Additional Supplementary Materials for "Adaptive Bayesian SLOPE – High-dimensional Model Selection with Missing Values"

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

August 29, 2019

Abstract

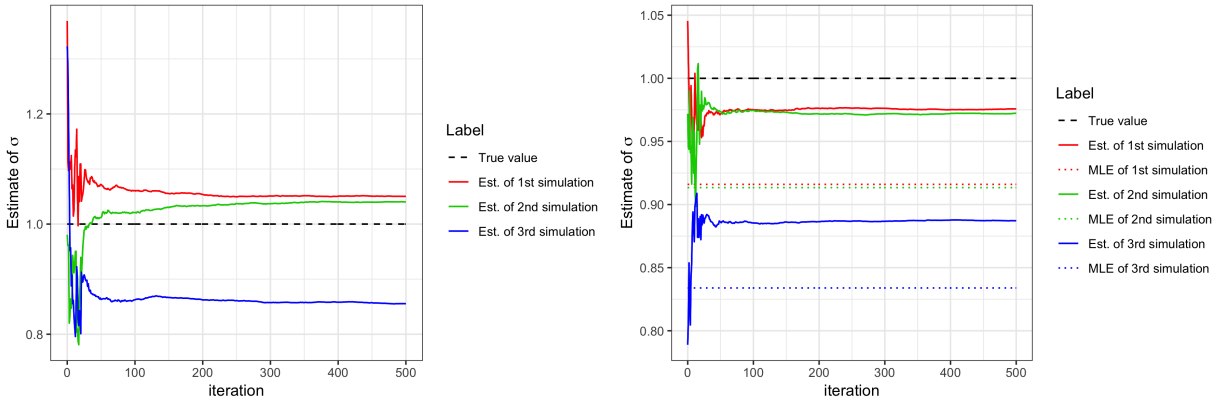
This document presents some supplementary simulation results for the paper "Adaptive Bayesian SLOPE – High-dimensional Model Selection with Missing Values" (Jiang et al., 2019).

Contents

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

1 Convergence of SAEM: σ

Following the simulation study in Subsection 4.1 (Jiang et al., 2019), 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

