Koh-Lanta

Missing Data Imputation in Supervised Context

Hadrien Lorenzo[1,2] & Olivier Cloarec[3] & Jérôme Saracco[1,2,4]

hadrien.lorenzo@u-bordeaux.fr

[1] *ASTRAL Team, Inria, Talence*
[2] *OptimAI Team, IMB, CNRS UMR 5251*
[3] *Corporate Research Advanced Data Analytics, Sartorius*
[4] *Bordeaux INP*

Tuesday June 7th 2022

A review

Mainly $(\mathbf{x}, \mathbf{y})$ a $(p+q)$-dimensional random vector divided in $\mathbf{x}_{\text{obs}}$ for observed values and $\mathbf{x}_{\text{mis}}$ for missing values in $\mathbf{x}$.

Hypothesis

$(\mathbf{x}, \mathbf{y})$ of parametric density of parameter θ.

Objective of the user

Estimate θ based solely on the observed values of a sample $\mathcal{D}_n = (\mathbf{x}_{i,\text{obs}}, \mathbf{y}_i)_{i=1\dots n}$, through the maximization of the likelihood (its logarithm)

$$\ell(\theta|\mathcal{D}_n) = \sum_{i=1}^{n} \ln\ p(\mathbf{x}_{i,\text{obs}}, \mathbf{y}_i|\theta) \tag{1}$$

A review

Problem

But the likelihood is not writable since values are missing.

A solution: the EM algorithm Dempster, Laird, and Rubin 1977

Use the EM algorithm which works in two steps, based on an initial model $\theta^{(0)}$, $\forall j > 0$

- $\forall i = 1 \dots n$: Evaluate $\mathbf{x}_{i,\text{mis}}$ thanks to the $\mathbf{x}_{i,\text{obs}}$, $\mathbf{y}_i$ and $\theta^{(j-1)}$. The data-set is completed in $\mathcal{D}_n^{(c)}$.
- Evaluate $\ell(\theta|\mathcal{D}_n^{(c)})$, the completed likelihood.
- Maximize $\ell(\theta|\mathcal{D}_n^{(c)})$ with respect to $\theta \implies \theta^{(j)}$

The EM algorithm converges to a local maximum.

In practice

Situation

Imputation, the procedure that evaluates the missing values, is based on conditional model such as

$$x_1 | x_2, x_3, \mathbf{y}, \theta \tag{2}$$

if missing values are in x_1 but not in x_2, x_3 and $\mathbf{y}$.

Law of x_1 based on the other variables

$$x_1 | x_2, x_3, \mathbf{y}, \theta \sim f_{1|2,3,\mathbf{y}}(x_2, x_3, \mathbf{y}, \theta), \tag{3}$$

$f_{1|2,3,\mathbf{y}}(\cdot)$ is untractable in high dimension...

Example: Gaussian Multivariate Normal (GMN)

$$\left[\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} x_1 \\ \mathbf{x}_{-1} \end{pmatrix}\right] \Big| \mu, \mathbf{\Sigma} \sim \mathcal{N}(\mu, \mathbf{\Sigma})$$

$$\text{where } \mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \mathbf{\Sigma}_{(1)(-1)} \\ \mathbf{\Sigma}_{(1)(-1)}^\top & \mathbf{\Sigma}_{-1} \end{pmatrix},$$

$$x_1 | \mathbf{x}_{-1}, \mu, \mathbf{\Sigma} \sim \mathcal{N}\left(\mu_{1|-1}, \sigma^2_{1|-1}\right),$$

with the Schur complement :

$$\mu_{1|-1} = \mu_1 + \mathbf{\Sigma}_{(1)(-1)} (\mathbf{\Sigma}_{-1})^{-1} \left[(\mathbf{x}_{-1})^\top - \mu_{-1}\right],$$

$$\sigma^2_{1|-1} = \sigma_1^2 - \mathbf{\Sigma}_{(1)(-1)} (\mathbf{\Sigma}_{-1})^{-1} \mathbf{\Sigma}_{(1)(-1)}^\top.$$

High dimensional context, $\mathbf{\Sigma}_{-1}$ singular

A solution: *R*-dimensional structure

Additional hypothesis on the structure of the data.
For example the Probabilistic PCA (PPCA) Tipping and Bishop 1999 such as

$$\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \mu + \mathbf{W}\mathbf{t} + \varepsilon, \tag{4}$$

- $\mu \in \mathbb{R}^{p+q}$ is the mean vector,
- $\mathbf{W} \in \mathbb{R}^{(p+q)\times R}$ is a deterministic matrix,
- $\mathbf{t} \in \mathbb{R}^{R}$ is a random vector such as $\mathbf{t} \sim \mathcal{N}(0, \mathbb{I}_R)$,
- $R << \min(p+q, n)$ is the underlying number of components,
- $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_{p+q})$.

