Université Paris-Saclay | École doctorale de mathématiques Hadamard (EDMH)

Statistical inference with incomplete and high-dimensional data—modeling polytraumatized patients

Wei Jiang

CMAP, École Polytechnique & XPOP, INRIA Saclay

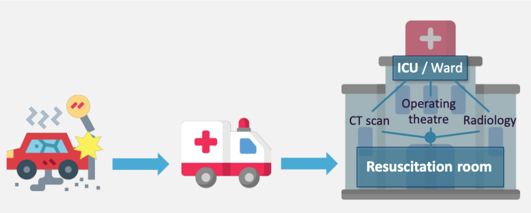
Advisors: Julie Josse, Marc Lavielle

21st Sep. 2020

TraumaBase project: decision support for patients

- 20000 trauma patients + 250 measurements variables

Center	Accident	Age	Sex	Lactactes	BP	**Shock**	**Platelet**	...
Beaujon	fall	54	m	NA	180	yes	292000	
Pitie	gun	26	m	NA	131	no	323000	
Beaujon	moto	63	m	3.9	NA	yes	318000	
Pitie	moto	30	f	NA	107	no	211000	
HEGP	knife	16	m	2.5	118	no	184000	
⋮								⋱



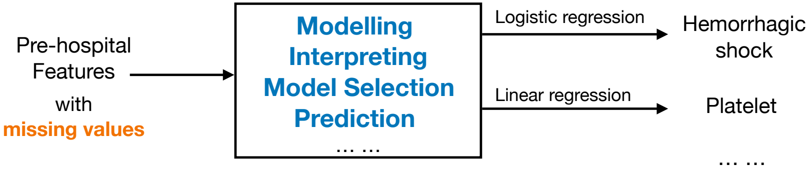
Management scheme of a traumatized patient.

TraumaBase project: decision support for patients

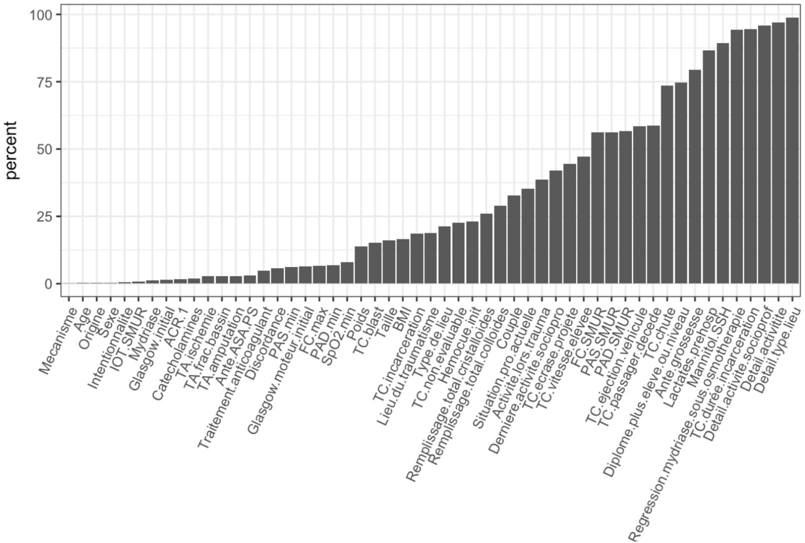
- 20000 trauma patients + 250 measurements variables

Center	Accident	Age	Sex	Lactactes	BP	**Shock**	**Platelet**	. . .
Beaujon	fall	54	m	NA	180	yes	292000	
Pitie	gun	26	m	NA	131	no	323000	
Beaujon	moto	63	m	3.9	NA	yes	318000	
Pitie	moto	30	f	NA	107	no	211000	
HEGP	knife	16	m	2.5	118	no	184000	
⋮								⋱

Objective: help the clinicians make decisions



TraumaBase: percentage of missing values



List-wise deletion ?

"One of the ironies of Big Data is that missing data play an ever more significant role." (Samworth, 2019)

Example

A $n \times p$ dataset, each entry has a probability 1% to be missing independently.

List-wise deletion ?

"One of the ironies of Big Data is that missing data play an ever more significant role." (Samworth, 2019)

Example

A $n \times p$ dataset, each entry has a probability 1% to be missing independently.

- $p = 5 \xRightarrow[\text{deletion}]{\text{List-wise}} 95\%$ rows kept
- $p = 300 \xRightarrow[\text{deletion}]{\text{List-wise}} 5\%$ rows kept

$\Rightarrow$ **List-wise deletion impossible**

Literature on missing values

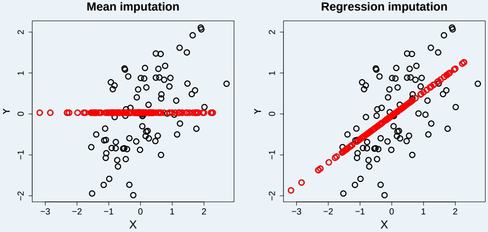
- R-miss-tastic: resource website for managing missing data, 150 packages (most based on imputation)
- Books: Schafer (2002), Little & Rubin (2019); Kim & Shao (2013); Carpenter & Kenward (2013); Stef van Buuren (2018)

Literature on missing values

- R-miss-tastic: resource website for managing missing data, 150 packages (most based on imputation)
- Books: Schafer (2002), Little & Rubin (2019); Kim & Shao (2013); Carpenter & Kenward (2013); Stef van Buuren (2018)

Single imputation

Example: $(x_i, y_i) \sim \mathcal{N}(\mu, \Sigma)$ *i.i.d.*, 70% missing entries on y randomly
Aim: Estimate parameters & their variance



	Mean imputation	Regression imputation
$\mu_y = 0$	0.01	0.01
$\sigma_y = 1$	0.5	0.72
$\rho = 0.6$	0.30	0.78

$\Rightarrow$ have bias & fail to evaluate the uncertainty caused by NA

Recommended method 1: multiple imputation

Example:

X_1	X_2	X_3	Y
NA	20	10	1
-6	45	**NA**	1
0	**NA**	30	0
NA	32	35	1
1	63	40	1
-2	**NA**	12	0

$\Rightarrow$ logistic regression with parameter β

Recommended method 1: multiple imputation

Example:

X_1	X_2	X_3	Y
NA	20	10	1
-6	45	NA	1
0	NA	30	0
NA	32	35	1
1	63	40	1
-2	NA	12	0

$\Rightarrow$ logistic regression with parameter β

1. Generate M plausible values for each missing entry

X_1	X_2	X_3	Y
3	20	10	1
-6	45	6	1
0	4	30	0
-4	32	35	1
1	63	40	1
-2	15	12	0

X_1	X_2	X_3	Y
-7	20	10	1
-6	45	9	1
0	12	30	0
13	32	35	1
1	63	40	1
-2	10	12	0

X_1	X_2	X_3	Y
7	20	10	1
-6	45	12	1
0	-5	30	0
2	32	35	1
1	63	40	1
-2	20	12	0

2. Perform the analysis on each imputed data set: $\hat{\beta}_m, \widehat{Var}\left(\hat{\beta}_m\right)$

3. Combine the results (Rubin's rules):

$$\hat{\beta} = \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}_m \qquad \hat{V} = \frac{1}{M}\sum_{m=1}^{M}\widehat{Var}\left(\hat{\beta}_m\right) + \frac{1+\frac{1}{M}}{M-1}\sum_{m=1}^{M}\left(\hat{\beta}_m - \hat{\beta}\right)^2$$

Recommended method 1: multiple imputation

Example:

X_1	X_2	X_3	Y
NA	20	10	1
-6	45	NA	1
0	NA	30	0
NA	32	35	1
1	63	40	1
-2	NA	12	0

$\Rightarrow$ logistic regression with parameter β

1. Generate M plausible values for each missing entry

X_1	X_2	X_3	Y
3	20	10	1
-6	45	6	1
0	4	30	0
-4	32	35	1
1	63	40	1
-2	15	12	0

X_1	X_2	X_3	Y
-7	20	10	1
-6	45	9	1
0	12	30	0
13	32	35	1
1	63	40	1
-2	10	12	0

X_1	X_2	X_3	Y
7	20	10	1
-6	45	12	1
0	-5	30	0
2	32	35	1
1	63	40	1
-2	20	12	0

2. Perform the analysis on each imputed data set: $\hat{\beta}_m, \widehat{Var}\left(\hat{\beta}_m\right)$

3. Combine the results (Rubin's rules):

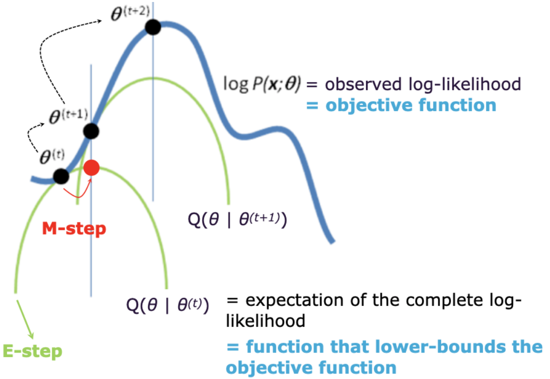
$$\hat{\beta} = \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}_m \qquad \hat{V} = \frac{1}{M}\sum_{m=1}^{M}\widehat{Var}\left(\hat{\beta}_m\right) + \frac{1+\frac{1}{M}}{M-1}\sum_{m=1}^{M}\left(\hat{\beta}_m - \hat{\beta}\right)^2$$

