Additional Supplementary Materials for "Adaptive Bayesian SLOPE – Model Selection with Incomplete Data"

Wei Jiang[1] Małgorzata Bogdan[2] Julie Josse[1] Szymon Majewski[3]
Błażej Miasojedow[4] Veronika Ročková[5] TraumaBase® Group[6]

March 31, 2021

Abstract

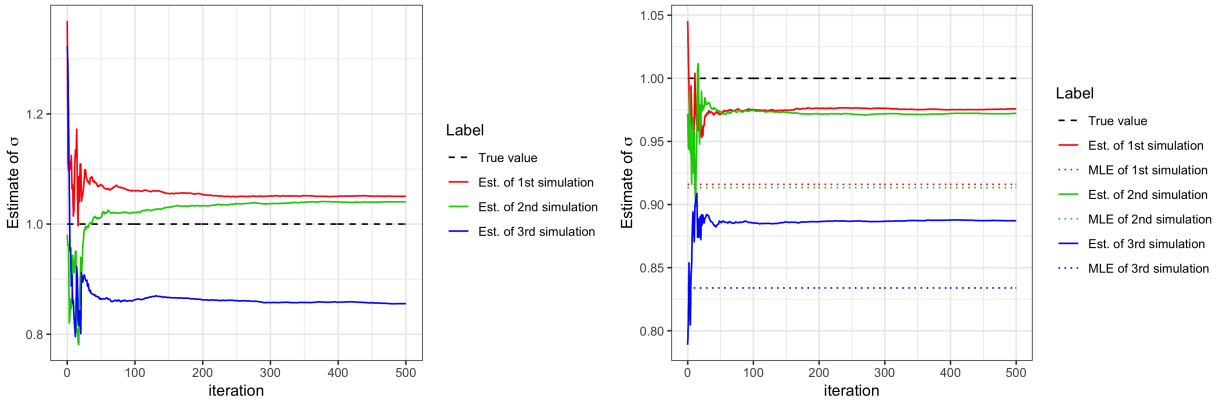
This document presents some supplementary simulation results for the paper "Adaptive Bayesian SLOPE – High-dimensional Model Selection with Missing Values" (Jiang et al., 2021).

Contents

[1]Inria XPOP and CMAP, École Polytechnique, France
[2]University of Wroclaw, Poland and Lund University, Sweden
[3]CMAP, École Polytechnique, France
[4]University of Warsaw, Poland
[5]University of Chicago Booth School of Business, USA
[6]Hôpital Beaujon, APHP, France

1 Convergence of SAEM: σ

Following the simulation study in Subsection 4.2 (Jiang et al., 2021), we represent the convergence curves for σ with *ABSLOPE* in Figure 1 (a). The behavior is the same as for the *beta* coefficients. We also represent convergence in the case without missing values in Figure 1 (b), in order to compare the estimate of σ by *ABSLOPE* (colored solid curves) to the biased MLE estimator without prior knowledge (colored dashed lines), *i.e.*, $\hat{\sigma}_{\text{MLE}} = \sqrt{\frac{RSS}{n}}$. We can see that the estimates of σ with both methods are biased downward, but since *ABSLOPE* has an additional correction term, it leads to a less biased estimator.



