How to deal with missing values in the high dimensional supervised context?

Hadrien Lorenzo & Olivier Cloarec & Jérôme Saracco

hadrien.lorenzo@u-bordeaux.fr [orator]

Inria, IMB, Sartorius

CMStatistics 2021, King's College London, UK

December 19, 2021

Why talking about supervised imputation ?

Vaccine research context

- Important Ebola vaccine data-set (see Rechtien et al. (2017)).
- $n = 20$, 4 blocks of high dimensions ($p_k = 20.000$, RNA-Seq).
- $\mathbf{y}$, 5-dimensional to be predicted.
- Many missing values: 30%.

Why talking about supervised imputation ?

Vaccine research context

- Important Ebola vaccine data-set (see Rechtien et al. (2017)).
- $n = 20$, 4 blocks of high dimensions ($p_k = 20.000$, RNA-Seq).
- **y**, 5-dimensional to be predicted.
- Many missing values: 30%.

Classical imputation

- Hypothesis: multivariate Gaussian.
- A few variables are imputed, let say $\mathcal{V}_{\text{imputation}}$.

Why talking about supervised imputation ?

Vaccine research context

- Important Ebola vaccine data-set (see Rechtien et al. (2017)).
- $n = 20$, 4 blocks of high dimensions ($p_k = 20.000$, RNA-Seq).
- $\mathbf{y}$, 5-dimensional to be predicted.
- Many missing values: 30%.

Classical imputation

- Hypothesis: multivariate Gaussian.
- A few variables are imputed, let say $\mathcal{V}_{\text{imputation}}$.

Then supervised analysis (sparse PLS)

A few variables are selected for prediction, let say $\mathcal{V}_{\text{prediction}}$.

Why talking about supervised imputation ? (II)

Observation

$$\mathcal{V}_{\text{imputation}} \bigcap \mathcal{V}_{\text{prediction}} = \emptyset.$$

Why talking about supervised imputation ? (II)

Observation

$$\mathcal{V}_{\text{imputation}} \bigcap \mathcal{V}_{\text{prediction}} = \varnothing.$$

Interpretation

No imputed variables are of interest to answer the question.

Why talking about supervised imputation ? (II)

Observation

$$\mathcal{V}_{\text{imputation}} \bigcap \mathcal{V}_{\text{prediction}} = \varnothing.$$

Interpretation

No imputed variables are of interest to answer the question.

Causation

Imputed variables are associated to each other but not to $\mathbf{y}$.

Why talking about supervised imputation ? (II)

Observation

$$\mathcal{V}_{\text{imputation}} \bigcap \mathcal{V}_{\text{prediction}} = \varnothing.$$

Interpretation

No imputed variables are of interest to answer the question.

Causation

Imputed variables are associated to each other but not to **y**.

Solution

Hide useless variables: regularization through supervised analysis.

Imputation and data augmentation

If $\mathbf{x}$ is a random vector of parameter $\boldsymbol{\theta}$, $\mathbf{x}^{(obs)}$ is observed and $\mathbf{x}^{(miss)}$ is missing, a prior on $\boldsymbol{\theta}$ is chosen by the user.

Imputation and data augmentation

If $\mathbf{x}$ is a random vector of parameter $\boldsymbol{\theta}$, $\mathbf{x}^{(obs)}$ is observed and $\mathbf{x}^{(miss)}$ is missing, a prior on $\boldsymbol{\theta}$ is chosen by the user.
The two posterior distributions $p(\mathbf{x}^{(miss)}|\mathbf{x}^{(obs)})$ and $p(\boldsymbol{\theta}|\mathbf{x}^{(obs)})$ verify

$$p(\mathbf{x}^{(miss)}|\mathbf{x}^{(obs)}) = \int_{\boldsymbol{\theta}\in\Theta} p(\mathbf{x}^{(miss)}|\boldsymbol{\theta},\mathbf{x}^{(obs)})p(\boldsymbol{\theta}|\mathbf{x}^{(obs)})d\boldsymbol{\theta},$$
$$p(\boldsymbol{\theta}|\mathbf{x}^{(obs)}) = \int_{\mathbf{x}^{(miss)}\in\mathbf{X}^{(miss)}} p(\boldsymbol{\theta}|\mathbf{x}^{(miss)},\mathbf{x}^{(obs)})p(\mathbf{x}^{(miss)}|\mathbf{x}^{(obs)})d\boldsymbol{\theta},$$