Back to the example

$$\left[\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} x_1 \\ \mathbf{x}_{-1} \end{pmatrix}\right] \Big| \mu, \mathbf{W}, \sigma^2 \sim \mathcal{N}\left(\mu, \mathbf{W}\mathbf{W}^\top + \sigma^2 \mathbb{I}_{p+q}\right),$$

$$\text{where } \mathbf{W}\mathbf{W}^\top + \sigma^2 \mathbb{I}_{p+q} = \begin{pmatrix} a_1 + \sigma^2 & \mathbf{A}_{(1)(-1)} \\ \mathbf{A}_{(1)(-1)}^\top & \mathbf{A}_{-1} + \sigma^2 \mathbb{I}_{p+q-1} \end{pmatrix},$$

$$x_1 | \mathbf{x}_{-1}, \mu, \mathbf{W}, \sigma^2 \sim \mathcal{N}\left(\mu_{1|-1}, \sigma^2_{1|-1}\right),$$

with the Schur complement :

$$\mu_{1|-1} = \mu_1 + \mathbf{A}_{(1)(-1)}\left(\mathbf{A}_{-1} + \sigma^2 \mathbb{I}_{p+q-1}\right)^{-1}\left[(\mathbf{x}_{-1})^\top - \mu_{-1}\right],$$

$$\sigma^2_{1|-1} = a_1 + \sigma^2 - \mathbf{A}_{(1)(-1)}\left(\mathbf{A}_{-1} + \sigma^2 \mathbb{I}_{p+q-1}\right)^{-1}\mathbf{A}_{(1)(-1)}^\top$$

Why estimate the joint distribution ?

Indeed, not interesting nor...

- ...to evaluate the missing values.
- ...for the research question: $\mathbf{y}|\mathbf{x}$.

Another solution: do not estimate the joint model: Fully Conditional Specifications (FCS)

$\forall j \in [\![1, p+q]\!]$, if $\mathbf{x}_j$ shows missing values:

- Draw $\tilde{\theta}_{j|-j}$ from $\theta_{j|-j}|\mathbf{x}_{j,obs}, \tilde{\mathbf{X}}_{-j}$.
- Draw $\tilde{\mathbf{x}}_{j,\text{mis}}$ from $\mathbf{x}_{j,\text{mis}}|\mathbf{x}_{j,obs}, \tilde{\theta}_{j|-j}$

Re-do for M cycles.

...Gibbs sampling.

Remarks on FCS

- Allows to specify a model per variable (efficient in presence of categorical variables).
- Converges to the joint distributions, for many model assumptions.
- More computations per iteration.
- Needs regularization techniques, Ridge for MICE (Buuren and Groothuis-Oudshoorn 2010).

Between JM and FCS

- JM and FCS evaluate many models useful for imputations.
- Many models are still useless for the research question: $\mathbf{y}|\mathbf{x}$.
- **Solution ?**
  - Draw $\hat{\theta}_{\mathbf{x}|\mathbf{y}}$ from $\theta_{\mathbf{x}|\mathbf{y}}|\tilde{\mathbf{X}}, \mathbf{Y}$.
  - Draw $\tilde{\mathbf{x}}_{\text{mis}}$ from $\mathbf{x}_{\text{mis}}|\mathbf{Y}, \hat{\theta}_{\mathbf{x}|\mathbf{y}}$

  Re-do until stabilization of $\hat{\theta}_{\mathbf{x}|\mathbf{y}}$.

