Statistical methods for reduction of dimension of ”fat” data sets

Malgorzata Bogdan
University of Wroclaw (Poland), Lund University (Sweden)

CASUS Webinar, 2020

9th of October, 2020

Outline

- Sorted L-One Penalized Estimator (SLOPE)
- Adaptive Bayesian SLOPE
- Varclust - a new algorithm for subspace clustering

Motivation: Paris Hospital, TraumaBase Group Data

- *Traumabase*® data:
  20000 major trauma patients × 250 measurements..

Accident type	**Age**	**Sex**	**Blood pressure**	**Lactate**	**Temperature**	**Platelet** (G/L)
Falling	50	M	140		35.6	150
Fire	28	F		4.8	36.7	250
Knife	30	M	120	1.2		270
Traffic accident	23	M	110	3.6	35.8	170
Knife	33	M	106		36.3	230
Traffic accident	58	F	150		38.2	400

- **Objective:**
  Develop models to help emergency doctors make decisions.
  Measurements $\xrightarrow{\text{Predict}}$ Platelet $\Rightarrow$ $X \xrightarrow{\text{Regression}} y$
- **Challenge :**

  How to **select** relevant measurements with **missing values**?

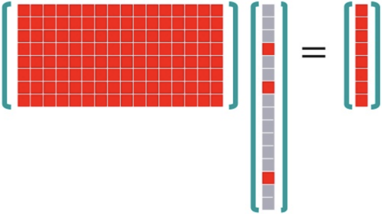
Model selection in high-dimension

Linear regression model: $y = X\beta + \varepsilon,$

- $y = (y_i)$: vector of response of length n
- $X = (X_{ij})$: a standardized design matrix of dimension $n \times p$
- $\beta = (\beta_j)$: regression coefficient of length p
- $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$

Assumptions:

- high-dimension: p large (including $p \geq n$)
- β is **sparse** with $k < n$ nonzero coefficients

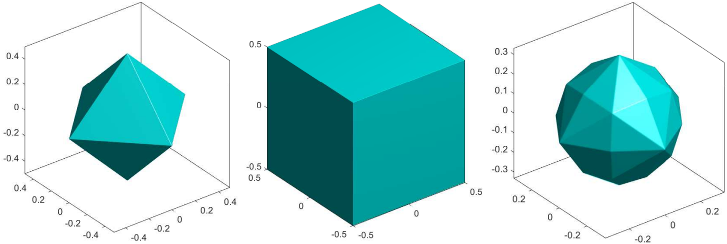


l_1 penalization methods

- LASSO (Tibshirani, 1996)
$$\hat{\beta}_{LASSO} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \frac{1}{2}\|y - X\beta\|^2 + \lambda\|\beta\|_1,$$
detects important variables with high probability but includes many **false positives.**

- SLOPE (B, van den Berg, Su, Candès, arxiv 2013, B,van den Berg, Sabatti, Su, Candès, AoAS , 2015) penalizes larger coefficients more stringently
$$\hat{\beta}_{SLOPE} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \frac{1}{2}\|y - X\beta\|^2 + \sigma \sum_{j=1}^{p} \lambda_j |\beta|_{(j)},$$
where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ and $|\beta|_{(1)} \geq |\beta|_{(2)} \geq \cdots \geq |\beta|_{(p)}$.

Unit balls for different SLOPE sequences by D.Brzyski



(a) (2,2,2) (b) (2,0,0) (c) (3,2,1)

Clustering in the context of portfolio optimization - Kremmer, Lee, B and Paterlini ”Journal of Banking and Finance”, 2019

The class of models attainable by SLOPE - Schneider and Tardivel, arxiv 2020

False discovery rate (FDR) control

- Let $\widetilde{\beta}$ be estimate of β
- We define:
  - the number of all discoveries, $R := \left|\left\{i : \ \widetilde{\beta}_i \neq 0\right\}\right|$
  - the number of false discoveries, $V := \left|\left\{i : \ \beta_i = 0, \quad \widetilde{\beta}_i \neq 0\right\}\right|$
  - false discovery rate - expected proportion of false discoveries among all discoveries

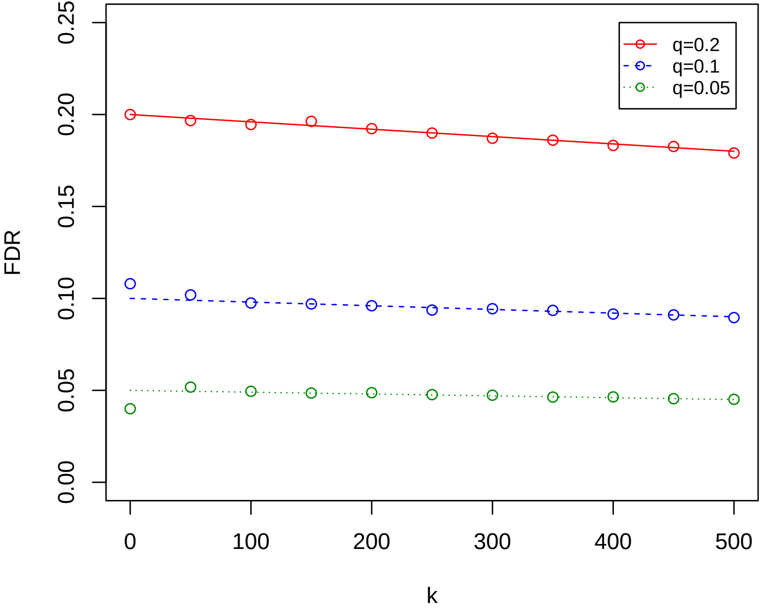
$$FDR := \mathbb{E}\left[\frac{V}{\max\{R, 1\}}\right]$$

Theorem (B,van den Berg, Su and Candès (2013))
When $X^T X = I$ SLOPE with

$$\lambda_i^{BH} := \sigma\Phi^{-1}\left(1 - i \cdot \frac{q}{2p}\right)$$

controls FDR at the level $q\frac{p_0}{p}$.

Orthogonal design, $n = p = 5000$



Optimality in prediction and estimation

Su and Candès (Annals of Statistics, 2016),

Bellec, Lecué, Tsybakov (Annals of Statistics, 2018):

SLOPE with the BH related sequence of tuning parameters adapts to the uknown sparsity and attains minimax prediction and estimation rates for the estimation error $||\hat{\beta} - \beta||^2$.

The selection of the optimal λ for LASSO depends on unknown sparsity k.

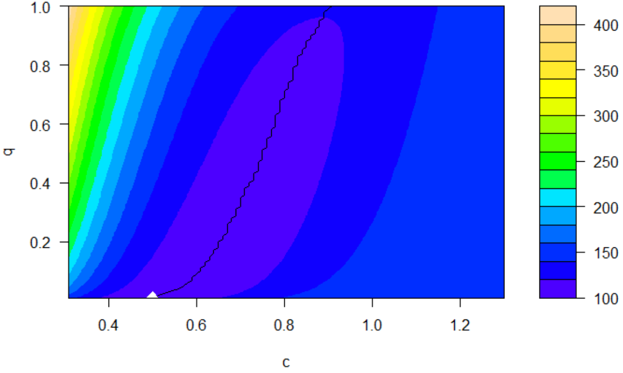
Extension to classification by logistic regression by Abramovich and Grinshtein (2018, IEEE Trans. Inf. Theory)