(a) With 10% missing values

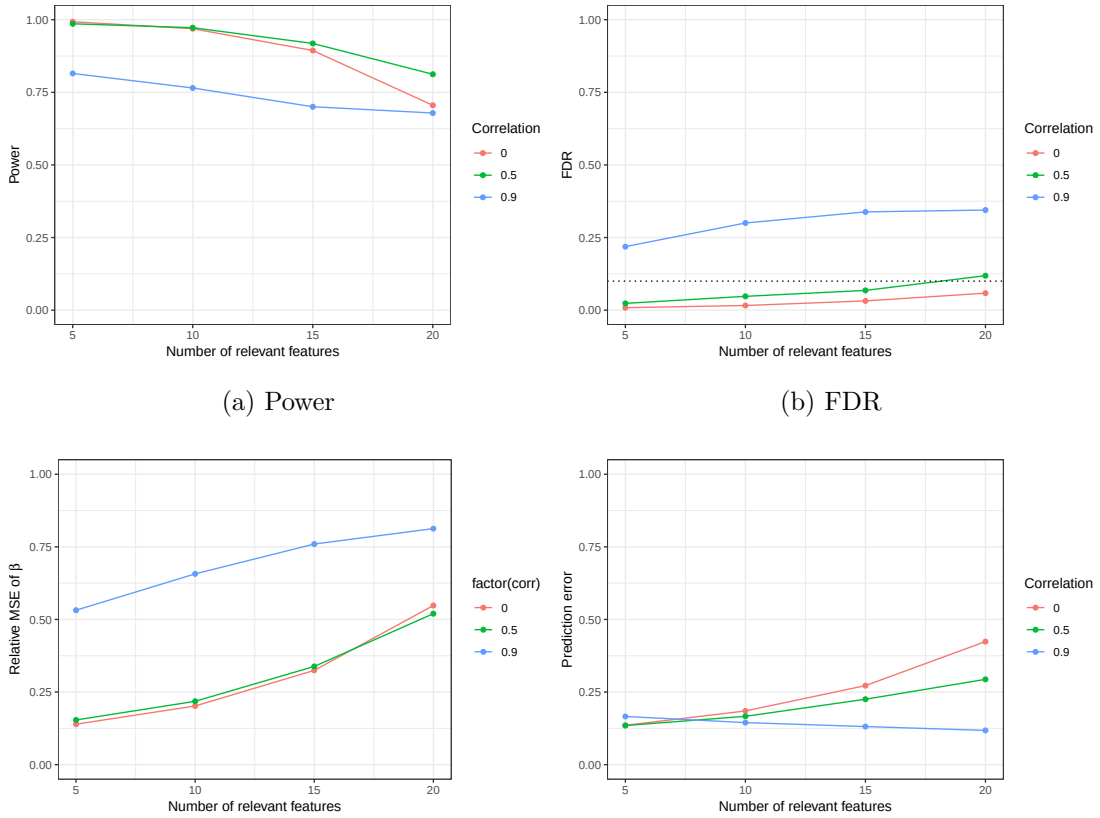
(b) Without missing values

Figure 1: Convergence plots for σ with *ABSLOPE* (colored solid curves). (a) Case with 10% missing values; (b) Case without missing values. Black dash line represents the true value for σ. In (b) Colored dash lines indicate the biased MLE $\hat{\sigma}_{\text{MLE}} = \sqrt{\frac{RSS}{n}}$. Estimates obtained with three different sets of simulated data are represented by three different colors.

2 Behavior of ABSLOPE: effect of correlation

Following the simulation study in Subsection 4.3 (Jiang et al., 2021), we consider additional scenarios varying correlation as follows.

2.1 $n = p = 100$, 10% missingness, strong signal - vary correlation



(a) Power (b) FDR

(c) Bias of β (d) Prediction error

Figure 2: Mean of power (a), FDR (b), bias of the estimate for β (c) and prediction error (d), as function of length of true signal, over the 200 simulations. Results for $n = p = 100$, with 10% missingness and strong signal.

We consider a small dataset $n = p = 100$. The signal strength is strong and equals to $3\sqrt{2 \log p}$ and the percentage of missingness is 10%. We then vary the sparsity and correlation. The results in Figure 2 show:

- When there is no or little correlation, the FDR is controlled to the desired level of

0.1, but in case of high correlation, the control of the FDR is lost.

- The existence of a correlation can give more power. On one hand, the generation of missing covariates depends on those observed; on the other hand, the structure among covariates improves the prediction performances.