Blocked Gibbs Sampling (J. S. Liu, Wong, and Kong 1994)

A latent vector model

The latent vector model

$$\mathbf{x} = \boldsymbol{\mu}_x + \mathbf{Pt} + \boldsymbol{\epsilon}_x, \quad (5)$$

$$\mathbf{y} = \boldsymbol{\mu}_y + \mathbf{Ct} + \boldsymbol{\epsilon}_y. \quad (6)$$

- $(\boldsymbol{\mu}_x^\top, \boldsymbol{\mu}_y^\top)^\top \in \mathbb{R}^{p+q}$ is the mean vector,
- $\mathbf{P} \in \mathbb{R}^{p\times R}$ and $\mathbf{C} \in \mathbb{R}^{q\times R}$ are deterministic matrix,
- $\mathbf{t} \in \mathbb{R}^R$ is a random vector such as $\mathbf{t} \sim \mathcal{N}(0, \mathbb{I}_R)$,
- $R << \min(p+q, n)$ is the underlying number of components,
- $\boldsymbol{\epsilon}_x \sim \mathcal{N}(0, \mathbf{D}_x)$ and $\boldsymbol{\epsilon}_y \sim \mathcal{N}(0, \mathbf{D}_y)$, $\mathbf{D}_x$ and $\mathbf{D}_y$ diagonals.

Estimation of the Partial Least Squares (PLS) model plus the matrix $\mathbf{B}$ such as

$$\mathbf{y} \approx \mathbf{B}^\top \mathbf{x} \quad (7)$$

NIPALS to deal with missing values

How does it work ?
Preda, Saporta, and Hadj Mbarek 2010
Tenenhaus 1998

(a), **(c)**, **(d)** and **(e)** computed on the observations $\neq$ `NA`:

$$w_j \propto \sum_{i=1}^{n} x_{i,j} u_i \, \delta_{x_{i,j} \neq \mathtt{NA}}, \quad t_i \propto \sum_{j=1}^{p} x_{i,j} w_j \, \delta_{x_{i,j} \neq \mathtt{NA}}, \tag{8}$$

$$c_k \propto \sum_{i=1}^{n} y_{i,k} t_i \, \delta_{y_{i,k} \neq \mathtt{NA}}, \quad u_i \propto \sum_{k=1}^{q} y_{i,k} c_k \, \delta_{y_{i,k} \neq \mathtt{NA}}. \tag{9}$$

Hypothesis, ... poorly translated from Bastien 2008:

"*The missing values do not modify the slopes of the regression lines of the clouds* $(\mathbf{Y}^{(r)}, \mathbf{X}^{(r)})$ *estimating* $\mathbf{w}^{(r)}$".

NIPALS, why and why not ?

Remark

- Robust to missing values in both **x** and **y** parts.
- Does not impute the missing values, different from the EM algorithm spirit.
- Strong hypothesis in the high dimensional context (n low for example).

An alternative to NIPALS

PLS-MI of Bastien 2008

Application of the Data Augmentation (DA) of Tanner and Wong 1987 to the PLS context, such as

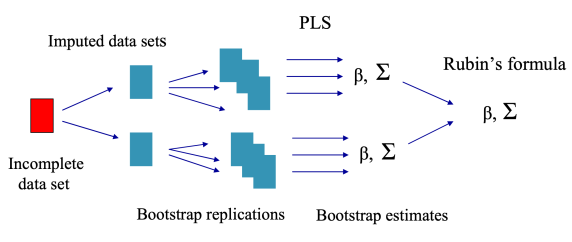


Figure: From Bastien 2008

Remark: The posterior distribution $(\mathbf{x}, \mathbf{y})$ is evaluated: not adapted to high dimension context.

The Koh-Lanta algorithm

For a number of times $M > 1$ and the sample $\mathcal{D}_n$, do

The Koh-Lanta algorithm

For a number of times $M > 1$ and the sample $\mathcal{D}_n$, do

- $\mathcal{D}_n^\star$: bootstrap $\mathcal{D}_n$, $\hookrightarrow$ Sample variability

The Koh-Lanta algorithm

For a number of times $M > 1$ and the sample $\mathcal{D}_n$, do

- $\mathcal{D}_n^\star$: bootstrap $\mathcal{D}_n$, $\hookrightarrow$ Sample variability
- $\hat{\theta}_{\mathbf{x}|\mathbf{y}}$: estimation of $\theta_{\mathbf{x}|\mathbf{y}}$ on $\mathcal{D}_n^\star$ (iterative procedure),