Heat Maps of $MSE(X\hat{\beta})$ by D. Nowakowski

Independent predictors

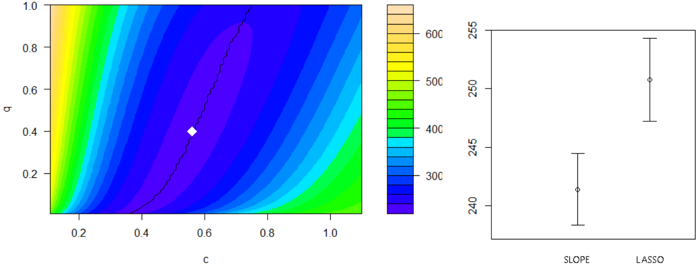
$$\lambda_i = c\Phi\left(1 - \frac{iq}{2p}\right), \quad n = p = 1000, k = 20$$

$$\text{for} \;\; i \in S, \quad \beta_i = \sqrt{2\log\frac{p}{k}}$$



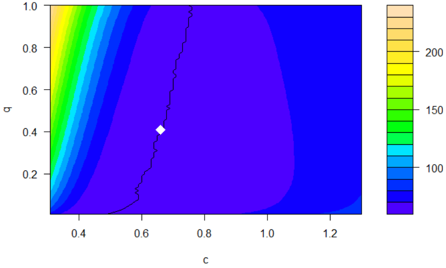
Independent predictors

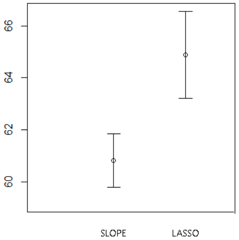
$$n = p = 1000, k = 100$$



Correlated predictors

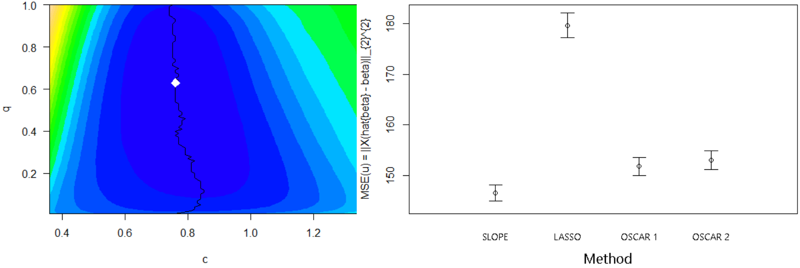
$$n = p = 1000, k = 20, \rho(X_i, X_j) = 0.5 \text{ for } i \neq j$$





Correlated predictors

$$n = p = 1000, k = 100$$



Extensions and Applications

- Lee, Brzyski, B. *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, JMLR:W and CP* **vol.51**, 780–789, 2016 - FDR control with Generalized Dantzig Selector.
- Brzyski, Peterson, Sobczyk, Candès, B., Sabatti, Ćontrolling the rate of GWAS false discoveries", *Genetics*, **205**, 61–75, 2017, *geneSLOPE* package in R by P. Sobczyk.
- Virouleau, Guilloux, Gaiffas, B., arXiv:1712.02640, 2017 - Robust regression and outliers detection using the mean-shift model with SLOPE.
- Kos, B., arXiv:1908.08791, 2019 - Consistency and asymptotic FDR control in high-dimensional multiple regression.
- Kos, PhD thesis, 2019 - Asymptotic FDR control in low dimensional multiple and logistic regression.
- Kremer, Brzyski, B., Paterlini, SSRN 3412061, 2019 - application for index tracking.
- Kremer, Lee, B., Paterlini, *Journal of Banking and Finance* 110, 105687, 2020 - application for portfolio selection.
- Brzyski, Gossmann, Su, B., *Journal of the American Statistical Association*, 114(525), 419–433, 2019 - group SLOPE for selection of groups of predictors, *grpSLOPE* package in R by A. Gossmann.
- Lee, Sobczyk, M.Bogdan, "Structure Learning of Gaussian Markov Random Fields with False Discovery Rate Control", *Symmetry* 11 (10), 1311, 2019 - application for gaussian graphical models using neighborhood selection strategy.
- P.Sobczyk (PhD Thesis), M.Makowski (MSc Thesis) - application for graphical models using the joint likelihood function, first prize for M. Makowski in the "National competition for the best master's thesis regarding machine learning or data analysis" in the category "Methods and Algorithms"
- Larrson,B.,Wallin, arxiv 2020 - strong rule for discarding predictors, speeding up the SLOPE algorithm; accepted for *NeurIPS, 2020*.

Packages on CRAN

- *SLOPE*, SLOPE for Generalized Linear Models, a novel strong screening rule for SLOPE, maintained by Johan Larsson (Lund University)
- *geneSLOPE* - SLOPE for Genome Wide Association Studies, selection of clusters of correlated SNPs, maintained by Piotr Sobczyk (OLX group)
- *grpSLOPE* - SLOPE for selection of groups of predictors, maintained by Alexey Gossman (FDA, USA)

Robust regression with SLOPE

A.Virouleau, A.Guilloux, S.Gaiffas, M.Bogdan (arxive, 2017)

Mean-shift model for robust regression

Candes and Randall (2006), Gannaz (2006) and McCann and Welsch (CSDA, 2007) ,

$$y = X\beta + I\mu + \varepsilon \tag{1}$$

$\mu \in R^n$ is the sparse vector of “outliers” and $\varepsilon \sim N(0, \sigma^2 I)$

She and Owen (IPOD, JASA, 2012) and Nguyen and Tran (E-lasso, IEEE Trans. Inf. Th., 2013) use L_1 penalty for μ and β

Virouleau, Guilloux, Gaiffas, B (2017) use SLOPE penalties:

$$\min_{\beta \in {}^p, \mu \in {}^n} \left\{ \|y - X\beta - \mu\|_2^2 + 2\rho_1 J_{\tilde{\lambda}}(\beta) + 2\rho_2 J_{\lambda}(\mu) \right\}$$

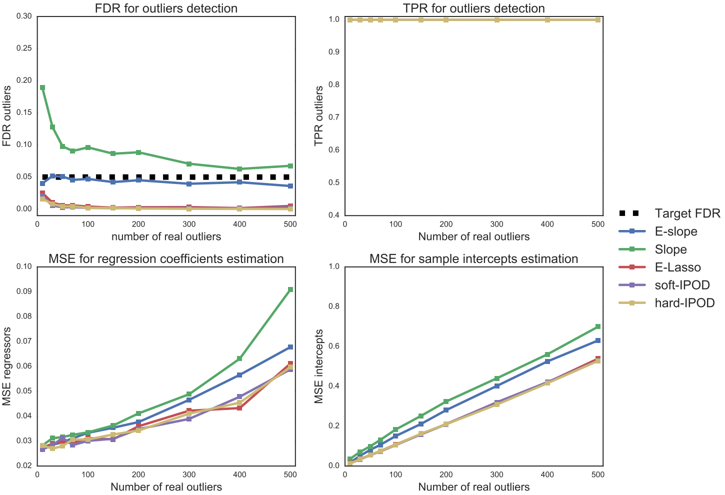
$$\lambda_i(\beta) = \sigma\sqrt{\log\left(\frac{2p}{i}\right)}, \quad \lambda_i(\mu) = \sigma\sqrt{\log\left(\frac{2n}{i}\right)}$$

Estimation properties and model selection properties

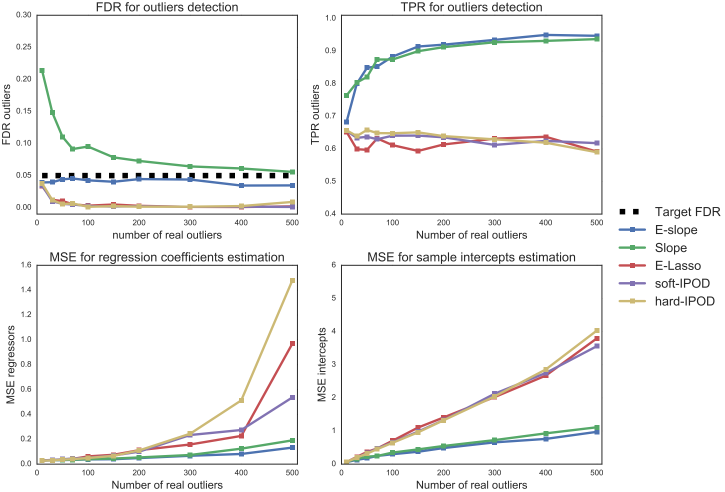
When $k \log\left(p/k\right) \leq s \log\left(n/s\right)$ then the mean-shift version of SLOPE retains asymptotic estimation and prediction optimality.

