Assessment of maximum likelihood PCA missing data imputation

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

[1]Dep. de Estadística e Investigación Operativa Aplicadas y Calidad, Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain.

[2]Dep. of Biostatistics and Investigation, Universidad Católica de Valencia San Vicente Mártir, C/Quevedo 2. 46001 Valencia, Spain.

Maximum likelihood principal component analysis (MLPCA) was originally proposed to incorporate measurement error variance information in principal component analysis (PCA) models. MLPCA can be used to fit PCA models in the presence of missing data, simply by assigning very large variances to the non-measured values. An assessment of maximum likelihood missing data imputation is performed in this paper, analysing the algorithm of MLPCA and adapting several methods for PCA model building with missing data to its maximum likelihood version. In this way, known data regression (KDR), KDR with principal component regression (PCR), KDR with partial least squares regression (PLS), and trimmed scores regression (TSR) methods are implemented within the MLPCA method to work as different imputation steps. Six data sets are analysed using several percentages of missing data, comparing the performance of the original algorithm, and its adapted regression-based methods, with other state-of-the-art methods.

Keywords: maximum likelihood principal component analysis, missing data, regression-based methods, PCA model building, trimmed scores regression

1. INTRODUCTION

Principal component analysis[1] (PCA) is one of the most applied methods for data understanding. The original variables are projected onto the latent space, where data most vary, and a new set of

[*] Correspondence to: A. Folch-Fortuny (abfolfor@upv.es)

uncorrelated variables are obtained, the principal components (PCs), summarizing the most relevant features of data. Wentzell *et al.*[2] proposed in 1997 a new PCA approach, based on maximum likelihood fitting, called maximum likelihood PCA (MLPCA). This methodology allows incorporating information about the measurement errors in the model. MLPCA has been widely applied in several works within chemistry and biology, *e.g.* to analyse reflectance Fourier transformed infrared (FTIR) microspectroscopic data[3] and ion mass spectroscopic data[4], to fault detection in process industry[5], to the characterization of measurement errors in nuclear magnetic resonance (NMR) data[6] and gene expression data[7], to determine the appropriate number of reactions in stoichiometric modelling[8], and as a useful preprocessing tool for metabolomic, proteomic, transcriptomic[9] and environmental[10] data analysis.

Shortly after the publication of the original MLPCA algorithm, an application of this method was proposed addressing the missing data (MD) problem in PCA model building (PCA-MB)[11]. MLPCA deals with the missing values by assigning them large variances prior to implementing the method, which guides the algorithm to fit a PCA model disregarding these data points. The MLPCA approach for MD has been applied successfully in the literature to fluorescent, chromatographic, near-infrared spectroscopic[11], spectrophotometric[12], and environmental[13] data.

Folch-Fortuny *et al.*[14] address the problem of PCA-MB with missing data. In this work, several methods originally proposed for PCA model exploitation (PCA-ME)[15,16], *i.e.* when a fixed PCA model is used to infer missing values in new incomplete observations, are adapted to the model building context. Basically, the idea was to adapt the known iterative algorithm (IA)[17] for PCA-MB by replacing the prediction of the PCA model to that resulting when we treat each incomplete row in the data set as a new observation with missing values, and applying the projection to the model plane (PMP) method for PCA-ME[18]. This adaptation arose from the fact that PMP is, under some general conditions, equivalent to IA and the minimization of the squared prediction error (SPE) for PCA-ME, as proved in [15]. Thus, the aim in [14] was to assess whether this equivalence held in the PCA-MB context.

The regression-based methods, proposed in [15], were also adapted, jointly with PMP, to the PCA-MB context in [14]. These methods are: known data regression (KDR), KDR with principal component regression (PCR), KDR with partial least squares regression (PLS) and trimmed scores regression (TSR). All these methods impute the missing values in a data set by fitting different regression-based schemes between the available data and the missing positions. Several other methods were compared to the previous ones in [14], including the modified NIPALS algorithm[19], the nonlinear programming approach[20] and multiple imputation by data augmentation[21]. The conclusion was that TSR represents a good compromise solution between prediction quality, robustness against data structure and computation time[14]; outperforming other approaches implemented in commercial software as ProSensus[22], SIMCA[23] and PLS Toolbox[24]. TSR and most of the other approaches compared in [14] are now implemented in a freely available user-friendly MATLAB toolbox[25] (http://mseg.webs.upv.es).

Nelson[26] showed the equivalence between the scores calculation by columns in MLPCA and the PMP algorithm for PCA-ME. Here, we are going to prove the equivalence between the imputation step by columns in MLPCA algorithm and the adapted PMP method for PCA-MB.

The aim of this paper is, thus, to answer three questions that arise from the aforementioned equivalence:

i) Once the algorithms converge, are the imputed values of MLPCA and PMP for PCA-MB equal?

ii) Since TSR outperforms PMP, as proven in [14], if the imputation step in MLPCA is substituted by a TSR-based imputation, does the imputation outperform the original MLPCA?

iii) In any case, does MLPCA, or its adapted version with TSR, outperform the original TSR algorithm?

To answer these research questions, we propose here to adapt the regression-based methods[14] (KDR, KDR with PCR, KDR with PLS and TSR) to work as different imputation steps within the MLPCA algorithm, providing a framework for maximum likelihood missing data imputation. The performance of these methods is compared to PMP and TSR methods using six data sets, actual and simulated ones, from different research areas.

The rest of the paper is organised as follows. Section 2 proves the equivalence between the imputation step by columns of MLPCA and the PMP method for PCA model building, and describes how the regression-based methods are adapted to its maximum likelihood (ML) version. Section 3 describes the data sets used in this study, as well as how the comparative study is performed. Section 4 shows the results of the ML regression-based methods, jointly with the original PMP and TSR algorithms. Finally, the conclusions are highlighted in Section 5.

2. METHODOLOGY

Let $\mathbf{X}$ be an N by K matrix, $\mathbf{x}_i^{\mathrm{T}}$ its i^{th} row and $\mathbf{y}_{\mathrm{k}}$ its k^{th} column. Each row represents a point in the K-dimensional space of the $\mathbf{X}$ observations, and each column a point in the N-dimensional space of the $\mathbf{X}$ variables. Row i can be decomposed in $\mathbf{x}_i^{\mathrm{T}} = \mathbf{x}_i^{0,\mathrm{T}} + \boldsymbol{\varepsilon}_i^{\mathrm{T}}$, where $\mathbf{x}_i^{0,\mathrm{T}}$ are the true values and $\boldsymbol{\varepsilon}_i^{\mathrm{T}}$ are their measurement errors[26]. As well, column k can be decomposed in its true and error parts: $\mathbf{y}_k = \mathbf{y}_k^0 + \boldsymbol{\eta}_k$. Both errors are assumed normally distributed in each of the K and N dimensions, respectively.