Imputation and data augmentation

If $\mathbf{x}$ is a random vector of parameter $\boldsymbol{\theta}$, $\mathbf{x}^{(obs)}$ is observed and $\mathbf{x}^{(miss)}$ is missing, a prior on $\boldsymbol{\theta}$ is chosen by the user.
The two posterior distributions $p(\mathbf{x}^{(miss)}|\mathbf{x}^{(obs)})$ and $p(\boldsymbol{\theta}|\mathbf{x}^{(obs)})$ verify

$$p(\mathbf{x}^{(miss)}|\mathbf{x}^{(obs)}) = \int_{\boldsymbol{\theta}\in\Theta} p(\mathbf{x}^{(miss)}|\boldsymbol{\theta},\mathbf{x}^{(obs)})p(\boldsymbol{\theta}|\mathbf{x}^{(obs)})d\boldsymbol{\theta},$$
$$p(\boldsymbol{\theta}|\mathbf{x}^{(obs)}) = \int_{\mathbf{x}^{(miss)}\in\mathbf{x}^{(miss)}} p(\boldsymbol{\theta}|\mathbf{x}^{(miss)},\mathbf{x}^{(obs)})p(\mathbf{x}^{(miss)}|\mathbf{x}^{(obs)})d\boldsymbol{\theta},$$

leading to stationary equations, if $g(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathbf{x}^{(obs)})$:

$$g(\boldsymbol{\theta}) = \int_{\phi\in\Theta} K(\boldsymbol{\theta},\phi)g(\phi)d\phi = \mathcal{T}g(\boldsymbol{\theta})$$

where $K(\boldsymbol{\theta},\phi) = \int_{z\in X^{(miss)}} p(\boldsymbol{\theta}|z,\mathbf{x}^{(obs)})p(z|\mathbf{x}^{(obs)},\phi)dz$.
See for example Tanner and Wong (1987) on data augmentation and Rubin (1996) for global review.

Algorithms

Use Markov Chain Monte Carlo methods to converge to the posteriors such as at each iteration

$$\boldsymbol{\theta}^\star \sim p(\boldsymbol{\theta}|\mathbf{x}^{(obs)}, \mathbf{x}^{(miss),\star}),$$
$$\mathbf{x}^{(miss),\star} \sim p(\mathbf{x}^{(miss)}|\boldsymbol{\theta}^\star, \mathbf{x}^{(obs)}),$$

and at ∞, draws follow $p(\mathbf{x}^{(miss)}, \boldsymbol{\theta}|\mathbf{x}^{(obs)})$.

Weakness?

Necessity to define a joint model. Hard for mixed data.
Solution? FAMD (PCA for mixed data, see Audigier, Husson, and Josse (2016)) treats categorical variables through dummy variables.
Other solution? Define conditional models such as $p(\mathbf{x}_1|\mathbf{x}_2, \cdots, \mathbf{x}_p, \boldsymbol{\theta}_1)$ where variables can be qualitative/quantitative.
Iterate per coordinate in cycles until convergence.

Chained Equations/Fully Conditional Specification (FCS)

And for supervised problems ?

Idea: Partially (?) Conditional Specification: cycles of

$$\begin{array}{lcl}
\boldsymbol{\theta}_x^{\star(t)} & \sim & p(\boldsymbol{\theta}_x|\mathbf{x}^{\text{obs}},\mathbf{y}^{\text{obs}},\mathbf{y}^{miss\star(t-1)}),\\
\mathbf{x}^{miss\star(t)} & \sim & p(\boldsymbol{x}^{(miss)}|\mathbf{x}^{\text{obs}},\mathbf{y}^{\text{obs}},\mathbf{y}^{miss\star(t-1)},\boldsymbol{\theta}_x^{\star(t)}),\\
\boldsymbol{\theta}_y^{\star(t)} & \sim & p(\boldsymbol{\theta}_y|\mathbf{x}^{\text{obs}},\mathbf{x}^{miss\star(t)},\mathbf{y}^{\text{obs}}),\\
\mathbf{y}^{miss\star(t)} & \sim & p(\boldsymbol{y}^{(miss)}|\mathbf{x}^{\text{obs}},\mathbf{x}^{miss\star(t)},\mathbf{y}^{\text{obs}},\boldsymbol{\theta}_y^{\star(t)}),
\end{array} \tag{1}$$

And for supervised problems ?