+ **Variability** of missing values is taken into account

– **Aggregating different models** from multiple imputed data is complex

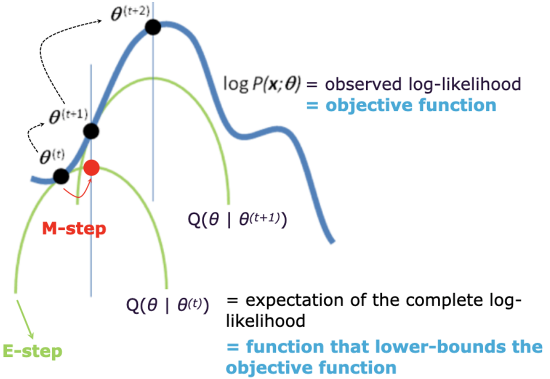
Recommended method 2: EM algorithm

Modify the estimation process to deal with missing values.
Maximum observed likelihood: EM algorithm to obtain point estimates + Supplemented EM (Meng & Rubin, 1991) for their variability



Recommended method 2: EM algorithm

Modify the estimation process to deal with missing values.
Maximum observed likelihood: EM algorithm to obtain point estimates + Supplemented EM (Meng & Rubin, 1991) for their variability



+ Perfectly dedicated toward the problem (ML estimates)
– One specific algorithm for each statistical method
– Not many implementations even for simple models (*e.g. logistic regression*)

– Not a complete methodology

Objectives and contributions

- Complete methodologies for **estimation, model selection and prediction** (few competitors) with **missing data**
  - Classical setting ($n > p$): logistic regression (SAEM)
  - High dimension ($p > n$): parametric & non-parametric regression (FDR control)
- Software packages
  - Implementation of **R packages**
  - Numerical experiments
- Application to the **medical dataset**—TraumaBase
  - Predict the risk of hemorrhagic shock
  - Predict platelet levels

Contribution 1:

Logistic regression with missing covariates

(Jiang , Josse, Lavielle, TraumaBase, 2020)

ELSEVIER

Computational Statistics & Data Analysis

Volume 145, May 2020, 106907

Logistic regression with missing covariates —Parameter estimation, model selection and prediction within a joint-modeling framework

Wei Jiang [a], Julie Josse [a], Marc Lavielle [a], TraumaBase Group [b]

Logistic regression model

$X = (x_{ij})$ a $n \times p$ matrix of quantitative covariates
$y = (y_i)$ an n-vector of binary responses $\{0, 1\}$

Logistic regression model

$$\mathbb{P}\left(y_i = 1 | X_i; \beta\right) = \frac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}$$

Covariates

$$X_i \underset{\text{i.i.d.}}{\sim} \mathcal{N}_p(\mu, \Sigma)$$

Log-likelihood for complete-data with the set of parameters $\theta = (\mu, \Sigma, \beta)$

$$\ell(\theta; X, y) = \sum_{i=1}^{n} \Big(\log(\mathtt{p}(y_i | X_i; \beta)) + \log(\mathtt{p}(X_i; \mu, \Sigma)) \Big).$$

Missing data mechanisms

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

Pattern of missingness: R with $R_{ij} = \begin{cases} 1, & \text{if } X_{ij} \text{ is observed;} \\ 0, & \text{otherwise.} \end{cases}$

Missing data mechanisms

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

Pattern of missingness: R with $R_{ij} = \begin{cases} 1, & \text{if } X_{ij} \text{ is observed;} \\ 0, & \text{otherwise.} \end{cases}$

Missing completely at random (MCAR)

$\mathrm{p}(R \mid X) = \mathrm{p}(R)$ *e.g.* Data lost when merging databases

Missing at random (MAR)

$\mathrm{p}(R \mid X) = \mathrm{p}(R \mid X_{\text{obs}})$ *e.g.* Blood pressure not collected at larger probability in traffic accident.

Missing not at random (MNAR)

$\mathrm{p}(R \mid X) = \mathrm{p}(R \mid X_{\text{obs}}, X_{\text{mis}})$ *e.g.* Blood pressure not collected at larger probability when its value < 90 mmHg.

Missing data mechanisms

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

Pattern of missingness: R with $R_{ij} = \begin{cases} 1, & \text{if } X_{ij} \text{ is observed;} \\ 0, & \text{otherwise.} \end{cases}$

Missing completely at random (MCAR)

$\mathrm{p}(R \mid X) = \mathrm{p}(R)$ *e.g.* Data lost when merging databases

Missing at random (MAR)

$\mathrm{p}(R \mid X) = \mathrm{p}(R \mid X_{\text{obs}})$ *e.g.* Blood pressure not collected at larger probability in traffic accident.

Missing not at random (MNAR)

$\mathrm{p}(R \mid X) = \mathrm{p}(R \mid X_{\text{obs}}, X_{\text{mis}})$ *e.g.* Blood pressure not collected at larger probability when its value < 90 mmHg.

Assumption: Missing data are **Missing at Random**

$\Rightarrow$ Ignore modeling missing mechanism

EM algorithm with missing data

Observed-data likelihood

Aim: $\arg\max_\theta \underline{\ell(\theta; X_{\text{obs}}, y)} = \int \ell(\theta; X, y) dX_{\text{mis}}$.

EM:

Complete-data likelihood

- **E-step:** Evaluate the quantity

$$\begin{aligned} Q_k(\theta) &= \mathbb{E}[\ell(\theta; X, y) | X_{\text{obs}}, y; \theta_{k-1}] \\ &= \int \ell(\theta; X, y) \mathsf{p}(X_{\text{mis}} | X_{\text{obs}}, y; \theta_{k-1}) dX_{\text{mis}}. \end{aligned}$$

- **M-step:** $\theta_k = \arg\max_\theta Q_k(\theta)$.

EM algorithm with missing data

Aim: $\arg\max_\theta \ell(\theta; X_{\text{obs}}, y) = \int \ell(\theta; X, y) dX_{\text{mis}}$.
EM:

- **E-step:** Evaluate the quantity

$$\begin{aligned} Q_k(\theta) &= \mathbb{E}[\ell(\theta; X, y) | X_{\text{obs}}, y; \theta_{k-1}] \\ &= \int \ell(\theta; X, y) \mathtt{p}(X_{\text{mis}} | X_{\text{obs}}, y; \theta_{k-1}) dX_{\text{mis}}. \end{aligned}$$

- **M-step:** $\theta_k = \arg\max_\theta Q_k(\theta)$.

Unfeasible computation of expectation!

MCEM (Wei & Tanner 1990): Generate a large set of samples of missing data from $\mathtt{p}(X_{\text{mis}} | X_{\text{obs}}, y; \theta_{k-1})$ and replaces the expectation by an empirical mean.

EM algorithm with missing data

Aim: $\arg\max_\theta \ell(\theta; X_{\text{obs}}, y) = \int \ell(\theta; X, y) dX_{\text{mis}}$.
EM:

- **E-step:** Evaluate the quantity

$$\begin{aligned} Q_k(\theta) &= \mathbb{E}[\ell(\theta; X, y) | X_{\text{obs}}, y; \theta_{k-1}] \\ &= \int \ell(\theta; X, y) \mathtt{p}(X_{\text{mis}} | X_{\text{obs}}, y; \theta_{k-1}) dX_{\text{mis}}. \end{aligned}$$

- **M-step:** $\theta_k = \arg\max_\theta Q_k(\theta)$.

Unfeasible computation of expectation!

MCEM (Wei & Tanner 1990): Generate a large set of samples of missing data from $\mathtt{p}(X_{\text{mis}} | X_{\text{obs}}, y; \theta_{k-1})$ and replaces the expectation by an empirical mean.

Require a huge number of samples to converge!

Stochastic Approximation EM

(book, Lavielle 2014) Starting from an initial guess θ_0, the kth iteration consists of three steps:

- **Simulation:** For $i = 1, 2, \cdots, n$, draw one sample $X_{i,\text{mis}}^{(k)}$ from

$$\mathtt{p}(X_{i,\text{mis}} | X_{i,\text{obs}}, y_i; \theta_{k-1}).$$

- **Stochastic approximation:** Update the function Q

$$Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k \left(\ell(\theta; X_{\text{obs}}, X_{\text{mis}}^{(k)}, y) - Q_{k-1}(\theta) \right),$$

  where (γ_k) is a decreasing sequence of positive numbers.

- **Maximization:** $\theta_k = \arg\max_\theta Q_k(\theta)$.

Stochastic Approximation EM

(book, Lavielle 2014) Starting from an initial guess θ_0, the kth iteration consists of three steps:

- **Simulation:** For $i = 1, 2, \cdots, n$, draw one sample $X_{i,\text{mis}}^{(k)}$ from
$$\mathtt{p}(X_{i,\text{mis}} | X_{i,\text{obs}}, y_i; \theta_{k-1}).$$
- **Stochastic approximation:** Update the function Q
$$Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k \left(\ell(\theta; X_{\text{obs}}, X_{\text{mis}}^{(k)}, y) - Q_{k-1}(\theta) \right),$$
where (γ_k) is a decreasing sequence of positive numbers.
- **Maximization:** $\theta_k = \arg\max_\theta Q_k(\theta)$.

Convergence: (Allassonniere et al. 2010)
The choice of the sequence (γ_k) is important for ensuring the almost sure convergence of SAEM to a MLE.