The maximisation of the likelihood is obtained by minimising the following objective function:

$$S^2 = \sum_{i=1}^{N} (\mathbf{x}_i^{\mathrm{T}} - \hat{\mathbf{x}}_i^{\mathrm{T}})\, \boldsymbol{\Sigma}_i^{-1} (\mathbf{x}_k - \hat{\mathbf{x}}_k) = \sum_{k=1}^{K} (\mathbf{y}_i^{\mathrm{T}} - \hat{\mathbf{y}}_i^{\mathrm{T}})\, \boldsymbol{\Psi}_k^{-1} (\mathbf{y}_k - \hat{\mathbf{y}}_k) \tag{1}$$

where $\boldsymbol{\Sigma}_i$ is the covariance matrix of the errors $\boldsymbol{\varepsilon}_i^{\mathrm{T}}$ of observation $\hat{\mathbf{x}}_i^{\mathrm{T}}$, and $\boldsymbol{\Psi}_k$ is the covariance matrix of the errors $\boldsymbol{\eta}_k$ of variable $\hat{\mathbf{y}}_{\mathrm{k}}$. The estimation of both vectors arise from:

$$\hat{\mathbf{x}}_{\mathrm{i}} = \hat{\mathbf{P}}(\hat{\mathbf{P}}^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1} \hat{\mathbf{P}})^{-1} \hat{\mathbf{P}}^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1} \mathbf{x}_i \tag{2}$$

$$\hat{\mathbf{y}}_k = \hat{\mathbf{U}}(\hat{\mathbf{U}}^{\mathrm{T}} \boldsymbol{\Psi}_k^{-1} \hat{\mathbf{U}})^{-1} \hat{\mathbf{U}}^{\mathrm{T}} \boldsymbol{\Psi}_k^{-1} \mathbf{y}_k \tag{3}$$

where $\hat{\mathbf{U}}$ $(N\times A)$, $\hat{\mathbf{D}}$ $(A\times A)$ and $\hat{\mathbf{P}}$ $(K\times A)$ represent the singular value decomposition of $\hat{\mathbf{X}} = \hat{\mathbf{U}}\hat{\mathbf{D}}\hat{\mathbf{P}}^{\mathrm{T}} = [\hat{\mathbf{y}}_1 \ldots \hat{\mathbf{y}}_K] = [\hat{\mathbf{x}}_1 \ldots \hat{\mathbf{x}}_N]^{\mathrm{T}}$, using A dimensions or components.

MLPCA algorithm is an alternating least squares procedure that starts imputing initial guesses for $\hat{\mathbf{U}}$ and $\hat{\mathbf{P}}$ based on the SVD decomposition of $\mathbf{X}$. At each iteration, the algorithm has two steps. The first one consists of projecting the rows $\mathbf{x}_i^{\mathrm{T}}$ on the columns of $\hat{\mathbf{P}}$, computing the objective function, and recalculating $\hat{\mathbf{U}}$ and $\hat{\mathbf{P}}$ from an SVD using the estimations. The second step consists of projecting the columns $\mathbf{y}_k$ on the columns of $\widehat{\mathbf{U}}$, computing also the objective function, and finally recalculating again $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{P}}$ from an SVD. Convergence is achieved when the difference between the estimations of the observations are below a specified threshold[27].

The adaptation of MLPCA to model building with missing data assumes uncorrelated errors for both objects $\mathbf{x}_i^{\mathrm{T}}$ and variables $\mathbf{y}_k$, therefore matrices $\boldsymbol{\Sigma}_i$ and $\boldsymbol{\Psi}_k$ are diagonal[26,27]. In this algorithm, large variances (10^{10}) are assigned to the missing measurements, and ones to the available ones. Therefore, the inversion of matrices $\boldsymbol{\Sigma}_i$ and $\boldsymbol{\Psi}_k$ produces diagonal matrices with 1s and 0s. The ones serve to fit these specific measurements in the PCA and the 0s to disregard the missing measurements in the multivariate model.

Let us assume that row i has missing values. The values in this vector can be rearranged to have the missing values in its first R_i positions without loss of generality, and the remaining $K - R_i$ available values at the end. This partition in $\mathbf{x}_i^{\mathrm{T}}$, induces a partition in the $\mathbf{X}$ data set, being $\mathbf{X}^{\#}$ $(N \times R_i)$ the missing part and $\mathbf{X}^{*}$ $(N \times (K - R_i))$ the available part, according to row i. Additionally, this partition can be transferred into a SVD (or PCA) model, $\mathbf{X} = \mathbf{UDP}^{\mathrm{T}}$, being $\mathbf{P}^{\#}$ $(R_i \times A)$ the missing part of the loadings matrix, and $\mathbf{P}^{*}$ $((K - R_i) \times A)$ the available part. Figure 1 shows a scheme of this notation.

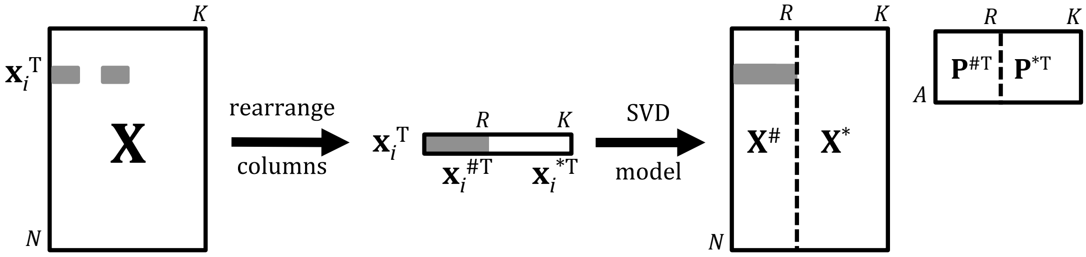

Figure 1. Partition induced in $\mathbf{X}$ matrix by the missing data in its i^{th} row. Grey squares denote missing positions in the data set.

Using this partition, the inverse of matrix $\boldsymbol{\Sigma}_i$ can be written as:

$$\boldsymbol{\Sigma}_i^{-1} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{K\text{-}R_i} \end{bmatrix} \tag{4}$$

where $\mathbf{I}_{K-R_i}$ is the identity matrix with $K - R_i$ rows/columns, according to the missing data pattern in $\mathbf{x}_i^{\mathrm{T}}$.

Substituting this expression in Equation 2, observation $\hat{\mathbf{x}}_i^{\mathrm{T}}$ can be computed as:

$$\begin{aligned}\hat{\mathbf{x}}_i = \begin{bmatrix}\hat{\mathbf{x}}_i^{\#}\\ \hat{\mathbf{x}}_i^{*}\end{bmatrix} = \begin{bmatrix}\widehat{\mathbf{P}}^{\#}\\ \widehat{\mathbf{P}}^{*}\end{bmatrix}\left(\begin{bmatrix}\widehat{\mathbf{P}}^{\#\mathrm{T}} & \widehat{\mathbf{P}}^{*\mathrm{T}}\end{bmatrix}\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_{K\text{-}R_i}\end{bmatrix}\begin{bmatrix}\widehat{\mathbf{P}}^{\#}\\ \widehat{\mathbf{P}}^{*}\end{bmatrix}\right)^{-1}\begin{bmatrix}\widehat{\mathbf{P}}^{\#\mathrm{T}} & \widehat{\mathbf{P}}^{*\mathrm{T}}\end{bmatrix}\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_{K\text{-}R_i}\end{bmatrix}\begin{bmatrix}\mathbf{0}\\ \mathbf{x}_i^{*}\end{bmatrix} = \\ = \begin{bmatrix}\widehat{\mathbf{P}}^{\#}\\ \widehat{\mathbf{P}}^{*}\end{bmatrix}\left(\widehat{\mathbf{P}}^{*\mathrm{T}}\widehat{\mathbf{P}}^{*}\right)^{-1}\widehat{\mathbf{P}}^{*\mathrm{T}}\mathbf{x}_i^{*} = \begin{bmatrix}\widehat{\mathbf{P}}^{\#}\left(\widehat{\mathbf{P}}^{*\mathrm{T}}\widehat{\mathbf{P}}^{*}\right)^{-1}\widehat{\mathbf{P}}^{*\mathrm{T}}\mathbf{x}_i^{*}\\ \widehat{\mathbf{P}}^{*}\left(\widehat{\mathbf{P}}^{*\mathrm{T}}\widehat{\mathbf{P}}^{*}\right)^{-1}\widehat{\mathbf{P}}^{*\mathrm{T}}\mathbf{x}_i^{*}\end{bmatrix}\end{aligned} \qquad (5)$$