Idea: Partially (?) Conditional Specification: cycles of

$$
\begin{array}{lcl}
\boldsymbol{\theta}_x^{\star(t)} & \sim & p(\boldsymbol{\theta}_x|\mathbf{x}^{\text{obs}},\mathbf{y}^{\text{obs}},\mathbf{y}^{miss\star(t-1)}), \\
\mathbf{x}^{miss\star(t)} & \sim & p(\mathbf{x}^{(miss)}|\mathbf{x}^{\text{obs}},\mathbf{y}^{\text{obs}},\mathbf{y}^{miss\star(t-1)},\boldsymbol{\theta}_x^{\star(t)}), \\
\boldsymbol{\theta}_y^{\star(t)} & \sim & p(\boldsymbol{\theta}_y|\mathbf{x}^{\text{obs}},\mathbf{x}^{miss\star(t)},\mathbf{y}^{\text{obs}}), \\
\mathbf{y}^{miss\star(t)} & \sim & p(\mathbf{y}^{(miss)}|\mathbf{x}^{\text{obs}},\mathbf{x}^{miss\star(t)},\mathbf{y}^{\text{obs}},\boldsymbol{\theta}_y^{\star(t)}),
\end{array}
\qquad (1)
$$

Statistical model, close to Trygg and Wold (2003)

$$
\begin{array}{rcl}
\mathbf{x} & = & f_{\boldsymbol{\theta}_x}(\mathbf{x}_y,\mathbf{x}_x), \\
\mathbf{y} & = & g_{\boldsymbol{\theta}_y}(\mathbf{y}_x,\mathbf{y}_y), \\
\mathbf{x} & \perp\!\!\!\perp & \mathbf{y}|\mathbf{x}_y,\boldsymbol{\theta}_x, \\
\mathbf{y} & \perp\!\!\!\perp & \mathbf{x}|\mathbf{y}_x,\boldsymbol{\theta}_y,
\end{array}
\qquad (2)
$$

where $\mathbf{x}_x \perp\!\!\!\perp \mathbf{y}$, $\mathbf{x}_x \perp\!\!\!\perp \mathbf{x}_y$, $\mathbf{y}_y \perp\!\!\!\perp \mathbf{x}$ and $\mathbf{y}_y \perp\!\!\!\perp \mathbf{y}_x$.

And for supervised problems ? (II)

Observation: Impute $\mathbf{x}$ (resp. $\mathbf{y}$) uses information from:

- $\mathbf{x}$ (resp. $\mathbf{y}$) through $\mathbf{x}_x$ (resp. $\mathbf{y}_y$) –> JM part,
- $\mathbf{y}$ (resp. $\mathbf{x}$) through $\mathbf{x}_y$ (resp. $\mathbf{y}_x$) –> FCS part.

Remark: If $\mathbf{y}$ is univariate, $\mathbf{y}_y$ does not exist.

And for supervised problems ? (II)

Observation: Impute $\mathbf{x}$ (resp. $\mathbf{y}$) uses information from:

- $\mathbf{x}$ (resp. $\mathbf{y}$) through $\mathbf{x}_x$ (resp. $\mathbf{y}_y$) –> JM part,
- $\mathbf{y}$ (resp. $\mathbf{x}$) through $\mathbf{x}_y$ (resp. $\mathbf{y}_x$) –> FCS part.

Remark: If $\mathbf{y}$ is univariate, $\mathbf{y}_y$ does not exist.

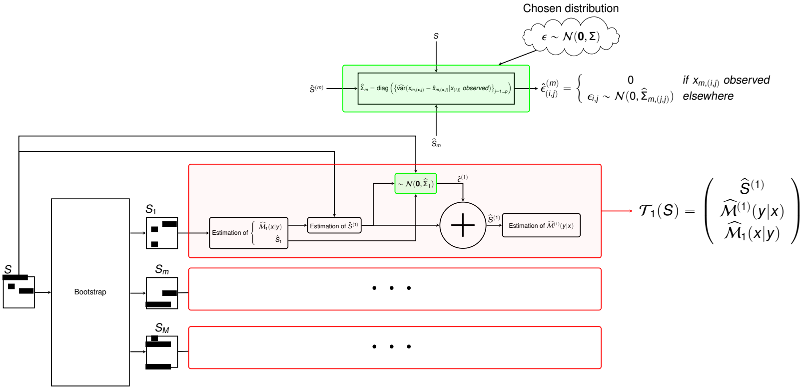
Koh-Lanta

Hypothesis for this presentation: No missing values in $\mathbf{y}$.
Parameter variability: Bootstrap of the input dataset.
Distribution hypothesis: Multivariate Gaussian.
Simplifying assumption: $\mathbf{x}_x$ and $\mathbf{y}_y$ with respective diagonal variance matrices: only noise, not really realistic.