Under some sparsity assumptions the mean-shift SLOPE asymptotically controls the False Discovery Rate in terms of outliers detection

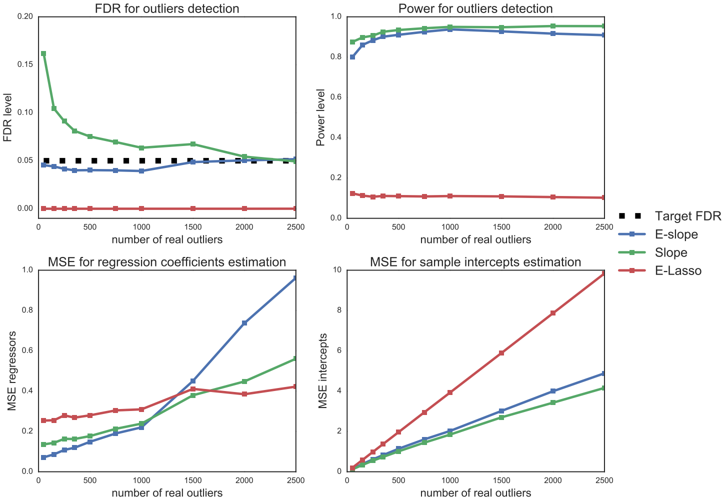
Low dimensional set-up; large outliers



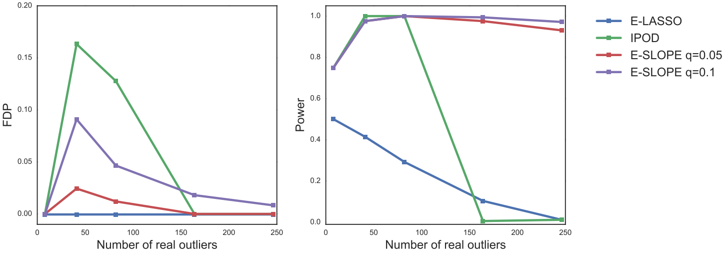
Low dimensional set-up; small outliers



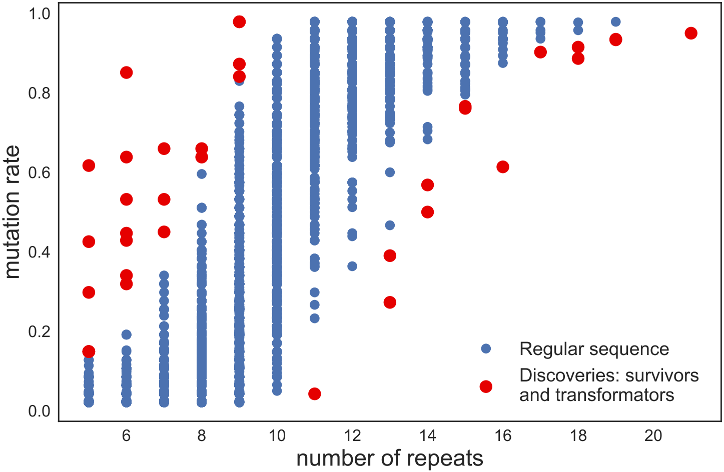
High dimensional set-up; small outliers



Simulated Outliers for the Retail Sales Data

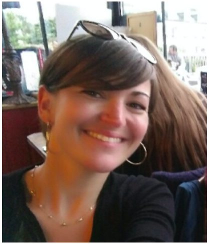


Mutation Rates in Colorectal Cancer



Adaptive SLOPE with missing values (1)

W. Jiang, MB, J.Josse, B.Miasojedow, V.Rockova, TraumaBase Group (2019)



traumabase.eu

Problems with LASSO and SLOPE

The same parameter λ is used for shrinkage and selection

Elimination of false discoveries leads to a large bias of important predictors

Unexplained effect of important predictors is taken over by non-important variables

Identification of the true model is possible only under very restrictive assumptions on the signal sparsity and the correlations between predictors

Solution - adaptive versions, use smaller λ for predictors which seem to be important [prior knowledge or iterations of the algorithm]

Spike and Slab LASSO

V.Rockova, E. George, JASA 2018

LASSO has a Bayesian interpretation as a posterior mode under the Laplace prior

$$\pi(\beta) = C(\lambda) \prod_{i=1}^{n} e^{-|\beta_i|\lambda}$$

Spike and Slab LASSO uses a spike and slab Laplace prior:

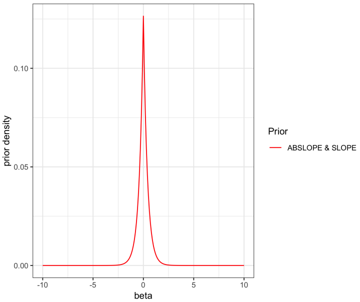
$$\gamma = (\gamma_1, \ldots, \gamma_p)$$

$\gamma_i = 1$ if β_i is "large" and $\gamma_i = 0$ if β_i is "small"

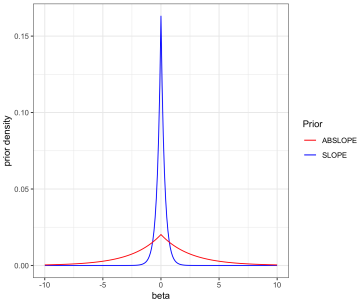
$$\pi(\beta|\lambda, \gamma) \propto c^{\sum_{i=1}^{p} 1(\gamma_i = 1)} \prod_{i=1}^{p} e^{-w_i|\beta_i|\lambda_0},$$

where $w_i = 1$ if $\gamma_i = 0$ and $w_i = c \in (0, 1)$ if $\gamma_i = 1$.

Spike and Slab Prior



(d) Null β



(e) Non-null β

Spike and Slab LASSO (2)

The maximum aposteriori rule is given by reweighted LASSO

$$\hat{\beta}(\gamma) = argmin_{b \in R^p} \frac{1}{2} ||y - Xb||_2^2 + \lambda_0 \sum_{i=1}^{p} w_i |b_i|$$

$$w_i = c\gamma_i + (1 - \gamma_i)$$

Prior for γ: $\gamma_1, \dots, \gamma_p$ are iid such that

$$P(\gamma_i = 1) = \theta = 1 - P(\gamma_i = 0)$$

In consecutive iterations γ_i is replaced with

$$\pi_i^t = P(\gamma_i = 1 | \beta^t, c) = \frac{c\theta e^{-c|\beta_i^t|\lambda_0}}{c\theta e^{-c|\beta_i^t|\lambda_0} + (1-\theta) e^{-|\beta_i^t|\lambda_0}}$$

and then a new estimate $\hat{\beta}^{t+1}$ is calculated by solving reweighted LASSO with the vector γ replaced with the vector π^t.

Borrowing information

When updating i^{th} variable θ is replaced by $E(\theta|\beta_{-i})$

$\lambda_1 = c\lambda_0$ - fixed at some small value

SSL package creates the path of SSL solutions for the sequence of 100 λ_0 values

Bayesian SLOPE

SLOPE estimate $=$ MAP of a Bayesian regression with SLOPE prior.

$$\hat{\beta}_{SLOPE} = \arg\max_{\beta} \mathtt{p}(y \mid X, \beta, \sigma^2; \lambda) \propto \mathtt{p}(y \mid X, \beta)\mathtt{p}(\beta \mid \sigma^2; \lambda)$$

where the SLOPE prior:

$$\mathtt{p}(\beta \mid \sigma^2; \lambda) \propto \prod_{j=1}^{p} \exp\left(-\frac{1}{\sigma}\lambda_j |\beta|_{(j)}\right)$$

Adaptive Bayesian SLOPE

We propose an adaptive version of Bayesian SLOPE (ABSLOPE). After standardizing X so each column has a unit L_2 norm, the prior for β is

$$\mathrm{p}(\beta \mid \gamma, c, \sigma^2; \lambda) \propto c^{\sum_{j=1}^{p} \mathbb{I}(\gamma_j = 1)} \prod_j \exp\left\{ -w_j |\beta_j| \frac{1}{\sigma} \lambda_{r(W\beta, j)} \right\},$$

Interpretation of the model:

- β_j from the slab component ⇒ **true signal**; from the spike component ⇒ noise.
- $\gamma_j \in \{0, 1\}$ signal indicator. $\gamma_j | \theta \sim \mathit{Bernoulli}(\theta)$ and θ the **sparsity**.
- $1/c \in [1, \infty)$: proportional to the **average signal magnitude**.
- $W = \mathrm{diag}(w_1, w_2, \cdots, w_p)$ and its diagonal element:

$$w_j = c\gamma_j + (1 - \gamma_j) = \begin{cases} c, & \gamma_j = 1 \\ 1, & \gamma_j = 0 \end{cases}.$$

Spike and Slab LASSO (Rockova and George, 2018), vs ABSLOPE

ABSLOPE spike prior models the effects which are not distinguishable from the noise, which allows for FDR control

Slab component is ”estimated” via the estimation of the average signal magnitude

Model selection with missing values

Decomposition: $X = (X_{\text{obs}}, X_{\text{mis}})$

Pattern: matrix M with $M_{ij} = \begin{cases} 1, & \text{if } X_{ij} \text{ is observed} \\ 0, & \text{otherwise} \end{cases}$

Assumption 1: Missing at random (MAR)

$\mathrm{p}(M \mid X_{\text{obs}}, X_{\text{mis}}) = \mathrm{p}(M \mid X_{\text{obs}}) \quad \Rightarrow \quad$ ignorable missing patterns
e.g. People at older age didn't tell his income at larger probability.

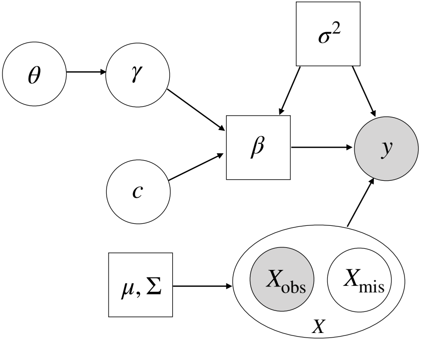
Assumption 2: Distribution of covariates

$X_i \sim_{\text{i.i.d.}} \mathcal{N}_p(\mu, \Sigma), \quad i = 1, \cdots, n.$

Problem: With NA, only a few methods are available to select a model, and their performances are limited. For example,

- (Claeskens and Consentino, 2008) adapts AIC to missing values ⇒ Impossible to deal with high dimensional analysis.
- (Loh and Wainwright, 2012) LASSO with NA
  ⇒ Non-convex optimization; requires to know bound of $\|\beta\|_1$
  ⇒ difficult in practice

ABSLOPE with missingness: Summary



$$\begin{aligned}\ell_{\text{comp}} &= \log \mathrm{p}(y, X, \gamma, c, \beta, \theta, \sigma^2) \\ &= \log \left\{ \mathrm{p}(y \mid X, \beta, \sigma^2)\, \mathrm{p}(X \mid \mu, \Sigma)\, \mathrm{p}(\beta \mid \gamma, c)\, \mathrm{p}(\gamma \mid \theta)\, \mathrm{p}(c, \sigma, \theta) \right\}\end{aligned}$$

Objective: Maximize $\ell_{\text{obs}} = \iiint \ell_{\text{comp}}\, dX_{\text{mis}}\, dc\, d\theta\, d\gamma.$

EM algorithm

- *E step:* evaluate

$$Q^t = \mathbb{E}(\ell_{\text{comp}}) \quad \text{wrt} \quad \mathrm{p}(X_{\text{mis}}, \gamma, c, \theta \mid y, X_{\text{obs}}, \beta^t, \sigma^t, \mu^t, \Sigma^t).$$

- *M step:* update

$$\beta^t, \sigma^t, \mu^t, \Sigma^t = \arg\max Q^t$$

Problem: The function Q is not tractable. $\Rightarrow$

1. Monte Carlo EM ? (Wei and Tanner 1990) ~~Monte Carlo EM ?~~

   Expensive to generate a large number of samples.
2. Stochastic Approximation EM (book, Lavielle 2014)
   - One sample in each iteration;

Adapted SAEM algorithm

- *E step:*
  $Q^t = \mathbb{E}(\ell_{\text{comp}}) \quad \text{wrt} \quad \mathrm{p}(X_{\text{mis}}, \gamma, c, \theta \mid y, X_{\text{obs}}, \beta^t, \sigma^t, \mu^t, \Sigma^t).$
  - *Simulation:* draw one sample $(X_{\text{mis}}^t, \gamma^t, c^t, \theta^t)$ from

    $$\mathrm{p}(X_{\text{mis}}, \gamma, c, \theta \mid y, X_{\text{obs}}, \beta^{t-1}, \sigma^{t-1}, \mu^{t-1}, \Sigma^{t-1});$$

    [Gibbs sampling]
  - *Stochastic approximation:* update function Q with

    $$Q^t = Q^{t-1} + \xi_t \left(\ell_{\text{comp}}(X_{\text{mis}}^t, \gamma^t, c^t, \theta^t) - Q^{t-1}\right), \quad \text{where} \quad \xi_t \in (0,1].$$

- *M step:* $\beta^{t+1}, \sigma^{t+1}, \mu^{t+1}, \Sigma^{t+1} = \arg\max Q^{t+1}$.
  [When $\eta_t = 1$: Reweighted SLOPE, Shrinkage of covariance]

Details of initialization, generating samples and optimization are in the draft (arXiv:1909.06631)

SLOBE

Instead of using Gibbs sampling γ and *c* are replaced with the approximation to their conditional expectations given data, β and σ

R package: ABSLOPE

Install package:

```
library(devtools)
install_github("wjiang94/ABSLOPE")
```

Main algorithm:

```
lambda = create_lambda_bhq(ncol(X),fdr=0.10)
list.res = ABSLOPE(X, y, lambda)
```

A fast and simplified algorithm (Rcpp):

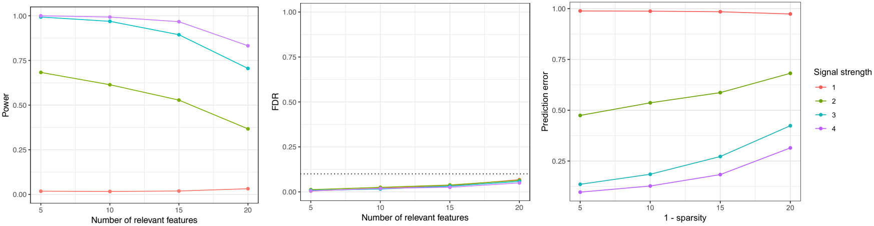
```
list.res.slobe = SLOBE(X, y, lambda)
```

Values:

```
list.res$beta
list.res$gamma
```

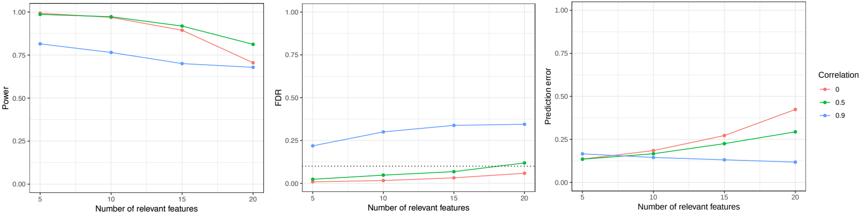
Simulation study (200 rep. $\Rightarrow$ average)