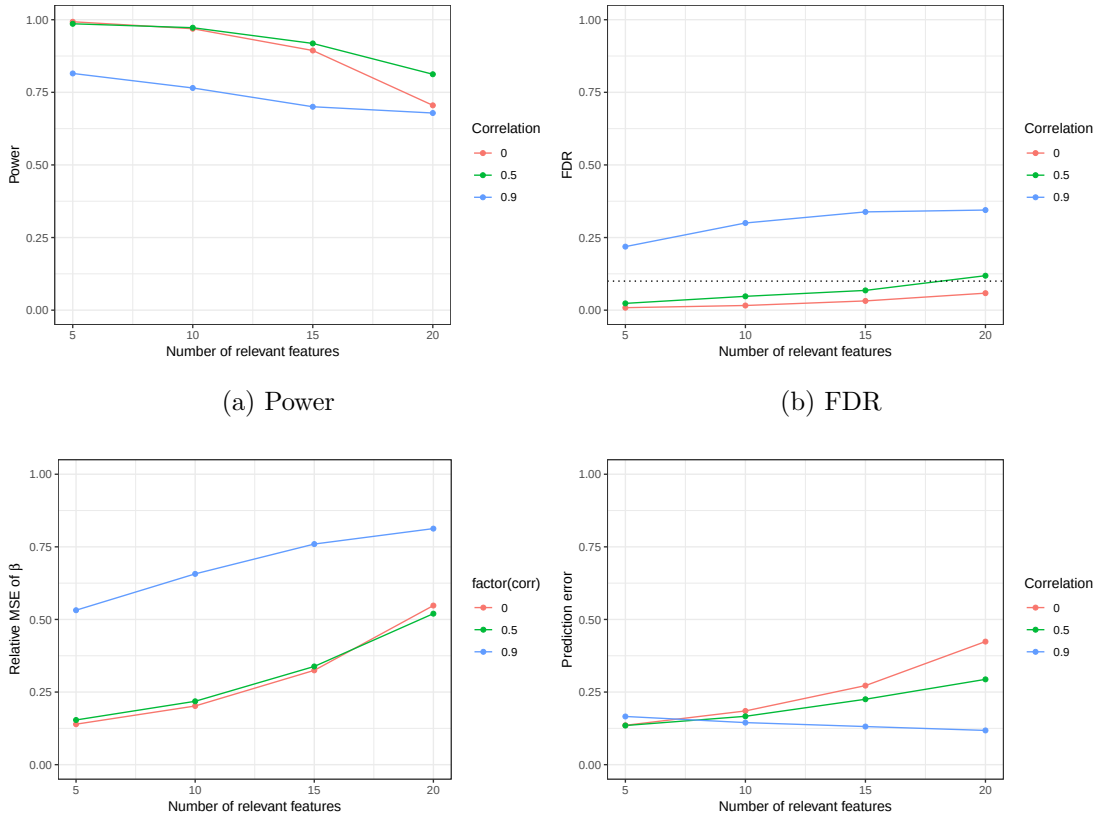
(b) Without missing values

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., 2019), 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.

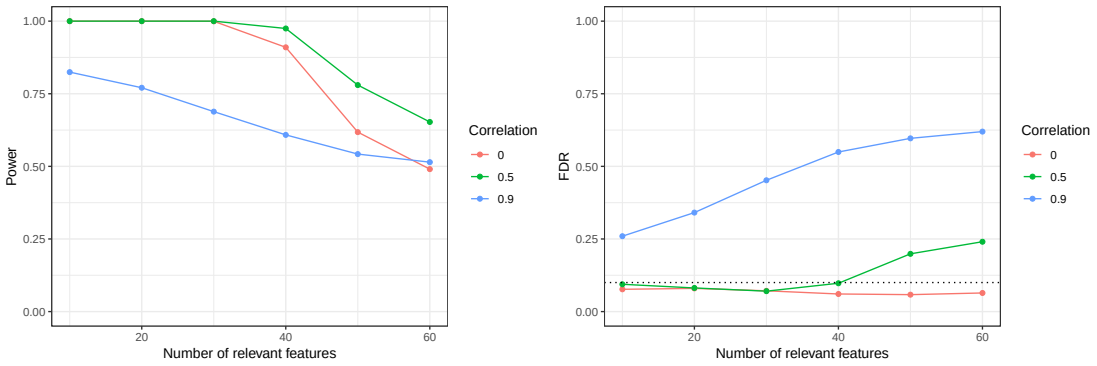
We consider a small dataset $n = p = 100$. The signal strength is strong and equals to $3\sqrt{2 \log p}$ and the percentage of missingness is 10%. We then vary the sparsity and correlation. The results in Figure 2 show:

- 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.

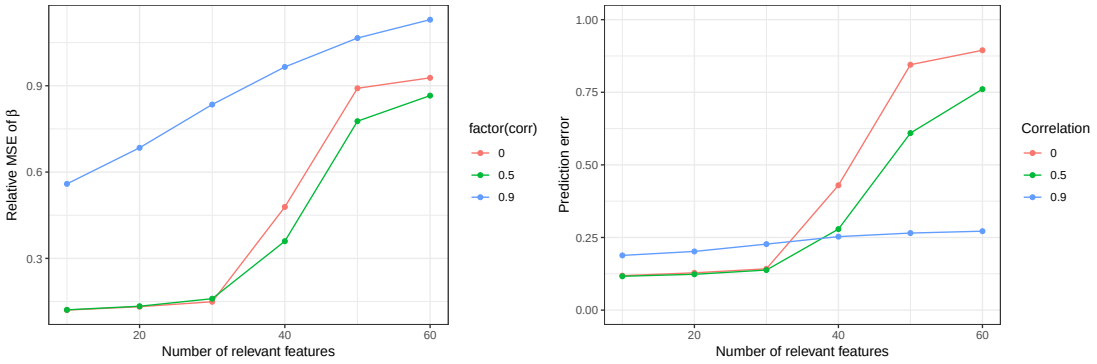
- The existence of a correlation can give more power. On on hand, the generation of missing covariates depends on those observed; on the other hand, the structure among covariates improves the prediction performances.