Alternatively, the inverse of matrix $\mathbf{\Psi}_k$ can be written as:

$$\mathbf{\Psi}_k^{-1} = \begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_{N\text{-}R_k}\end{bmatrix} \qquad (6)$$

where $\mathbf{I}_{N-R_k}$ is the identity matrix with $N - R_k$ rows/columns, according to column $\mathbf{y}_k$. Following Equation 3, $\hat{\mathbf{y}}_k$ is therefore computed as:

$$\begin{aligned}\hat{\mathbf{y}}_k = \begin{bmatrix}\mathbf{y}_k^{\#}\\ \hat{\mathbf{y}}_k^{*}\end{bmatrix} = \begin{bmatrix}\widehat{\mathbf{U}}^{\#}\\ \widehat{\mathbf{U}}^{*}\end{bmatrix}\left(\begin{bmatrix}\widehat{\mathbf{U}}^{\#\mathrm{T}} & \widehat{\mathbf{U}}^{*\mathrm{T}}\end{bmatrix}\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_{N\text{-}R_k}\end{bmatrix}\begin{bmatrix}\widehat{\mathbf{U}}^{\#}\\ \widehat{\mathbf{U}}^{*}\end{bmatrix}\right)^{-1}\begin{bmatrix}\widehat{\mathbf{U}}^{\#\mathrm{T}} & \widehat{\mathbf{U}}^{*\mathrm{T}}\end{bmatrix}\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_{N\text{-}R_k}\end{bmatrix}\begin{bmatrix}\mathbf{0}\\ \mathbf{y}_k^{*}\end{bmatrix} = \\ = \begin{bmatrix}\widehat{\mathbf{U}}^{\#}\\ \widehat{\mathbf{U}}^{*}\end{bmatrix}\left(\widehat{\mathbf{U}}^{*\mathrm{T}}\widehat{\mathbf{U}}^{*}\right)^{-1}\widehat{\mathbf{U}}^{*\mathrm{T}}\mathbf{y}_k^{*} = \begin{bmatrix}\widehat{\mathbf{U}}^{\#}\left(\widehat{\mathbf{U}}^{*\mathrm{T}}\widehat{\mathbf{U}}^{*}\right)^{-1}\widehat{\mathbf{U}}^{*\mathrm{T}}\mathbf{y}_k^{*}\\ \widehat{\mathbf{U}}^{*}\left(\widehat{\mathbf{U}}^{*\mathrm{T}}\mathbf{U}^{*}\right)^{-1}\widehat{\mathbf{U}}^{*\mathrm{T}}\mathbf{y}_k^{*}\end{bmatrix}\end{aligned} \qquad (7)$$

where $\widehat{\mathbf{U}}^{\#}$ $(R_k \times A)$ and $\widehat{\mathbf{U}}^{*}$ $((N - R_k) \times A)$ are the missing and available parts of $\widehat{\mathbf{U}}$.

The MLPCA imputation step of the missing values $\mathbf{x}_i^{\#\mathrm{T}}$ is the same as the PMP method for PCA model building presented recently in [14]. The main difference between MLPCA algorithm and PMP is that the former performs the imputation iteratively first by observations and then by variables, instead of only by observations, as PMP does. And additionally, the convergence in PMP is achieved based only on the imputed missing values, instead of the imputation of the available measurements, as it is in MLPCA.

In [14] it was shown that the imputation step in the adapted PMP algorithm for PCA-MB could be substituted by the regression-based methods presented in [15] (KDR and its variants, and TSR). Most of these methods showed a superior performance than PMP across several case studies. So, the idea here consists of adapting the alternating imputation of MLPCA algorithm to include the imputation step of the regression-based methods, thus proposing a maximum likelihood (ML) framework: ML-KDR, ML-KDR with PCR, ML-KDR with PLS and ML-TSR.

The imputation step of the regression-based missing data methods is:

$$\hat{\mathbf{x}}_i = \begin{bmatrix}\hat{\mathbf{x}}_i^{\#}\\ \hat{\mathbf{x}}_i^{*}\end{bmatrix} = \begin{bmatrix}\hat{\mathbf{S}}^{\#*}\mathbf{L}_i\left(\mathbf{L}_i^{\mathrm{T}}\hat{\mathbf{S}}^{**}\mathbf{L}_i\right)^{-1}\mathbf{L}_i^{\mathrm{T}}\mathbf{x}_i^{*}\\ \hat{\mathbf{S}}^{**}\mathbf{L}_i\left(\mathbf{L}_i^{\mathrm{T}}\hat{\mathbf{S}}^{**}\mathbf{L}_i\right)^{-1}\mathbf{L}_i^{\mathrm{T}}\mathbf{x}_i^{*}\end{bmatrix} \tag{8}$$

where $\hat{\mathbf{S}}$ is the covariance matrix of $\hat{\mathbf{X}}$, and:

$$\hat{\mathbf{S}} = [\hat{\mathbf{X}}^{\#}\hat{\mathbf{X}}^{*}]^{\mathrm{T}}[\hat{\mathbf{X}}^{\#}\hat{\mathbf{X}}^{*}]/(N-1) = \begin{bmatrix}\hat{\mathbf{X}}^{\#\mathrm{T}}\hat{\mathbf{X}}^{\#} & \hat{\mathbf{X}}^{\#\mathrm{T}}\hat{\mathbf{X}}^{*}\\ \hat{\mathbf{X}}^{*\mathrm{T}}\hat{\mathbf{X}}^{\#} & \hat{\mathbf{X}}^{*\mathrm{T}}\hat{\mathbf{X}}^{*}\end{bmatrix}/(N-1) = \begin{bmatrix}\hat{\mathbf{S}}^{\#\#} & \hat{\mathbf{S}}^{\#*}\\ \hat{\mathbf{S}}^{*\#} & \hat{\mathbf{S}}^{**}\end{bmatrix} \tag{9}$$

The key matrix $\mathbf{L}$ in Equation 8 particularises which method of the framework is being used for the imputation: $\mathbf{L} = \mathbf{I}$ for KDR; $\mathbf{L} = \hat{\mathbf{V}}_{1:\rho}$ for KDR with PCR, where $\hat{\mathbf{V}}_{1:\rho}$ is the eigenvector matrix of $\hat{\mathbf{S}}^{**}$ and $\rho \leq \mathrm{rank}(\hat{\mathbf{S}}^{**})$; $\mathbf{L} = \hat{\mathbf{W}}^{*}$ for KDR with PLS, where $\hat{\mathbf{W}}^{*}$ is the loadings matrix of the PLS model $\hat{\mathbf{T}}_{PLS} = \hat{\mathbf{X}}^{*\mathrm{T}}\hat{\mathbf{W}}^{*}$; and $\mathbf{L} = \hat{\mathbf{P}}^{*}$ for TSR.