$$\begin{aligned} \mathbf{x} &= \mathbf{M}\mathbf{y} + \boldsymbol{\epsilon}, \\ \mathbf{y} &= \mathbf{N}\mathbf{x} + \boldsymbol{\xi}. \end{aligned} \tag{3}$$

$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma} = \mathrm{diag}(\sigma_j^2)_{j=1..p})$

The Koh-Lanta algorithm for imputation



Which statistical model ?

Our approach uses only FCS part equivalents: needs of prediction models only.
We have chosen **ddsPLS** (see Lorenzo et al. (2021)):

- PLS based approach,
- Single penalization for sparsity per component (sparse both in **x** and **y**),
- Automatic fix the number of components (R) through bootstrap operations and minimizing $\bar{R}^2_B - \bar{Q}^2_B$ while $\bar{Q}^2_B > 0$,
- "**Koh-Lanta (in ddsPLS)**" : adapt to important noise context (see Cai and Liu (2011)),
- **LD** mode for **L**ow **D**imensional contexts: "**Koh-Lanta (in ddsPLS LD)**"

Competitive approaches

Two types of imputation:

- Single imputation: "**MEAN-PLS**" , "**NIPALS**" .
- Multiple imputation (MI): "**MI-NIPALS**" , "**missMDA-PLS**" .

Competitive approaches

Two types of imputation:

- Single imputation: "**MEAN-PLS**" , "**NIPALS**" .
- Multiple imputation (MI): "**MI-NIPALS**" , "**missMDA-PLS**" .

• "**NIPALS**" performs PLS analysis estimating covariance matrices only on the observed values (no real imputation, generalizable ?).

Competitive approaches

Two types of imputation:

- Single imputation: "**MEAN-PLS**" , "**NIPALS**" .
- Multiple imputation (MI): "**MI-NIPALS**" , "**missMDA-PLS**" .

• "**NIPALS**" performs PLS analysis estimating covariance matrices only on the observed values (no real imputation, generalizable ?).

• "**missMDA-PLS**" (see Josse, Pagès, and Husson (2011)) performs regularized PCA, function `MIPCA`. *R* is chosen through the function `estim_ncp` (cross-validation and minimizing MSE).

Model

Let us suppose the following latent variable model

$$x_j = \begin{cases} \sqrt{1-\sigma^2}\phi_1 + \sigma\varepsilon_j & \text{for} \quad j = 1\dots p_1 \\ \sqrt{1-\sigma^2}\phi_2 + \sigma\varepsilon_j & \text{for} \quad j = p_1+1\dots 2p_1 \\ \sqrt{1-\sigma^2}\phi_3 + \sigma\varepsilon_j & \text{for} \quad j = 2p_1+1\dots 2p_1+p_2 \\ \varepsilon_j & \text{for} \quad j = 2p_1+p_2+1\dots 2p_1+p_2+p_3 \end{cases} \tag{4}$$

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

where $\sigma = \sqrt{0.1} \approx 0.316$ and

$$(\phi_1, \phi_2, \phi_3, \boldsymbol{\varepsilon}'_{1\dots p}, \boldsymbol{\xi}'_{1\dots 3})' \sim \mathcal{N}\left(\mathbf{0}_{3+p+3}, \mathbb{I}_{3+p+3}\right).$$

Note that $\sqrt{\mathbb{E}(y_j - y_j^\star)^2} = \begin{cases} \sigma \approx 0.316 & \text{if } j = 1,2 \\ 1 & \text{if } j = 3 \end{cases}$

Simulation parameters

Idea: test both the impacts of high noise (structured and unstructured) and of sample size.

Simulation parameters

Idea: test both the impacts of high noise (structured and unstructured) and of sample size.

- $p_1 = 10$: information variance,
- $(p_2, p_3) \in \{(1, 1), (100, 500)\}$: low noise and high noise variances (structured and unstructured),
- $n \in \{100, 50, 20\}$: hard to very hard problems.