2.2 $n = p = 500$, 10% missingness, strong signal - vary correlation



(a) Power

(b) FDR

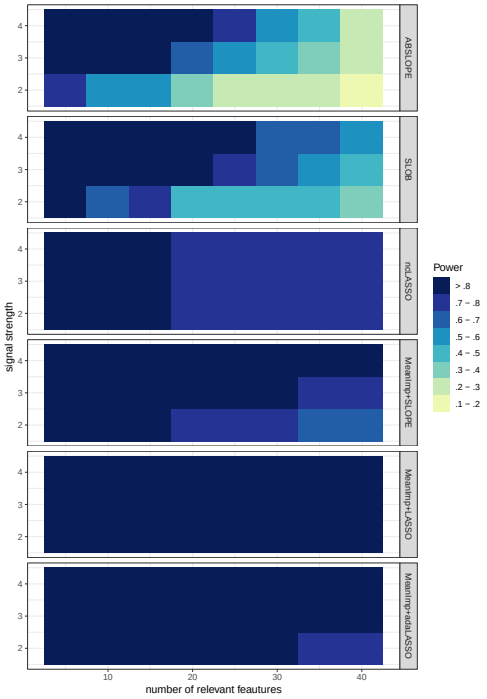


(c) Bias of β

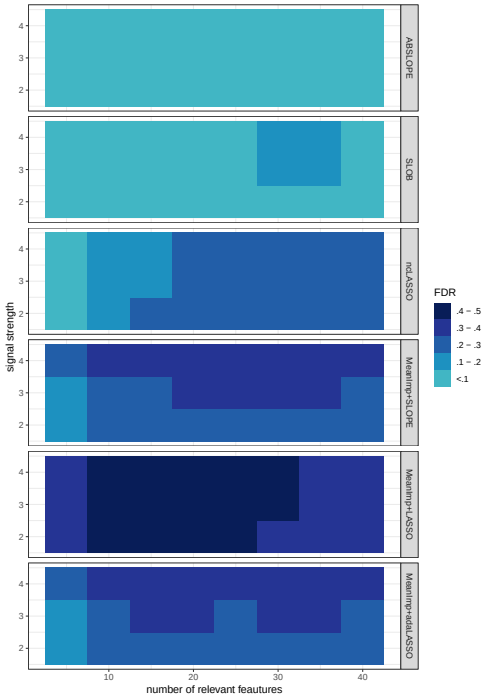
(d) Prediction error

Figure 3: 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 $n = p = 500$, with 10% missingness and strong signal.

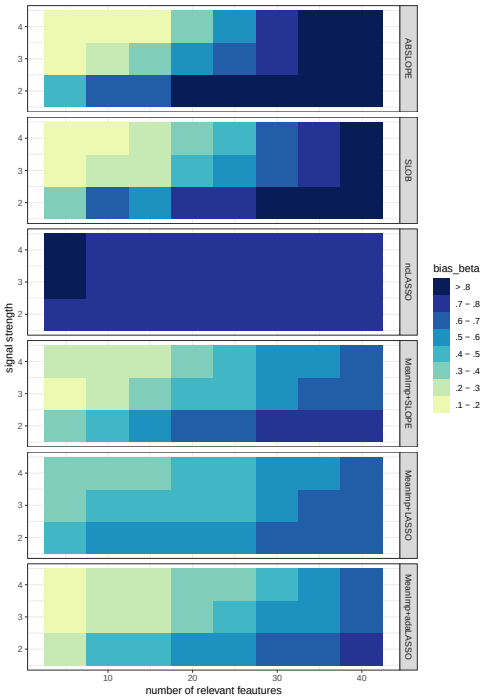
We consider a larger dataset $n = p = 500$ while the other parameters same as before. We then vary the sparsity and correlation. The results in Figure 3 show the same phenomenon as Figure 2 for the effect of correlation on FDR control.