3 Behavior of ABSLOPE: robustness to the Gaussian assumption for covariates

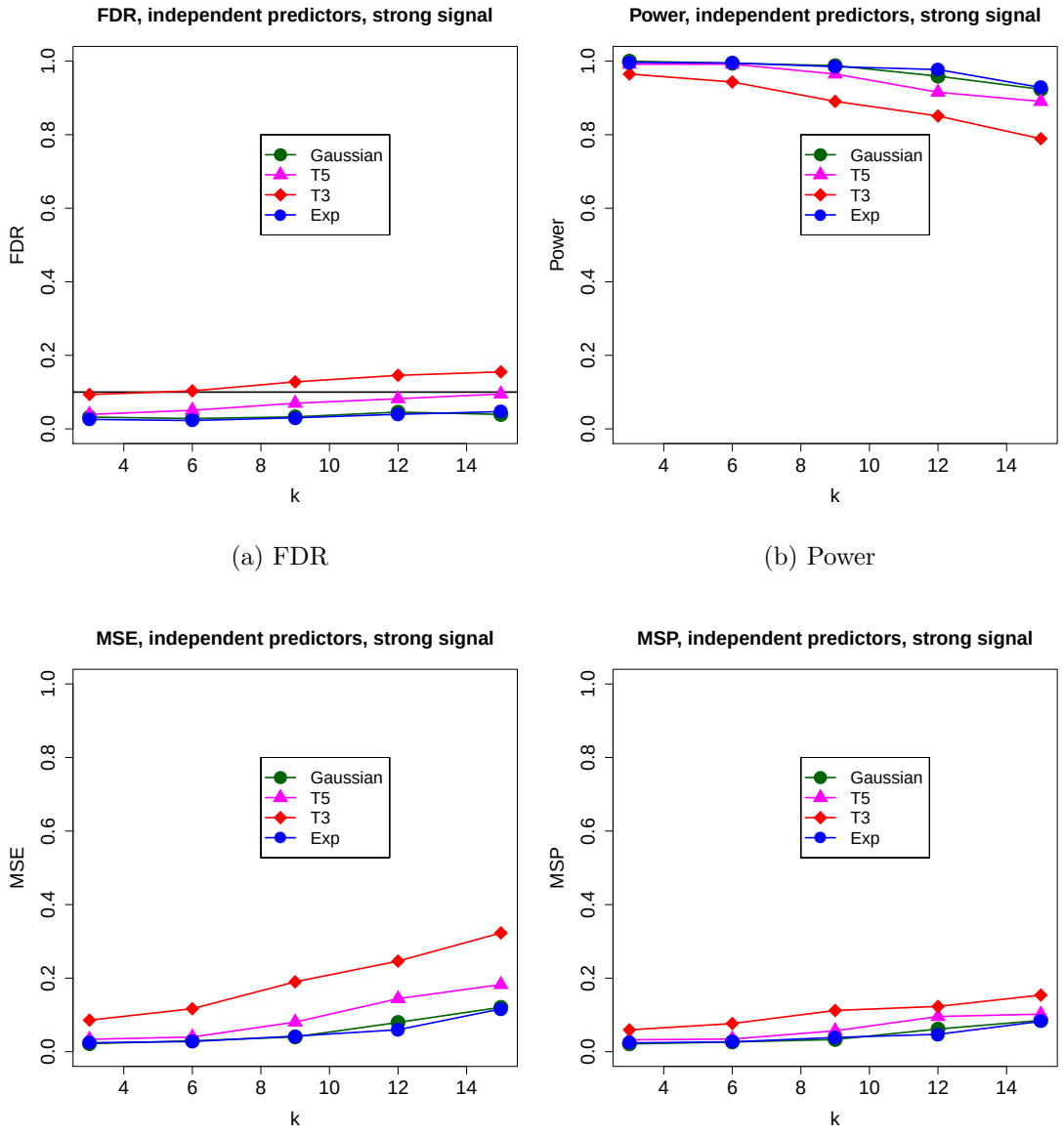
Following the simulation study in Subsection 4.3 (Jiang et al., 2021), we consider additional scenarios under misspecified models, *i.e.*, when the covariates don't follow a Gaussian assumption. In the case with miss-pecified models, we generated a design matrix of size $n = 100, p = 100$ by drawing independent observations from

- Student distribution with 3 degrees of freedom
- Student distribution with 5 degrees of freedom
- Exponential distribution with $\lambda = 1$

The results in Figure 3 show that the power, FDR, estimation bias for regression coefficients, and prediction error with the proposed ABSLOPE based on the Gaussian assumptions are preserved when the covariates are generated from the exponential distribution. These characteristics are more sensitive to the heavy-tailed t-distribution, but FDR is still controlled very close to the nominal level and the power does not suffer much from the violations of the normality assumption.

4 SLOBE vs SLOBE with pre-estimated σ

Figures 4 - 7 compare the performance of SLOBE and SSLASSO algorithms using the build-in procedures for estimating σ and relying on the initial estimate by cross-validated LASSO provided in equation (19) of (Jiang et al., 2021). In case of SSLASSO this estimator is used also as the starting point for the version of the algorithm with unknown σ.



(a) FDR

(b) Power

(c) MSE

(d) MSP

Figure 3: Estimated FDR (a), Power (b), relative mean squared error of β estimation (c) relative mean square prediction error (d), as a function of the number of true nonzero regression coefficients. Results for $n = p = 100$, with 10% missingness and strong signal, under misspecified models, averaged over 200 replicates.