The Koh-Lanta algorithm

For a number of times $M > 1$ and the sample $\mathcal{D}_n$, do

- $\mathcal{D}_n^\star$: bootstrap $\mathcal{D}_n$, $\hookrightarrow$ Sample variability
- $\hat{\theta}_{\mathbf{x}|\mathbf{y}}$: estimation of $\theta_{\mathbf{x}|\mathbf{y}}$ on $\mathcal{D}_n^\star$ (iterative procedure),
- $\tilde{\mathbf{X}}$: proper imputation of $\mathbf{X}$ based on $\hat{\theta}_{\mathbf{x}|\mathbf{y}}$, $\hookrightarrow$ Model variability

The Koh-Lanta algorithm

For a number of times $M > 1$ and the sample $\mathcal{D}_n$, do

- $\mathcal{D}_n^\star$: bootstrap $\mathcal{D}_n$, $\hookrightarrow$ Sample variability
- $\hat{\theta}_{\mathbf{x}|\mathbf{y}}$: estimation of $\theta_{\mathbf{x}|\mathbf{y}}$ on $\mathcal{D}_n^\star$ (iterative procedure),
- $\tilde{\mathbf{X}}$: proper imputation of $\mathbf{X}$ based on $\hat{\theta}_{\mathbf{x}|\mathbf{y}}$, $\hookrightarrow$ Model variability
- $\hat{\theta}_{\mathbf{y}|\mathbf{x}}$: estimation of $\theta_{\mathbf{y}|\mathbf{x}}$ on $\tilde{\mathcal{D}}_n = (\tilde{\mathbf{X}}, \mathbf{Y})$,
- return $\left(\hat{\theta}_{\mathbf{x}|\mathbf{y}}, \tilde{\mathbf{X}}, \hat{\theta}_{\mathbf{y}|\mathbf{x}}\right)$.

The Koh-Lanta algorithm

For a number of times $M > 1$ and the sample $\mathcal{D}_n$, do

- $\mathcal{D}_n^\star$: bootstrap $\mathcal{D}_n$, $\hookrightarrow$ Sample variability
- $\hat{\theta}_{\mathbf{x}|\mathbf{y}}$: estimation of $\theta_{\mathbf{x}|\mathbf{y}}$ on $\mathcal{D}_n^\star$ (iterative procedure),
- $\tilde{\mathbf{X}}$: proper imputation of $\mathbf{X}$ based on $\hat{\theta}_{\mathbf{x}|\mathbf{y}}$, $\hookrightarrow$ Model variability
- $\hat{\theta}_{\mathbf{y}|\mathbf{x}}$: estimation of $\theta_{\mathbf{y}|\mathbf{x}}$ on $\tilde{\mathcal{D}}_n = (\tilde{\mathbf{X}}, \mathbf{Y})$,
- return $\left(\hat{\theta}_{\mathbf{x}|\mathbf{y}}, \tilde{\mathbf{X}}, \hat{\theta}_{\mathbf{y}|\mathbf{x}}\right)$.

data-driven sparse PLS (**ddsPLS**, Lorenzo et al. 2022) is used:

- **y** multivariate,
- a few hyper-parameters.

data driven Sparse PLS (ddsPLS)

Equivalent to NIPALS but covariance matrix is estimated with

$$S_{\lambda^{(r)}}\left(\mathbf{Y}^{(r)\top}\mathbf{X}^{(r)}/n\right),$$

where the soft-thresholding operator is

$$S_\lambda(x_{i,j}) = \text{sign}(x_{i,j})\max(0, |x_{i,j}| - \lambda).$$

- Regularization and variable selection in **x** and **y**.
- R and $(\lambda^{(1)}, ..., \lambda^{(R)})$ fixed by bootstrap.

ddsPLS in Koh-Lanta

- Use ddsPLS to estimate $\theta_{\mathbf{x}|\mathbf{y}}$ and $\theta_{\mathbf{y}|\mathbf{x}}$.
- Automatic fix of R and λ using validation approaches through bootstrap:
  - "**Koh-Lanta (in ddsPLS)**" : minimizes $\bar{R}^2_B - \bar{Q}^2_B$ and restricts the grid of λ to avoid small values (Cai and W. Liu 2011).
  - "**Koh-Lanta (in ddsPLS LD)**" : maximizes $\bar{Q}^2_B$.