Metropolis-Hastings algorithm

Target distribution

$$\begin{aligned} f_i(X_{i,\mathrm{mis}}) &= \mathtt{p}(X_{i,\mathrm{mis}}|X_{i,\mathrm{obs}}, y_i; \theta) \\ &\propto \mathtt{p}(y_i|X_i; \beta)\, \mathtt{p}(X_{i,\mathrm{mis}}|X_{i,\mathrm{obs}}; \mu, \Sigma). \end{aligned}$$

Metropolis-Hastings algorithm

Target distribution

$$
\begin{aligned}
f_i(X_{i,\text{mis}}) &= \mathtt{p}(X_{i,\text{mis}}|X_{i,\text{obs}}, y_i; \theta) \\
&\propto \mathtt{p}(y_i|X_i; \beta)\, \mathtt{p}(X_{i,\text{mis}}|X_{i,\text{obs}}; \mu, \Sigma).
\end{aligned}
$$

Proposal distribution $g_i(X_{i,\text{mis}}) = \mathtt{p}(X_{i,\text{mis}}|X_{i,\text{obs}}; \mu, \Sigma)$

$$X_{i,\text{mis}}|X_{i,\text{obs}} \sim \mathcal{N}_p(\mu_i, \Sigma_i)$$

$$
\begin{aligned}
\mu_i &= \mu_{i,\text{mis}} + \Sigma_{i,\text{mis},\text{obs}} \Sigma_{i,\text{obs},\text{obs}}^{-1} (X_{i,\text{obs}} - \mu_{i,\text{obs}}), \\
\Sigma_i &= \Sigma_{i,\text{mis},\text{mis}} - \Sigma_{i,\text{mis},\text{obs}} \Sigma_{i,\text{obs},\text{obs}}^{-1} \Sigma_{i,\text{obs},\text{mis}},
\end{aligned}
$$

Metropolis-Hastings algorithm

Target distribution

$$f_i(X_{i,\text{mis}}) = \mathtt{p}(X_{i,\text{mis}}|X_{i,\text{obs}}, y_i; \theta)$$
$$\propto \mathtt{p}(y_i|X_i; \beta)\, \mathtt{p}(X_{i,\text{mis}}|X_{i,\text{obs}}; \mu, \Sigma).$$

Proposal distribution $g_i(X_{i,\text{mis}}) = \mathtt{p}(X_{i,\text{mis}}|X_{i,\text{obs}}; \mu, \Sigma)$

$$X_{i,\text{mis}}|X_{i,\text{obs}} \sim \mathcal{N}_p(\mu_i, \Sigma_i)$$

$$\mu_i = \mu_{i,\text{mis}} + \Sigma_{i,\text{mis},\text{obs}}\Sigma^{-1}_{i,\text{obs},\text{obs}}(X_{i,\text{obs}} - \mu_{i,\text{obs}}),$$
$$\Sigma_i = \Sigma_{i,\text{mis},\text{mis}} - \Sigma_{i,\text{mis},\text{obs}}\Sigma^{-1}_{i,\text{obs},\text{obs}}\Sigma_{i,\text{obs},\text{mis}},$$

Metropolis:

1. $\boldsymbol{z}_{im}^{(k)} \sim g_i(\boldsymbol{x}_{i,mis})$, $u \sim \mathcal{U}[0,1]$
2. $r = \frac{f_i(\boldsymbol{z}_{im}^{(k)})/g_i(\boldsymbol{z}_{im}^{(k)})}{f_i(\boldsymbol{z}_{i,m-1}^{(k)})/g_i(\boldsymbol{z}_{i,m-1}^{(k)})}$
3. If $u < r$, accept $\boldsymbol{z}_{im}^{(k)}$

Only need a few steps of Markov chains in each iteration of SAEM.

Metropolis-Hastings algorithm

Target distribution

$$f_i(X_{i,\text{mis}}) = \mathtt{p}(X_{i,\text{mis}}|X_{i,\text{obs}}, y_i; \theta)$$
$$\propto \mathtt{p}(y_i|X_i; \beta)\, \mathtt{p}(X_{i,\text{mis}}|X_{i,\text{obs}}; \mu, \Sigma).$$

Proposal distribution $g_i(X_{i,\text{mis}}) = \mathtt{p}(X_{i,\text{mis}}|X_{i,\text{obs}}; \mu, \Sigma)$

$$X_{i,\text{mis}}|X_{i,\text{obs}} \sim \mathcal{N}_p(\mu_i, \Sigma_i)$$

$$\mu_i = \mu_{i,\text{mis}} + \Sigma_{i,\text{mis},\text{obs}}\Sigma^{-1}_{i,\text{obs},\text{obs}}(X_{i,\text{obs}} - \mu_{i,\text{obs}}),$$
$$\Sigma_i = \Sigma_{i,\text{mis},\text{mis}} - \Sigma_{i,\text{mis},\text{obs}}\Sigma^{-1}_{i,\text{obs},\text{obs}}\Sigma_{i,\text{obs},\text{mis}},$$

Metropolis:

1. $\boldsymbol{z}_{im}^{(k)} \sim g_i(\boldsymbol{x}_{i,mis}), u \sim \mathcal{U}[0,1]$
2. $r = \dfrac{f_i(\boldsymbol{z}_{im}^{(k)})/g_i(\boldsymbol{z}_{im}^{(k)})}{f_i(\boldsymbol{z}_{i,m-1}^{(k)})/g_i(\boldsymbol{z}_{i,m-1}^{(k)})}$
3. If $u < r$, accept $\boldsymbol{z}_{im}^{(k)}$

Only need a few steps of Markov chains in each iteration of SAEM.

Variance estimation:

Given the MH samples of unobserved data, and the SAEM estimate
$\Rightarrow$ Estimate **observed Fisher information** by empirical means.

Model selection: criterion BIC

With $\tilde{p}_{\boldsymbol{\theta}}$ the number of estimated parameters in a given model $\mathcal{M}$, model selection criterion (**penalized likelihood**) :

$$\mathrm{BIC}(\mathcal{M}) = -2\ell(\hat{\theta}_{\mathcal{M}}; X_{\mathrm{obs}}, y) + \log(n)d(\mathcal{M}),$$

How to estimate **observed likelihood** ?

Model selection: criterion BIC

With $\tilde{p}_{\boldsymbol{\theta}}$ the number of estimated parameters in a given model $\mathcal{M}$, model selection criterion (**penalized likelihood**) :

$$\mathrm{BIC}(\mathcal{M}) = -2\ell(\hat{\theta}_{\mathcal{M}}; X_{\mathrm{obs}}, y) + \log(n)d(\mathcal{M}),$$

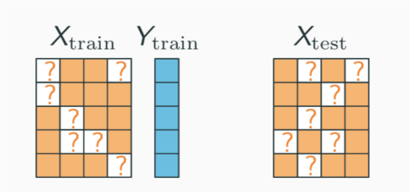
How to estimate **observed likelihood** ?

$$\begin{aligned}
\mathtt{p}(y_i, X_{i,\mathrm{obs}}; \theta) &= \int \mathtt{p}(y_i, X_{i,\mathrm{obs}} | X_{i,\mathrm{mis}}; \theta) \mathtt{p}(X_{i,\mathrm{mis}}; \theta) dX_{i,\mathrm{mis}} \\
&= \int \mathtt{p}(y_i, X_{i,\mathrm{obs}} | X_{i,\mathrm{mis}}; \theta) \frac{\mathtt{p}(X_{i,\mathrm{mis}}; \theta)}{g_i(X_{i,\mathrm{mis}})} g_i(X_{i,\mathrm{mis}}) dX_{i,\mathrm{mis}} \\
&= \mathbb{E}_{g_i}\left(\mathtt{p}(y_i, X_{i,\mathrm{obs}} | X_{i,\mathrm{mis}}; \theta) \frac{\mathtt{p}(X_{i,\mathrm{mis}}; \theta)}{g_i(X_{i,\mathrm{mis}})} \right).
\end{aligned}$$

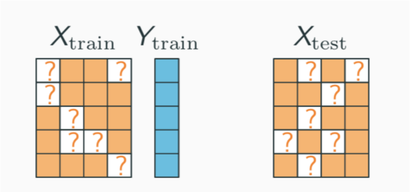
Sample from g_i (the proposal distribution in SAEM)

$\Rightarrow$ Empirical mean.

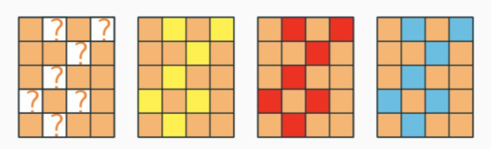
Prediction: missing values in test set



Prediction: missing values in test set



$$x_{\text{mis}}^{(1)},\ x_{\text{mis}}^{(2)},\ \cdots\ x_{\text{mis}}^{(M)} \sim \mathrm{p}\left(x_{\text{mis}} \mid x_{\text{obs}}\right) \Downarrow$$



$$p_m(y) = \mathrm{p}\left(y|x_{\text{obs}}, x_{\text{mis}}^{(m)}\right): \qquad p_1 \qquad p_2 \quad \cdots \quad p_M$$

$$\hat{y} = \arg\max_y \mathrm{p}\left(y|x_{\text{obs}}\right) = \arg\max_y \sum_{m=1}^{M} p_m(y)$$

Method comparison: estimates & coverage

x: $p = 5$, $n = 10\,000$; $y \in \{0, 1\}$
percentage of missingness = 10%
1000 replicates

Figure: Estimation bias of $\hat{\beta}_3$.

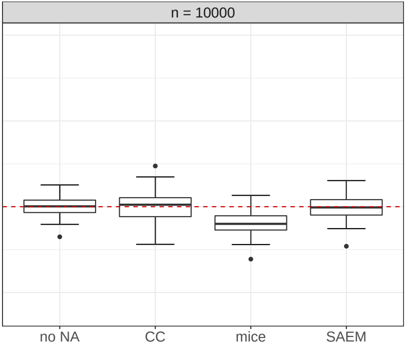


Table: Coverage of confidence interval.

	no NA	CC	mice	SAEM
β_0	95.2	94.4	95.2	94.9
β_1	96.0	94.7	93.9	95.1
β_2	95.5	94.6	94.0	94.3
β_3	94.9	94.3	**86.5**	94.7
β_4	94.6	94.2	96.2	95.4
β_5	95.9	94.4	**89.6**	94.7

Method comparison: estimates & coverage

x: $p = 5$, $n = 10\,000$; $y \in \{0, 1\}$
percentage of missingness = 10%
1000 replicates

Figure: Estimation bias of $\hat{\beta}_3$.

