Imputation of mixed data with multilevel singular value decomposition

François Husson, Julie Josse, Balasubramanian Narasimhan,
Geneviève Robin

February 28, 2018

Abstract

Statistical analysis of large data sets offers new opportunities to better understand many processes. Yet, data accumulation often implies relaxing acquisition procedures or compounding diverse sources. As a consequence, such data sets often contain mixed data, *i.e.* both quantitative and qualitative and many missing values. Furthermore, aggregated data present a natural *multilevel* structure, where individuals or samples are nested within different sites, such as countries or hospitals. Imputation of multilevel data has therefore drawn some attention recently, but current solutions are not designed to handle mixed data, and suffer from important drawbacks such as their computational cost. In this article, we propose a single imputation method for multilevel data, which can be used to complete either quantitative, categorical or mixed data. The method is based on multilevel singular value decomposition (SVD), which consists in decomposing the variability of the data into two components, the between and within groups variability, and performing SVD on both parts. We show on a simulation study that in comparison to competitors, the method has the great advantages of handling data sets of various size, and being computationally faster. Furthermore, it is the first so far to handle mixed data. We apply the method to impute a medical data set resulting from the aggregation of several data sets coming from different hospitals. This application falls in the framework of a larger project on Trauma patients. To overcome obstacles associated to the aggregation of medical data, we turn to distributed computation. The method is implemented in an R package.

Keywords: matrix completion, systematically and sporadically missing values, distributed computation.

1 Introduction

Consider a dataset $Y \in \mathbb{R}^{n \times p}$ which is naturally the row concatenation of K smaller datasets $Y_k \in \mathbb{R}^{n_k \times p}$, $k \in \{1, ..., K\}$. Y collects the measurements of p variables across a population of n individuals categorized in K groups, such that the k-th group contains n_k individuals and $\sum_{k=1}^{K} n_k = n$:

$$Y = \begin{pmatrix} \underline{Y_1} \\ \underline{Y_2} \\ \vdots \\ Y_K \end{pmatrix} \begin{matrix} \updownarrow n_1 \\ \updownarrow n_2 \\ \vdots \\ \updownarrow n_K \end{matrix} .$$

For a group $k \in \{1, ..., K\}$, an individual of the k-th group $i_k \in \{1, ..., n_k\}$ and a variable $j \in \{1, ..., p\}$, we denote by $y_{k,i_k,j}$ the value of variable j taken by individual i_k in group k. Such structure is often called *multilevel structure*, and occurs in many fields of applications. Famous examples include pupils nested within schools or patients within hospitals. Throughout this article, we focus on this latter example with a running application in public health. If some entries of Y are missing, we denote by M the indicator matrix of observations, with $M_{k,i_k,j} = 1$ if $y_{k,i_k,j}$ is observed and $M_{k,i_k,j} = 0$ otherwise. To handle missing values, corresponding to $M_{k,i_k,j} = 0$, a popular approach consists in imputing them, *i.e.* replacing the missing entries with plausible values to get a completed data set.

To do so, several approaches have been developed, and a complete overview of state of the art methods is available in Audigier et al. (2018). Latest proposals have focused on handling both sporadically missing values, which correspond to some entries missing for some variables, and systematically missing values where some variables are completely unobserved in one or more groups. To take into account the hierarchical structure of

the data, most imputation methods are based on random effects regression models, such as suggested by Resche-Rigon and White (2016) and Quartagno and Carpenter (2016). However, current solutions suffer from important gaps that deserve further development. In particular, they are not designed to handle mixed data (quantitative and categorical), struggle with large dimensions and are extremely costly in terms of computations.

In the same time, imputation by iterative SVD algorithms have proven excellent imputation capacities for quantitative (Hastie et al., 2015), qualitative (Audigier et al., 2017) and mixed data (Audigier et al., 2016). This can be explained in part because they assume an underlying low-rank structure for the data which is plausible for many large data sets, as discussed in Udell and Townsend (2017). These methods behave particularly well compared to competitors in terms of prediction of the missing values, in particular when the number of observations is small with respect to the number of variables, and when the qualitative variables have many categories and some of them are rare. In addition, they are often competitive in terms of execution time. However, these methods are not dedicated to the multilevel data we address in this paper. The work we present here can be casted as an extension of imputation methods based on SVD to the multilevel framework.

The paper is organized as follows. In Section 2, we start by presenting multilevel component methods to analyze quantitative, categorical and mixed data in the complete case where all entries are observed. We begin in Section 2.1 by reviewing the multilevel simultaneous component analysis (MLSCA) of Timmerman (2006), dedicated to quantitative data, which operates by estimating principal directions of variability for both levels of variability, *i.e.* for the between groups variability and for the within groups variability. Then, our first main contribution is to derive in Section 2.2 and Section 2.3 two multilevel component methods to analyze qualitative and mixed data respectively. To the best of our knowledge, we are the first to propose such methods. Our second main contribution is to propose in Section 3 multilevel single imputation methods to impute categorical and mixed variables with a multilevel structure. In Section 4 we show on synthetic data that our methods have smaller prediction errors than competitors when the data are generated with a multilevel model. Finally, in Section 5, we illustrate the methods with the imputation of a large register from Paris hospitals. The methods are implemented in the R (R Core Team, 2017) package missMDA (Josse and Husson, 2016).

2   Multilevel component methods

2.1   Multilevel Principal Component Analysis (MLPCA)

For sake of clarity, we start by recalling the multilevel extension of principal component analysis (PCA, Pearson (1901)) described in Timmerman (2006). Assume the data set Y contains quantitative variables only. The measured values can be decomposed, for a group $k \in \{1, ..., K\}$, an individual $i_k \in \{1, ..., n_k\}$ in the k-th group and a variable $j \in \{1, ..., p\}$, as

$$y_{k,i_k,j} = \underbrace{y_{.,.,j}}_{\text{offset}} + \underbrace{y_{k,.,j} - y_{.,.,j}}_{\text{between}} + \underbrace{y_{k,i_k,j} - y_{.,k,j}}_{\text{within}}.$$

Here,

$$y_{.,.,j} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i_k=1}^{n_k} y_{k,i_k,j}$$

is the overall mean of variable j and

$$y_{k,.,j} = \frac{1}{n_k} \sum_{i_k=1}^{n_k} y_{k,i_k,j}$$

is the mean of variable j among individuals of group k. Then, $(y_{k,.,j} - y_{.,.,j})$ is the deviation of group k to the overall mean of variable j, and $(y_{k,i_k,j} - y_{k,.,j})$ is the deviation of individual i_k to the mean of variable j in group k. Written in matrix form, this gives

$$Y = 1_n m^\top + Y_b + Y_w,$$

where 1_n is the $n \times 1$ vector of ones and m is the $p \times 1$ vector containing the overall means of the p variables, Y_b contains the variable means per group minus the overall means, and Y_w contains the residuals. Similarly to what is done in analysis of variance, we can split the sum of squares for each variable j as

$$\sum_{k=1}^{K} \sum_{i_k=1}^{n_k} y_{k,i_k,j}^2 = \sum_{k=1}^{K} n_k y_{.,.,j}^2 + \sum_{k=1}^{K} n_k (y_{k,.,j} - y_{.,.,j})^2 + \sum_{k=1}^{K} \sum_{i_k=1}^{n_k} (y_{k,i_k,j} - y_{k,.,j})^2.$$