Other approaches

- "**MI-NIPALS-PLS**" uses NIPALS algorithm. $R \in [\![1, 5]\!]$ fixed to minimize the leave-one-out prediction error.
- "**NIPALS-PLS**" uses the NIPALS algorithm for simple imputation.
- "**MEAN-PLS**" . Imputes missing values to mean. Then build PLS model.
- "**missMDA-PLS**" . JM approach, proper MI with PPCA, implementation in missMDA (Josse and Husson 2016). A PLS model is then computed on the completed data set.

Simulation structure

- $p = 2\mathbf{p_1} + \mathbf{p_2} + \mathbf{p_3}$ (where $\mathbf{p_1}$, $\mathbf{p_2}$ and $\mathbf{p_3}$ described below)
- $q = 3$

The latent variables $\phi_j,\ j = 1, \ldots, 3$, $\sigma = \sqrt{0.1}$

$$
x_j = \begin{cases} \sqrt{1-\sigma^2}\phi_1 + \sigma\epsilon_j & \text{for} \quad j = 1, \ldots, \mathbf{p_1} \\ \sqrt{1-\sigma^2}\phi_2 + \sigma\epsilon_j & \text{for} \quad j = \mathbf{p_1} + 1, \ldots, 2\mathbf{p_1} \\ \sqrt{1-\sigma^2}\phi_3 + \sigma\epsilon_j & \text{for} \quad j = 2\mathbf{p_1} + 1, \ldots, 2\mathbf{p_1} + \mathbf{p_2} \\ \epsilon_j & \text{for} \quad j = 2\mathbf{p_1} + \mathbf{p_2} + 1, \ldots, 2\mathbf{p_1} + \mathbf{p_2} + \mathbf{p_3} \end{cases}
$$

$$
\begin{cases} y_1 = \sqrt{1-\sigma^2}\phi_1 + \sigma\xi_1 \\ y_2 = \sqrt{1-\sigma^2}(\phi_1 + 2\phi_2)/\sqrt{5} + \sigma\xi_2 \\ y_3 = \xi_3 \end{cases} \tag{10}
$$

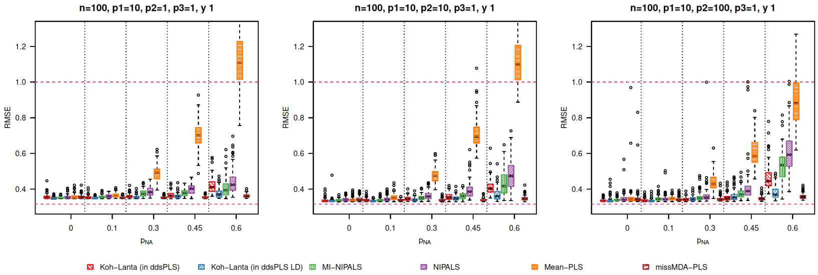
where $(\phi^\top, \epsilon_{1\ldots p}^\top, \xi_{1\ldots 3}^\top)^\top \sim \mathcal{N}_{3+p+q}\left(0_{3+p+q}, \mathbb{I}_{3+p+q}\right)$.
Only variables x_1 to $x_{2\mathbf{p_1}}$ would be selected.

$\mathbf{p_1}$	$\mathbf{p_2}$	$\mathbf{p_3}$	n	p_{NA}
10	$\in \{1, 10, 100\}$	$\in \{1, 100, 500\}$	$\in \{20, 50, 100\}$	$\in \{0, 0.1, 0.3, 0.6\}$