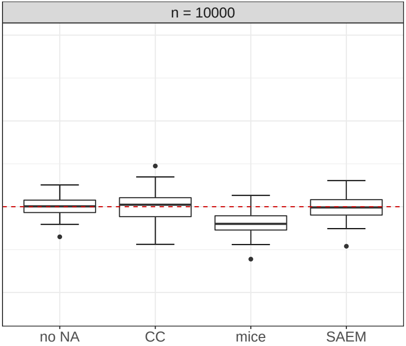


Table: Coverage of confidence interval.

	no NA	CC	mice	SAEM
β_0	95.2	94.4	95.2	94.9
β_1	96.0	94.7	93.9	95.1
β_2	95.5	94.6	94.0	94.3
β_3	94.9	94.3	**86.5**	94.7
β_4	94.6	94.2	96.2	95.4
β_5	95.9	94.4	**89.6**	94.7

Extended simulations:

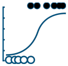
- Robustness (model-misspecification)
- Percentage of missingness
- Separability of classes

Application on TraumaBase

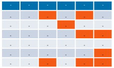
Variables
Age
Weight
Height
BMI
Glasgow
Motor Glasgow
Pulse Pressure min
Pulse Pressure at arrival
Heart Rate max
Heart Rate at arrival
Hb Hemocue
SpO_2
Volume Expander colloids
Volume Expander crystalloids.

- 6384 patients
- 14 continuous variables

Logistic regression with missing values



+



Hemorrhagic shock

$P\left(y = 1 \,\middle|\, X; \hat{\beta}\right)$?

Application on TraumaBase

Variables	Effect	Estimate (std error)
Age	+	0.011 (0.0033)
~~Weight~~		
~~Height~~		
~~BMI~~		
~~Glasgow~~		
Motor Glasgow	-	-0.16 (0.036)
Pulse Pressure min	-	-0.025 (0.0050)
Pulse Pressure at arrival	-	-0.021 (0.0056)
Heart Rate max	+	0.026 (0.0025)
~~Heart Rate at arrival~~		
Hb Hemocue	-	-0.23 (0.031)
~~SpO_2~~		
Volume Expander colloids	+	0.0019 (0.00021)
Volume Expander crystalloids.	+	0.00090 (0.00010)

« - » effects

- A low Glasgow score means one makes no motor response, often in the case of hemorrhagic shock.

Hemorrhagic shock

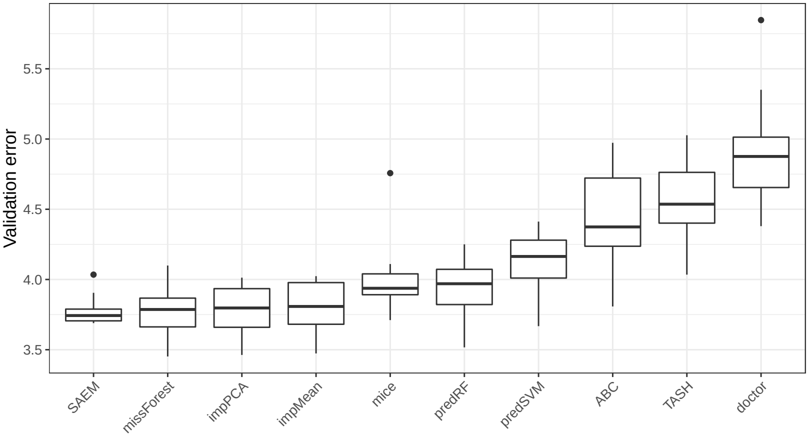
$$P\left(y = 1 \mid X; \hat{\beta}\right) \; ?$$

« + » effects

- Older people tend to have a larger possibility to suffer from hemorrhagic shock.

Predictive performance

Random split : training set (70%) + test set (30%) (repeated 15 times)



False Negative costs 10 times more than False Positive ⇒ Threshold

R package: misaem (Available in CRAN)

misaem: Linear Regression and Logistic Regression with Missing Covariates

Estimate parameters of linear regression and logistic regression with missing covariates with missing data, perform model selection and prediction, using EM-type algorithms.

Version:	1.0.0
Depends:	R (≥ 3.4.0)

CRAN
Mirrors

Parameter estimation:

```
miss.glist = miss.glm(y~., data = df, maxruns = 500)
summary(miss.glist)
```

misaem: Linear Regression and Logistic Regression with Missing Covariates

Estimate parameters of linear regression and logistic regression with missing covariates with missing data, perform model selection and prediction, using EM-type algorithms.

Version:	1.0.0
Depends:	R (≥ 3.4.0)

CRAN
Mirrors

Parameter estimation:

```
miss.glist = miss.glm(y~., data = df, maxruns = 500)
summary(miss.glist)
```

Model selection with BIC:

```
miss.model = miss.glm.model.select(y, X)
print(miss.model)
```

R package: misaem (Available in CRAN)

misaem: Linear Regression and Logistic Regression with Missing Covariates

Estimate parameters of linear regression and logistic regression with missing covariates with missing data, perform model selection and prediction, using EM-type algorithms.

Version:	1.0.0
Depends:	R (≥ 3.4.0)

CRAN
Mirrors

Parameter estimation:

```
miss.glist = miss.glm(y~., data = df, maxruns = 500)
summary(miss.glist)
```

Model selection with BIC:

```
miss.model = miss.glm.model.select(y, X)
print(miss.model)
```

Prediction on (incomplete) test set:

```
pr.saem <- predict(miss.model, X.test)
```

R package: misaem (Available in CRAN)

misaem: Linear Regression and Logistic Regression with Missing Covariates

Estimate parameters of linear regression and logistic regression with missing covariates with missing data, perform model selection and prediction, using EM-type algorithms.

Version:	1.0.0
Depends:	R (≥ 3.4.0)

CRAN
Mirrors

Parameter estimation:

```
miss.glist = miss.glm(y~., data = df, maxruns = 500)
summary(miss.glist)
```

Model selection with BIC:

```
miss.model = miss.glm.model.select(y, X)
print(miss.model)
```

Prediction on (incomplete) test set:

```
pr.saem <- predict(miss.model, X.test)
```

Also provide solutions for **linear regression with missing values:**

```
miss.list = miss.lm(y~., data = df)
```

Contribution 2:

Variable selection for high-dimensional incomplete data

(Jiang, Bogdan, Josse, Miasojedow, Rockova, 2019)

arXiv | Cornell University

Statistics > Methodology

arXiv:1909.06631 (stat)

[Submitted on 14 Sep 2019 (v1), last revised 6 Nov 2019 (this version, v2)]

Adaptive Bayesian SLOPE -- High-dimensional Model Selection with Missing Values

Wei Jiang, Malgorzata Bogdan, Julie Josse, Blazej Miasojedow, Veronika Rockova, TraumaBase Group

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

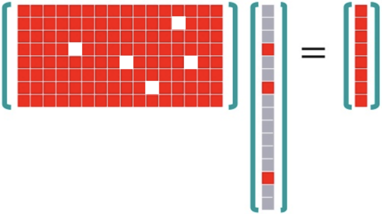
$$y \in \mathbb{R}^n, \quad X \in \mathbb{R}^{n \times p}, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$$

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

$$y \in \mathbb{R}^n, \quad X \in \mathbb{R}^{n \times p}, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$$

Assumptions:

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



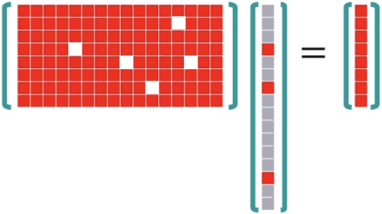
Model selection in high-dimension

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

$$y \in \mathbb{R}^n, \quad X \in \mathbb{R}^{n\times p}, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$$

Assumptions:

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



Aims:

- Model selection with FDR control
- Parameter estimation with less bias
- Managing missing values

- LASSO (Tibshirani, 1996)

$$\hat{\beta}_{LASSO} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|^2 + \lambda\|\beta\|_1,$$

detects important variables with high probability but includes many **false positives**.

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

- SLOPE (Bogdan et al., 2015) penalizes larger coefficients more stringently
$$\hat{\beta}_{SLOPE} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|^2 + \sigma \sum_{j=1}^{p} \lambda_j |\beta|_{(j)},$$
where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ and $|\beta|_{(1)} \geq |\beta|_{(2)} \geq \cdots \geq |\beta|_{(p)}$.

l_1 penalization methods (complete data)

- 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 (Bogdan et al., 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 $|\beta|_{(1)} \geq |\beta|_{(2)} \geq \cdots \geq |\beta|_{(p)}$.

  To control **False Discovery Rate (FDR)** at level q:
  $\lambda_{BH}(j) = \phi^{-1}(1 - q_j), \quad q_j = \frac{jq}{2p}, \quad X^T X = I, \quad$ then
$$FDR = \mathbb{E}\left[\frac{\#\text{False rejections}}{\#\text{Rejections}}\right] \leq q$$

Problem: λ for SLOPE leading to FDR control are typically large.
SLOPE often returns **an inconsistent estimation.**

$\Rightarrow$ improve?