In the classical framework where there is no multilevel structure, PCA yields the best fixed rank estimator of Y in terms of the least squares criterion. The multilevel extension naturally leads, for $(k, i_k, j) \in \{1, \ldots, K\} \times$

$\{1, \ldots, n_k\} \times \{1, \ldots, p\}$, to modelling the offsets, the between and within terms separately by explaining as well as possible both the between and within sum of squares. Therefore, multilevel PCA (MLPCA) consists in assuming two low-rank models, for the between matrix $Y_b = (y_{k,.,j} - y_{.,.,j})_{k,j}$ - that we approximate by a matrix of rank Q_b, and for the within matrix $Y_w = (y_{k,i_k,j} - y_{k,.,j})_{k,i_k,j}$ - that we approximate by a matrix of rank Q_w. This yields the following decomposition:

$$Y = 1_n m^\top + F_b V_b^\top + F_w V_w^\top + E, \tag{1}$$

where F_b is the matrix of size $n \times Q_b$ containing the between component scores:

$$F_b = \begin{pmatrix} \frac{F_{b,1}}{F_{b,2}} \\ \frac{\vdots}{F_{b,K}} \end{pmatrix}, \tag{2}$$

where for all $k \in \{1, \ldots, K\}$, $F_{b,k}$ is row-wise constant, with $f_{b,k}$ repeated on every row. Let $I_k \in \{0, 1\}^n$ be the indicator vector of group k such that the i-th entry $I_{k,i} = 1$ if individual i belongs to group k and 0 otherwise. Representation (2) is equivalent to

$$F_b = \sum_{k=1}^K I_k f_{b,k}^\top.$$

Finally, V_b is the $p \times Q_b$ between loadings matrix, F_w ($n \times Q_w$) denotes the within component scores, V_w ($p \times Q_w$) denotes the within loadings matrix, and E ($n \times p$) denotes the matrix of residuals. Note that in this model, the within loadings matrix V_w is constrained to be constant across groups. Model (1) is called multilevel simultaneous component analysis (MLSCA) in Timmerman (2006). We keep the name MLPCA for simplicity.

In terms of interpretation, the low rank structure on the between part implies that there are dimensions of variability to describe the hospitals: for instance the first dimension could oppose hospitals that resort to a large extent to pelvic and chest X-ray to hospitals where those examinations are not usually performed. The low rank structure on the within part implies that there are dimensions of variability to describe the patients: for instance the first dimension opposes patients with a head trauma (taking specific values for variables related to head trauma) to other patients. The constraint that the within loading matrix is the same across hospitals means that this dimension is the same from one hospital to the other but the strength of the dimension, *i.e.* the variability of patients on the dimension, can differ from one group to the other. This constraint also leads to less parameters to estimate.

The model is fitted by solving the least squares problem with respect to the parameter $\Theta = (m, F_b, V_b, F_w, V_w)$:

$$\begin{aligned} \hat{\Theta} \in \mathrm{argmin}_\Theta \quad & \left\| Y - \left(1_n m^\top + F_b V_b^\top + F_w V_w^\top \right) \right\|^2, \\ \text{such that} \quad & F_b = \sum_{k=1}^K I_k f_{b,k}^\top, \ \sum_{k=1}^K n_k f_{b,k} = 0_{Q_b}, \ 1_n^\top F_w = 0_{Q_w}, \end{aligned} \tag{3}$$

where the last two constraints serve for identifiability. The problem is separable, and the solution is obtained in Timmerman (2006) by computing the variables means to estimate m, the matrix of means per group centered by the overall mean Y_b and the within matrix Y_w of the data centered per group. Then, the truncated singular value decomposition (SVD) of $Y_b = U_b \Lambda_b^{1/2} V_b^\top$ at rank Q_b and of $Y_w = U_w \Lambda_w^{1/2} V_w^\top$ at rank Q_w are performed to estimate the parameters $(F_b = U_b \Lambda_b^{1/2}, V_b, F_w = U_w \Lambda_w^{1/2}, V_w)$. Such a solution is in agreement with the rationale of performing an SVD on the matrix of means per group to study the differences between groups and a SVD of the matrix centered by groups to study the differences between patients after discarding the hospital effects.

2.2 Multilevel Multiple Correspondence Analysis (MLMCA)

We now propose a new counterpart of MLPCA to analyse categorical variables. Our method is based on multiple correspondence analysis (MCA, Greenacre and Blasius (2006); Husson et al. (2017)), that we extend to handle multilevel structures. MCA is considered to be the counterpart of PCA for categorical data analysis, and has been successfully applied in many fields of applications, such as survey data analysis, to visualize associations between categories. More precisely, categorical data are coded as a complete disjunctive table Z where all categories of all variables are represented as indicator vectors. In other words $z_{ic} = 1$ if individual i takes the

category c and 0 otherwise. For example, if there are $p = 2$ variables with 2 and 3 levels respectively, we have the following equivalent codings:

$$Y = \begin{pmatrix} 1 & 1 \\ 2 & 3 \\ 1 & 2 \\ 2 & 3 \\ 2 & 2 \\ 2 & 2 \end{pmatrix} \iff Z = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \end{pmatrix}.$$

For $1 \le j \le p$ we denote by C_j the number of categories of variable j, and $C = \sum_{j=1}^{p} C_j$ the total number of categories. For $1 \le c \le C$, $Z_{,c}$ is the c-th column of Z corresponding to the indicator of category c. We define $\pi_c = n^{-1} 1_n^\top Z_{,c}$ the proportion of observations in category c, $\pi = (\pi_1, \ldots, \pi_C)^\top$ and D_π the $C \times C$ diagonal matrix with π on its diagonal. Multiple correspondence analysis (MCA) is defined as the SVD of the matrix

$$A = \frac{1}{np} \left(Z - 1_n \pi^\top \right) D_\pi^{-1/2}. \tag{4}$$

This specific transformation endows MCA with many properties: the distances between the rows and columns in the transformed matrix A coincide with the chi-squared distances, the first principal component (the scores) is the quantitative variable most related to the categorical variables in the sense of the η^2 coefficient of analysis of variance (Husson et al., 2017, Section 3). This latter property justifies why MCA is considered as the equivalent of PCA for categorical data. MCA is also strongly connected to a multilogit-bilinear model (Fithian and Josse, 2017).