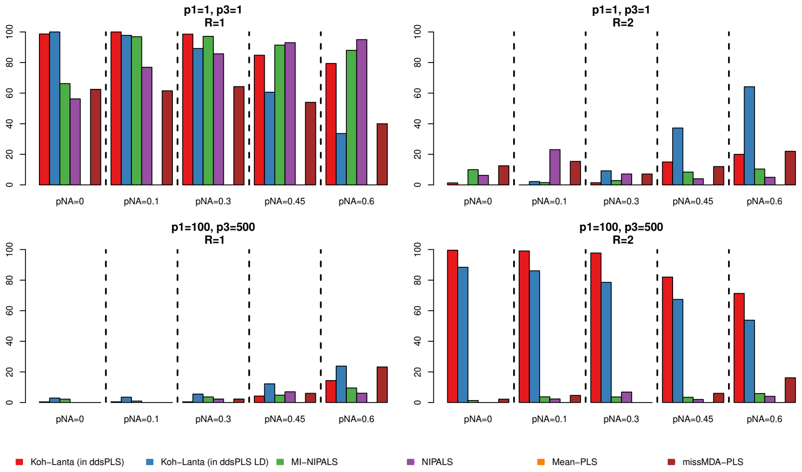
Simulation parameters

Idea: test both the impacts of high noise (structured and unstructured) and of sample size.

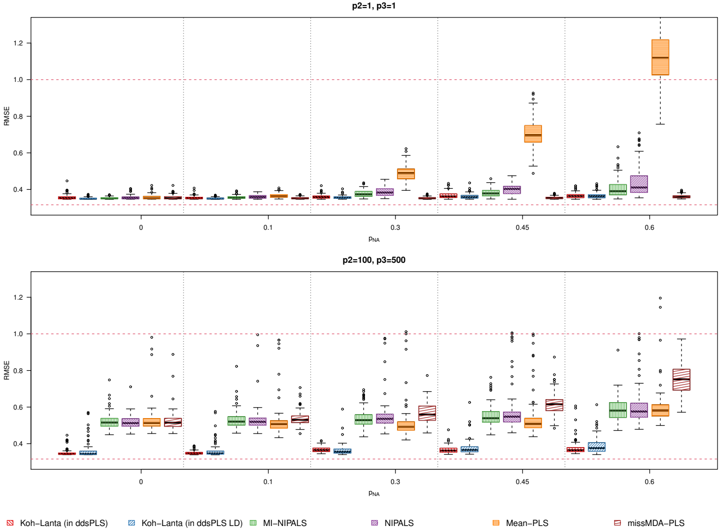
- $p_1 = 10$: information variance,
- $(p_2, p_3) \in \{(1, 1), (100, 500)\}$: low noise and high noise variances (structured and unstructured),
- $n \in \{100, 50, 20\}$: hard to very hard problems.

Remark: Use of factorial linear methods, the number of components R should be equal to $R = 2$.

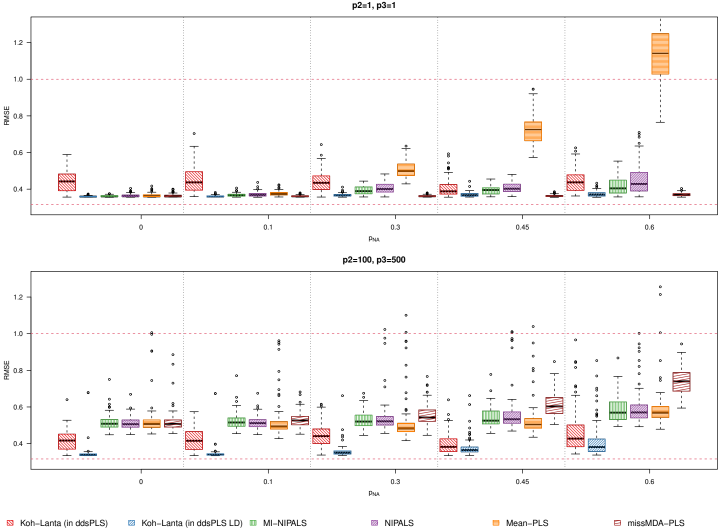
$n = 100$, R


p1=1, p3=1
R=1
p1=1, p3=1
R=2
p1=100, p3=500
R=1
p1=100, p3=500
R=2
0
20
40
60
80
100
pNA=0
pNA=0.1
pNA=0.3
pNA=0.45
pNA=0.6
Koh−Lanta (in ddsPLS)
Koh−Lanta (in ddsPLS LD)
MI−NIPALS
NIPALS
Mean−PLS
missMDA−PLS