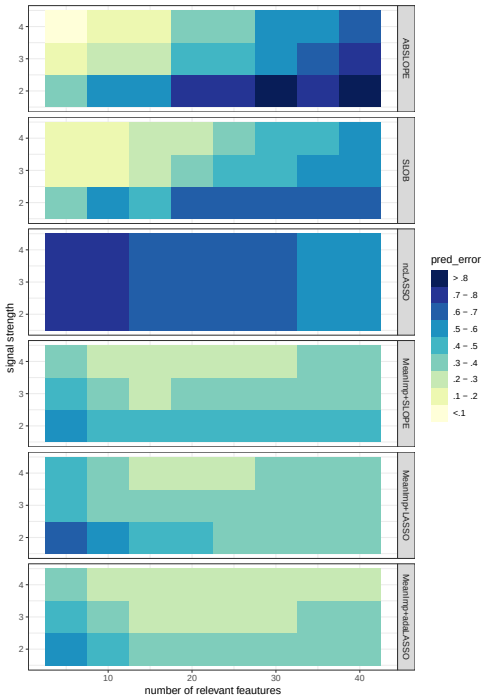
(a) Power



(b) FDR



(c) Bias of β



(d) Prediction error

Figure 4: Comparison of power (a), FDR (b), bias of β (c) and prediction error (d) with varying sparsity and signal strength, with 10% missingness over 200 simulations in the case without correlation.

3 Comparison with competitors: $n = p = 100$

Following the simulation study in Subsection 4.4 (Jiang et al., 2019), we compare the proposed methodology with its competitors as follows. Figure 4 summarizes the result for the case $n = p = 100$, 10% missingness and without correlation. Lighter colors indicate smaller values.

- *ABSLOPE* and *SLOB* both have a strong power and an accurate prediction when the sparsity is large and the signal strength is strong enough;
- FDR is always controlled with *ABSLOPE* or *SLOB*. Other methods pay a price in FDR control to achieve good power.

4 Variables in the TraumaBase dataset and preprocessing

Following the introduction of TraumaBase dataset in Subsection 5.1 (Jiang et al., 2019), 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.