We introduce the following strategy for multilevel MCA (MLMCA). From the indicator matrix of dummy variables Z, we start by defining a between part and a within part. MCA, in the sense of the SVD of a transformed matrix (4), will then be applied on each part. For $k \in \{1, ..., K\}$, define Z_k the sub-matrix of Z containing all categories and the rows corresponding to individuals of group k. The between part is defined block-wise as the mean of the indicator matrix per group k with the following $n_k \times p$ matrices, stacked below one another:

$$Z_{b,k} = n_k^{-1} 1_{n_k} 1_{n_k}^\top Z_k.$$

The entries of $Z_{b,k}$ contain the proportion of observations taking each category in group k (n_{c_k}/n_k) (for instance the proportion of individuals carrying some disease in a particular hospital). Finally

$$Z_b = \begin{pmatrix} \dfrac{Z_{b,1}}{Z_{b,2}} \\ \hline \vdots \\ \hline Z_{b,K} \end{pmatrix}.$$

MCA (4) is afterwards applied to the fuzzy indicator matrix Z_b, i.e. SVD is applied to

$$(Z_b - 1_n \pi^\top) D_\pi^{-1/2}.$$

This results in obtaining between component scores $F_b \in \mathbb{R}^{n \times Q_b}$ and between loadings $V_b \in \mathbb{R}^{n \times Q_b}$. The estimated between matrix is then $\hat{Z}_b = F_b V_b^\top D_\pi^{1/2} + 1_n \pi^\top$. As for the within part, MCA is applied to the data where the between part has been swept out, i.e. SVD is applied to the following matrix:

$$(Z - Z_b) D_\pi^{-1/2}. \tag{5}$$

Weighting by the inverse square root of the margins of the categories implies that more weight is given to categories which are rare over all groups (for instance a rare disease). We obtain within component scores $F_w \in \mathbb{R}^{n \times Q_w}$, within loadings $V_w \in \mathbb{R}^{n \times Q_w}$, and the estimated within matrix $\hat{Z}_w = F_w V_w^\top D_\pi^{1/2}$.

Finally, we estimate Z by $\hat{Z} = \hat{Z}_b + \hat{Z}_w$. As with MCA (Josse et al., 2012), the reconstructed fuzzy indicator matrix $\hat{Z} = \hat{Z}_b + \hat{Z}_w$ has the property that the sum of values for one individual and one variable is equal to one. Consequently, the estimated values can be considered as degrees of membership to the categories. This property will prove useful for the imputation.

Remark Another approach to define MLMCA would have been to directly apply MLPCA on the matrix A (4). It turns out that the two strategies are equivalent which strengthens this method to be defined as Multilevel MCA.

2.3 Multilevel Factorial Analysis of Mixed Data (MLFAMD)

Consider now a mixed data set $Y = (Y_q, Y_c)$, where Y_q is a submatrix containing p_q quantitative variables, and Y_c a submatrix containing p_c categories:

$$Y = \begin{pmatrix} 0.3 & -3.4 & 0.1 & 0 & 1 & 0 & 1 & 0 \\ 1.4 & 0.4 & -2.8 & 1 & 0 & 0 & 0 & 1 \\ 9.2 & 1.8 & 7.1 & 0 & 1 & 1 & 0 & 1 \end{pmatrix}.$$
$$\underbrace{\hphantom{0.3 \quad -3.4 \quad 0.1}}_{Y_q} \qquad \underbrace{\hphantom{0 \quad 1 \quad 0 \quad 1 \quad 0}}_{Y_c}$$

In the same flavour, we define a multilevel method for mixed data by extending a counterpart of PCA for mixed data, namely factorial analysis for mixed data (FAMD), presented in Pagès (2014). FAMD consists in transforming the categorical variables as in MCA (4) and concatenating them with the quantitative variables. Then, each quantitative variable is standardized (centered and divided by its standard deviation). Finally, SVD is applied to this weighted matrix. This specific weighting ensures that all quantitative and categorical variables play the same role in the analysis. More precisely, the principal components, denoted F_q for $q = 1, ..., Q$ maximize the link between the quantitative and categorical variables in the following sense:

$$F_q = \arg\max_{F_q \in \mathbb{R}^n} \sum_{j=1}^{p_q} r^2(F_q, Y_j) + \sum_{jc=1}^{p_c} \eta^2(F_q, Y_{jc}),$$

with the constraint that F_q is orthogonal to $F_{q'}$ for all $q' < q$ and with Y_j being the variable j, r^2 the square of the correlation coefficient and η^2 the square of the correlation ratio. This formulation highlights that FAMD can be seen as the counterpart of PCA for mixed data. More details about the method are given in Pagès (2014).

The extension to a multilevel structure, named MLFAMD, is now straightforward following what is done for MCA and categorical data in the previous section. Denote C the number of categories, $\pi \in (0,1)^C$ the vector of categories proportions and D_π the $C \times C$ diagonal matrix containing π on its diagonal. Denote $m \in \mathbb{R}^{p_q}$ the vector of means of the quantitative variables, and $\Sigma \in \mathbb{R}^{p_q \times p_q}$ the diagonal matrix containing the standard deviations of Y_q. MLFAMD consists in doing the following transformations.

$$W \in \mathbb{R}^{n \times (p_q + p_c)} \leftarrow \left((Y_q - 1_n m^\top) \Sigma^{-1}, \frac{1}{np} \left(Y_c - 1_n \pi^\top \right) D_\pi^{-1/2} \right). \tag{6}$$

Then, multilevel SVD is performed on the matrix W. This boils down to computing the between and within part, and performing SVD on both separately:

$$W_b = \sum_{k=1}^{K} 1_{n_k} 1_{n_k}^\top W, \quad W_w = W - W_b. \tag{7}$$

3 Multilevel imputation

3.1 Imputation with MLPCA

We now focus on the case where some values in Y are missing. Recall that M is the $n \times p$ indicator matrix of observations with $M_{k,i_k,j} = 1$ if $y_{k,i_k,j}$ is observed and $M_{k,i_k,j} = 0$ otherwise. We denote by M_k the restriction of matrix M to the rows belonging to group $k \in \{1, ..., K\}$. Consider a Missing (Completely) At Random (M(C)AR) setting (Little and Rubin, 2002) where the process that generated the missing values can be ignored. Low rank matrix estimation was extended for incomplete data sets (Hastie et al., 2015) by weighting the original least squares criterion. Using the notation $\Theta = (m, F_b, V_b, F_w, V_w)$, the optimization problem is the following, with $\odot$ denoting the Hadamard product:

$$\Theta \in \text{argmin}_\Theta \quad \left\| M \odot \left(Y - (1_n m^\top + F_b V_b^\top + F_w V_w^\top) \right) \right\|_2^2,$$
$$\text{such that} \quad F_b = \sum_{k=1}^{K} I_k f_{b,k}^\top, \ \sum_{k=1}^{K} n_k f_{b,k} = 0_{Q_b}, \ 1_n^\top F_w = 0_{Q_w}. \tag{8}$$

In Josse et al. (2013), the authors solved such a program using an iterative imputation algorithm. Note that the aim in Josse et al. (2013) was to perform MLPCA with missing values, *i.e. estimate the parameters*, and not to *impute multilevel data*. The distinction may appear tenuous as the algorithm involves an underlying imputation of the missing entries, but the quality of this imputation was never evaluated in itself. Let m_0 be the mean vector of the non-missing entries. The algorithm works iteratively as described in Algorithm 1. The detailed

---
Algorithm 1: Iterative MLPCA
---

0. **Initialize** missing values: $\hat{Y} = Y \odot M + 1_n m_0^\top \odot (1 - M)$.