Therefore, to adapt the MLPCA original algorithm[11] to use the regression-based methods, we have to substitute the imputation step (Equations 2-3) by:

$$\hat{\mathbf{x}}_{\mathrm{i}} = \hat{\mathbf{S}}\mathbf{\Lambda}_i\mathbf{L}_i(\mathbf{L}_i^{\mathrm{T}}\mathbf{\Lambda}_i^{\mathrm{T}}\hat{\mathbf{S}}\mathbf{\Lambda}_i\mathbf{L}_i)^{-1}\mathbf{L}_i^{\mathrm{T}}\mathbf{\Lambda}_i^{\mathrm{T}}\mathbf{x}_i \tag{10}$$

$$\hat{\mathbf{y}}_{\mathrm{k}} = \hat{\mathbf{S}}\mathbf{\Phi}_k\mathbf{L}_k(\mathbf{L}_k^{\mathrm{T}}\mathbf{\Phi}_k^{\mathrm{T}}\hat{\mathbf{S}}\mathbf{\Phi}_k\mathbf{L}_k)^{-1}\mathbf{L}_k^{\mathrm{T}}\mathbf{\Phi}_k^{\mathrm{T}}\mathbf{y}_k \tag{11}$$

where $\mathbf{L}$ matrix is the same as in the regression-based framework, particularising for the missing data pattern in row i or column k. And:

$$\mathbf{\Lambda}_i = \begin{bmatrix}\mathbf{0}\\ \mathbf{I}_{K\text{-}R_i}\end{bmatrix} \tag{12}$$

$$\mathbf{\Phi}_k = \begin{bmatrix}\mathbf{0}\\ \mathbf{I}_{N\text{-}R_k}\end{bmatrix} \tag{13}$$

The equivalence between Equations 10 and 8 is shown in Appendix B. For more details on MLPCA, readers are referred to [26,27] and the original paper[11]. The Matlab source code of the algorithm for PCA model building with missing data is reproduced here in Appendix B with slight changes to introduce the imputation step of the regression-based methods.

3. DATA SETS AND COMPARATIVE STUDY

Six data sets are used in the present study to compare the results of the different imputation methods included in the framework. The first data set contains FTIR miscroscopy spectra of a polymer laminate consisting of three layers: polyethylene (PE), isophtalic polyester (IPE, presence originally unknown), and polyethylene terephthalate (PET). The polymer was scanned in a seventeen point transect across the different layers, obtaining measurements from 81 wavelengths[28-30]. The second case study consists of a set of measured and inferred fluxes from *Pichia pastoris* cultures on heterogeneous culture media[31]. The measured fluxes were collected from a literature review; later on, a grey modelling approach was applied to infer the intracellular fluxes according to the observed extracellular ones. From the original data set with 3600 scenarios and 45 fluxes, a representative sample of 105 individuals is selected for the present comparative study. This data set has 3 biologically relevant PCs. Finally, a simulated data set, with 100 observations and 10 variables, is used to compare the performance of the different maximum likelihood methods[32,33]. This data set has 4 eigenvalues (3, 2.5, 2 and 1.5) explaining 90% of the variance in data.

Three additional data sets are analysed here, taken from [14], where the adaptation of the regression based methods to PCA-MB was proposed. The first one consists of the percentage composition of eight fatty acids in 75 olive oils of South Apulia[34]. The second one is a set of NIR spectra (750-1550nm in 2nm intervals) measured on 40 diesel fuels[35]. And the last one is a 100×10 simulated data set with 3 components explaining 40, 30 and 20% of variance[32,33].

Two performance criteria are used to compare the results of the different methods. The first one is the mean squared prediction error (MSPE):

$$MSPE(Method) = \frac{\sum_{i=1}^{N}\sum_{j=1}^{K}\left(\hat{x}_{ij} - \hat{x}_{ij}^{Method}\right)^2}{NK} \qquad (14)$$

where $\hat{x}_{ij}$ is the predicted value for the j^{th} variable of the i^{th} observation in the prediction matrix $\widehat{\mathbf{X}} = \widehat{\mathbf{T}}\widehat{\mathbf{P}}^{\mathrm{T}}$ obtained from the complete data set; $\hat{x}_{ij}^{\text{Method}}$ the analogous prediction obtained after applying the corresponding method on the incomplete data set; and N and K are the number of rows and columns in the data set, respectively. The original regression-based framework methods use, as convergence criterion, the difference between consecutive imputations of missing values. Instead, MLPCA use the difference between the available measurements and their predictions from the current PCA model. For this, we decided to show the MSPEs for the available measurements and the missing ones separately, using Equation 14.

The second criterion consists of the cosine between the first loading vector obtained using the full data matrix and its corresponding one from the incomplete data set. The cosines of further PCs are not shown, since their values are strongly affected by the deviations of the first PC[14].

Six different levels of missing values are generated for all data sets, ranging from 10% to 60% of missing data. Also, 50 different MD patterns are generated for each percentage of missing data, in order to build confidence intervals for the *MSPEs*. The intervals are built based on the LSD significance of a three-factor mixed-effects ANOVA, where method and percentage of MD are fixed-effect factors (and also their interaction), and the replicates is the random-effect factor (nested to percentage). Given the positive skewness of *MSPE*, a logarithmic transformation is used to ease the visualization of the plots.

4. RESULTS

In this section the results of the comparative study are presented. However, we decided to exclude the results of ML-KDR, ML-KDR with PCR and ML-KDR with PLS due to large computation times, something already observed in [14], and due to the instability of some of them, especially ML-KDR (also observed in [14] with KDR) and ML-KDR with PLS. Therefore, the results of MLPCA, ML-TSR, TSR and PMP are shown, in order to answer the three research questions posed in the Introduction.

4.1. FTIR microspectroscopy

Figure 2 shows the results of the first case study. The two upper plots, A) and B), show the logarithm of the MSPEs and the cosines of PC#1, respectively. Figures 2C-2D show also the logarithm of the MSPEs but considering only the measured values and the imputed values separately. Regarding Figure 2A, there exist no statistical differences between MLPCA and ML-TSR in all percentages of missing data. TSR and PMP statistically outperform both ML approaches for low percentages of missing data (10-20%). From 50% onwards, TSR is superior to PMP, MLPCA and ML-TSR. The cosines shown in Figure 2B are coherent with the results of the *MSPEs*, having TSR the highest cosines from 30% to 60%.