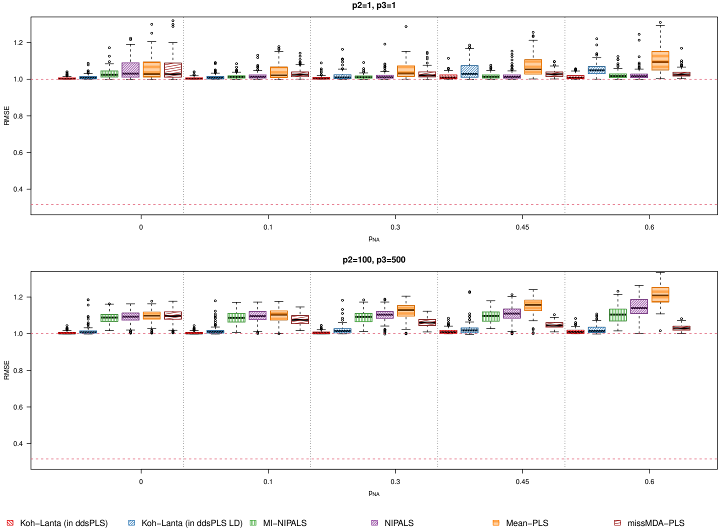
$n = 100$, RMSE for y_1



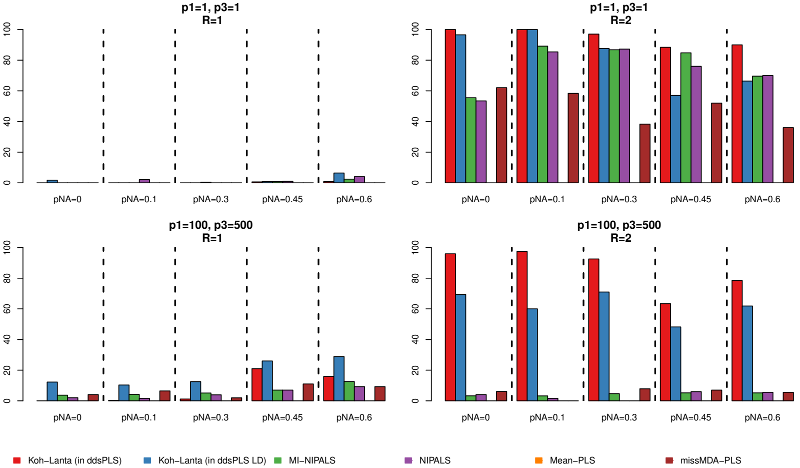
$n = 100$, RMSE for y_2



$n = 100$, RMSE for y_3

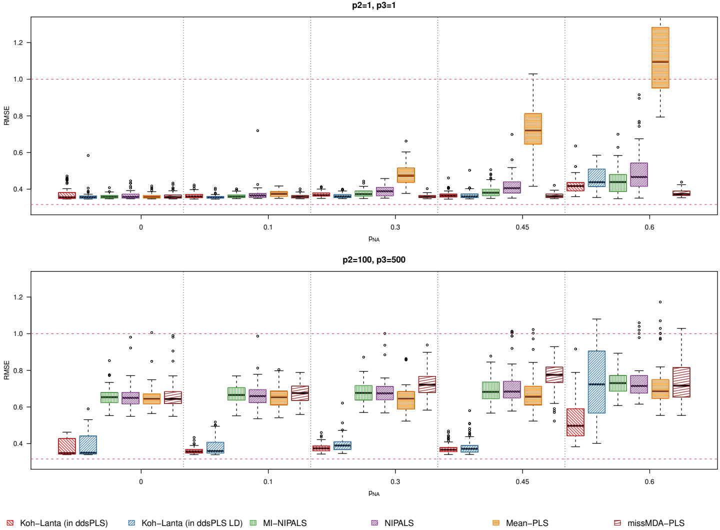


$n = 50$, R

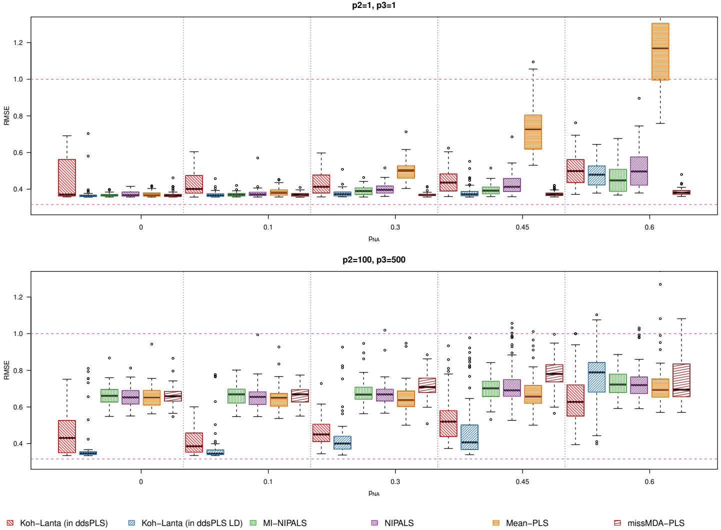


p1=1, p3=1
R=1
p1=1, p3=1
R=2
p1=100, p3=500
R=1
p1=100, p3=500
R=2
0
20
40
60
80
100
pNA=0
pNA=0.1
pNA=0.3
pNA=0.45
pNA=0.6
Koh−Lanta (in ddsPLS)
Koh−Lanta (in ddsPLS LD)
MI−NIPALS
NIPALS
Mean−PLS
missMDA−PLS