$n = p = 100$, no correlation and 10% missingness



(f) Power (g) FDR (h) Prediction error

$n = p = 100$, with 10% missingness and strong signal



(i) Power (j) FDR (k) Prediction error

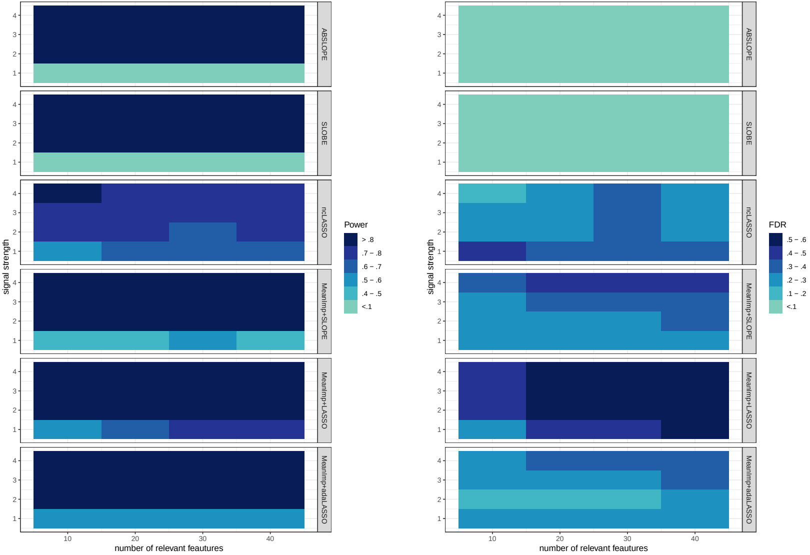
Method comparison

- **ABSLOPE** and **SLOBE**
- **ncLASSO:** non convex LASSO (Loh and Wainwright, 2012)
- **MeanImp + SLOPE:** Mean imputation followed by SLOPE with known σ
- **MeanImp + LASSO:** Mean imputation followed by LASSO, with λ tuned by cross validation
- **MeanImp + adaLASSO:** Mean imputation followed by adaptive LASSO (Zou, 2006)

In the SLOPE type methods, λ = BH sequence which controls the FDR at level **0.1**

Method comparison (200 rep. ⇒ average)

500×500 dataset, 10% missingness, with correlation



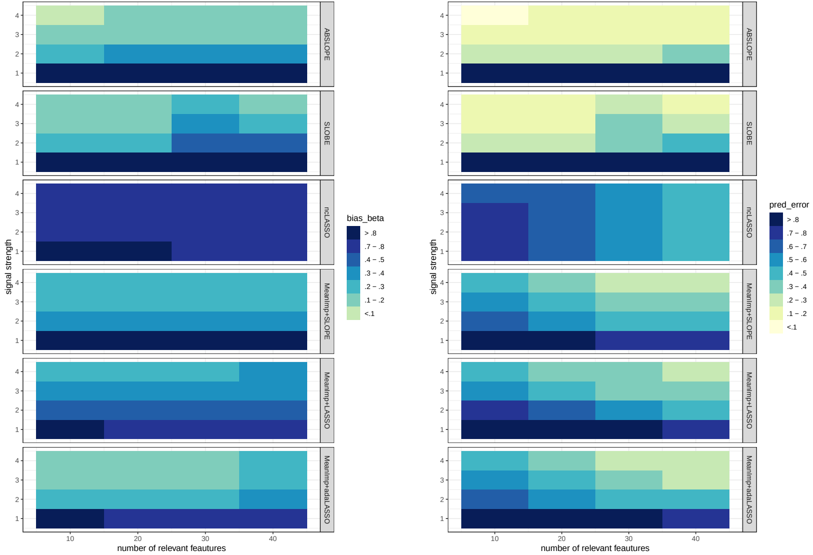
(l) Power

(m) FDR

Rysunek: Comparison of power (a), FDR (b), bias of β (c) and

Method comparison (200 rep. ⇒ average)

500×500 dataset, 10% missingness, with correlation



(a) Bias of β

(b) Prediction error

Rysunek: Comparison of power (a), FDR (b), bias of β (c) and

Variables in the TraumaBase data set (APHP)

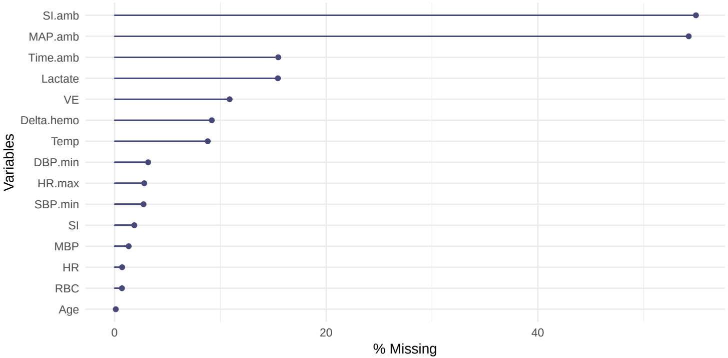
Goal - quick prediction of the level of platelets

- *Age:* Age
- *SI:* Shock index indicates level of occult shock based on heart rate (FC) and systolic blood pressure (PAS). $SI = \frac{FC}{PAS}$. Evaluated on arrival of hospital.
- *PAM:* Mean arterial pressure is an average blood pressure in an individual during a single cardiac cycle, based on systolic blood pressure (PAS) and diastolic blood pressure (PAD). $PAM = \frac{2PAD+PAS}{3}$. Evaluated on arrival of hospital.
- *delta_Hemocue:* The difference between the hemoglobin on arrival at hospital and that in the ambulance.
- *Temps.lieux.hop:* Time spent in hospital *i.e.*, medicalization time, in minutes.
- *Lactates:* The conjugate base of lactic acid.
- *Temperature:* Patient's body temperature.

Variables

- *FC:* heart rate measured on arrival of hospital.
- *Remplissage:* A volume expander is a type of intravenous therapy that has the function of providing volume for the circulatory system.
- *CGR.dechoc:* A binary index which indicates whether the transfusion of Red Blood Cells Concentrates is performed.
- *SI.SMUR:* Shock index measured on ambulance.
- *PAM.SMUR:* Mean arterial pressure measured in the ambulance.
- *FC.max:* Maximum value of measured heart rate in the ambulance.
- *PAS.min:* Minimum value of measured systolic blood pressure in the ambulance.
- *PAD.min:* Minimum value of measured diastolic blood pressure in the ambulance.

Percentage of missing values



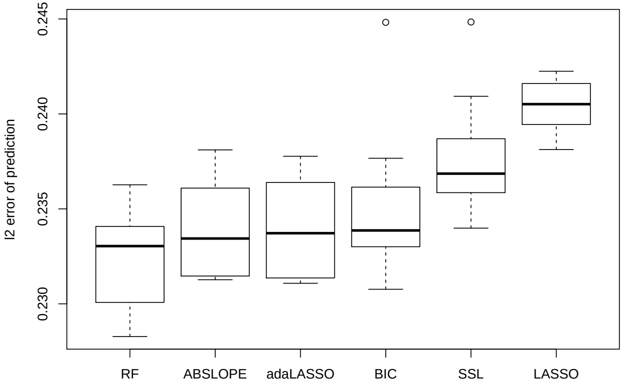
Rysunek: Percentage of missing values in each pre-selected variable from TraumaBase.

Rysunek: Number of times that each variable is selected over 10 replications. Bold numbers indicate which variables are included in the model selected by ABSLOPE.

Variable	ABSLOPE	SLOPE	LASSO	adaLASSO	BIC
Age	**10**	10	4	10	10
SI	**10**	2	0	0	9
MBP	1	10	1	10	1
Delta.hemo	**10**	10	8	10	10
Time.amb	2	6	0	4	0
Lactate	**10**	10	10	10	10
Temp	2	10	0	0	0
HR	**10**	10	1	10	10
VE	**10**	10	2	10	10
RBC	**10**	10	10	10	10
SI.amb	0	0	0	0	0
MBP.amb	0	0	0	0	0
HR.max	3	9	0	1	0
SBP.min	5	10	10	10	8

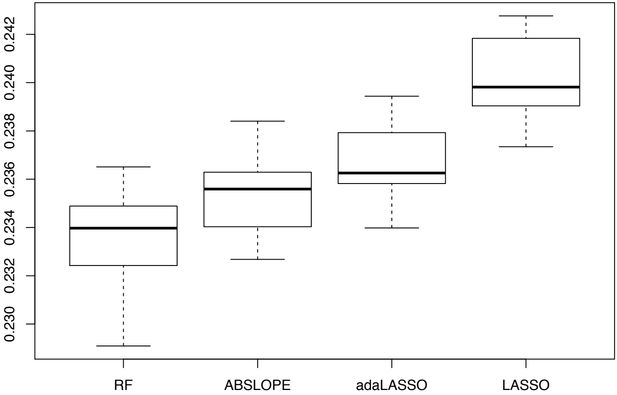
More on the real data...