1. Estimate F_b, V_b, F_w, V_w with multilevel PCA (3);

2. Impute $\hat{Y} = Y \odot M + (1_n m^\top + F_b V_b^\top + F_w V_w^\top) \odot (1 - M)$;

3. Update means $m = n^{-1} 1_n^\top \hat{Y}$.

Repeat steps 1, 2, 3 until empirical stabilization of the prediction.

---

algorithm for iterative MLPCA with missing values is given in Algorithm 2 below.

In the end, the algorithm outputs both the between and within scores and loadings obtained from the incomplete dataset, as well as a dataset imputed with the MLPCA model (1). Thus, it is a single imputation method (Schafer, 1997; Little and Rubin, 2002) which takes into account the multilevel structure of the data. Note also that the algorithm corresponds to an expectation-maximization (EM) algorithm of the multilevel model (1) assuming gaussian noise. To prevent overfitting, the SVD step is replaced by regularized SVD, *i.e.* where the singular values are shrunk (Verbanck et al., 2013; Josse et al., 2017), as discussed in Section 3.2.

3.2 Imputation with MLMCA and MLFAMD

We define iterative imputation algorithms for multilevel MCA and multilevel FAMD in Algorithm 3. Note that we use an accelerated version of the algorithm where each part, the between and the within parts, are not updated simultaneously but one at a time. This corresponds to what is known as a generalized EM step, where the least-squares criterion is decreased at each step, but not minimized.

Note also that these methods require to select two tuning parameters, namely the number of between and within components Q_b and Q_w from an incomplete data set. This is far from being trivial especially in the case of categorical variables as, even in the complete case and without multilevel structure, not many options are available. Consequently, we advocate the use of cross-validation to select these components.

Furthermore, to prevent overfitting we actually perform a regularized SVD where singular values are shrunk. Let $\lambda_l, 1 \le l \le Q_b$, and $\nu_q, 1 \le q \le Q_w$, be the ordered singular values of W_b and W_w, defined in (7), respectively. Let $\hat{\sigma}_b^2 = 1/(K - Q_b) \sum_{s=Q_b+1}^K \lambda_s$ and $\hat{\sigma}_w^2 = 1/(p - Q_w) \sum_{s=Q_w+1}^p \nu_s$. We shrink the singular values as follows:

$$(\lambda_1, \ldots, \lambda_{Q_b}) \leftarrow \left(\frac{\lambda_1 - \hat{\sigma}_b^2}{\sqrt{\lambda_1}}, \ldots, \frac{\lambda_{Q_b} - \hat{\sigma}_b^2}{\sqrt{\lambda_{Q_b}}} \right),$$

$$(\nu_1, \ldots, \nu_{Q_w}) \leftarrow \left(\frac{\nu_1 - \hat{\sigma}_w^2}{\sqrt{\nu_1}}, \ldots, \frac{\nu_{Q_w} - \hat{\sigma}_w^2}{\sqrt{\nu_{Q_w}}} \right).$$

Finally, the algorithms we present in this paper can be implemented in parallel across groups, providing that groups share their mean values, standard deviations, sample sizes, and right singular vectors. The procedure is described in Appendix A.

4 Simulation study

4.1 Imputation of multilevel quantitative data

We conducted a comparative simulation study to contrast the performances of the multilevel imputation with PCA (MLPCA) to other single imputation methods, namely

1. mean imputation which consists in imputing by the mean of each variable, used as a benchmark method;

2. a separate PCA imputation where each group is imputed independently, using the R package missMDA (Josse and Husson, 2016);

3. a global imputation by PCA (which ignores the multilevel structure and the group variable) using the R package missMDA (Josse and Husson, 2016);

4. imputation with iterative conditional random effects regression models as implemented in the R package mice (van Buuren, 2012);

Algorithm 2: Multilevel PCA with missing values

Data: $Y = (Y_{\text{obs}}, Y_{\text{mis}}) \in \mathbb{R}^{N \times p}$, Q_b, Q_w
Result: $Q \in \mathbb{R}^{n \times k}$, $R \in \mathbb{R}^{p \times k}$, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k$
Initialization: $\hat{m} = (1/K)1_K'Y$; $W = \text{diag}(\sqrt{K_j})$.

    for $(i,j) \in \{1, n\} \times \{1, p\}$ **do**
        if $M_{ij} = 0$ **then**
            $Y_{ij} \leftarrow \hat{m}_j^0$
        end
    end

1 **repeat**
2   **Estimation of the between structure**
3 $Y^b = \sum_{k=1}^{N} N_k^{-1} I_k \left(1'_{N_k} Y_k - \hat{m}^{0'} \right)$
4 $WY^b = FDB'$
5 $U_b \leftarrow W^{-1}F[,1:Q_b]D[1:Q_b,1:Q_b]$; $V_b \leftarrow B[,1:Q_b]$
6 $\hat{Y}^b = U_b V_b'$
7   **Estimation of the within structure**
8 $Y^w = Y - 1_n\hat{m}' - Y^b$
9 $Y^w = FDB'$
10 $U_w \leftarrow F[,1:Q_w]D[1:Q_w,1:Q_w]$; $V_w \leftarrow B[,1:Q_w]$
11 $\hat{Y}^w = U_w V_w'$
12   **Imputation of the missing values**
13 $\hat{Y} = 1_n\hat{m}'\hat{Y}^b + \hat{Y}^w$
14 $Y \leftarrow M \odot Y + (1_n 1'_p - M) \odot \hat{Y}$
15 $\hat{m} = n^{-1}1'_n Y$
16 **until** *convergence*;

---

5. imputation by a joint model based on random effects models as implemented in the R package jomo (Quartagno and Carpenter, 2017);

6. imputation with iterative random forest (RF) as implemented in the R package missForest, (Stekhoven and Bühlmann, 2012).

Note that methods 4 and 5 are considered as the references to impute multilevel quantitative data (Audigier et al., 2018). The imputation based on random forests is not specifically designed to handle a multilevel structure, but is expected to perform well in such a hierarchical setting. Indeed, random forests can account for interactions between variables, and therefore in particular for interactions between the categorical variable indicating the group and the other variables. This is another way of handling the multilevel structure. Finally, even though we focus here on quantitative variables, we also added imputation method for mixed data with FAMD (Audigier et al., 2016), where the group membership is used as a categorical variable. This is a way to take into account the hierarchical structure of the data.