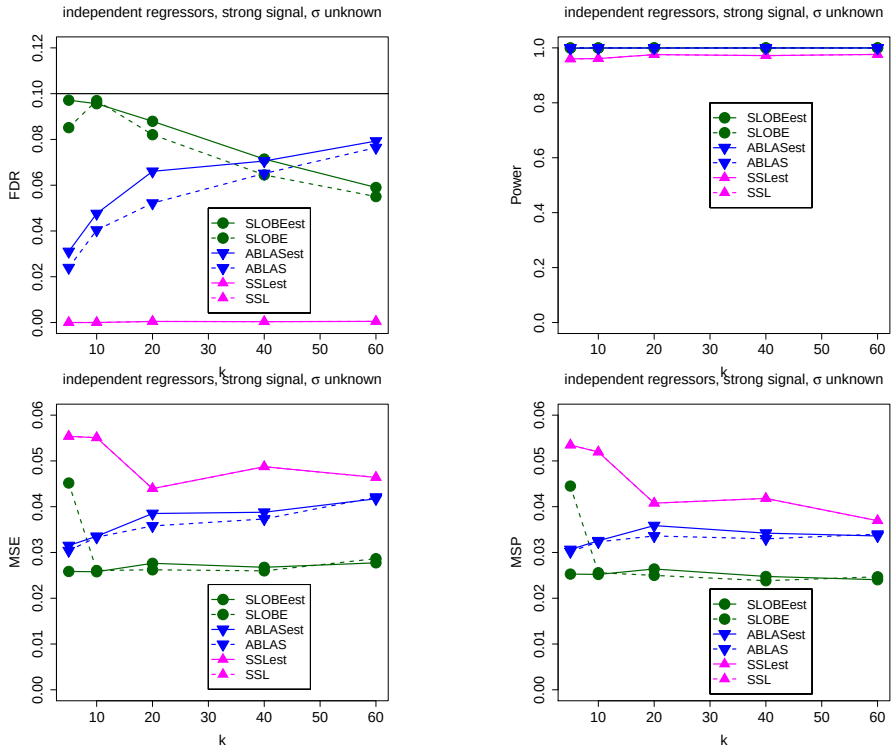


Figure 4: Different performance measures as the function of the number of true nonzero regression coefficients for complete data, independent regressors and strong signals. Extension 'est' refers to procedures using a preliminary estimator of σ provided in the formula (17) in (Jiang et al., 2021).

Figures 4 - 7 show that the two versions of SSLASSO yield exactly the same results. In case of SLOBE and ABLAS we observe that the procedures using build-in functions for σ estimation are slightly more conservative than the procedures using the estimator from the cross-validated LASSO. The largest difference occurs for the independent regressors and weak signals, where SLOBE based on build-in estimator systematically has FDR below the nominal level, while SLOBE using the estimator of σ is more similar to its version using the true σ value and has FDR slightly above the nominal level.

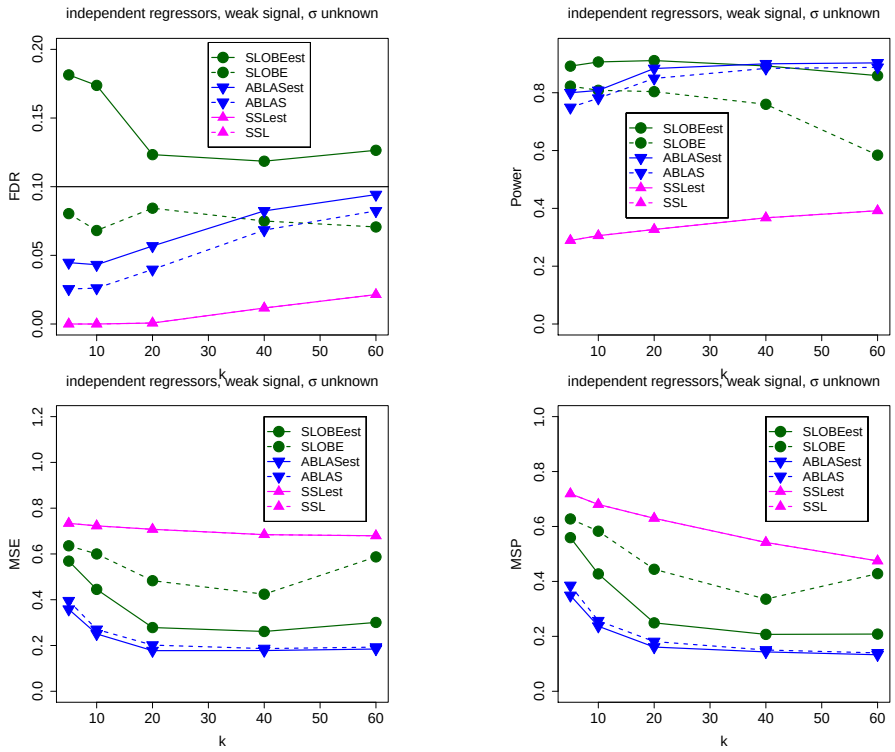


Figure 5: Different performance measures as the function of the number of true nonzero regression coefficients for complete data, independent regressors and weak signals. Extension 'est' refers to procedures using a preliminary estimator of σ provided in the formula (17) in (Jiang et al., 2021).

