Estimation and imputation in PPCA with MNAR data

Aude Sportisse, Claire Boyer, Julie Josse

02 août 2019

Contents

This R Notebook aims to present the code of our paper (Sportisse, Boyer, and Josse 2019).

```
library("FactoMineR")
library("softImpute")
library("MASS")
library("gtools")
library("ggplot2")
library("gridExtra")
library("pracma")
```

```
source("PPCAMNAR_mainfunction.R")
source("PPCAMNAR_auxiliaryfunctions.R")
```

Let consider a simple case where $p = 10, r = 2$ and in which seven variables can be missing, fixed to be $Y_{.j}, j \in \{1, \ldots, 7\}$, under a self-masked MNAR mechanism. The noise level is fixed to $\sigma = 0.1$.

The PPCA model can be written as:

$$\begin{pmatrix} Y_{.1} & \ldots & Y_{.10} \end{pmatrix} = \mathbf{1} \begin{pmatrix} \alpha_1 & \ldots & \alpha_{10} \end{pmatrix} + \begin{pmatrix} W_{.1} & W_{.2} \end{pmatrix} B + \epsilon,$$

with $B \in \mathbb{R}^{2 \times 10}$ and $\epsilon \in \mathbb{R}^{n \times 10}$.

For the simulations, the mean vector $\begin{pmatrix} \alpha_1 & \ldots & \alpha_{10} \end{pmatrix}$ is set to zero.

```
set.seed(23) #change seed
n <- 1000 #number of observations.
p <- 10 #number of variables.
r <- 2 #number of latent variables.
sigma <- 0.1 #noise level (standart deviation).
mean_theo <- 0 #vector for the theoretical mean of the data matrix.
indcolNA <- c(1,2,3,4,5,9,10) #indexes of the missing variables.

#coefficient matrix.
B <- matrix(rnorm(p * r), nrow = r, ncol = p) #random

# Z = matrix(rnorm(p * r), nrow = p, ncol = r)  #uniform in the family of r orthogonal vectors in dim p
# sqrtZ = sqrtm(t(Z)%*%Z)
# inv_sqrtZ = inv(sqrtZ$Binv)
# ZZ = Z%*%inv_sqrtZ
# B = t(ZZ)
```

MNAR missing values are introduced using a logistic regression.

```
modmecha <- "Logistic"
```

The method proposed in our paper involve the selection of observed variables on which the regression will be performed leading to two approaches: * aggregation (agg): in which the final estimator is provided by computing the median of intermediate mean or variance estimators corresponding to several possible combinations of the observed variables. * random (noagg): the final estimator is built upon only one choice of fully observed variables, uniformly randomly drawn among all combinations of observed variables.

```
agg <- 1 #aggregation with median
noagg <- 0 #at random
```

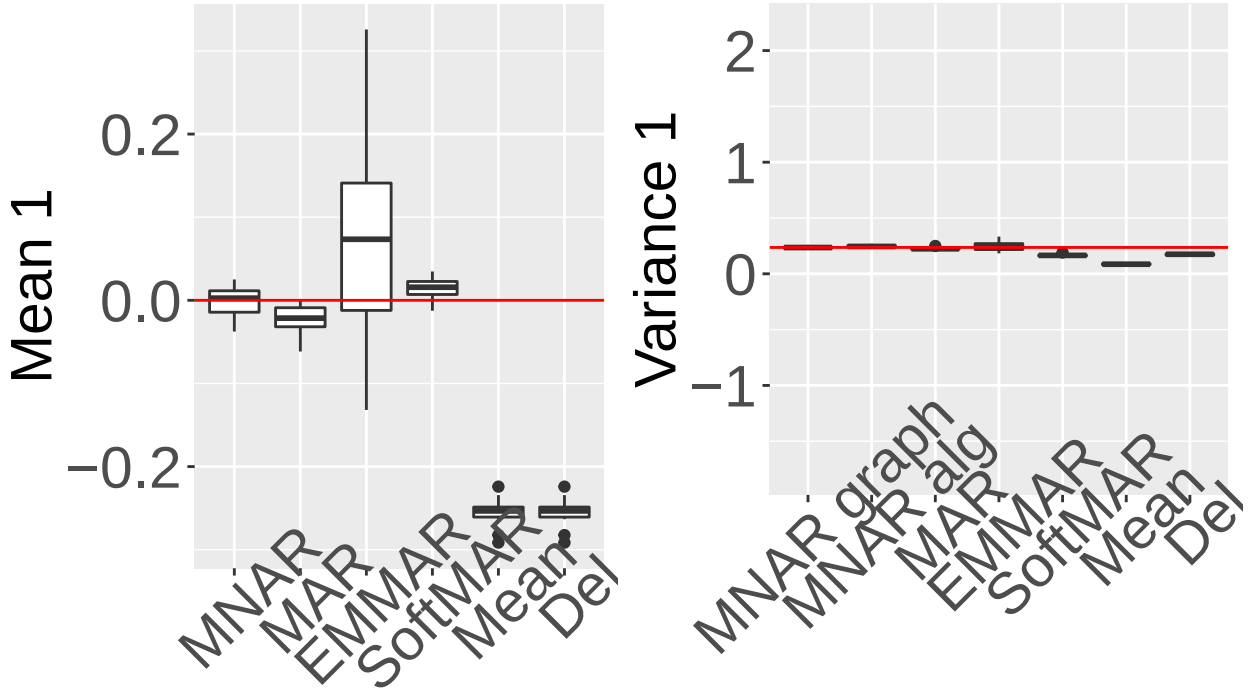
The main function is **ComparMethods_PPCA_iteration**, which gives the estimations for the mean, the variance and the covariances associated to each MNAR missing variable of the data matrix generated under the PPCA model using the coefficient matrix and the noise level. The main arguments are:

