment of missing measurements in PCA and PLS models. Ph.D. Dissertation. Department of Chemical Engineering, McMaster University. Hamilton, Ontario, Canada 2002.

[10] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm (with discussion), Journal of the Royal Statistical Society Series B 39 (1977) 1-38.

[11] J.L. Schafer, Analysis of Incomplete Multivariate Data, CRC Press, New York, 1997.

[12] P.D. Allison, Missing Data, Sage, Thousand Oaks, 2001.

[13] M.A. Tanner, W.H. Wong, The calculation of posterior distribution by data augmentation (with discussion), Journal of the American Statistical Association 82 (1987) 528-550.

[14] R. López-Negrete de la Fuente, S. García-Muñoz, L.T. Biegler, An efficient nonlinear programming strategy for PCA models with incomplete data sets, Journal of Chemometrics 24 (2010) 301–311.

[15] Y. Liu, S.D. Brown, Comparison of five iterative imputation methods for multivariate clasification, Chemometrics and Intelligent Laboratory Systems 120 (2013) 106-115.

[16] W.J. Krzanowski, Missing value imputation in multivariate data using the singular value decomposition matrix, Biometrical Letters 25 (1988) 31-39.

[17] R.E. Dear, A principal-component missing-data method for multiple regression models, reprint, System Development Corp., Santa Monica, CA, 1959.

[18] I.R. White, P. Royston, A.M. Wood, Multiple imputation using chained equations: issues and guidance for practice, Statistics in Medicine 30 (2011) 377–399.

[19] T. Schneider, Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values, Journal of Climate 14 (2001) 853-871.

[20] R.D. Fierro, G.H. Golub, P.C. Hansen, D.P. O'Leary, Regularization by truncated total least squares, SIAM Journal on Scientific Computing 18 (1997) 1223–1241.

[21] M.P. Gómez-Carracedo, J.M. Andrade, P. López-Mahía, S. Muniategui, D. Prada, A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets, Chemometrics and Intelligent Laboratory Systems 134 (2014) 23-33.

[22] J. Karhunen, Robust PCA methods for complete and missing data, Neural Network World 5(11) (2011) 357-392.

[23] H. Xu, C. Caramanis, S. Sanghavi, Robust PCA via Outlier Pursuit, IEEE Transactions on Information Theory 58(5) (2012) 3047-3064.

[24] I. Stanimirova, M. Daszykowski, B. Walczak, Dealing with missing values and outliers in principal component analysis, Talanta 72 (2007) 172-178.

[25] S. Seernels, T. Verdonck, Principal Component Analysis for data containing outliers and missing elements, Computational Statistics and Data Analysis 52(3) (2008) 1712-1727.

[26] D.B. Rubin, Multiple Imputation for Nonresponse in Surveys, Wiley & Sons, 1987.

[27] A. Wächter, L.T. Biegler, On the implementation of a primal-dual interior point filter line search algorithm for large-scale nonlinear programming, Mathematical Programming 106(1) (2006) 25–57.

[28] M. Forina, C. Armanino, S. Lanteri, E. Tiscornia, Classification of Olive Oils from their Fatty Acid Composition, in H. Martens, H.Jr Russwurm Eds, Food Research and Data Analysis, Applied Science Pub, London, 1983, pp. 189-214

[29] S.A. Hutzler and G.B. Bessee, Remote Near-Infrared Fuel Monitoring System, Interim Report, U.S. Army TARDEC Fuels and Lubricants Research Facility, Southwest Research Institute, San Antonio, United States, 1997.

[30] F. Arteaga, A. Ferrer, How to simulate normal data sets with the desired correlation structure, Chemometrics and Intelligent Laboratory Systems 101 (2010) 38-42.

[31] F. Arteaga, A. Ferrer, Building covariance matrices with the desired structure, Chemometrics and Intelligent Laboratory Systems 127 (2013) 80-88.

[32] J. Camacho, Visualizing Big Data with Compressed Score Plots: Approach and research challenges, Chemometrics and Intelligent Laboratory Systems 135 (2014) 110-125.

[33] ProSensus MultiVariate, ProSensus Inc. (http://www.prosensus.ca) (2015)

[34] SIMCA-P, Umetrics (http://www.umetrics.com) (2015)

[35] PLS Toolbox, Eigenvector Research Inc. (http://www.eigenvector.com) (2015)