Bayesian SLOPE (complete data)

Problem: λ for SLOPE leading to FDR control are typically large.
SLOPE often returns **an inconsistent estimation.**

$$\Rightarrow \text{improve?}$$

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 (complete data)

We propose an adaptive version of Bayesian SLOPE (ABSLOPE), with the prior for β as

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

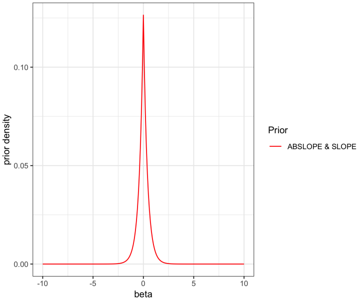
Interpretation of the model:

- β_j is large enough $\Rightarrow$ **true signal**; 0 $\Rightarrow$ noise.
- $\gamma_j \in \{0, 1\}$ signal indicator. $\gamma_j | \theta \sim Bernoulli(\theta)$ and θ the **sparsity**.
- $c \in [0, 1]$: the inverse of **average signal magnitude**.
- $W = \text{diag}(w_1, w_2, \cdots, w_p)$ and its diagonal element:

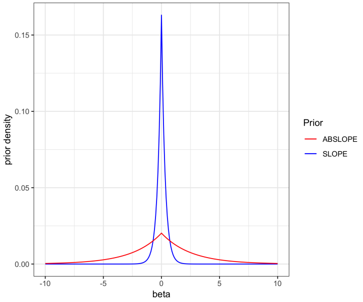
$$w_j = c\gamma_j + (1 - \gamma_j) = \begin{cases} c, & \gamma_j = 1 \\ 1, & \gamma_j = 0 \end{cases}.$$

Advantage of introducing W:

- when $\gamma_j = 0$, $w_j = 1$, i.e., the null variables are treated with the regular SLOPE penalty
- when $\gamma_j = 1$, $w_j = c < 1$, i.e, **smaller penalty** $\lambda_{r(W\beta,j)}$ for true predictors than the regular SLOPE one



Null β

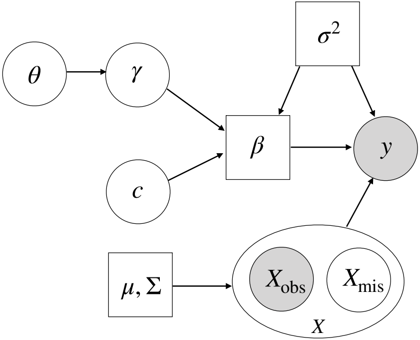


Non-null β

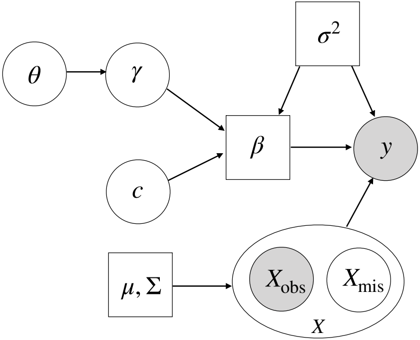
Figure: comparison of SLOPE prior and ABSLOPE prior

Modeling with missingness

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



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



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

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

Adapted SAEM algorithm

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

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

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

    $$Q^t = Q^{t-1} + \eta_t \left(\ell_{\text{comp}} \Big|_{X^t_{\text{mis}}, \gamma^t, c^t, \theta^t} - Q^{t-1} \right).$$

- *M step:* $\beta^t, \sigma^t, \mu^t, \Sigma^t = \arg\max Q^t$.
  [**Proximal gradient descent, Shrinkage of covariance**]

Details of initialization, generating samples and optimization are in arXiv:1909.06631

Shrinkage of covariance matrix

Estimation of covariance matrix Σ in high-dimension:

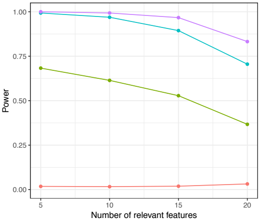
- In some special case, Σ is known.
- If given sparseness $\Rightarrow$ graphical lasso
- But no additional knowledge of $\Sigma \Rightarrow$ shrinkage estimation.
  Optimal linear shrinkage (Ledoit and Wolf, 2012):

$$\hat{\Sigma} = \rho_1 I + \rho_2 S, \qquad \text{where } \rho_1, \rho_2 = \arg\min_{\rho_1,\rho_2} \mathbb{E}\|\hat{\Sigma} - \Sigma\|^2.$$

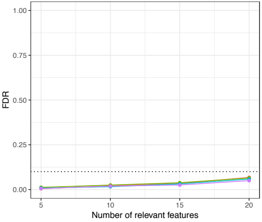
$\Rightarrow$ shrink the empirical eigenvalues towards their mean;
ρ_1 and ρ_2 chosen by asymptotically uniformly minimum quadratic risk.

Simulation study (200 rep. ⇒ average)

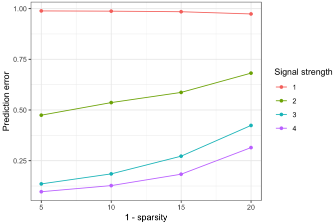
$n = p = 100$, no correlation and 10% missingness



Power



FDR

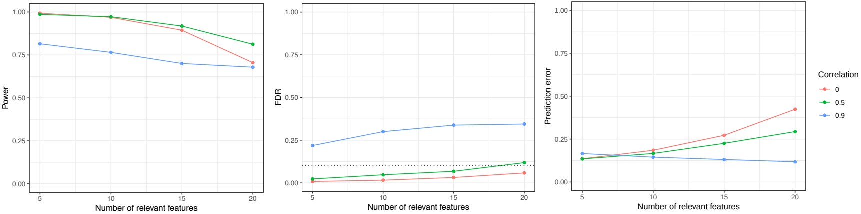


Prediction error

- FDR controlled at expected level 0.1.
- Power increases and estimation bias decreases if larger sparsity or stronger signal.

Simulation study (200 rep. ⇒ average)

with correlation



Power FDR Prediction error

- FDR controlled with small correlation.
- Existence of correlation increases the prediction accuracy.

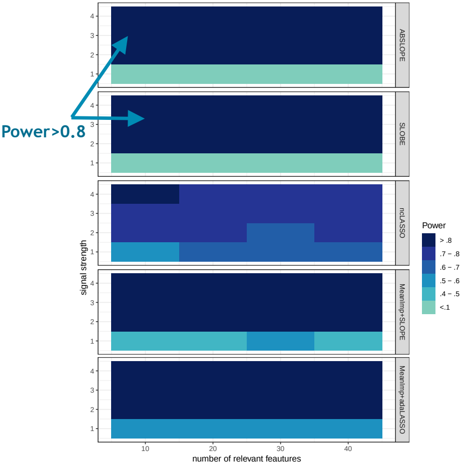
Method comparison (few competitors)

- **ABSLOPE**
- **SLOBE**: simplified version (conditional expectation instead of generating samples of latent variables)
- **ncLASSO** (Loh and Wainwright, 2012): LASSO with NA
  $\Rightarrow$ Non-convex optimisation
  requires to know bound of $\|\beta\|_1 \Rightarrow$ difficult in practice
- Mean imputation followed by
  - SLOPE with known σ
  - adaptive LASSO (Zou, 2006)

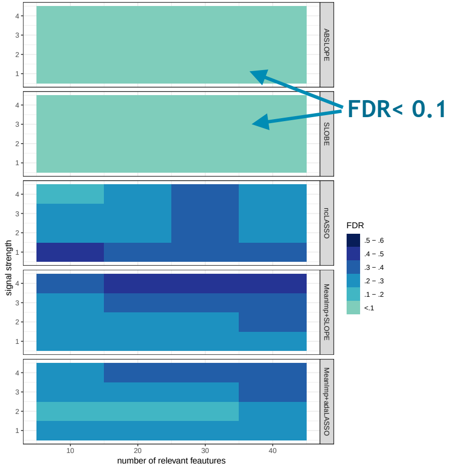
In the SLOPE type methods, λ = BH sequence which controls the FDR at level **0.1**

Method comparison (200 rep. ⇒ average)

500×500 dataset, 10% missingness, with correlation
darker color = larger value.



Power



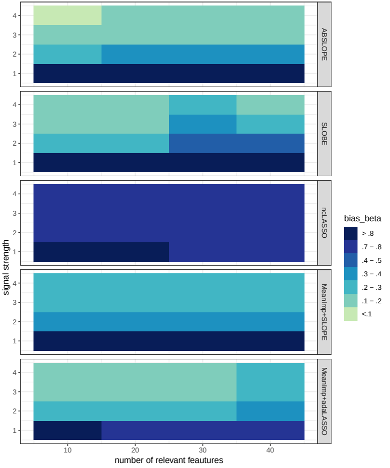
FDR

- ABSLOPE & SLOBE: FDR control (<0.1) when signal strength >1
- Others: sacrifice FDR to achieve good power

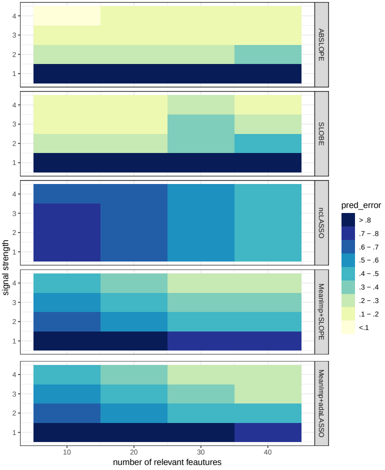
Method comparison (200 rep. ⇒ average)

500×500 dataset, 10% missingness, with correlation

darker color = larger value.



Bias of β



Prediction error

- ABSLOPE: good performance, especially with larger sparsity and stronger signal strength.

Computational cost

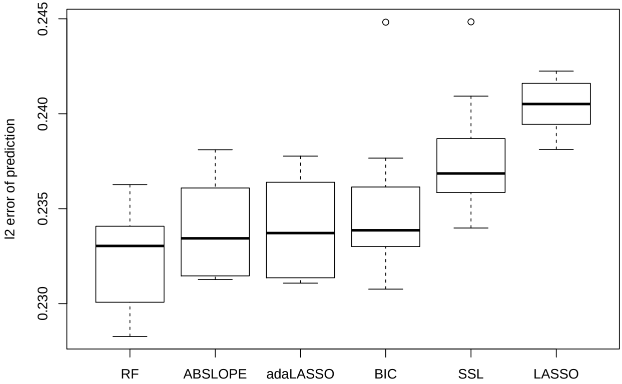
Execution time (seconds) for one simulation	$n = p = 100$			$n = p = 500$		
	min	mean	max	min	mean	max
ABSLOPE	12.83	14.33	20.98	646.53	696.09	975.73
SLOBE	**0.31**	**0.34**	**0.66**	**14.23**	**15.07**	**29.52**
ncLASSO	16.38	20.89	51.35	91.90	100.71	171.00
MeanImp + SLOPE	0.01	0.02	0.09	0.24	0.28	0.53
MeanImp + LASSO	0.10	0.14	0.32	1.75	1.85	3.06

[**Fast implementation:** Parallel computing + Rcpp (C++)]

More on the real data

TraumaBase: Measurements $\xrightarrow{\text{Predict}}$ Platelet

Cross-validation: random splits to training and test sets $\times$ 10



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

R package: ABSLOPE

ABSLOPE

R Package for "Adaptive Bayesian SLOPE --- High-dimensional Model Selection with Missing Values"

(2019, Bogdan M., Jiang W., Josse J., Miasojedow B., Rockova V.)