- seed.num: to fix the random number generator for each .
- n: number of observations.
- p: number of variables.
- r: number of latent variables.
- B: coefficient matrix.
- mean_theo: vector for the theoretical mean of the data matrix.
- sigma: noise level (standart deviation).
- indcolNA: indexes of the missing variables.
- nbNA: number of missing values (if modmecha=="Censor").
- modmecha: model for the MNAR mechanism: "Logistic" or "Censor".
- agg: 1 (default) to return the results for the graphical and algebraic MNAR methods by choosing the aggregation of the combinations of observable variables for the regressions, 0 otherwise.
- noagg: 1 to return the results for the graphical and algebraic MNAR methods by randomly choosing a combination of observable variables for the regressions, 0 (default) otherwise.
- simplify: FALSE if all the the combinations of observable variables shoud be considered for the graphical and algebraic MNAR methods, TRUE (default) if only a part of them is considered.
- r_reg: if simplify=TRUE, numnber of observable variables to use.

```
Nbit = 20
result = lapply(1:Nbit,ComparMethods_PPCA_iteration,n=n,
                  p=p,
                  r=r,
                  B=B,
                  mean_theo=mean_theo,
                  sigma=sigma,
                  indcolNA=indcolNA,
                  modmecha=modmecha,
                  agg = agg,
                  noagg = noagg,
                  simplify = FALSE)
```

Means and variances estimation

```
CovTheo <- t(B) %*% B + sigma ^ 2 * diag(1, ncol = p, nrow = p) #covariance matrix (theoretical) comput
j=1 #index of the missing variable for which we will obtain the graphics.
```