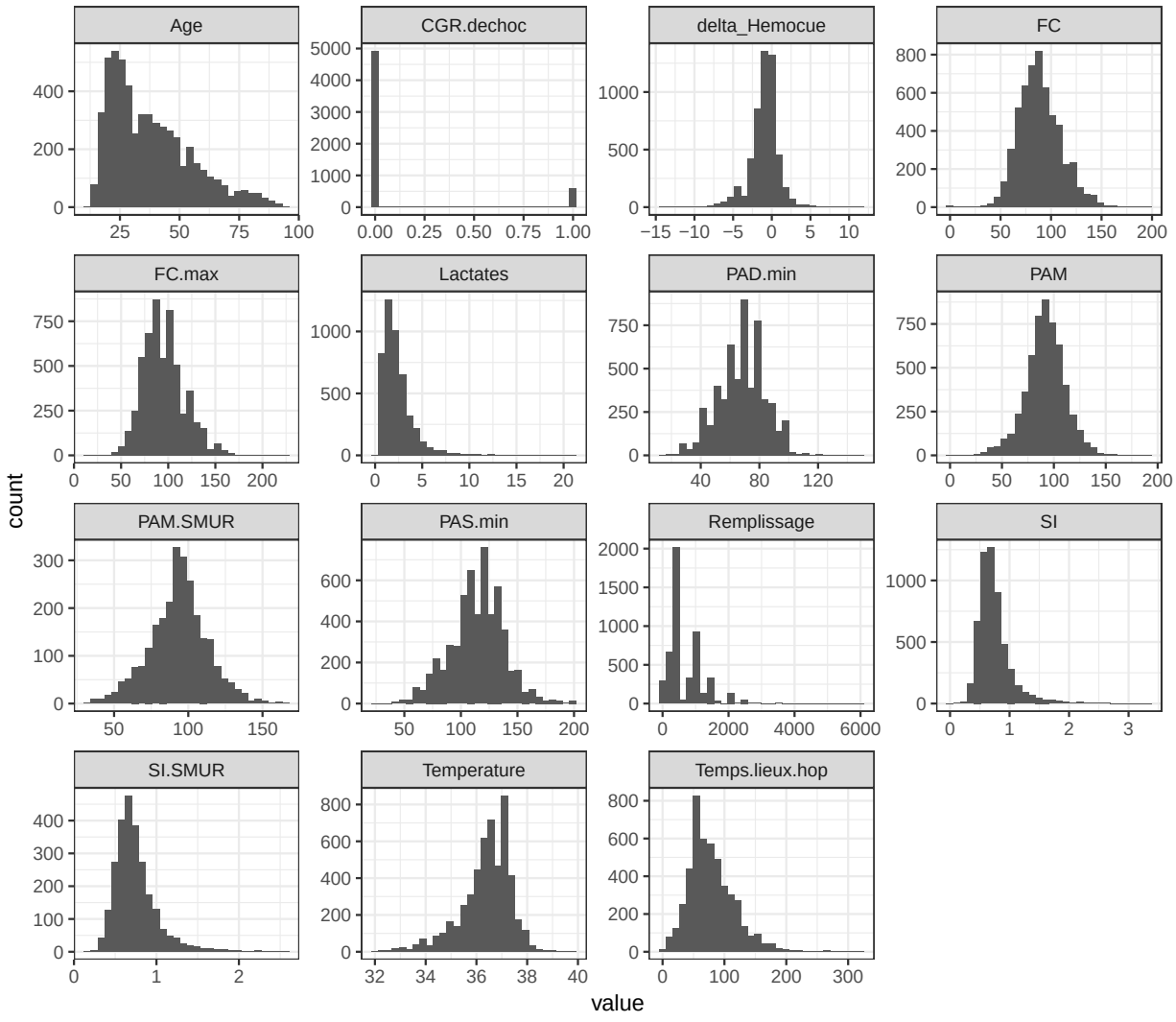
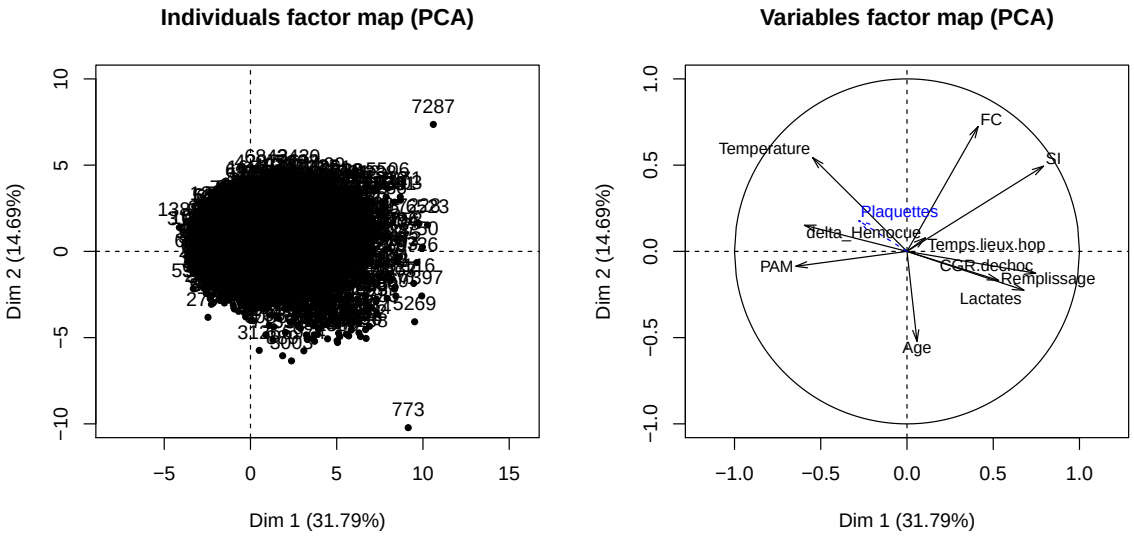