Languages

R 90.2%

C++ 9.8%

Main algorithm:

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

R package: ABSLOPE

ABSLOPE

R Package for "Adaptive Bayesian SLOPE --- High-dimensional Model Selection with Missing Values"

(2019, Bogdan M., Jiang W., Josse J., Miasojedow B., Rockova V.)

Languages

- R 90.2%
- C++ 9.8%

Main algorithm:

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

A fast and simplified algorithm (C++):

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

R package: ABSLOPE

ABSLOPE

R Package for "Adaptive Bayesian SLOPE --- High-dimensional Model Selection with Missing Values"

(2019, Bogdan M., Jiang W., Josse J., Miasojedow B., Rockova V.)

Languages

- R 90.2%
- C++ 9.8%

Main algorithm:

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

A fast and simplified algorithm (C++):

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

Coefficient and support recovery:

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

Contribution 3:

Controlled model selection with non-parametric regression model

(preprint, 2020)

MISSKNOCKOFF: CONTROLLED VARIABLE SELECTION WITH MISSING VALUES

WEI JIANG[1], SZYMON MAJEWSKI[2], MALGORZATA BOGDAN[3], JULIE JOSSE[1], ASAF WEINSTEIN[4]

1. CMAP, ECOLE POLYTECHNIQUE & INRIA XPOP, FRANCE 2. UNIVERSITY OF WARSAW, POLAND
3. UNIVERSITY OF WROCLAW, POLAND & LUND UNIVERSITY, SWEDEN 4. THE HEBREW UNIVERSITY OF JERUSALEM

Model-X assumption (complete data)

Similar setting (High-dimensional sparse regression) and aim (FDR control) as ABSLOPE:

$$n \textit{ i.i.d.} \text{ samples } (X_{i1}, X_{i2}, \cdots, X_{ip}, y_i)_{i=1}^n$$

$$y_i \mid (X_{i1}, \ldots, X_{ip}) \overset{\text{ind.}}{\sim} P_{y|X}, \quad i = 1, \ldots, n$$

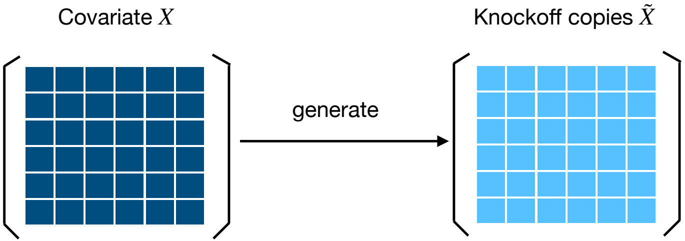
but:

- Conditional distribution $P_{y|X}$ not specified (non-parametric)
- Distribution of X is known (model-X)

Knockoff method (complete data)

Non-parametric model selection with **knockoff** (Candes et al., 2018)

1. Generate "fake" variables (without looking at y)



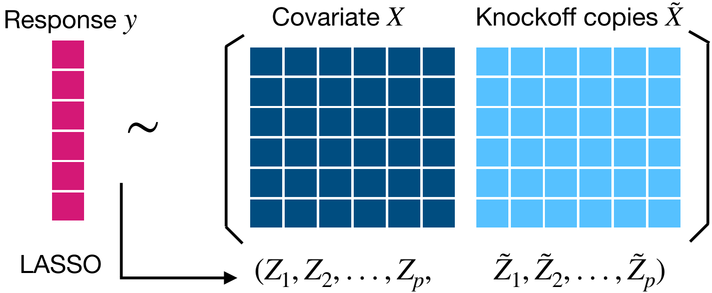
- Correlation between $\tilde{X}_j$ and $\tilde{X}_k$
  $=$ Correlation between X_j and X_k $(j \neq k)$
- Correlation between X_j and $\tilde{X}_k$
  $=$ Correlation between X_j and X_k $(j \neq k)$

$\Rightarrow$ Knockoffs have same structure but all null.

Knockoff method (complete data)

Non-parametric model selection with **knockoff** (Candes et al., 2018)

1. Generate "fake" variables (without looking at y)
2. **Measure variable importance**



$$(Z_1, Z_2, \ldots, Z_p, \quad \tilde{Z}_1, \tilde{Z}_2, \ldots, \tilde{Z}_p)$$

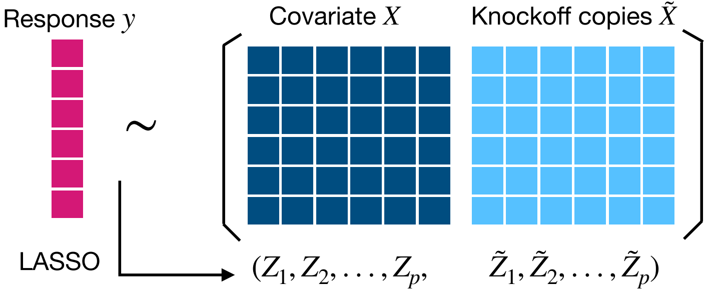
$$= (|\beta_1|, |\beta_2|, \ldots, |\beta_p|, \quad |\tilde{\beta}_1|, |\tilde{\beta}_2|, \ldots, |\tilde{\beta}_p|)$$

- Null variable: $Z_j \overset{\mathrm{d}}{=} \tilde{Z}_j$
- Important variable: $Z_j >> \tilde{Z}_j$

Knockoff method (complete data)

Non-parametric model selection with **knockoff** (Candes et al., 2018)

1. Generate "fake" variables (without looking at y)
2. Measure variable importance

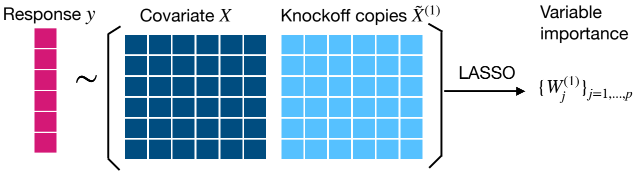


3. **Select** variables more important than their knockoff copies:
   - Large $W_j = Z_j - \tilde{Z}_j$
   - $W_j \geq \tau$ a threshold to control FDR at q:

$$\tau = \min\left\{ t > 0 : \frac{1 + \#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\}} \right\}$$

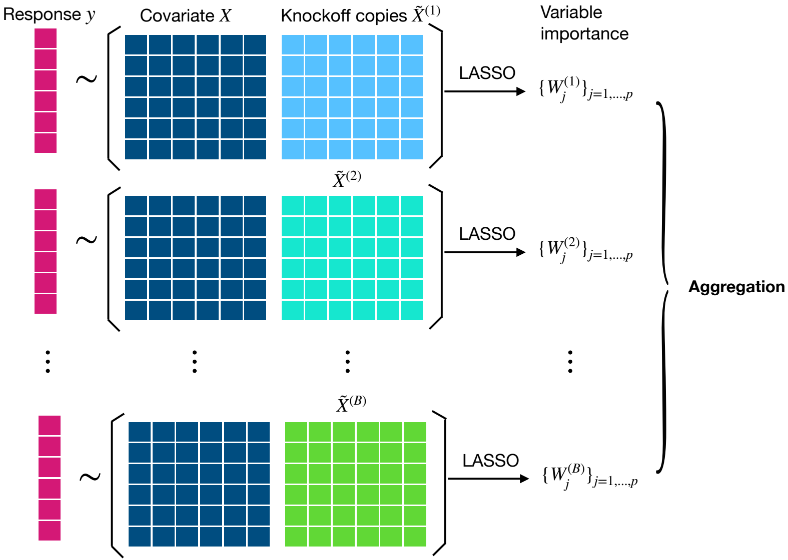
Multiple knockoffs (complete data)

Single knockoff $\rightarrow$ instability $\Rightarrow$ Multiple knockoffs



Multiple knockoffs (complete data)

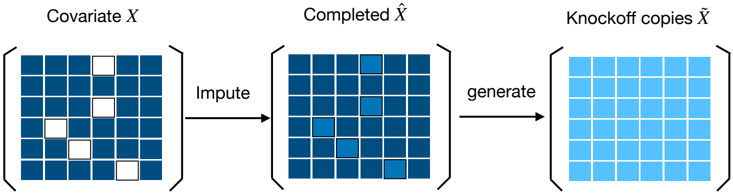
Single knockoff $\rightarrow$ instability $\Rightarrow$ Multiple knockoffs



missKnockoff: single imputation

Contribution:

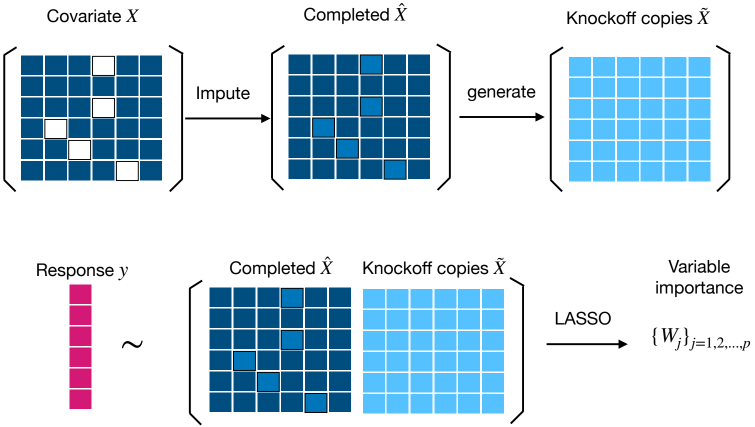
- Combine single knockoff with single imputation



missKnockoff: single imputation

Contribution:

- Combine single knockoff with single imputation



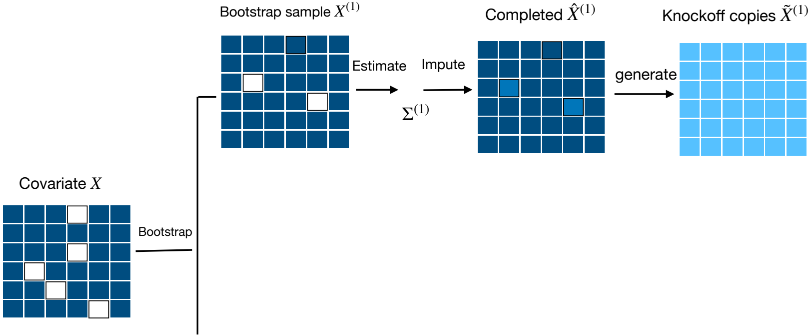
missKnockoff: multiple imputation

Contributions:

- Multiple imputation $\Rightarrow$ single knockoff on each imputed dataset values
- Suggest new aggregation rules (inspired by multiple knockoffs)

+ take variability into account

Step1: Bootstrap B times



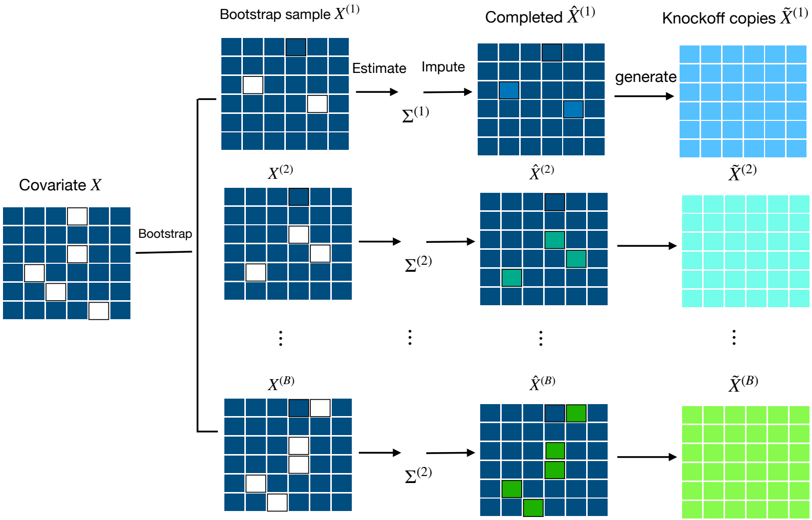
On each bootstrap sample, estimate the covariance (Schneider, 2001; Lounici et al., 2014):

$$\Sigma^{(b)} = \left(\delta^{-1} - \delta^{-2}\right) \operatorname{diag}\left(\Sigma_n\right) + \delta^{-2}\Sigma_n \quad \Rightarrow \quad \text{impute } \mathrm{p}(X_{\text{mis}}|X_{\text{obs}})$$

δ: the proportion of observed entries
Σ_n: the linear shrinkage estimation on empirical covariance of initially imputed dataset by 0.

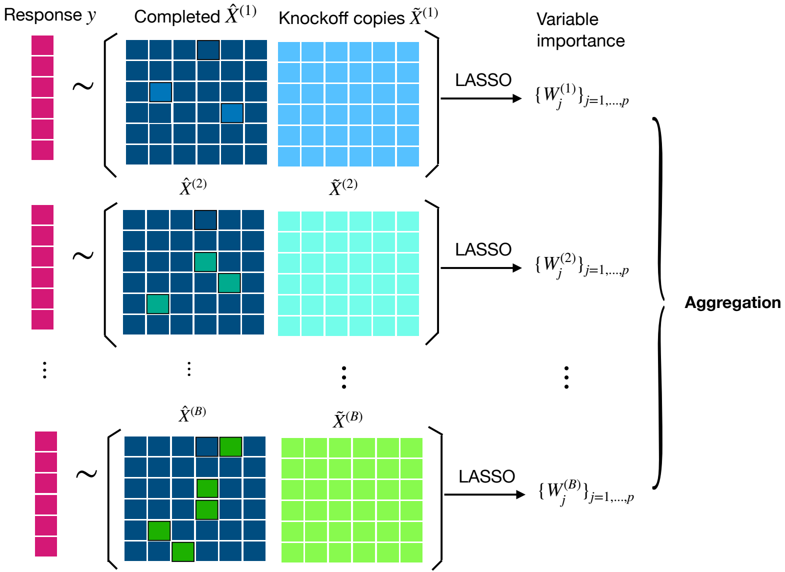
missKnockoff: multiple imputation

Step1: Bootstrap B times



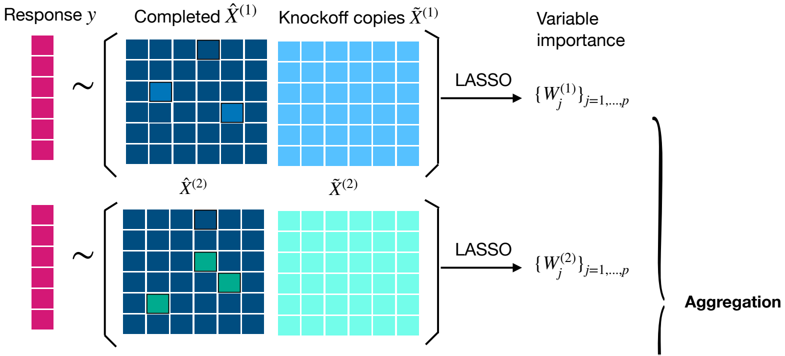
missKnockoff: multiple imputation

Step2: Measure variable importance



missKnockoff: multiple imputation

Step3: Aggregation by averaging the cases



1. Estimate the knockoff threshold:

$$\tau = \min\left\{ t : \frac{1}{B}\sum_{b=1}^{B} \frac{\#\left\{j : W_j^{(b)} \le -t\right\} + 1}{\#\left\{j : W_j^{(b)} \ge t\right\}} \le q \right\}.$$

2. Calculate the median of $\{W_j^{(b)}\}$ over $b = 1, 2, \cdots, B$ to obtain $\bar{W}_j$.

If $\bar{W}_j > \tau \Rightarrow$ **Select j-th variable.**

Theoretical result

Theorem (FDR control for single missKnockoff)

missKnockoff procedure with single imputation from $\mathtt{p}(X_{mis}|X_{obs})$ *controls FDR at level* q.

Theoretical result

Theorem (FDR control for single missKnockoff)

missKnockoff procedure with single imputation from $\mathtt{p}(X_{mis}|X_{obs})$ *controls FDR at level* q.

Theorem (FDR estimation for multiple missKnockoff)

Consider the ***single missKnockoff*** *procedure, which rejects* $H_{0j} : \beta_j = 0$ *if* $W_j > t$*, and let*

$$FDR(t) = \mathbb{E}\left[\frac{\#\{j \in H_0 : W_j \geq t\}}{\#\{j : W_j \geq t\}}\right] .$$

Then for the ***multiple missKnockoffs*** *procedure with variable importance statistics* $\{W_j^b\}$:

$$\underbrace{\mathbb{E}\left(\frac{1}{B}\sum_{b=1}^{B}\frac{\#\left\{j : W_j^{(b)} \leq -t\right\}}{\#\left\{j : W_j^{(b)} \geq t\right\}}\right)}_{\widehat{\boldsymbol{FDR}}_B(t)} \geq FDR(t) .$$

Theoretical result

Theorem (FDR control for single missKnockoff)

missKnockoff procedure with single imputation from $p(X_{mis}|X_{obs})$ *controls FDR at level* q.

Theorem (FDR estimation for multiple missKnockoff)

*Consider the **single missKnockoff** procedure, which rejects* $H_{0j} : \beta_j = 0$ *if* $W_j > t$*, and let*

$$FDR(t) = \mathbb{E}\left[\frac{\#\{j \in H_0 : W_j \geq t\}}{\#\{j : W_j \geq t\}}\right] .$$

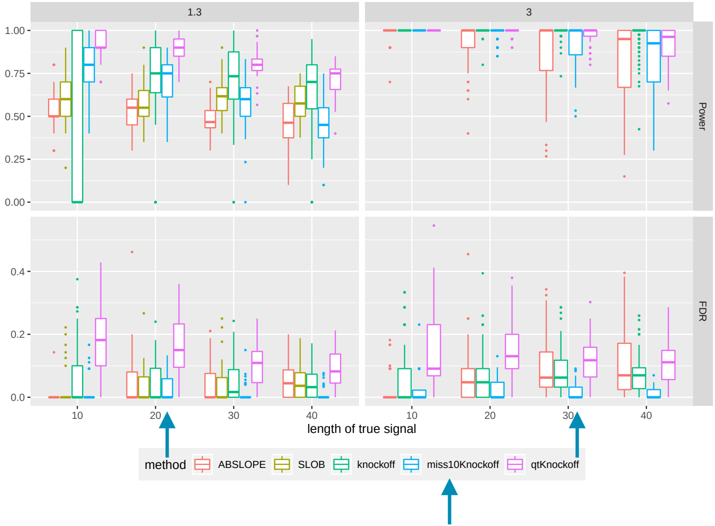
*Then for the **multiple missKnockoffs** procedure with variable importance statistics* $\{W_j^b\}$:

$$\underbrace{\mathbb{E}\left(\frac{1}{B}\sum_{b=1}^{B}\frac{\#\left\{j : W_j^{(b)} \leq -t\right\}}{\#\left\{j : W_j^{(b)} \geq t\right\}}\right)}_{\widehat{\boldsymbol{FDR}}_B(t)} \geq FDR(t) .$$

- $\widehat{\boldsymbol{FDR}}_B(t)$ for missKnockoff with B bootstrap is an upwards biased estimator of $FDR\ (t)$, with variance which diminishes with B (for $t > 0$ and $B > 1$).
- It holds almost surely that $\lim_{B\to\infty} \widehat{\boldsymbol{FDR}}_B(t) = \mathbb{E}\left[\widehat{FDR}(t)|X_{\text{obs}}, y\right]$, the right side = the conditional expectation of estimated false discovery proportion provided by the single missKnockoff procedure.