TraumaBase: Measurements $\xrightarrow{\text{Predict}}$ Platelet
Cross-validation: random splits to training and test sets × 10



- Comparable to random forest
- Interpretable model selection and estimation results

With interactions



Method	Variables selected
ABSLOPE	Age * MBP.amb, Delta.hemo * Lactate
	Lactate * RBC, HR * SBP.min
LASSO	RBC, SBP.min
	Age * Lactate, Age * VE
	Delta.hemo * Lactate, Delta.hemo * VE
	Lactate * VE, Lactate * RBC
adaLASSO	Age * Time.amb, Age * HR
	Age * MBP.amb, Age * SBP.min
	MBP * HR, Delta.hemo * VE
	Lactate * VE,HR * HR.max
	HR * SBP.min, VE * RBC

Conclusion & Future research

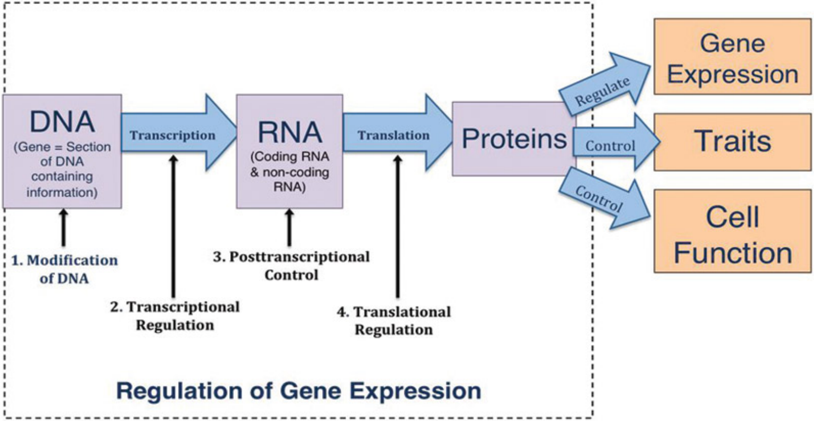
Conclusion:

- ABSLOPE reduces the estimation bias of large regression coefficients.
- This allows for
  1. Improved estimation and prediction properties
  2. FDR control under much wider range of scenarios than for regular SLOPE
- Modeling in a Bayesian framework allows for the estimation of the structure of predictors such as the **signal sparsity** and the **signal strength**;

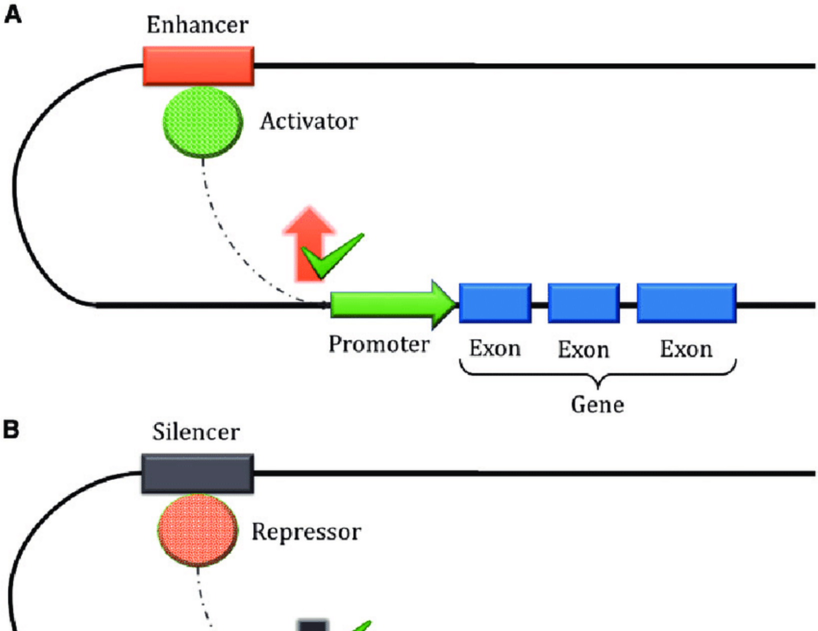
Future research:

- Deal with other missing mechanisms
- Application for other statistical models (e.g. GLM or Gaussian Graphical Models)
- Theoretical analysis of statistical properties (asymptotic FDR control, minimaxity)
- **Speeding the SLOPE algorithm, see e.g. Larsson, B., Wallin, ”The Strong Screening Rule for SLOPE”, arXiv:2005.03730,**

Varclust Motivation - Gene Expression


DNA
(Gene = Section of DNA containing information)
Transcription
RNA
(Coding RNA & non-coding RNA)
Translation
Proteins
Regulate
Control
Control
Gene Expression
Traits
Cell Function
1. Modification of DNA
2. Transcriptional Regulation
3. Posttranscriptional Control
4. Translational Regulation
Regulation of Gene Expression

Transcription factors



PCA - reduction of dimensionality of ”omics” data

$X_{n \times p}$ - data matrix (e.g. gene expressions), $n = k \times 100$, $p \approx 20000$ - number of genes

Assumptions : $X = M + E$, where M is of a low rank and E is a random noise

We usually assume that $e_{ij} \sim N(0, \sigma)$

Mathematical goal - recovering M, separation of the signal from noise

Practical goal - data compression, several basis vectors [Principal Components] may contain most of the information and be applied for prediction (of the patient’s response to the therapy)

Principal Components Analysis (2)

Method - Singular Value Decomposition:

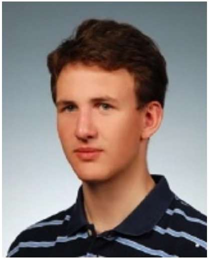
$$X = U_{n \times l} D_{l \times l} V_{l \times p}^{T} \ \ ,$$

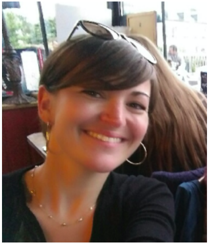
$$U^T U = I_{l \times l}, V^T V = I_{l \times l}, \ \ l = min\{n, p\}$$

Statistical Goal - determining rank *k* of matrix *M*

PESEL (PEnalized SEmi-integrated Likelihood)

Sobczyk, Bogdan, Josse, Journal of Computational Graphical Statistics, 2017





Bayesian Information Criterion (BIC) (1)

$A_1 \in A_2 \in A_3 \dots$ - nested sequence of statistical models

In our example A_k - $rank(M) \leq k$

θ - vector of parameters of A_k:

eleements of $U_k \in S_{k,n}$, $V_k \in S_{k,p}, D_k$, i σ

$S_{k,n}$ - Stiefel manifold of orthonormal matrices of dimension $n \times k$

$l(X, \theta)$ - likelihood function (density of the distribution describing the data)

Bayesian Information Criterion (BIC) (2)

In general situation BIC suggests selecting the model maximizing

$$max_{\theta \in A_k} \log l(X, \theta) - 1/2 dim(A_k) \log N$$

where N is the number of independent observations.

BIC is justified (consistent) where $dim(A_k) = const$ when $N \to \infty$

In our case $N = np$, so $dim(A_k)$ increases with n and p

Idea - reduction of the number of parameters by integrating them out with respect to some prior distribution