We first simulate data according to the multilevel model (1) with Gaussian noise and set the number of between and within components to 2. For MLPCA, global PCA and FAMD, we select the number of components resulting in the smallest errors. This corresponded to $Q_b = 2$ and $Q_w = 2$ for MLPCA, and to 4 dimensions for the global PCA and global FAMD. We start with $n_k = 200$ observations per group k and we vary the number of groups K (3, 5), the number of variables J (5, 10, 30), the intensity of the noise ($\sigma = 1, 2$) and the percentage of missing values (10%, 20%, 40%), which are missing completely at random (MCAR). The detail is available in the associated code provided as supplementary material. We then compute the mean squared error (MSE) of prediction, and repeat the process 100 times. We use default parameters for the other methods. Figure 1 is representative of many results where multilevel imputation MLPCA improves both on global PCA imputation and separate PCA imputation but also on competitors. We have not included the results from the package mice as, using the default parameters, we encountered too many errors. It may be explained by the size of the data set, as the method does not behave well when there are not too many variables. More tuning is surely required to use the mice package seamlessly.

Algorithm 3: Iterative MLMCA and iterative MLFAMD
---

(0) **Initialization**

    (a) Initialize missing values: mean imputation for quantitative data, proportion imputation for dummy variables.

    (b) Compute weights, standard deviations and column margins.

(1) **Repeat until convergence:**

    (a) Estimate parameters (with MLFAMD or MLMCA)

    (b) Impute the missing entries with fitted values

    (c) Update means, standard deviations, column margins.[a]

---

    [a]After each imputations, the means and standard deviations are modified. Hence we need to recenter and rescale the data.
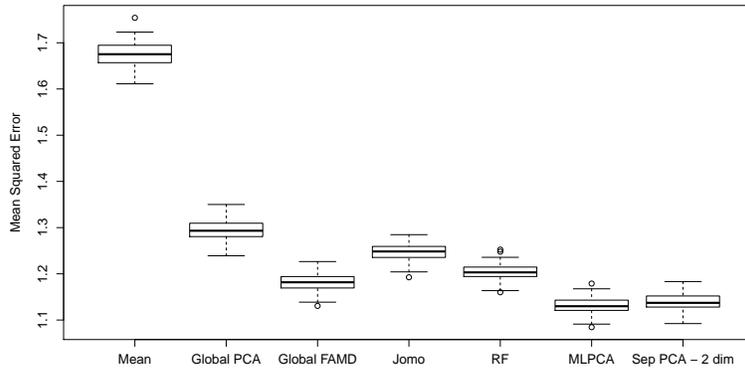

Figure 1: MSE of prediction for a data with $J = 10$ variables, $K = 5$ groups, $n_k = 200$ observations per group and 20 % of missing values completely at random. MLPCA is performed with $Q_b = 2$ and $Q_w = 2$.

We summarize here our main findings with respect to all the simulations carried out. Imputations with random forests and FAMD often perform similarly with a slight advantage for FAMD especially when the percentage of missing values is larger. Imputation with jomo encounters many difficulties when the number of variables increases as well as when the noise increases. Finally imputation based on separate PCA collapses when the percentage of missing values increases. The multilevel imputation is always the most accurate. This is expected, but still reassuring, as the data are simulated according to a multilevel model. We simulated data without a multilevel structure and the performances of multilevel PCA are only slightly lower than the ones of the global PCA.

All the methods have of course their strengths and weaknesses and the properties of an imputation method depend on the inherent characteristics of the method, an imputation method based on linear relationships provides good prediction for data with strong linear relationships, contrary to imputation using random forests which are designed for non-linear relationships. Since the structure of the data is not known in advance, one could use cross-validation and select the method which best predicts the removed entries.

Figure 2 represents the differences, for each group, between imputing with a separate PCA and with MLPCA. The improvement of a multilevel imputation over a separate imputation differs from one study to the other but still groups have interest in using a multilevel imputation. Indeed, the results presented in Figure 2 reveal that in terms of predicting the missing entries, multilevel PCA yields better results that separate PCA for every group, thus showing that as far as imputation is concerned, all groups benefit from participating in the study. This justifies the use of distributed multilevel methods in contexts where there are confidentiality issues at stake, by quantifying how much the different centers gain in terms of imputation accuracy, as further discussed in Appendix A.
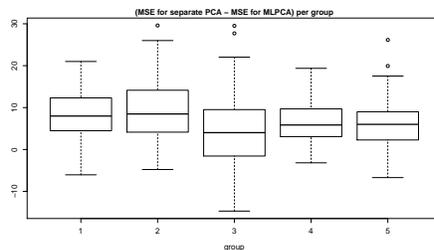
Figure 2: Difference between MSE obtained with separate PCA and with MLPCA for each group.

4.2 Imputation of multilevel mixed data

To simulate mixed data, we use the same design as for quantitative variables but cut some of the variables into categories. We vary the same parameters as for the quantitative variables but also the ratio of the number of quantitative over the number of categorical variables. Note that the methods implemented in the packages mice and jomo can handle mixed data when categorical variables are binary, but not when variables have more than two categories. This is why they are not included in the simulations. The global FAMD imputation is performed with 2, 4 and 6 dimensions whereas we also vary Q_b and Q_w for the multilevel method between 2 and 4. We display only the number of components which resulted in the lowest prediction error for each of the methods concerned. Figure 3 shows again that imputing with the multilevel method gives better results
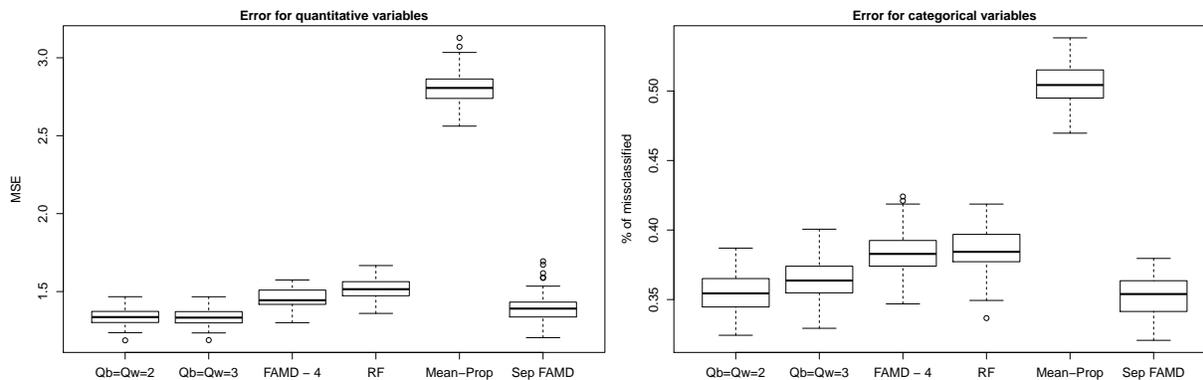

Figure 3: Data set with $J = 10$ variables, 5 quantitative and 5 categorical variables, 20% of missing values, $K = 5$ groups: on the left MSE for the quantitative variables; on the right percentage of misclassified for categorical variables. Multilevel FAMD is represented for different values of Q_b and Q_w whereas Global FAMD with 4 dimensions. RF is the imputation with random forest, Mean-Prop means that the imputation is done by the mean for quantitative variables and the proportion for categorical ones, and sep FAMD gives the results when separate FAMD are performed on each group.

than imputing with global FAMD or with random forests. This is true for the quantitative and the categorical variables. Note that imputation with multilevel FAMD is quite stable with respect to the number of between and within components. As far as the computational time is concerned, we compare in Table 1 the performances of the different approaches. There is a clear advantage of using SVD based imputation rather than other methods.