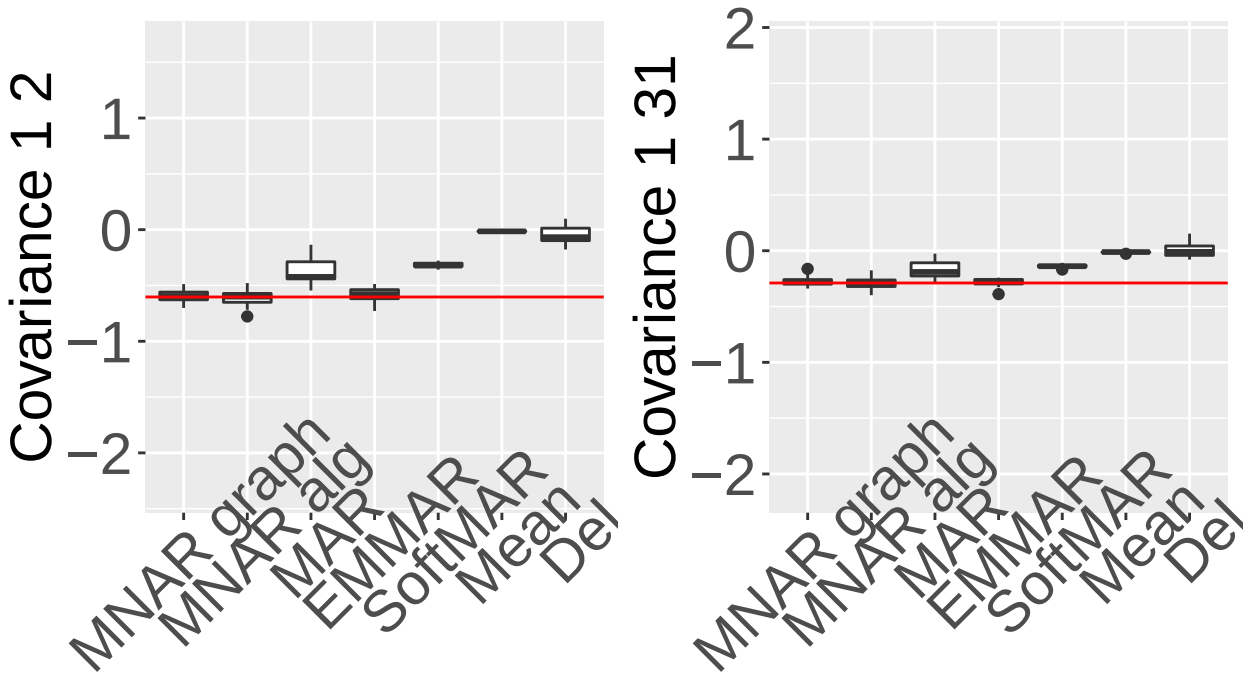


Covariances estimation

```
j=1 #index of the missing variable for which we will obtain the graphics.
l1=2 #index of the variable (observable or missing) for which we will obtain the graphics.
l2=3 #index of the ariable (observable or missing) for which we will obtain the graphics.
```

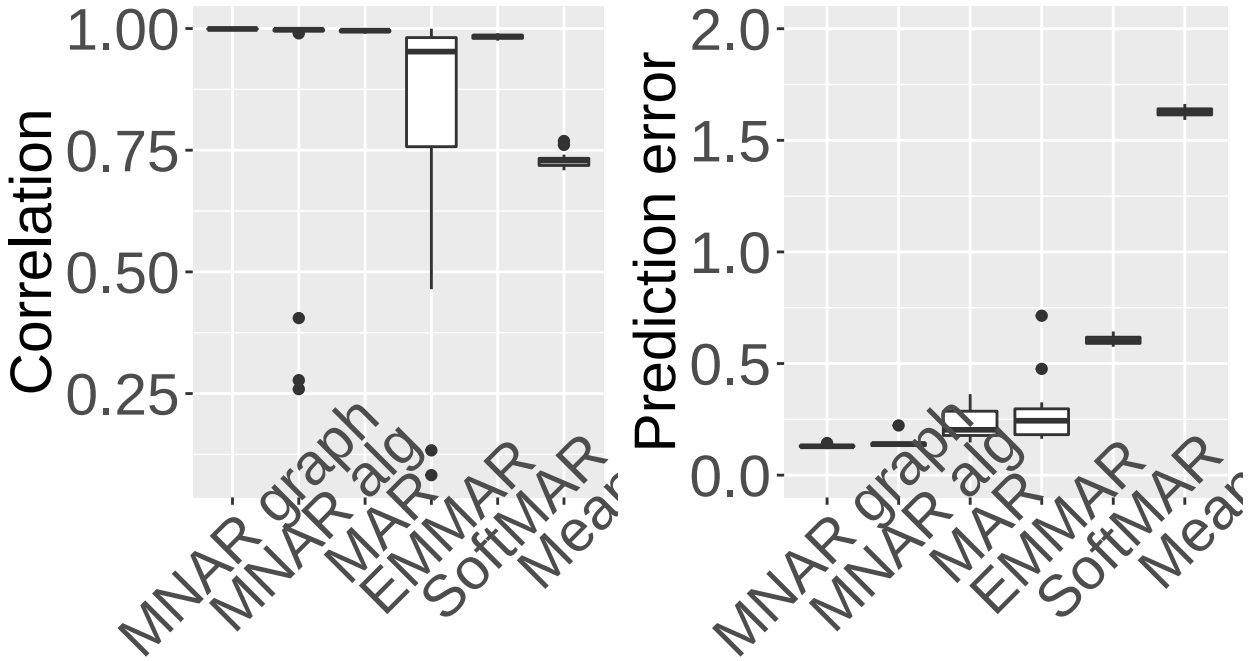
```
## Warning: Removed 6 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 6 rows containing non-finite values (stat_boxplot).
```



Estimation of the loading matrix and prediction error

```
## Warning: Removed 11 rows containing non-finite values (stat_boxplot).
```



Sportisse, Aude, Claire Boyer, and Julie Josse. 2019. "Estimation with Informative Missing Data in the Low-Rank Model with Random Effects." *arXiv Preprint arXiv:1906.02493*.