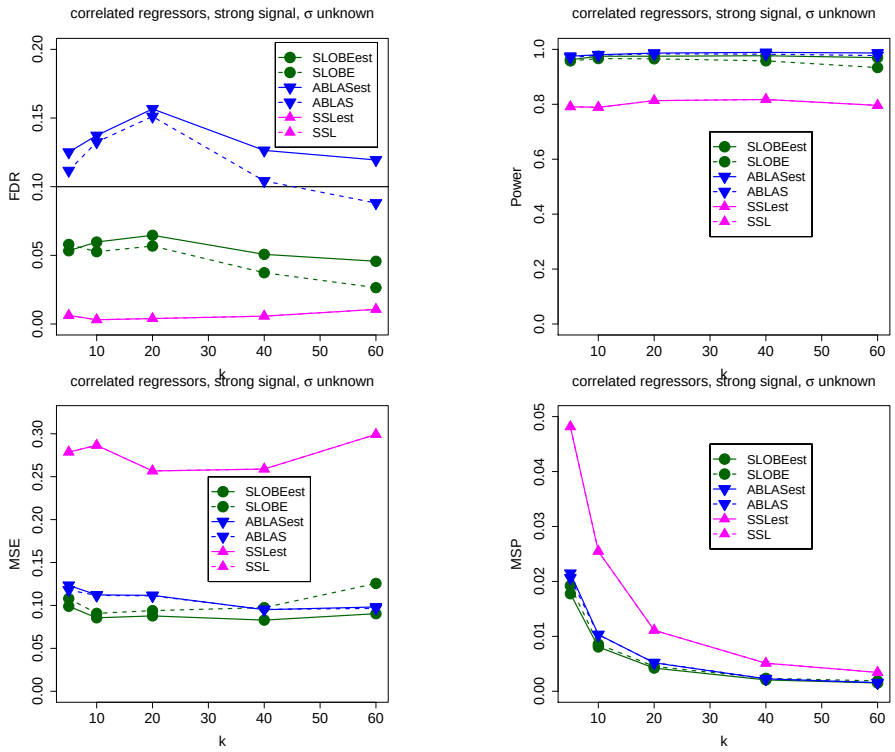


Figure 6: Different performance measures as the function of the number of true nonzero regression coefficients for complete data, correlated regressors and weak signals. Extension 'est' refers to procedures using a preliminary estimator of σ provided in the formula (17) in (Jiang et al., 2021).