Figure 5: Histograms of pre-selected variables from TraumaBase.

- *MAP.amb:* Mean arterial pressure measured in the ambulance.
- *HR.max:* Maximum value of measured heart rate in the ambulance.
- *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 5.

With PCA, we visualized the individual and variable factor map on the two first dimension. As shown on the left in Figure 6, 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.

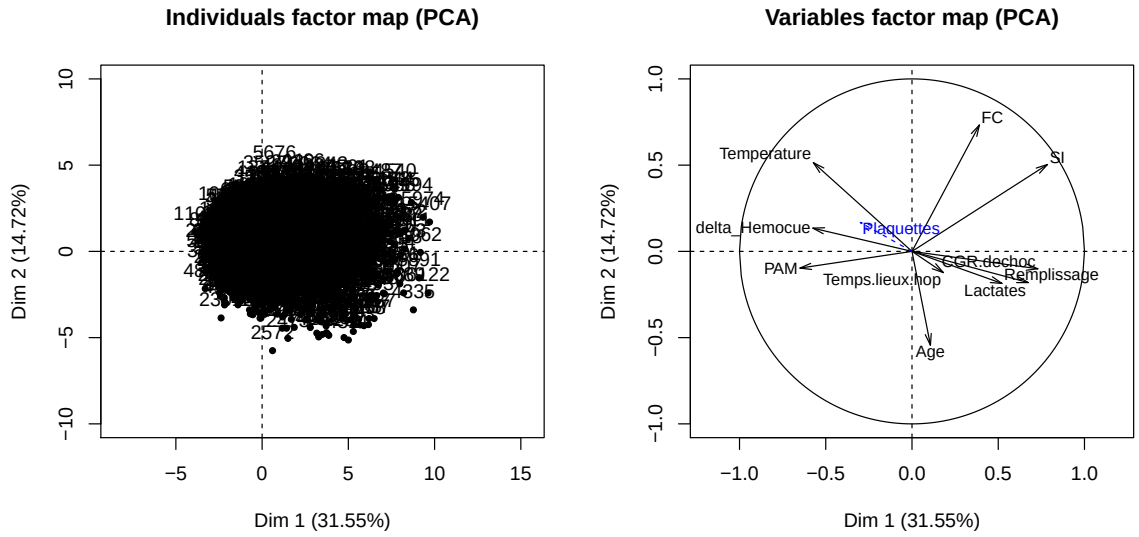


(a) Observation's factor map

(b) Variable's factor map

Figure 6: 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 7, 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 7: The factor maps from PCA after correction of wrongly recorded entries. (a) Observation's factor map (b) Variable's factor map.

References

Jiang, W., Bogdan, M., Josse, J., Miasojedow, B., Ročková, V., and Group, T. (2019). Adaptive Bayesian SLOPE – high-dimensional model selection with missing values. https://drive.google.com/open?id=1Y_FzqfmYZQHlapbm9AHQWZmcs49iwRU1.