PESEL for large p

Assume that $M = TW^T$, where
$T = [t_{i,l}]_{n \times k}$ is the matrix of ”hidden factors”,
$W = [w_{i,l}]_{p \times k}$ is the matrix of coefficients

prior distribution -

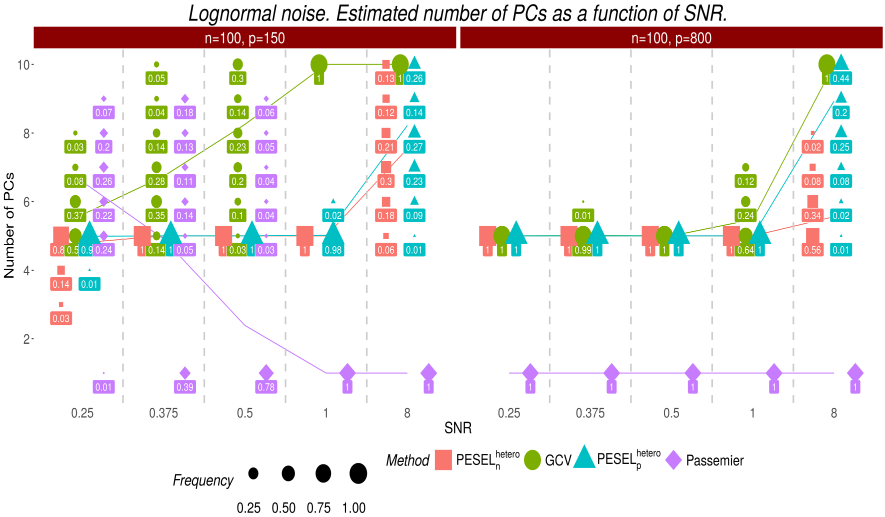
$$w_{j\cdot} \sim N(0, I_k) \quad ,$$

which implies, that $x_{\cdot 1}, \ldots, x_{\cdot p}$ są are iid random vectors from the distribution

$$x_{\cdot j} \sim N(0; TT^T + \sigma^2 I_n) \quad .$$

Now we have p independent vectors and the number of parameters does not depend on p - we can apply BIC if only $p >> n$

Errors from the log-normal distribution



Varclust

Multiple Latent Component Clustering, Sobczyk, Wilczyński, Bogdan, Josse

Awards for young scientists P. Sobczyk (Vienna workshop on simulation, 2015) i S. Wilczyński (International Conference on Biometrics and Bio-Pharmaceutical Statistics, Wiedeń 2017)

Goal: Identification of groups of co-regulated variables (genetic pathways) and selection of appropriate Principal Components.

Mathematics: clustering of variables into groups, such that each of them is spanned by just few of ”hidden” variables.

Package *varclust* by P. Sobczyk and S. Wilczyński- Algorithm K-medioids around PCs. Estimation of the number of clusters and their dimensions by modifications of BIC.

Methodology

K-centroids algorithm

Centers - PCs, distance - BIC

Estimation of clusters dimensions by PESEL

Repeat for different K and estimate K by mBIC

Informative prior distribution and mBIC

- The problem with BIC (non-informative prior)
- Prior distribution taking into account the number of clusters and maximal dimension of the subspace

$$P(M) = \frac{1}{K^p}\frac{1}{d^K}$$

$$mBIC = \sum_{i=1}^{K} \ln\left(\widehat{P}(X_i|M_i)\right) - p\ln(K) - K\ln(d)$$

Application

mBIC can be used to compare different models and to choose the number of clusters in the data.

Overview of the algorithm

Algorithm 1: Multiple Latent Clustering Components

```
Input: $n$ - number of individuals, $p$ - number of variables,
  $X_{n \times p} = (x_1, \ldots, x_p)$ - data set, $d$ - maximal subspace
  dimension, $N$ - number of runs of the algorithm
  Scale $X$ to have columns with mean 0 and unit variance
  for $i \in \{1, \ldots, N\}$ do
     Find the model using K-means and store its value of mBIC
  end for
  Choose the model with the highest value of mBIC and return
  the model (segmentation, mBIC, factors) as the result. =0
```

K-means step

1. Initialize clusters' centres
2. Until convergence or maximal number of iterations is reached repeat:
    - For every variable x_j and every cluster factors $F_{j'}$ fit a linear regression model without intercept $lm(x_j \sim F_{j'})$ and store BIC value as $BIC_{jj'}$
    - Assign variable x_j to the cluster M_q where

      $$q = \underset{j' \in \{1,\dots,K\}}{\arg\max} BIC_{jj'}$$

    - For every cluster M_i use PESEL to estimate its dimensionality k_i with an upper bound of d. Use PCA to compute the first k_i principal components and store them in F_i

Compared methods

1. **Sparse Subspace Clustering** (SSC)
2. Low Rank Subspace Clustering (LRSC)
3. **MLCC with random initialization** (MLCC)
4. **MLCC with initialization by the result of SSC** (MLCC$_{aSSC}$)
5. MLCC with initialization by sparse PCA (MLCC$_{sPCA}$)
6. ClustOfVar (COV)

Data generation - shared factors

```
Input: $n$, $SNR$, $K$, $p$, $d$
  Number of factors $m \leftarrow K\frac{d}{2}$
  Factors $F = (f_1, \dots, f_m)$, $f_i \sim N(0, I_n)$
  Draw subspaces' dimension $d_1, \dots d_K$ uniformly from
  $\{1, \dots, d\}$
  for $i = 1, \dots, K$ do
     $F_i \leftarrow$ sample of size $d_i$ from columns of $F$
     Draw matrix of coefficients $C_i$ from
     $U(0.1, 1) \cdot sgn(U(-1, 1))$
     Variables in the $i$-th subspace are $X_i \leftarrow F_i C_i$
  end for
  Scale matrix $X = (X_1, \dots, X_K)$ (columns with unit variance)
  return $X + Z$ where $Z \sim N(0, \frac{1}{SNR} I_n)$ =0
```

Data generation - independent subspaces

Remark

To generate data without shared factors we draw independently i-th subspaces basis F_i as sample of size d_i from standard multivariate normal distribution

Measures of effectiveness

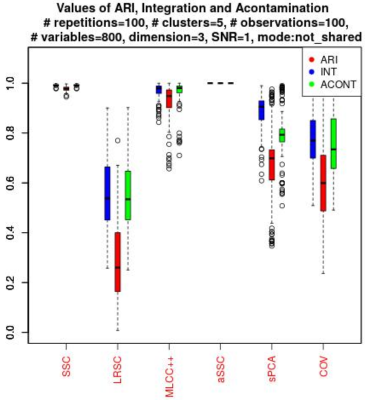
Compare two partitions $A = (A_1, \dots A_n),\ B = (B_1, \dots, B_m)$

- Adjusted Rand Index (ARI)
- Integration
- Acontamination
- ARI $\in [-1, 1]$, Integration, Acontamination $\in [0, 1]$.

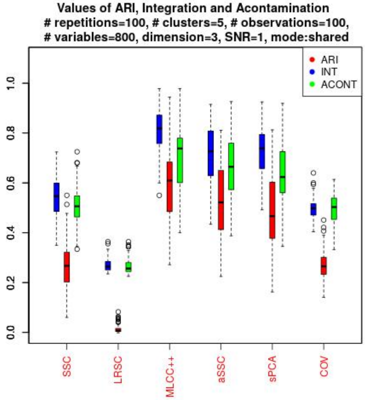
Remark

The bigger the indices, the better the clustering.

Mode

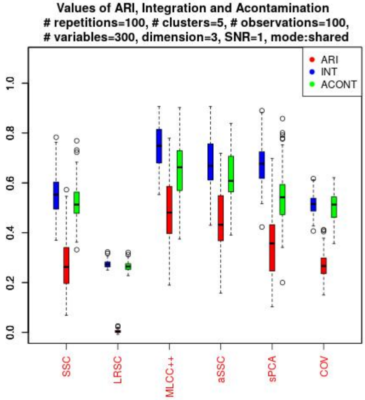


Values of ARI, Integration and Acontamination
# repetitions=100, # clusters=5, # observations=100,
# variables=800, dimension=3, SNR=1, mode:not_shared

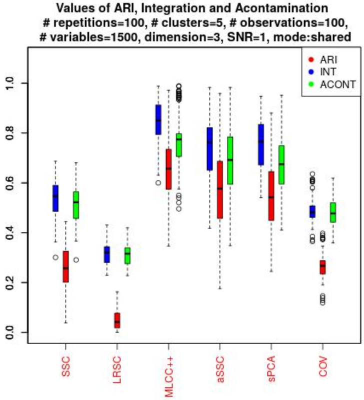


Values of ARI, Integration and Acontamination
# repetitions=100, # clusters=5, # observations=100,
# variables=800, dimension=3, SNR=1, mode:shared