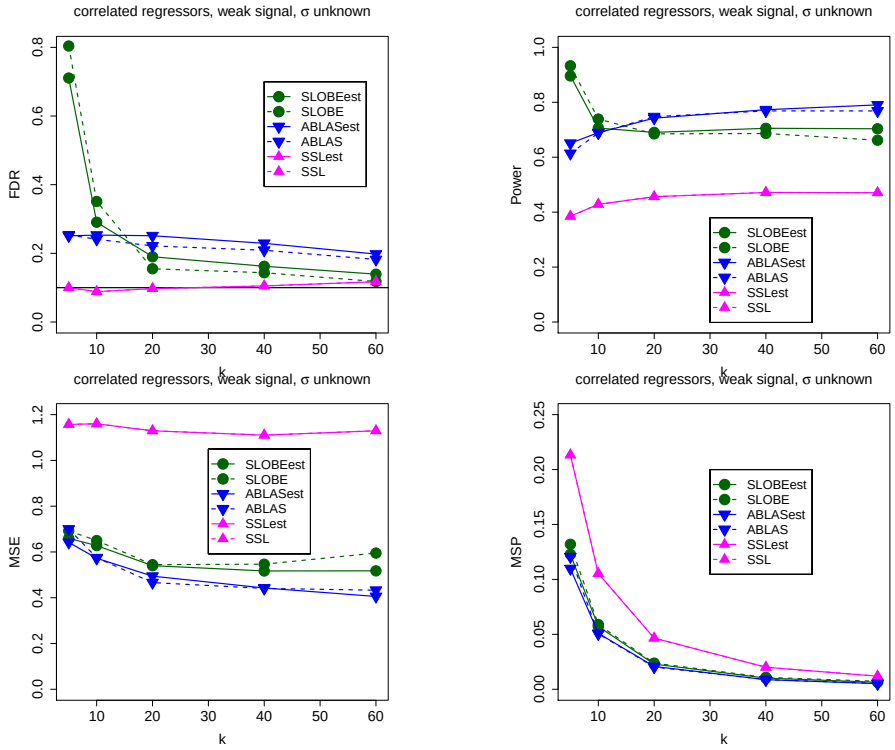


Figure 7: Different performance measures as the function of the number of true nonzero regression coefficients for complete data, correlated regressors and weak signals. Extension 'est' refers to procedures using a preliminary estimator of σ provided in the formula (17) in (Jiang et al., 2021).

5 Variables in the TraumaBase dataset and preprocessing

Following the introduction of TraumaBase dataset in Subsection 5.1 (Jiang et al., 2021), we give the detailed explanation of the variables in the TraumaBase dataset:

- *Age:* Age
- *SI:* Shock index indicates level of occult shock based on heart rate (HR) and systolic blood pressure (SBP). $SI = \frac{HR}{SBP}$. Evaluated on arrival of hospital.
- *MBP:* Mean arterial pressure is an average blood pressure in an individual during a single cardiac cycle, based on systolic blood pressure (SBP) and diastolic blood pressure (DBP). $MBP = \frac{2DBP+SBP}{3}$. Evaluated on arrival of hospital.
- *Delta.hemo:* The difference between the hemoglobin on arrival at hospital and that in the ambulance.
- *Time.amb:* Time spent in the ambulance *i.e.*, transportation time from accident site to hospital, in minutes.
- *Lactate:* The conjugate base of lactic acid.
- *Temp:* Patient's body temperature.
- *HR:* heart rate measured on arrival of hospital.
- *VE:* A volume expander is a type of intravenous therapy that has the function of providing volume for the circulatory system.
- *RBC:* A binary index which indicates whether the transfusion of Red Blood Cells Concentrates is performed.
- *SI.amb:* Shock index measured on ambulance.
- *MAP.amb:* Mean arterial pressure measured in the ambulance.
- *HR.max:* Maximum value of measured heart rate in the ambulance.