Interpretation

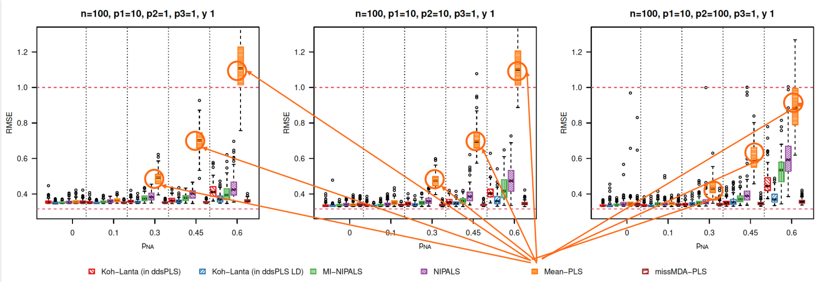
- $(\mathbf{p_1}, \mathbf{p_2}, \mathbf{p_3}) = (10, 1, 1)$ easy case.
- $(\mathbf{p_1}, \mathbf{p_2}, \mathbf{p_3}) = (10, 100, 500)$ hard case where
  - $\mathbf{p_3} = 500$ uncorrelated and useless variables are observed,
  - $\mathbf{p_2} = 100$ correlated and useless variables are observed,
  - only two times $\mathbf{p_1} = 10$ useful variables are observed.
- $n \in \{20, 50, 100\}$: hard context only.

$\mathbf{p_3}$ = 1, easy case for which method ?

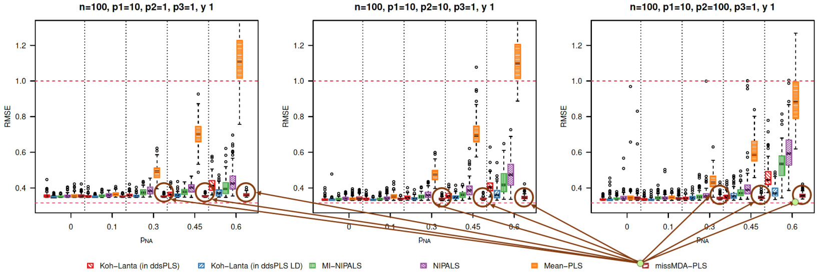


Mean imputation shows bad performances.

$p_3 = 1$, easy case for which method ?

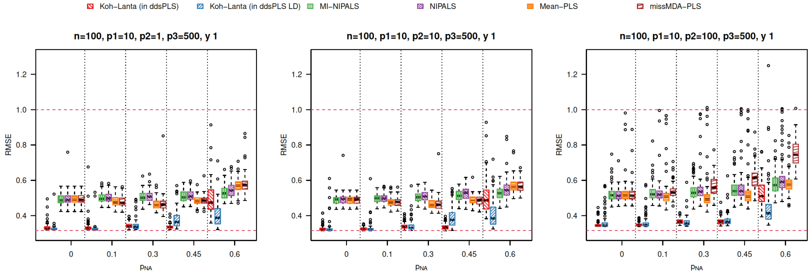


$p_3 = 1$, easy case for which method ?

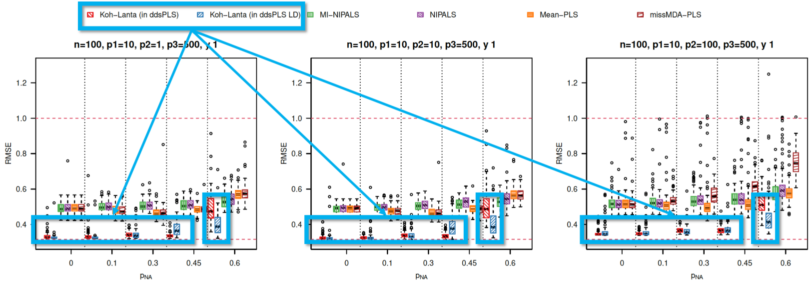


JM outperforms other approaches.

$\mathbf{p_3} = 500$, hard case for which method ?

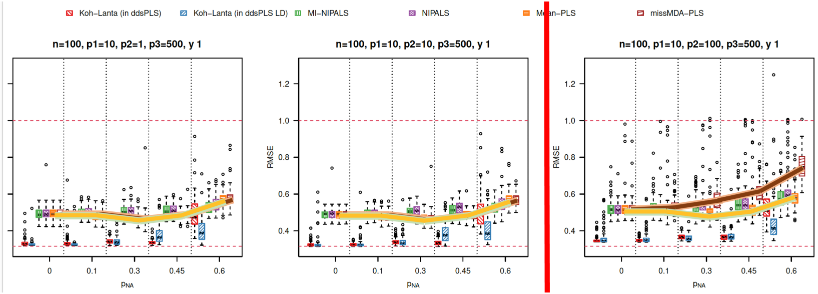


$\mathbf{p_3} = 500$, hard case for which method ?