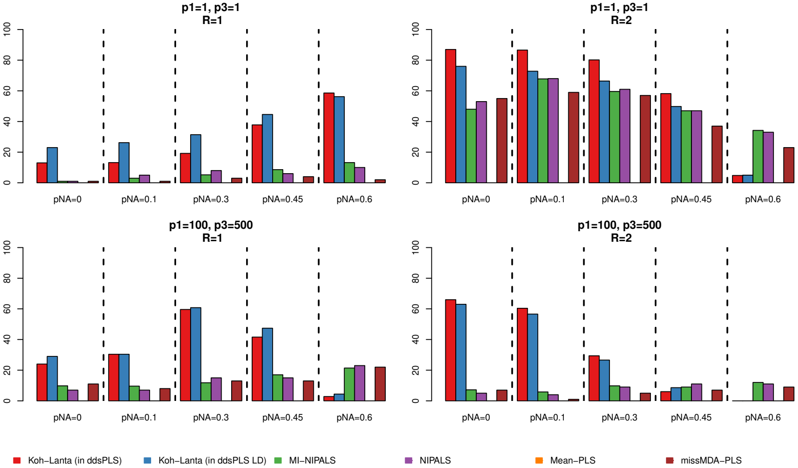
$n = 50$, RMSE for y_1



$n = 50$, RMSE for y_2

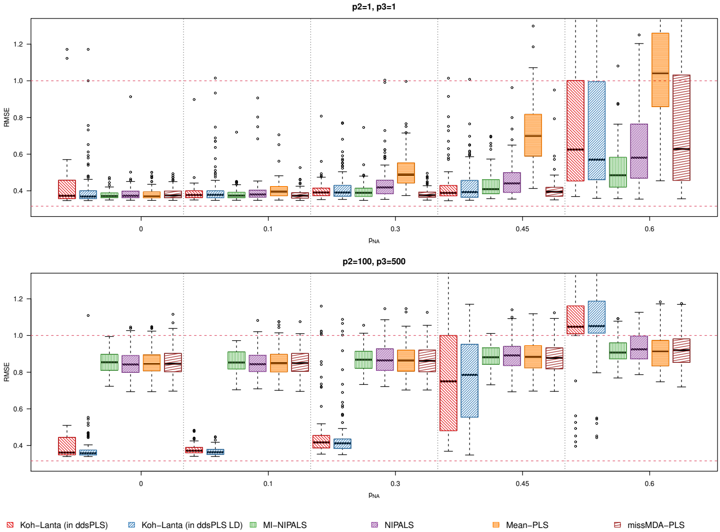


$n = 20$, R

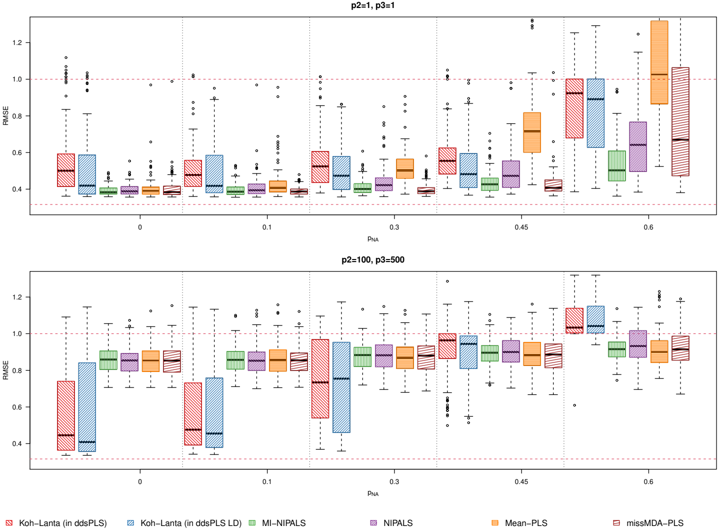


p1=1, p3=1
R=1
p1=1, p3=1
R=2
p1=100, p3=500
R=1
p1=100, p3=500
R=2
0
20
40
60
80
100
pNA=0
pNA=0.1
pNA=0.3
pNA=0.45
pNA=0.6
Koh−Lanta (in ddsPLS)
Koh−Lanta (in ddsPLS LD)
MI−NIPALS
NIPALS
Mean−PLS
missMDA−PLS

$n = 20$, RMSE for y_1



$n = 20$, RMSE for y_2



Summary

- "**MEAN-PLS**" suffers for large proportion of **NA**.

Summary

- “**MEAN-PLS**” suffers for large proportion of **NA**.
- “**missMDA-PLS**” is efficient for low noise but not efficient in high dimension.

Summary

- "**MEAN-PLS**" suffers for large proportion of **NA**.
- "**missMDA-PLS**" is efficient for low noise but not efficient in high dimension.
- "**Koh-Lanta (in ddsPLS)**" suffers for complex structures but "**Koh-Lanta (in ddsPLS LD)**" shows higher FPR in variable selection.

Summary

- "**MEAN-PLS**" suffers for large proportion of **NA**.
- "**missMDA-PLS**" is efficient for low noise but not efficient in high dimension.
- "**Koh-Lanta (in ddsPLS)**" suffers for complex structures but "**Koh-Lanta (in ddsPLS LD)**" shows higher FPR in variable selection.
- Competitors suffer more for high noise: no better than "**MEAN-PLS**" .

Summary

- “**MEAN-PLS**” suffers for large proportion of **NA**.
- “**missMDA-PLS**” is efficient for low noise but not efficient in high dimension.
- “**Koh-Lanta (in ddsPLS)**” suffers for complex structures but “**Koh-Lanta (in ddsPLS LD)**” shows higher FPR in variable selection.
- Competitors suffer more for high noise: no better than “**MEAN-PLS**” .
- $n = 20$ very hard but new methodologies suffer when they do not build models ($R = 0$).