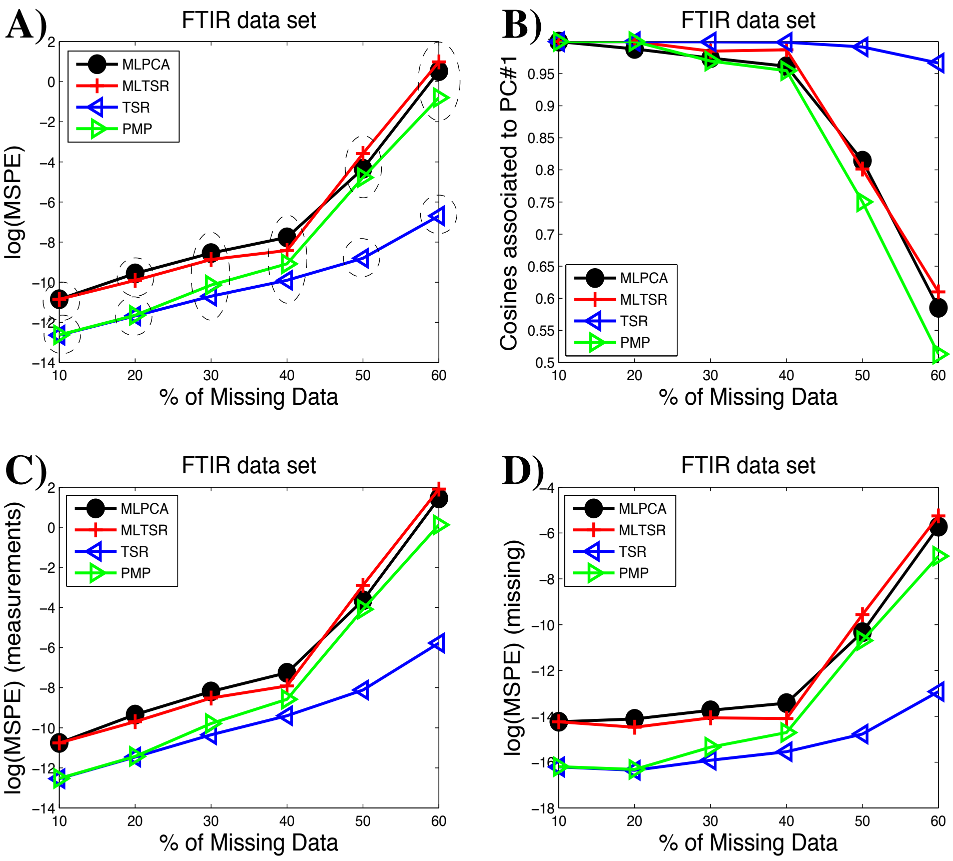

Figure 2. FTIR data set results. A) Logarithm of the MSPE for all measurements. B) Cosines associated to the first PC. C) Logarithm of the *MSPE* for the available measurements. D) Logarithm of the *MSPE* for the missing data. The dashed ellipses in a) mark the statistically significant differences between groups of methods. In A) TSR is statistically superior to MLPCA with 30-40% of MD. However, since there is no method statistically significant from all the rest, a single dashed ellipse encloses all of them.

The results in Figure 2C show that TSR and PMP are superior to the ML approaches in terms of the measured values, which implies that the PCA model fitted once the data is imputed with these methods is closer to the original one than using maximum likelihood estimations. Figure 2D is indeed very similar to Figure 2A, due to the fact that the errors in the imputed values between the true PCA model and the imputed one are way larger than in the measured values, as expected.

4.2. P. pastoris cultures on heterogeneous culture media

The results with the *P. pastoris* data set are similar to the previous ones, both in MSPEs and cosines (see Figure 3A-3B). TSR and PMP achieve the statistically best performance from 20%-40% of MD; and again, from 50% onwards, TSR becomes the best approach, being PMP superior to MLPCA and ML-TSR. The performances of TSR and PMP are indeed coherent with the results observed in [14], which confirm the results obtained in that paper. The cosines shown in Figure 3B are coherent with the *MSPE* values. The lower is the logarithm of the MSPE, the closer are the loading vectors of the reconstructed matrix to the actual ones.

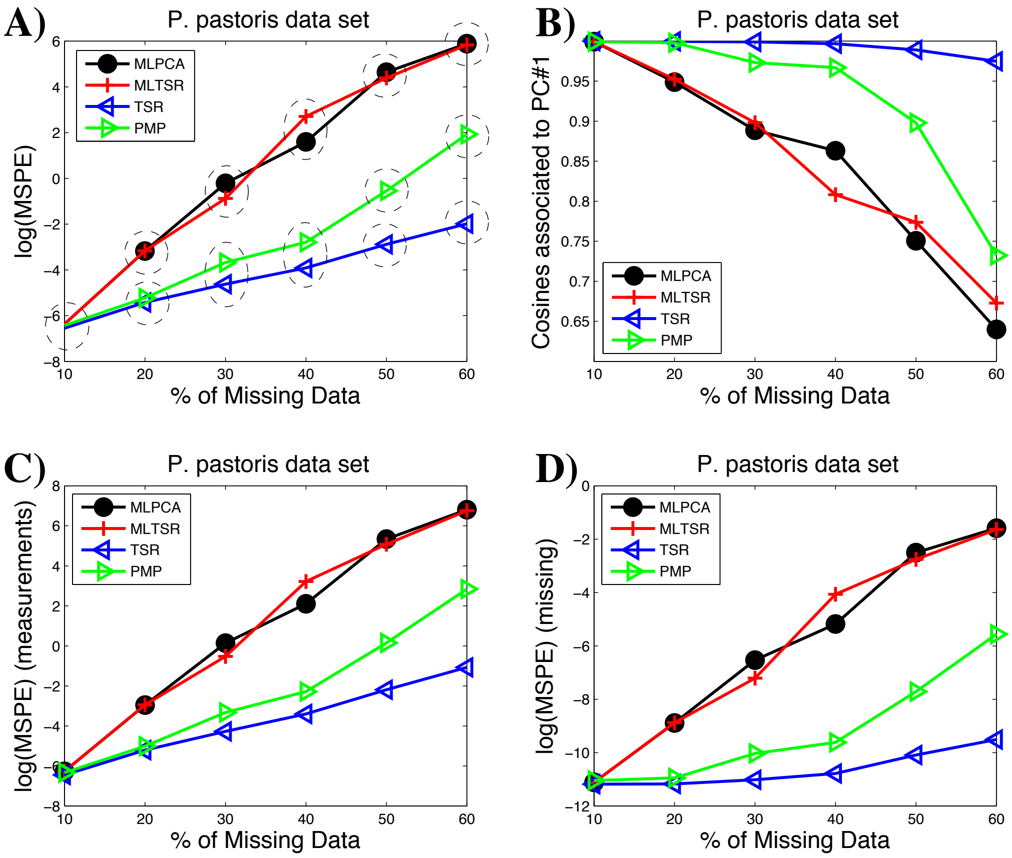

Figure 3. *P. pastoris* data set results. A) Logarithm of the *MSPE* for all measurements. B) Cosines associated to the first PC. C) Logarithm of the *MSPE* for the available measurements. D) Logarithm of the *MSPE* for the missing data. The dashed ellipses in A) mark the statistically significant differences between groups of methods.

Regarding Figures 3C-3D, the performance of all methods is also similar to the first example. For low percentages of MD, the differences among methods are smaller in the measured values, but from 30% of MD onwards, the PCA model obtained with TSR imputation resembles more to the real one.