5 Hospital data analysis

5.1 Traumabase

Our work is motivated by an application in public health on polytraumatized patients for the Traumabase[1] group at APHP (Public Assistance - Hospitals of Paris). Effective and timely management of major trauma patients is critical to improve outcomes and survival, given the high risks for the patient in case of delays or errors. With

---

[1] http://www.traumabase.eu/fr_FR

	$J = 10$	$J = 30$	$J = 15$	$J = 35$
Global PCA	0.09	0.3		
jomo	11	282		
Multilevel FAMD	1.5	1.2	2	7
Global FAMD	0.4	0.7	1	4
Random forest	59	200	27	246

Table 1: Time in seconds for a dataset with 20% of missing values, $K = 5$ groups and $n_k = 200$ observations per groups, with 10 and 30 quantitative variables for the two left columns and with additional 5 categorical variables for the two right columns.

the perspective of improving the decision-making process and the care of patients, 8 French Trauma centers have decided to collaborate to collect detailed high quality clinical data from the scene of the accident to the exit of the hospital. The resulting database, the Traumabase, has up to now gathered more than 7495 trauma admissions data, and is permanently updated.

The data are highly heterogeneous, multi-source, and contain many missing values. Furthermore, experts expect hospitals to have an influence on some of the variables, due to lack of practice standardization, and because the patients and their social status differ from one hospital to the other. We analyse a portion of the initial data set containing 8 features identified by physicians as prone to hospital effects. The data set of interest therefore consists in 5 qualitative and 3 quantitative variables measured over 7495 patients, and contains around 11% of missing values; furthermore, there is at least one missing entry for 49% of the patients. There are certainly different generation mechanisms at work: some variables (such as the type of accident and the hospital center) are completely observed whereas the patterns of missingness of other variables (such as pelvic and lung X-ray) are believed to depend on the hospital center. In first approximation, a Missing At Random (MAR) mechanism - where the probability of missingness is allowed to depend on the observed variables - seems satisfying.

We focus on imputing of the Traumabase data with iterative MLFAMD with two aims. First, the imputed data can be further analyzed with other statistical methods such as predictive models, to predict some outcome of interest. However, care must be taken when analysing an imputed data set, as discussed in Section 6. Secondly, as discussed in Appendix A, it can incite hospitals to share their data and potentially lead to better care for all the patients.

5.2 Simulated imputation of the Traumabase

To assess the quality of imputation and legitimate the use of iterative MLFAMD to impute the Traumabase, we first perform simulations by inserting an additional of 10% of missing values to the data set, predicting them with the different imputation methods described in Section 4, and computing the mean squared error of prediction for quantitative variables and the percentage of misclassification for categorical variables. Figure 4 presents the results over 100 replications of the experiment. In terms of prediction of quantitative variables, multilevel FAMD and global FAMD perform similarly and improve on the random forest imputation. We observe the same behavior for the categorical variables, with multilevel FAMD improving only slightly on global FAMD. Note that results are dependent on the number of selected components.

6 Conclusion

We proposed a method dedicated to the imputation of multilevel mixed data based on an iterative SVD algorithm. To the best of our knowledge this is the first multilevel method available for mixed data. Directions of future research include the development of an automated method to estimate the number of components Q_b and Q_w. A first approach is for now to select Q_b and Q_w with cross-validation. We are also eager to investigate a multiple imputation procedure based on this multilevel component method, in order to further analyse the Traumabase data set, with predictive model to study for instance the occurence of diagnosis errors based on patients profiles. Multiple imputation is important to reflect the uncertainty associated to the predicted values. We also believe the multilevel methods we have developed for mixed data can be useful for exploratory analysis and visualization. Finally, as discussed, the methods presented in this paper can be implemented in parallel across groups or sites. A following project we are currently involved in consists in exploiting this property to implement a real-time distributed and privacy preserving platform, dedicated to the imputation of health care data partitioned across several hospitals, without having to aggregate the data. One issue with the distribution technique described in Appendix A is that we use iterative procedures, therefore after N iterations each hospital has shared N summary
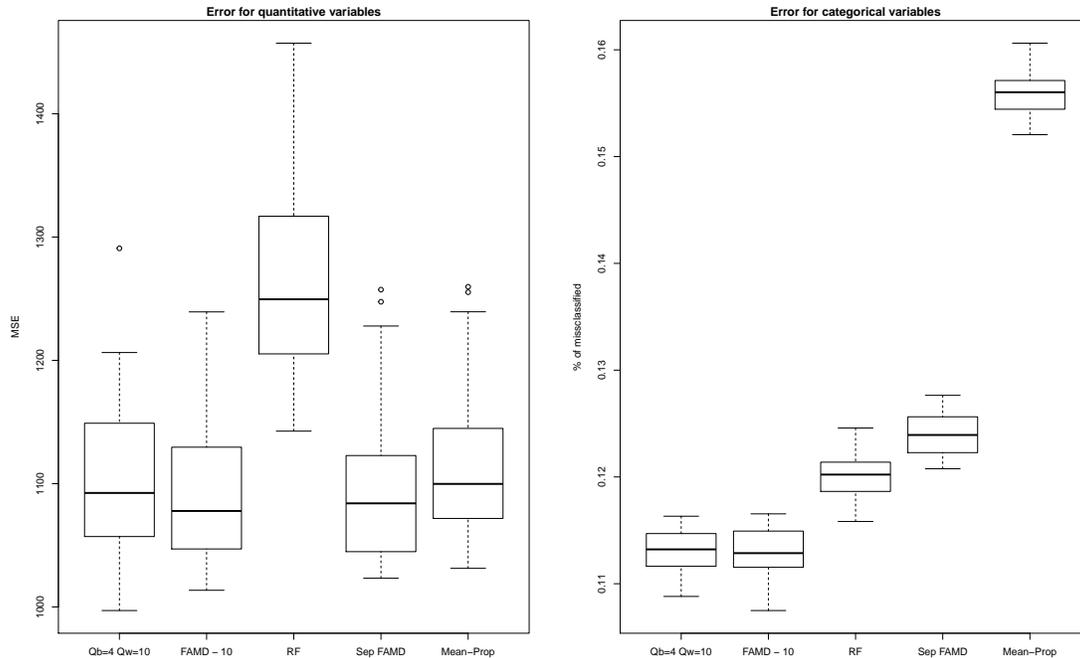
Figure 4: Traumabase: MSE of prediction and % of mis-classification.

statistics, which can lead to information leakage. A possible solution to this problem is to resort to homomorphic encryption (Gentry, 2009) which allows to perform computations on ecrypted data.