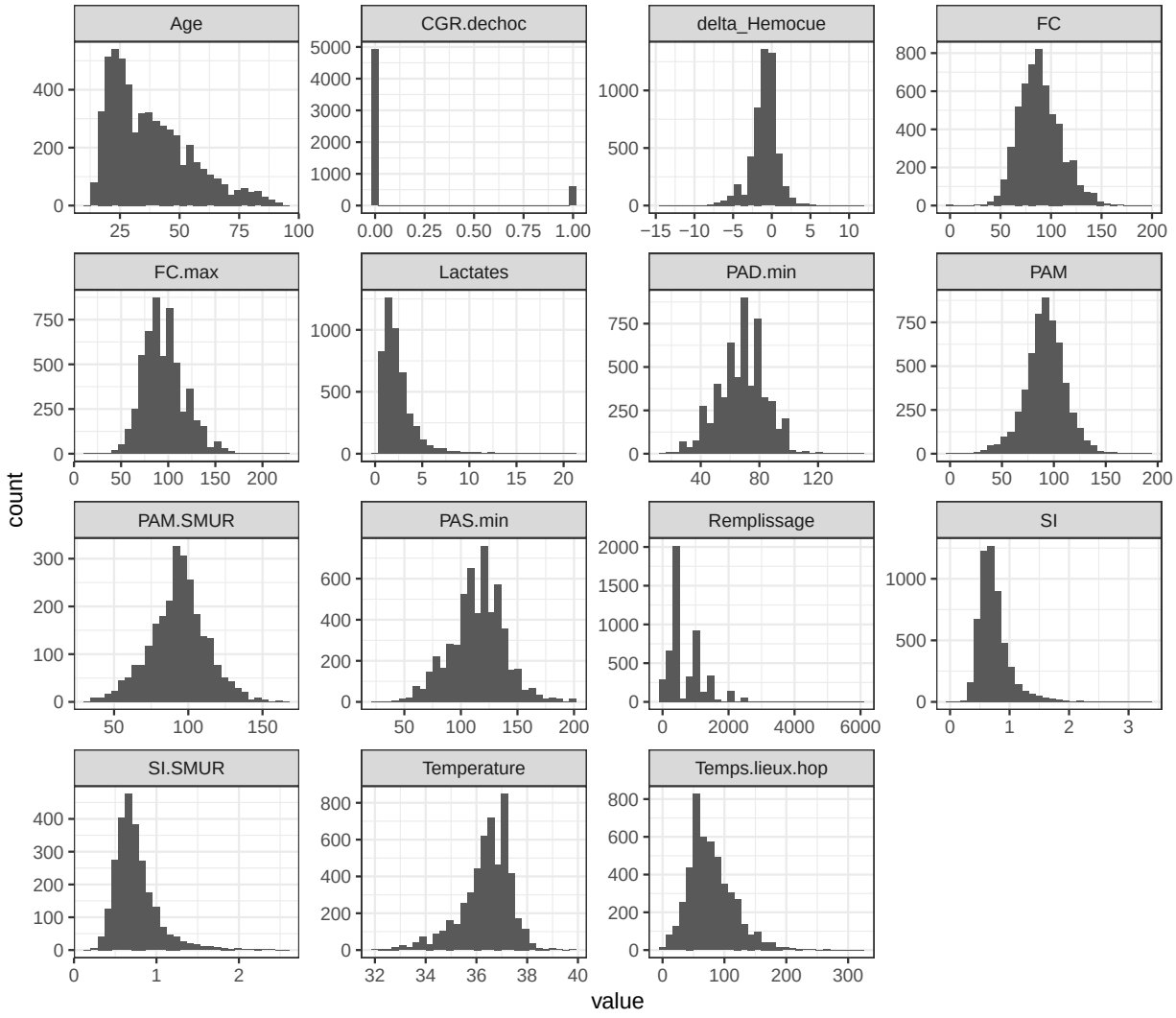
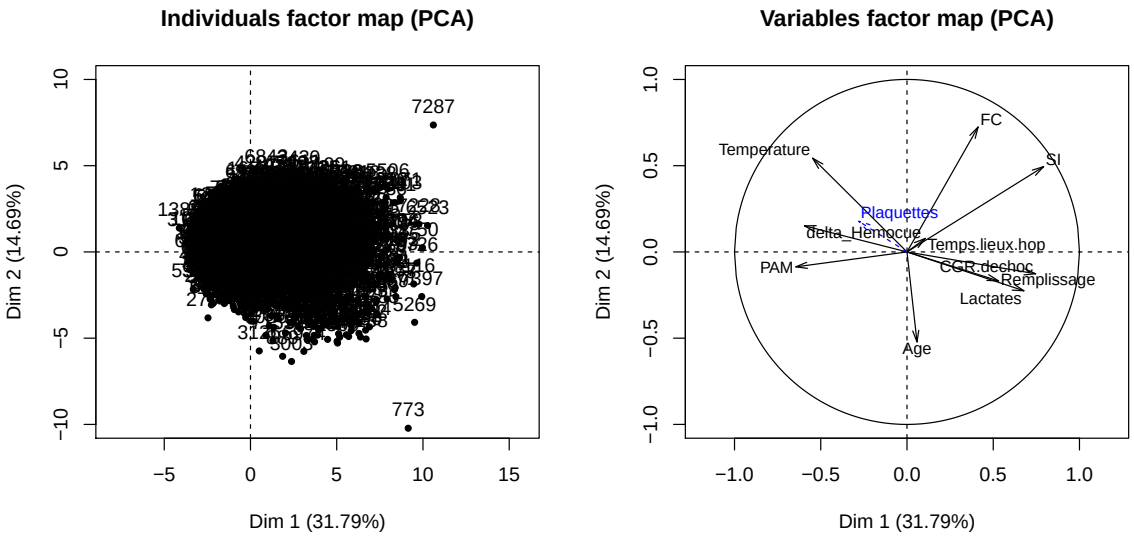


Figure 8: Histograms of pre-selected variables from TraumaBase.

- *SBP.min:* Minimum value of measured systolic blood pressure in the ambulance.
- *DBP.min:* Minimum value of measured diastolic blood pressure in the ambulance.

The distribution of each variable is displayed as Figure 8.

With PCA, we visualized the individual and variable factor map on the two first dimension. As shown on the left in Figure 9, there were two observations regarded as outliers. In details, the temperature of 773$^{\text{th}}$ patient was measured as 12.3, while the MBP of 7287$^{\text{th}}$ patient was only 38.33, which both stand for a mistake of record.