4.3 Simulated data set

In the Simulated data set with 4 PCs, the differences among TSR, ML-TSR, PMP and MLPCA are not statistically significant for low percentages of missing values (10-20%) (see Figure 4A). With 30%

of MD, TSR becomes statistically the best method and PMP the worst one. This is something that was observed in [14], also using a simulated data set[32,33]. The higher is the percentage of missing data, the more difficult is to impute properly for PMP. For higher percentages (30-60%), there are statistical differences among all methods: TSR maintains the best performance, followed by ML-TSR, MLPCA and PMP. This is the first case study where there exist differences between MLPCA and ML-TSR, being the latter statistically superior. These differences in the *MSPEs* can also be seen in Figure 4B, where all methods but TSR show huge deviations from the true principal coordinate of the data with low-medium percentages of MD (10-40%).

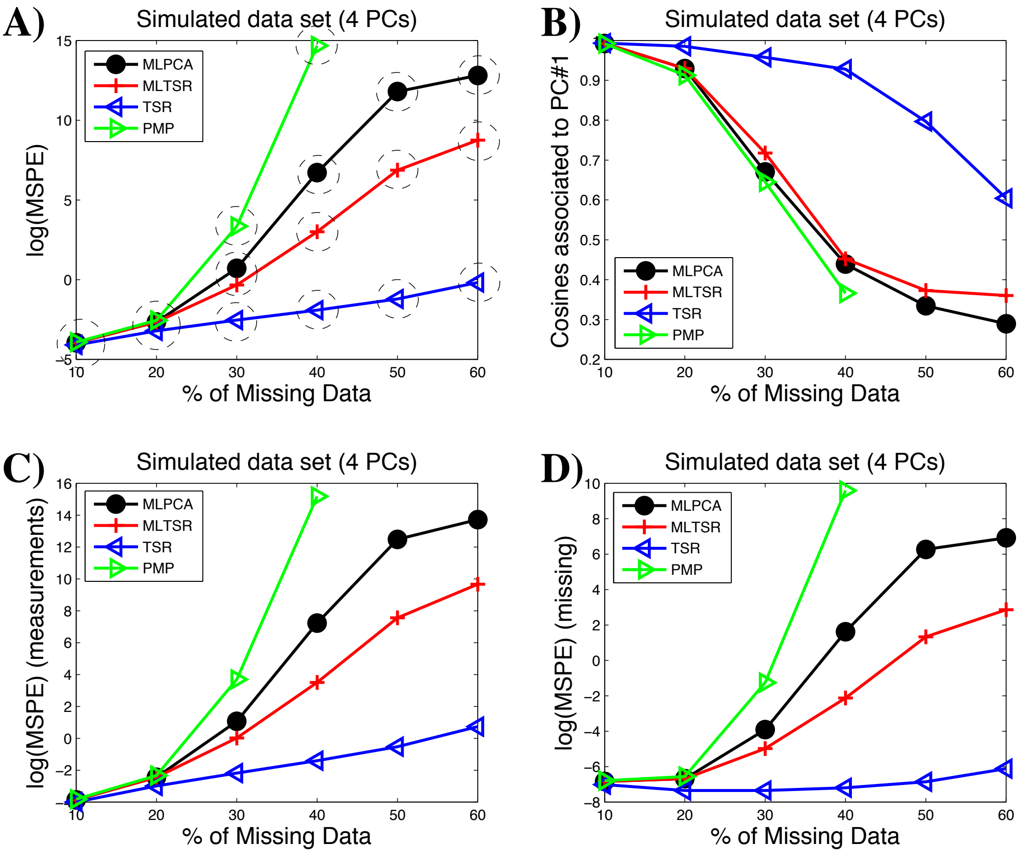

Figure 4. Simulated data set results. A) Logarithm of the MSPE for all measurements. B) Cosines associated to the first PC. C) Logarithm of the MSPE for the available measurements. D) Logarithm of the *MSPE* for the missing data. The dashed ellipses in A) mark the statistically significant differences between groups of methods.

In this third example the differences among methods regarding the measured values are narrower (see Figure 4C), but still showing the superiority of TSR.

4.4 Additional data sets

Three more data sets are used to compare the performance of the ML-based methods against PMP and TSR in its original form: the olive oil data set, the diesel NIR data set, and a 3-component simulated data set. The figures containing the logarithm of the MSPEs and the cosines associated to the first component are available as Supporting Information of this paper.

Summarizing the results, in these data sets the performance of TSR is statistically superior to PMP (as proven in [14]), and to MLPCA and ML-TSR for medium-high percentages (30-60%) and also for low percentages (10-20%) in the olive oil and diesel NIR data set. Also, the reconstruction of the available measurements with TSR is more similar to the PCA on complete data than the ML-based approaches in both data sets. These results are coherent with sections 4.1-4.2. Comparing ML-TSR and MLPCA, the former yields better results than MLPCA for high percentages of missing data (50-60%) in the 3-component simulated data set, as happened in section 4.3. with the 4-component simulated data set.

5. CONCLUSIONS

To conclude, it is worth to remember the research questions posed at the beginning of the paper:

- *Are the imputed values of MLPCA and PMP for model building equal?* The answer is no. The PMP imputation step performed alternatively by rows and columns in MLPCA drives the imputation in a different direction than performing it only by columns, as PMP does. Based on the six data sets analysed here, PMP, if converges, has better results than MLPCA. However, PMP suffered from convergence problems in some case studies, while MLPCA converge in all data sets and all MD percentages.

- *Does ML-TSR outperform the imputation of MLPCA?* The answer, based on the case studies analysed here, is that when the latent structure is complex, and the percentage of missing data is high, ML-TSR may outperform MLPCA. In other cases, the overall results have no statistically significant differences. However, MLPCA tends to be between 2-5 times faster than ML-TSR.
- *Does MLPCA or ML-TSR outperform the original TSR algorithm?* The answer is no. TSR outperforms the ML approaches for medium-high percentages of missing data. For low percentages, depending on the case study analysed, it is statistically superior or there exist no statistical difference compared to the other methods.

Finally, we recommend the use of trimmed score regression over MLPCA for PCA model building with missing data, since the both the reconstruction of the available and imputed values is statistically more accurate than using MLPCA or ML-TSR.

Appendix A. Regression-based imputation step in MLPCA.