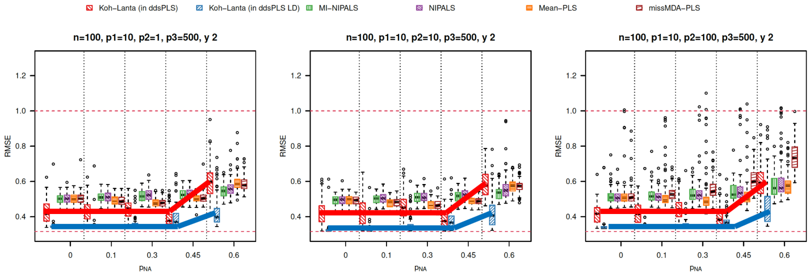
(ddsPLS)+(Koh-Lanta)'s capacity to deal with missing value+ high dimension. (?)

$\mathbf{p_3} = 500$, from low to high $\mathbf{p_2}$



JM overfits if $\mathbf{p_2}$ gets higher.

$p_3 = 500$, look at y_2, intricate response variable



ddsPLS's difficulty to deal with intricate variables.

Conclusion

- **Mean imputation** fails again,
- **JM imputation** seems to fail in high dimension,
- **Koh-Lanta** seems deal with `NA` in high dimension, but how to make the difference ?

References I

Bastien, Philippe (Mar. 2008). "RÉGRESSION PLS ET DONNÉES CENSURÉES". Theses. Conservatoire national des arts et metiers - CNAM. URL: https://tel.archives-ouvertes.fr/tel-00268344.

Buuren, S van and Karin Groothuis-Oudshoorn (2010). "mice: Multivariate imputation by chained equations in R". In: *Journal of statistical software*, pp. 1–68.

Cai, Tony and Weidong Liu (2011). "Adaptive Thresholding for Sparse Covariance Matrix Estimation". In: *Journal of the American Statistical Association* 106.494, pp. 672–684. DOI: 10.1198/jasa.2011.tm10560. eprint: https://doi.org/10.1198/jasa.2011.tm10560. URL: https://doi.org/10.1198/jasa.2011.tm10560.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). "Maximum likelihood from incomplete data via the EM algorithm". In: *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B* 39.1, pp. 1–38.

Josse, Julie and François Husson (2016). "missMDA: a package for handling missing values in multivariate data analysis". In: *Journal of Statistical Software* 70.1, pp. 1–31.

Liu, Jun S., Wing Hung Wong, and Augustine Kong (1994). "Covariance Structure of the Gibbs Sampler with Applications to the Comparisons of Estimators and Augmentation Schemes". In: *Biometrika* 81.1, pp. 27–40. ISSN: 00063444. URL: http://www.jstor.org/stable/2337047 (visited on 05/30/2022).

References II

Lorenzo, Hadrien et al. (2022). "Data-driven sparse partial least squares". In: *Statistical Analysis and Data Mining: The ASA Data Science Journal* 15.2, pp. 264–282.

Preda, Cristian, Gilbert Saporta, and Mohamed Hadj Mbarek (2010). "The NIPALS Algorithm for Missing Functional Data". In: *Revue roumaine de mathématiques pures et appliquées* 55.4, pp. 315–326. URL: https://hal.archives-ouvertes.fr/hal-01125940.

Tanner, Martin A. and Wing Hung Wong (1987). "The Calculation of Posterior Distributions by Data Augmentation". In: *Journal of the American Statistical Association* 82.398, pp. 528–540. ISSN: 01621459. URL: http://www.jstor.org/stable/2289457 (visited on 06/06/2022).

Tenenhaus, Michel (1998). *La régression PLS: théorie et pratique*. Editions technip.

Tipping, Michael E. and Christopher M. Bishop (1999). "Probabilistic Principal Component Analysis". In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 61.3, pp. 611–622. ISSN: 13697412, 14679868. URL: http://www.jstor.org/stable/2680726.