References

Audigier, V., Husson, F., and Josse, J. (2016). A principal component method to impute missing values for mixed data. *Advances in Data Analysis and Classification*, 10(1):5–26.

Audigier, V., Husson, F., and Josse, J. (2017). Mimca: multiple imputation for categorical variables with multiple correspondence analysis. *Statistics and Computing*, 27(2):501–518.

Audigier, V., White, I., Jolani, S., Debray, T., Quartagno, M., Carpenter, J., van Buuren, S., and Resche-Rigon, M. (2018). Multiple imputation for multilevel data with continuous and binary variables. *Statistcial Science*.

Fithian, W. and Josse, J. (2017). Multiple correspondence analysis & the multilogit bilinear model. *Journal of Multivariate Analysis*, 157:87–102.

Gentry, C. (2009). *A fully homomorphic encryption scheme*. PhD thesis, Stanford University.

Golub, G. H. and Van Loan, C. F. (1996). *Matrix Computations (3rd Ed.)*. Johns Hopkins University Press, Baltimore, MD, USA.

Greenacre, M. and Blasius, J. (2006). *Multiple Correspondence Analysis and Related Methods*. Chapman & Hall/CRC.

Hastie, T., Mazumder, R., Lee, J. D., and Zadeh, R. (2015). Matrix completion and low-rank svd via fast alternating least squares. *Journal of Machine Learning Research*, 16(1):3367–3402.

Husson, F., Lê, S., and Pagès, J. (2017). *Exploratory Multivariate Analysis by Example Using R*. Chapman & Hall/CRC, 2 edition.

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

Josse, J. and Husson, F. (2016). missMDA: A package for handling missing values in multivariate data analysis. *Journal of Statistical Software*, 70(1):1–31.

Josse, J., Timmerman, M. E., and Kiers, H. A. (2013). Missing values in multi-level simultaneous component analysis. *Chemometrics and Intelligent Laboratory Systems*, 129:21 – 32. Multiway and Multiset Methods.

Josse, J., Wager, S., and Sardy, S. (2017). denoiser: A package for low rank matrix estimation. *submitted*.

Little, R. J. A. and Rubin, D. B. (1987, 2002). *Statistical Analysis with Missing Data*. John Wiley & Sons series in probability and statistics, New-York.

Narasimhan, B., Rubin, L., Gross, S., Bendersky, M., and Lavori, P. W. (2017). Software for distributed computation on medical databases: A demonstration project. *Journal of Statistical Software*, 77(13):99–122.

Pagès, J. (2014). *Multiple Factor Analysis by Example Using R*. Chapman & Hall/CRC.

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Quartagno, M. and Carpenter, J. (2016). Multiple imputation for ipd metaanalysis: allowing for heterogeneity and studies with missing covariates. *Statistics in Medicine*, 35 (17):2938–2954.

Quartagno, M. and Carpenter, J. (2017). *jomo: A package for Multilevel Joint Modelling Multiple Imputation*.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Resche-Rigon, M. and White, I. R. (2016). Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Statistical Methods in Medical Research*.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall/CRC, London.

Stekhoven, D. and Bühlmann, P. (2012). Missforest - nonparametric missing value imputation for mixed-type data. *Bioinformatics*, 28:113–118.

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

Udell, M. and Townsend, A. (2017). Nice latent variable models have log-rank. *arXiv:1705.07474*.

van Buuren, S. (2012). *Flexible Imputation of Missing Data*. Chapman & Hall/CRC, Boca Raton.

Verbanck, M., Josse, J., and Husson, F. (2013). Regularised PCA to denoise and visualise data. *Statistics and Computing*, pages 1–16.

A   Distributed algorithm for iterative multilevel PCA

Having at disposal aggregated data (data from the different hospitals) is an opportunity to have more patients and to develop more relevant modelling. However, there are technical and social barriers to the aggregation of medical data. The size of combined databases often makes computations and storage intractable, while institutions are usually reluctant to share their data due to privacy concerns and proprietary attitudes. Both obstacles can be overcome by turning to distributed computations, which consists in leaving the data on sites and distributing the calculations, so that hospitals only share some intermediate results instead of the raw data (Narasimhan et al., 2017). Among other methods, SVD, which only involves inner products and sums, can be very straightforwardly implemented in a distributed manner. Consequently, one main advantage of the methods we present is that they can also be distributed across sites. More precisely, we consider a master-slave structure (Narasimhan et al., 2017) where a central server collects summary statistics computed locally on sites, as illustrated Figure 5.
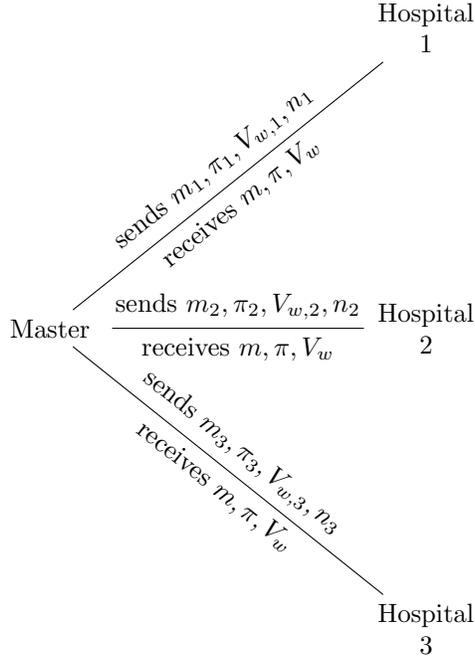

Figure 5: Master-slave distribution structure. The hospitals send their local means, proportions, sample size and right sigular vectors to the master. The master sends back the overall means, proportions, and right singular vectors to the hospitals.

We recall in Appendix B, how the power method (Golub and Van Loan, 1996), which computes the first left and right singular vectors of a matrix $Y \in \mathbb{R}^{n \times p}$, can be straightforwardly distributed over K different sites. This algorithm can then be used to perform a distributed rank-Q SVD, as shown in Algorithm 6. We take advantage of this property to develop a distributed version of the iterative PCA algorithm, presented in Algorithm 4. Plugged in Algorithm 1, Algorithm 4 leads to a distributed version of the iterative multilevel PCA algorithm.