Number of variables

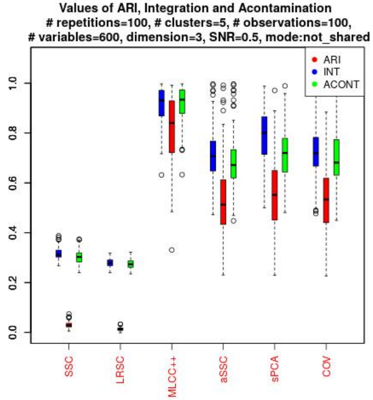


Values of ARI, Integration and Acontamination
# repetitions=100, # clusters=5, # observations=100,
# variables=300, dimension=3, SNR=1, mode:shared

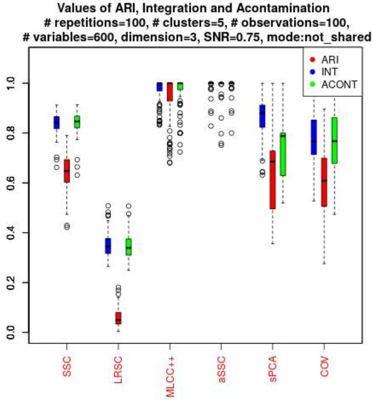


Values of ARI, Integration and Acontamination
# repetitions=100, # clusters=5, # observations=100,
# variables=1500, dimension=3, SNR=1, mode:shared

Signal to noise ratio

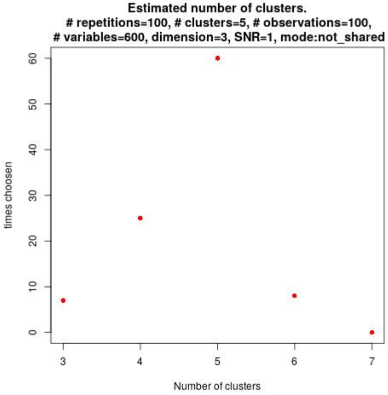


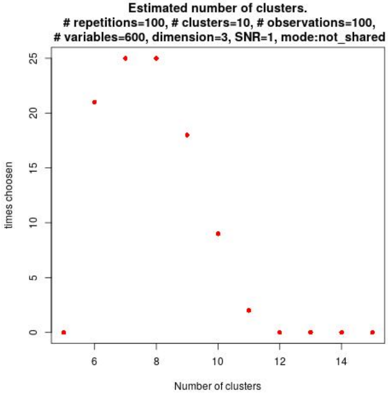
Values of ARI, Integration and Acontamination
# repetitions=100, # clusters=5, # observations=100,
# variables=600, dimension=3, SNR=0.5, mode:not_shared



Values of ARI, Integration and Acontamination
# repetitions=100, # clusters=5, # observations=100,
# variables=600, dimension=3, SNR=0.75, mode:not_shared

Estimation of the number of clusters


Estimated number of clusters.
# repetitions=100, # clusters=5, # observations=100,
# variables=600, dimension=3, SNR=1, mode:not_shared


Estimated number of clusters.
# repetitions=100, # clusters=10, # observations=100,
# variables=600, dimension=3, SNR=1, mode:not_shared

References (1)

- Bellec, Lecue, Tsybakov, ”Slope meets Lasso: Improved oracle bounds and optimality”, *Annals of Statistics*, **46** (6B), 3603-3642, 2018
- Bogdan, van den Berg, Sabatti, Su, Candès, ”SLOPE – Adaptive Variable Selection via Convex Optimization”, *Annals of Applied Statistics*, **9** (3), 1103–1140, 2015
- M. Bogdan, E. van den Berg, W. Su, E.J. Candès, Śtatistical estimation and testing via the ordered ℓ_1 norm”, arXiv:1310.1969, 2013.
- Brzyski, Gossmann, Su, Bogdan, ”Group SLOPE - adaptive selection of groups of predictors”, *Journal of the American Statistical Association*, 114(525), 419–433, 2019.
- Brzyski, Peterson, Sobczyk, Candès, Bogdan, Sabatti, Ćontrolling the rate of GWAS false discoveries”, *Genetics*, **205**, 61–75, 2017.
- Claeskens, Consentino, ”Variable Selection with Incomplete Covariate Data”, *Biometrics* **64**, 1062–1069,2008.
- Candes, Wakin, Boyd, ” Enhancing sparsity by reweighted l_1 minimization”, *J Fourier Anal Appl* **14**, 877–905, 2008.
- Donoho, Tanner,” Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing”, *Philosophical Trans. R. Soc. A*, **367** (1906), 4273-4293, 2009.
- Jiang, Bogdan, Josse, Miasojedow, Rockova, TB Group, ”Adaptive Bayesian SLOPE–High-dimensional Model Selection with Missing Values”, arXiv:1909.06631, 2019.
- Kos, M. ”Identification of Statistically Important Predictors in High-Dimensional Data. Theoretical Properties and Practical Applications.”, 2019, PhD thesis, University of Wroclaw, available upon request.
- Kos, Bogdan, ”On the asymptotic properties of SLOPE”, arXiv:1908.08791, 2019.
- Kremer, Brzyski, Bogdan, Paterlini, ”Sparse Index Clones via the Sorted L1-Norm”, SSRN 3412061, 2019.
- Kremer, Lee, Bogdan, Paterlini, ”Sparse portfolio selection via the sorted L1-Norm”, *Journal of Banking and Finance* 110, 105687, 2020.
- Larsson, Bogdan, Wallin, ”The Strong Screening Rule for SLOPE”, arXiv:2005.03730, 2020.
- Lavielle, ”Mixed Effects Models for the Population Approach: Models, Tasks, Methods and Tools”, Chapman and Hall/CRC, 2014.

References (2)

- Lee, Brzyski, B., ”Fast Saddle-Point Algorithm for Generalized Dantzig Selector and FDR Control with the Ordered l_1-Norm”, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, JMLR:W and CP* **vol.51**, 780–789, 2016.
- Lee, Sobczyk, Bogdan, ”Structure Learning of Gaussian Markov Random Fields with False Discovery Rate Control”, *Symmetry* 11 (10), 1311, 2019.
- Loh, Wainwright, ”High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity”, *Annals of Statistics* **40** (3), 1637–1664, 2012.
- Makowski, M. ”Precision matrix estimation in Gaussian graphical models”, 2019, Master Thesis, University of Wroclaw, available upon request.
- Nowakowski, D. ”Prediction properties of regularization methods in multiple regression”, 2019, Bachelor Thesis, University of Wroclaw, in Polish, available upon request.
- Rockova, George, ”The Spike-and-Slab LASSO”, *Journal of the Americal Statistical Association*, **113**, 431-444, 2018.
- Schneider, Tardivel, ”The Geometry of Uniqueness and Model Selection of Penalized Estimators including SLOPE, LASSO, and Basis Pursuit”, arXiv:2004.09106, 2020.
- Sobczyk, P. ”Identifying low-dimensional structures through model selection in high-dimensional data”, 2019, PhD thesis, Wroclaw University of Science and Technology, available upon request.
- Su, Bogdan, Candès, ”False Discoveries Occur Early on the Lasso Path”, *Annals of Statistics*, **45** (5), 2133 – 2150, 2017.
- Su, Candes, ”SLOPE is adaptive to unknown sparsity and asymptotically minimax”, *Annals of Statistics*, **44** (3), 1038-1068, 2016.
- Tardivel, Bogdan, Ón the sign recovery by LASSO, thresholded LASSO and thresholded Basis Pursuit Denoising”,arXiv:1812.05723, 2018.
- Tibshirani, ”Regression shrinkage and selection via the lasso”, *Journal of the Royal Statistical Society*, Series B, 267–288, 1996.
- Wei, Tanner, ”A Monte Carlo Implementation of the EM Algorithm and the Poor Man’s Data Augmentation Algorithms”, *Journal of the American Statistical Association* **85** (411) 699-704, 1990.
- Weinstein, Su, Bogdan, Barber, Candès, ”A Power Analysis for Knockoffs with the Lasso Coefficient-Difference Statistic”, arXiv:2007.15346, 2020.
- Virouleau, Guilloux, Gaiffas, Bogdan ”High-dimensional robust regression and outliers detection with SLOPE”, arXiv:1712.02640, 2017.
- Zou, The Adaptive Lasso and Its Oracle Properties, *Journal of the American Statistical Association*, **101:476**, 1418-1429, 2006.