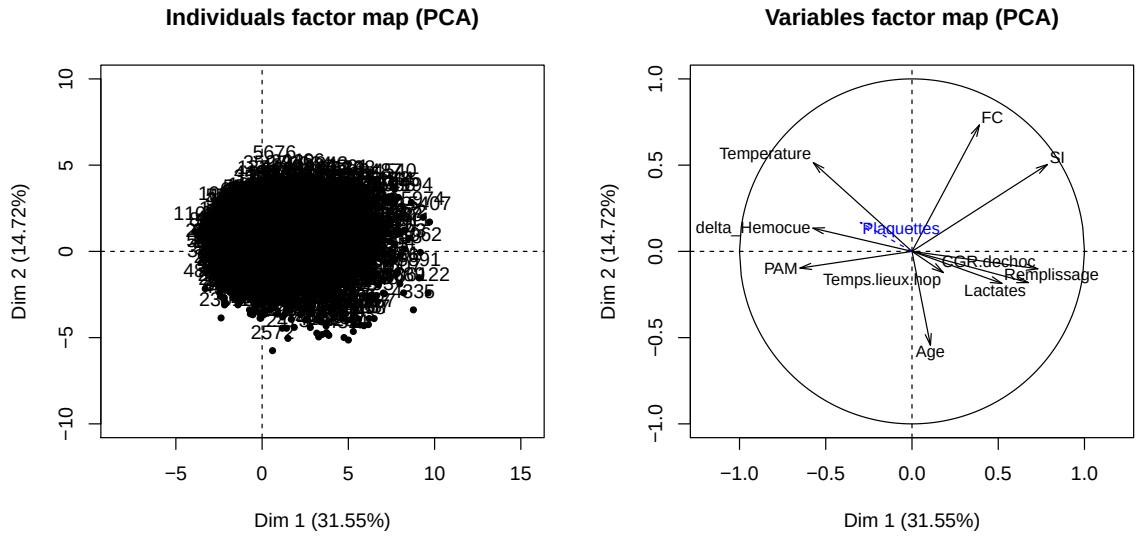


(a) Observation's factor map

(b) Variable's factor map

Figure 9: The factor maps from PCA before correction of wrongly recorded entries. (a) Observation's factor map (b) Variable's factor map.

We corrected all the mistakes in the records, for example, converting the value temperature smaller than 34 degree to *NA* and recalculating the MBP with the same unity for SBP. After that, we presented the factor maps from PCA in Figure 10, where the distribution of individuals in the principal dimensions were more homogeneous and the outliers disappeared .



(a) Observation's factor map

(b) Variable's factor map

Figure 10: The factor maps from PCA after correction of wrongly recorded entries. (a) Observation's factor map (b) Variable's factor map.

5.1 Comparison of computation time

Table 1 presents the execution time of the different methods that handle missing values considered in the simulation. For the methods based on imputation, we use the mean imputation as it is the quickest method. In addition, we have implemented our proposed SLOBE algorithm in C and integrated it within R. In the case $n = p = 100$, we observe that the most time consuming method is ABSLOPE, which takes on average 14 seconds for one run. Instead, the simplified version SLOBE reduced this cost to 0.3 seconds, which is shorter than the MeanImp + ALAS. When $n = p = 500$, the convergence of ABSLOPE requires much more time but SLOBE helps to simplify the complexity, which makes it capable of handling larger data sets.

References

Jiang, W., Bogdan, M., Josse, J., Majewski, S., Miasojedow, B., Ročková, V., and Group, T. (2021). Adaptive Bayesian SLOPE – model selection with incomplete data.

Table 1: Comparison of average execution time (in seconds) for one simulation, in the case without correlation and with 10% MCAR, for $n = p = 100$ and $n = p = 500$ calculated over 200 simulations. (MacBook Pro, 2.5 GHz, processor Intel Core i7)

Execution time (seconds)	$\boldsymbol{n = p = 100}$			$\boldsymbol{n = p = 500}$		
for one simulation	min	mean	max	min	mean	max
ABSLOPE	12.83	14.33	20.98	646.53	696.09	975.73
SLOBE	0.31	0.34	0.66	14.23	15.07	29.52
MeanImp + SLOPE	0.01	0.02	0.09	0.24	0.28	0.53
MeanImp + LCV	0.10	0.14	0.32	1.75	1.85	3.06
MeanImp + ALAS	0.45	0.58	1.12	45.06	47.20	71.24
MeanImp + SSL	0.05	0.06	0.11	0.56	0.62	1.06