B   Distributed rank-Q PCA

We start by reminding the power method (Golub and Van Loan, 1996), which computes the first left and right singular vectors of a matrix $Y \in \mathbb{R}^{n \times p}$. Without loss of generality, we assume $n \leq p$. Suppose $Y = U \Lambda^{1/2} V^\top$, $U = (u_1, \ldots, u_n)$, $V = (v_1, \ldots, v_n)$ and $\Lambda = \text{diag}(\lambda_1^2, \ldots, \lambda_n^2)$ $|\lambda_1| \geq |\lambda_2| \ldots \geq |\lambda_n|$. The power method is iterative and produces sequences of vectors $z^{(t)}$ and $q^{(t)}$ converging to u_1 and v_1 respectively, with iterations detailed in Algorithm 5. Let $q^{(0)}$ be a starting point satisfying $\|q^{(0)}\|_2 = 1$. The sequences $q^{(t)}$ and $z^{(t)}$ converge to u_1 and v_1 respectively, when $\langle q^{(0)}, u_1 \rangle \neq 0$ and $|\lambda_1| > |\lambda_2|$; the rate of convergence is dictated by the ratio $|\lambda_2|/|\lambda_1|$. This directly extends to the computation of the rank-Q SVD. One can actually estimate u_1, v_1 and λ_1, then f_2 by applying the same procedure to $Y - u_1 \lambda_1 v_1^\top$, and so on so forth. Moreover it is straightforward

---

Algorithm 4: Distributed iterative PCA

Data: $Y_k \in \mathbb{R}^{n_k \times p}$, Q_b, Q_w

Result: $\hat{m}$, F_b,V_b,F_w,V_w

1 Initialization: impute missing values with initial values;
  $(n \times p) = diag(\sqrt{n_k}).$;

2 $R = 0$, $\lambda = 0$

3 for $k = 1, \ldots K$ **do**

4 $\quad$ $F_k = 0$

5 $\quad$ transmit n_k to master

6 end

7 for $i = 1, \ldots, Q$ **do**

8 $\quad$ **for** $k = 1, \ldots, K$ **do**

9 $\quad\quad$ $q_k = (1, 1, \ldots, 1)$

10 $\quad$ **end**

11 $\quad$ $\|q\|_2 = \sqrt{\sum_{k=1}^{K} n_k}$

12 $\quad$ transmit $\|q\|_2$, V and λ to workers

13 $\quad$ **repeat**

14 $\quad\quad$ **for** $k = 1, \ldots, K$ **do**

15 $\quad\quad\quad$ $q_k = q_k / \|q\|_2$

16 $\quad\quad\quad$ $r_k = (Y_k - F_k V^\top)^\top q_k$

17 $\quad\quad\quad$ transmit r_k to master

18 $\quad\quad$ **end**

19 $\quad\quad$ $r = \sum_{k=1}^{K} r_k$ $r = r / \|r\|_2$ transmit r to workers

20 $\quad\quad$ **for** $k = 1, \ldots, K$ **do**

21 $\quad\quad\quad$ $q_k = Y_k r$

22 $\quad\quad\quad$ transmit $\|q_k\|_2$ to master

23 $\quad\quad$ **end**

24 $\quad\quad$ $\|q\|_2 = \sum_{k=1}^{K} \|q_k\|_2$

25 $\quad\quad$ transmit $\|q\|_2$ to workers

26 $\quad\quad$ $\lambda_i = \|q\|_2$

27 $\quad$ **until** *convergence*;

28 $\quad$ $V = $ combine by column $(V, \sqrt{\lambda_i} r)$

29 $\quad$ **for** $k = 1, \ldots, K$ **do**

30 $\quad\quad$ $F_k = $ combine by column (F_k, q_k)

31 $\quad$ **end**

32 end

---

to distribute this procedure when the data are grouped in K different sites with

$$Y = \begin{pmatrix} \dfrac{Y_1}{Y_2} \\ \vdots \\ \hline Y_K \end{pmatrix}.$$

Indeed, all the computations in Algorithm 5 can be done in parallel with a master-slave architecture, provided that the local right singular vectors v_j, $j \in \{1, \ldots, n\}$ are sent to the master. The corresponding algorithm is given in Algorithm 6, and leads exactly to applying the power method for rank-Q SVD to the entire data matrix Y. The procedure is implemented in the distcomp R package (Narasimhan et al., 2017). This algorithm can in turn be extended to perform distributed PCA with missing values, yielding the algorithm given in B.

Algorithm 5: Power method

---

1 for $t = 1, 2, \ldots$ **do**

2 $\quad z^{(t)} = Y^\top q^{(t-1)}$

3 $\quad z^{(t)} = z^{(t)} / \left\| z^{(t)} \right\|_2$

4 $\quad q^{(t)} = Y z^{(t)}$

5 $\quad \lambda^{(k)} = \left\| q^{(t)} \right\|_2$

6 $\quad q^{(t)} = q^{(t)} / \left\| q^{(t)} \right\|_2$

7 end

Algorithm 6: Distributed power method

Data: Workers private data $Y_k \in \mathbb{R}^{n_k \times p}$

Result: $F \in \mathbb{R}^{n \times Q}$, $V \in \mathbb{R}^{p \times Q}$, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_Q$

1 $F = 0$, $\lambda = 0$

2 for $k = 1, \ldots K$ **do**

3 $\quad F_k = 0$

4 $\quad$ transmit n_k to master

5 end

6 for $i = 1, \ldots, Q$ **do**

7 $\quad$ **for** $k = 1, \ldots K$ **do**

8 $\quad\quad q_k = (1, 1, \ldots, 1)$

9 $\quad$ **end**

10 $\quad \|q\|_2 = \sqrt{\sum_{k=1}^K n_k}$

11 $\quad$ transmit $\|q\|_2$, V and λ to workers

12 $\quad$ **repeat**

13 $\quad\quad$ **for** $k = 1, \ldots, K$ **do**

14 $\quad\quad\quad q_k = q^k / \|q\|_2$

15 $\quad\quad\quad r_k = (Y_k - F_k V^\top)^\top q_k$

16 $\quad\quad\quad$ transmit r_k to master

17 $\quad\quad$ **end**

18 $\quad\quad r = \sum_{k=1}^K r_k$

19 $\quad\quad r = r / \|r\|_2$

20 $\quad\quad$ transmit r to workers **for** $k = 1, \ldots, K$ **do**

21 $\quad\quad\quad q_k = Y_k r$

22 $\quad\quad\quad$ transmit $\|q_k\|_2$ to master

23 $\quad\quad$ **end**

24 $\quad\quad \|q\|_2 = \sum_{k=1}^K \|q_k\|_2$

25 $\quad\quad$ transmit $\|q\|_2$ to workers

26 $\quad\quad \lambda_i = \|q\|_2$

27 $\quad$ **until** *convergence*;

28 $\quad V = cbind(V, r)$

29 $\quad$ **for** $k = 1, \ldots, K$ **do**

30 $\quad\quad F_k = $ combine by column (F_k, q_k)

31 $\quad$ **end**

32 end