Conclusion & future works

- Even in imputation, regularization helps for hard settings (low n and/or large p).
- Dig in "Partially Conditional Specification".
- Koh-Lanta is presented though **ddsPLS** but can be generalized to any supervised context.

References I

[1] Vincent Audigier, François Husson, and Julie Josse. “A principal component method to impute missing values for mixed data”. In: *Advances in Data Analysis and Classification* 10.1 (2016), pp. 5–26.

[2] Tony Cai and Weidong Liu. “Adaptive Thresholding for Sparse Covariance Matrix Estimation”. In: *Journal of the American Statistical Association* 106.494 (2011), 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`.

[3] Julie Josse, Jérôme Pagès, and François Husson. “Multiple imputation in principal component analysis”. In: *Advances in data analysis and classification* 5.3 (2011), pp. 231–246.

[4] Hadrien Lorenzo et al. “Data-driven sparse partial least squares”. In: *Statistical Analysis and Data Mining: The ASA Data Science Journal* (2021).

[5] Anne Rechtien et al. “Systems Vaccinology Identifies an Early Innate Immune Signature as a Correlate of Antibody Responses to the Ebola Vaccine rVSV-ZEBOV”. In: *Cell Reports* 20.9 (Sept. 2017), pp. 2251–2261. ISSN: 2211-1247. DOI: `10.1016/j.celrep.2017.08.023`.

References II

[6] Donald B. Rubin. “Multiple Imputation After 18+ Years”. In: *Journal of the American Statistical Association* 91.434 (1996), pp. 473–489. ISSN: 01621459. URL: http://www.jstor.org/stable/2291635.

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

[8] Johan Trygg and Svante Wold. “O2-PLS, a two-block (X–Y) latent variable regression (LVR) method with an integral OSC filter”. In: *Journal of chemometrics* 17.1 (2003), pp. 53–64.