The equivalence between Equations 10 and 8 is proven here. Let us assume that we rearrange the values in row $\mathbf{x}_i^{\mathrm{T}}$ to have the R_i missing values, $\mathbf{x}_i^{\#\mathrm{T}}$, at the first positions, and the remaining $K - R_i$ available ones, $\mathbf{x}_i^{\#\mathrm{T}}$, at the end. We can use Equation 4 in Equation 10 to introduce the extension of the missing data partition, $\widehat{\mathbf{X}} = [\widehat{\mathbf{X}}^{\#}\widehat{\mathbf{X}}^{*}]$. Bearing in mind that the decomposition of the covariance matrix of $\widehat{\mathbf{X}}$ (see Equation 9), and matriX $\boldsymbol{\Lambda}_i$ (Equation 12), Equation 10 can be written as:

$$\hat{\mathbf{x}}_{\mathrm{i}} = \widehat{\mathbf{S}}\boldsymbol{\Lambda}_i\mathbf{L}_i(\mathbf{L}_i^{\mathrm{T}}\boldsymbol{\Lambda}_i^{\mathrm{T}}\widehat{\mathbf{S}}\boldsymbol{\Lambda}_i\mathbf{L}_i)^{-1}\mathbf{L}_i^{\mathrm{T}}\boldsymbol{\Lambda}_i^{\mathrm{T}}\mathbf{x}_i = \begin{bmatrix}\widehat{\mathbf{S}}^{\#\#} & \widehat{\mathbf{S}}^{\#*}\\ \widehat{\mathbf{S}}^{*\#} & \widehat{\mathbf{S}}^{**}\end{bmatrix}\begin{bmatrix}\mathbf{0}\\ \mathbf{I}_{K\text{-}R_i}\end{bmatrix}\mathbf{L}_i(\mathbf{L}_i^{\mathrm{T}}[\mathbf{0}\ \mathbf{I}_{K-R_i}]\begin{bmatrix}\widehat{\mathbf{S}}^{\#\#} & \widehat{\mathbf{S}}^{\#*}\\ \widehat{\mathbf{S}}^{*\#} & \widehat{\mathbf{S}}^{**}\end{bmatrix}\begin{bmatrix}\mathbf{0}\\ \mathbf{I}_{K\text{-}R_i}\end{bmatrix}\mathbf{L}_i)^{-1}\mathbf{L}_i^{\mathrm{T}}[\mathbf{0}\ \mathbf{I}_{K-R_i}]\mathbf{x}_i \tag{15}$$

$$= \begin{bmatrix}\widehat{\mathbf{S}}^{\#\#} & \widehat{\mathbf{S}}^{\#*}\\ \widehat{\mathbf{S}}^{*\#} & \widehat{\mathbf{S}}^{**}\end{bmatrix}\begin{bmatrix}\mathbf{0}\\ \mathbf{L}_i\end{bmatrix}([\mathbf{0}\ \mathbf{L}_i^{\mathrm{T}}]\begin{bmatrix}\widehat{\mathbf{S}}^{\#\#} & \widehat{\mathbf{S}}^{\#*}\\ \widehat{\mathbf{S}}^{*\#} & \widehat{\mathbf{S}}^{**}\end{bmatrix}\begin{bmatrix}\mathbf{0}\\ \mathbf{L}_i\end{bmatrix})^{-1}[\mathbf{0}\ \mathbf{L}_i^{\mathrm{T}}]\begin{bmatrix}\mathbf{0}\\ \mathbf{x}_i^{*}\end{bmatrix} = \begin{bmatrix}\widehat{\mathbf{S}}^{\#*}\mathbf{L}_i\left(\mathbf{L}_i^{\mathrm{T}}\widehat{\mathbf{S}}^{**}\mathbf{L}_i\right)^{-1}\mathbf{L}_i^{\mathrm{T}}\mathbf{x}_i^{*}\\ \widehat{\mathbf{S}}^{**}\mathbf{L}_i\left(\mathbf{L}_i^{\mathrm{T}}\widehat{\mathbf{S}}^{**}\mathbf{L}_i\right)^{-1}\mathbf{L}_i^{\mathrm{T}}\mathbf{x}_i^{*}\end{bmatrix}$$

The proof using Equation 11 is analogous; substituting $\hat{\mathbf{x}}_{\mathrm{i}}$ by $\hat{\mathbf{y}}_{\mathrm{k}}$, changing the subindices i by k and the matrix $\boldsymbol{\Lambda}_i$ by $\boldsymbol{\Phi}_k$, and bearing in mind that the $\mathbf{L}_k$ matrix is obtained using the missing data pattern of $\hat{\mathbf{y}}_{\mathrm{k}}$.

Appendix B. Matlab source code.

The source code for MLPCA is shown here. It consists of a modification of the original MLPCA algorithm[11] to perform the imputation step using the regression based framework methods. The function `mlpmp.m` correspond to the original imputation step of MLPCA, which is equivalent to the PMP method for PCA model building, as proved in this paper. Only the ML version of TSR is shown, since the other approaches are not competitive. The source code for the original TSR and PMP can be found in [14].

```
function [U,S,V,SOBJ,ErrFlag,count]=mlpca_generic(X,stdX,p,type)
%
%From Andrews and Wentzell (1997). Analytica Chimica Acta 350, 341-352
%
% This function performs MLPCA with missing data
%
% X     mxn matrix of observations.
% stdX  mxn matrix of standard deviations associated with X (zeros for
missing meassurements).
% p     is the model dimensionality.
% type  MD method chosen:
%       0 MLPCA, equivalent to PMP
%       1 MLTSR (maximum likelihood TSR)
%       2 MLPCR (maximum likelihood KDR with PCR)
%
% U,S,V     the pseudo-svd parameters.
% SOBJ      value of the objetive function.
% ErrFlag   indicates exit conditions: 0 = normal termination, 1 = max
iterations exceeded.
% count     number of iterations needed
```

```
%
% Initialization
%
convlim=1e-10;
maxiter=500;
XX=X;
varX=stdX.^2;
[i,j]=find(varX==0);
errmax=max(max(varX));
for k=1:length(i),
    varX(i(k),j(k))=1e10*errmax;
end
n=length(XX(1,:));
%
% Generate initial estimates
%
for i=1:length(X(:,1)),
    for j=1:length(X(:,1)),
        denom=min([nnz(X(i,:)) nnz(X(j,:))]);
        CV(i,j)=(X(i,:)*X(j,:)')/denom;
    end
end
[U,S,V]=svd(CV,0);
U0=U(:,1:p);
MLXaux=XX;
%
% Loop for alternating least squares
%
type
count=0;
```

```
Sold=0;
ErrFlag=-1;
while ErrFlag<0,
    count=count+1;   %%%%
    Sobj=0;
    MLX=zeros(size(XX));
    for i=1:n,
        % Method selection
        switch type
            case 0
                [MLX, Q]=mlpmp(XX,varX, U0, n, i, MLX);
            case 1
                [MLX, Q]=mltsr(XX, MLXaux', varX, U0, n, i, MLX);
            otherwise
                  error('Wrong method')
        end
        dx=XX(:,i)-MLX(:,i);
        Sobj=Sobj+dx'*Q*dx;
    end
    if rem(count,2)==1,
        abs(Sold-Sobj)/Sobj;
        if(abs(Sold-Sobj)/Sobj)<convlim,
            ErrFlag=0;
        elseif count>maxiter,
            ErrFlag=1;
        end
    end
    if ErrFlag<0,
        Sold=Sobj;
        [U,S,V]=svd(MLX,0);
        XX=XX';
        varX=varX';
        n=length(XX(1,:));
        U0=V(:,1:p);
        MLXaux=MLX';
    end
end
%
% Finished
%
[U,S,V]=svd(MLX,0);
U=U(:,1:p);
S=S(1:p,1:p);
V=V(:,1:p);
SOBJ=Sobj;


function [ MLX, Q] = mlpmp( XX, varX, U0, n, i, MLX )
% MLPCA imputation, equivalent to PMP
Q=diag(varX(:,i).^(-1));
F=inv(U0'*Q*U0);
MLX(:,i)=U0*F*U0'*Q*XX(:,i);
end


function [ MLX, Q] = mltsr( XX, MLXaux, varX, U0, n, i, MLX )
% MLPCA with TSR missing data imputation
Q=diag(varX(:,i).^(-1));
CV=cov(MLXaux);
MLX(:,i)= CV*Q*U0*pinv(U0'*Q*CV*Q*U0)*U0'*Q*XX(:,i);
end
```

Supporting Information

Three additional figures with the results of the comparative study using the last three data sets are available online.

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-C04-02 and DPI2014-55276-C5-1R, and the Spanish Ministry of Economy and Competitiveness through grant ECO2013-43353-R.

REFERENCES

1. Jolliffe IT, Principal Component Analysis. Springer-Verlag, NY, USA, 1986.

2. Wentzell PD, Andrews DT, Hamilton DC, Faber K, Kowalski BR, Maximum likelihood principal component analysis, *J. Chemometr.* 1997, **11**, 339-366.

3. Ristolainen M, Alén R, Malkavaara P, Pere J, Reflectance FTIR microspectroscopy for studying effect of Xylan removal on unbleached and bleached birch kraft pulps, *Holzforschung* 2002, **56**(5), 513-521.

4. Keenan MR, Maximum likelihood principal component analysis of time-of-flight secondary ion mass spectrometry spectral images, *J. Vac. Sci. Technol. A* 2005, **23**(4), 746-750.

5. Sang WC, Martin EB, Morris AJ, Lee I-B, Fault detection based on a maximum-likelihood principal component analysis (PCA) mixture, *Ind. Eng. Chem. Res.* 2005, **44**(7), 2316-2327.

6. Karakach TK, Wentzell PD, Walter JA, Characterization of the measurement error structure in 1D 1H NMR data for metabolomics studies, *Anal. Chim. Acta* 2009, **636**(2), 163-174.

7. Wentzell PD, Hou S, Exploratory data analysis with noisy measurements, *J. Chemometrics* 2012, **26**(6), 264-281.

8. Mailier J, Remy M, Vande Wouwer A, Stoichiometric identification with maximum likelihood principal component analysis, *J. Math. Biol.* 2013, **67**(4), 739-765.

9. Hoefsloot HCJ, Verouden MPH, Westerhuis JA, Smilde AK, Maximum likelihood scaling (MALS), *J. Chemometr.* 2006, **20**(3-4), 120-127.

10. Dadashi M, Abdollahi H, Tauler R, Maximum Likelihood Principal Component Analysis as initial projection step in Multivariate Curve Resolution analysis of noisy data, *Chem Intell. Lab.* 2012, **118**, 33-40.

11. Andrews DT, Wentzell PD, Applications of maximum likelihood principal component analysis: incomplete data sets and calibration transfer, *Anal. Chim. Acta* 1997, **350**(3), 341-352.

12. Ho P, Silva MCM, Hogg TA, Multiple imputation and maximum likelihood principal component analysis of incomplete multivariate data from a study of the ageing of port, *Chemom. Intell. Lab.* 2001, **55**, 1-11.

13. Stanimirova I, Practical approaches to principal component analysis for simultaneously dealing with missing and censored elements in chemical data, *Anal. Chim. Acta* 2013, **796**, 27-37.

14. Folch-Fortuny A, Arteaga F, Ferrer A, PCA model building with missing data: new proposals and a comparative study, *Chemom. Intell. Lab.* 2015, **146**, 77-88.

15. Arteaga F, Ferrer A, Dealing with missing data in MSPC: several methods, different interpretations, some examples, *J. Chemometr.* 2002, **16**, 408-418.

16. Arteaga F, Ferrer A, Framework for regression-based missing data imputation methods in on-line MSPC, *J. Chemometr.* 2005, **19**, 439-447.

17. Walczak B, Massart DL, Dealing with missing data Part I, *Chemom. Intell. Lab.* 2001, **58**, 15-27.

18. Nelson PRC, Taylor PA, MacGregor JF, Missing data methods in PCA and PLS: Score calculations with incomplete observations, *Chemom. Intell. Lab.* 1996, **35**, 45-65.

19. Wold S, Albano C, Dunn WJ, Esbensen K, Hellberg S, Johansson E, Sjöström M, Pattern recognition: finding and using regularities in multivariate data, in: Martens H, Russwurm H (Jr.) (Eds.), *Food Research and Data Analysis*, vol. 3, Applied Science Pub: London, UK, 1983, 183–185.

20. López-Negrete de la Fuente R, García-Muñoz S, Biegler LT, An efficient nonlinear programming strategy for PCA models with incomplete data sets, *J. Chemom.*, 2010, **24**, 301–311.

21. Schafer JL, *Analysis of Incomplete Multivariate Data*, CRC Press: New York, USA, 1997.

22. ProSensus MultiVariate release 15.02, ProSensus Inc, Ancaster, Ontario, Canada, 2015. (http://www.prosensus.com).

23. SIMCA release 14, Umetrics, Umea, Sweden, 2015. (http://www.umetrics.com).

24. PLS Toolbox release 7.9.5, Eigenvector Research Inc, Manson, Washington, USA, 2015. (http://www.eigenvector.com).

25. Folch-Fortuny A, Arteaga F, Ferrer A, Missing Data Imputation Toolbox for MATLAB, submitted.

26. Nelson PRC, Treatment of missing measurements in PCA and PLS models, Ph.D. Dissertation, Department of Chemical Engineering, McMaster University, Hamilton, Ontario, Canada, 2002.

27 Arteaga F, Control estadístico multivariante de procesos con datos faltantes mediante análisis de componentes principales, Ph.D. thesis Universidad Politécnica de Valencia, 2003.

28. Guilment J, Markel S, Windig W, Infrared chemical micro-imaging assisted by interactive self-modeling multivariate analysis, *Appl. Spectr* 1994, **48**, 320-326.

29. Windig W, Markel S, Simple-to-use interactive self-modeling mixture analysis of FTIR microscopy data, *J. Mol. Struct.* 1993, **292**, 161-170.

30. Windig W, Spectral data files for self-modeling curve resolution with examples using the Simplisma approach, *Chemom. Intell. Lab.* 1997, **36**, 3-16.

31. González-Martínez JM, Folch-Fortuny A, Llaneras F, Tortajada M, Picó J, Ferrer A, Metabolic flux understanding of *Pichia pastoris* grown on heterogeneous culture media, *Chemom. Intell. Lab.* 2014, **134**, 89-99

32. Arteaga F, Ferrer A, How to simulate normal data sets with the desired correlation structure, *Chemom. Intell. Lab.* 2010, **101**, 38-42.

33. Arteaga F, Ferrer A, Building covariance matrices with the desired structure, *Chemom. Intell. Lab.* 2013, **127**, 80-88.

34. Forina M, Armanino C, Lanteri S, Tiscornia E, Classification of olive oils from their fatty acid composition, in: Martens H, Russwurm Jr H. (Eds.), Food Research and Data Analysis, Applied Science Pub, London 1983, pp. 189–214.

35. Hutzler SA, Bessee GB, 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.