Simulation results (few competitors)

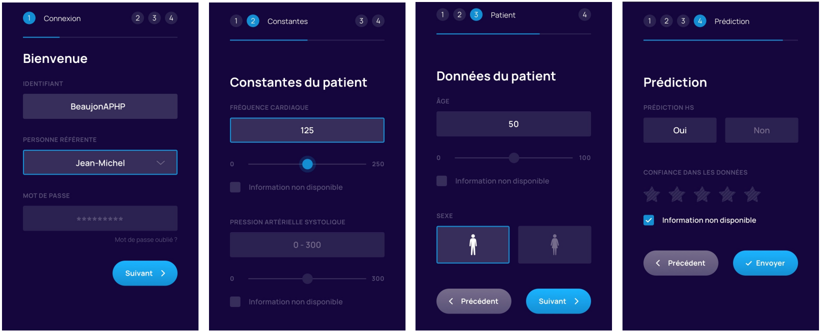
$n = p = 500$
Signal strength $1.3\sqrt{2\log p}$ (left) / strong $3\sqrt{2\log p}$ (right).



General conclusion

- Comprehensive framework for dealing with missing values from estimation to model selection for logistic regression model
  - Methodology, algorithm, simulations
  - R package misaem
- New methods for high-dimensional model selection with FDR control (parametric/ non-parametric)
  - Methodology, algorithm, theoretical results, simulations
  - R package ABSLOPE
- Analysis of hospital dataset (TraumaBase)
  - Improve health care (interpretability, transparency)
  - Results presented at French Society of Anesthesia & Intensive Care Medicine (SFAR) meeting
  - TraumaBase mobile application under development

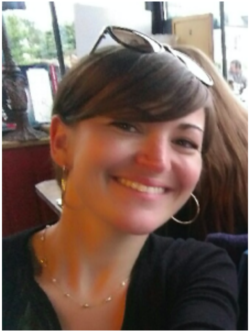
Screenshots of TraumaBase application



Perspectives

- Extension to deal with mixed incomplete covariates with both continuous and categorical, ordinal and binary data (ongoing)
  - General location model (Zhao and Udell, 2019)
  - Gaussian copula (Zhao and Udell, 2019)
- Extension of ABSLOPE (ordered l_1 penalty) in generalized linear models
- Extension to another missing mechanism (MNAR)
- Testing unconditional independence (Candes et al., 2018) with missing values (to improve the power for missKnockoff)

Acknowledgment



Julie Josse

Marc Lavielle

Acknowledgment

Malgorzata Bogdan

LABEX Mathématique Hadamard.

Fondation mathématique FMJH Jacques Hadamard

JUNIOR SCIENTIFIC VISIBILITY PROGRAM

Uniwersytet Wrocławski

Acknowledgment


Blazej Miasojedow


Asaf Weinstein


Veronika Rockova


Sophie Hamada

traumabase.eu


Tobias Gauss

Thanks for your attention!

Merci

Appendix 1:
Logistic regression with missing covariates

Variance estimation

Observed Fisher information matrix (FIM) *wrt* β

$$\mathcal{I}(\theta) = -\frac{\partial^2 \ell(\theta; X_{\text{obs}}, y)}{\partial\theta\partial\theta^T}.$$

Variance estimation

Observed Fisher information matrix (FIM) *wrt* β

$$\mathcal{I}(\theta) = -\frac{\partial^2 \ell(\theta; X_{\text{obs}}, y)}{\partial\theta\partial\theta^T}.$$

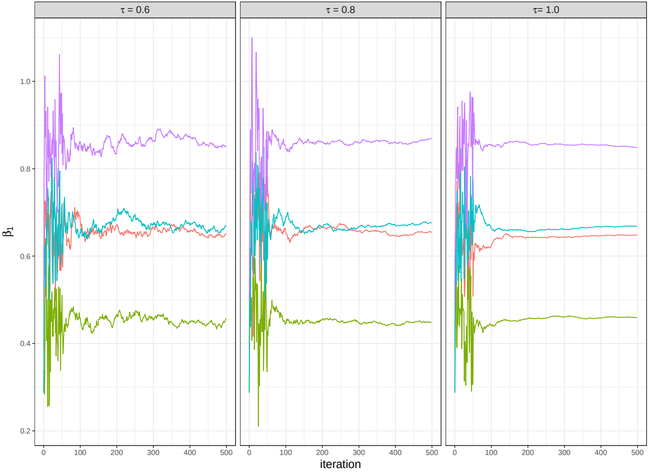
Louis formula

$$\begin{aligned}\mathcal{I}(\theta) = &-\mathbb{E}\left(\frac{\partial^2 \ell(\theta; X, y)}{\partial\theta\partial\theta^T}\middle| X_{\text{obs}}, y; \theta\right)\\ &-\mathbb{E}\left(\frac{\partial \ell(\theta; X, y)}{\partial\theta}\frac{\partial \ell(\theta; X, y)^T}{\partial\theta}\middle| X_{\text{obs}}, y; \theta\right)\\ &+\mathbb{E}\left(\frac{\partial \ell(\theta; X, y)}{\partial\theta}\middle| X_{\text{obs}}, y; \theta\right)\mathbb{E}\left(\frac{\partial \ell(\theta; X, y)}{\partial\theta}\middle| X_{\text{obs}}, y; \theta\right)^T.\end{aligned}$$

Given the MH samples of unobserved data $(X_{i,\text{mis}}^{(m)}, 1 \leq i \leq n, 1 \leq m \leq M)$, and the SAEM estimate $\hat{\boldsymbol{\theta}}$
$\Rightarrow$ Estimate FIM by empirical means.

Simulation study: SAEM behavior

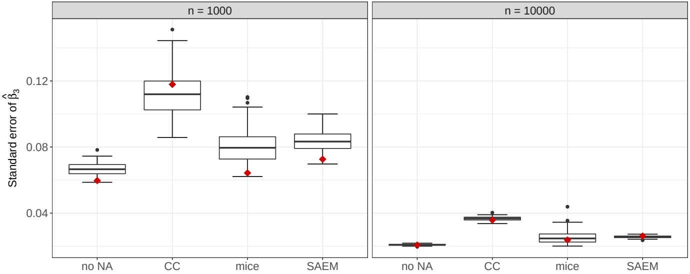
Step size : $\gamma_k = (k - k_1)^{-\tau}$
$k_1 = 50$ and $\tau = (0.6, 0.8, 1.0)$.
$N = 1000$, $p = 5$, percentage of missingness$= 10\%$
4 repetitions of simulations and 500 iterations:



Method comparison: coverage

Table: Coverage (%) for $n = 10\,000$, calculated over 1000 simulations.

parameter	no NA	CC	mice	SAEM
β_0	95.2	94.4	95.2	94.9
β_1	96.0	94.7	93.9	95.1
β_2	95.5	94.6	94.0	94.3
β_3	94.9	94.3	**86.5**	94.7
β_4	94.6	94.2	96.2	95.4
β_5	95.9	94.4	**89.6**	94.7



Model selection results

Table: For data with or without correlation, percentage of times that different criterion selects the correct true model (C), overfit (O), i.e. select more variables, and underfit (U) select less variables.

	Non-Correlated			Correlated		
Criterion	C	O	U	C	O	U
AIC_{obs}	60	40	0	65	32	3
AIC_{orig}	73	27	0	75	20	5
AIC_{cc}	67	32	1	77	16	7
BIC_{obs}	**92**	3	5	**94**	2	4
BIC_{orig}	96	2	2	93	0	7
BIC_{cc}	79	1	20	91	0	9

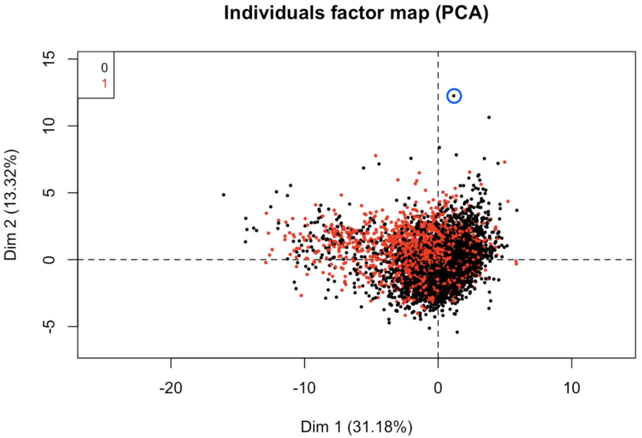
Method comparison: execution time

Table: Comparison of execution time between no NA, MCEM, mice, and SAEM with $n = 1000$ calculated over 1000 simulations.

Execution time (seconds)	no NA	MCEM	mice	SAEM
min	2.87×10^{-3}	492	0.64	9.96
mean	4.65×10^{-3}	773	0.70	13.50
max	43.50×10^{-3}	1077	0.76	16.79

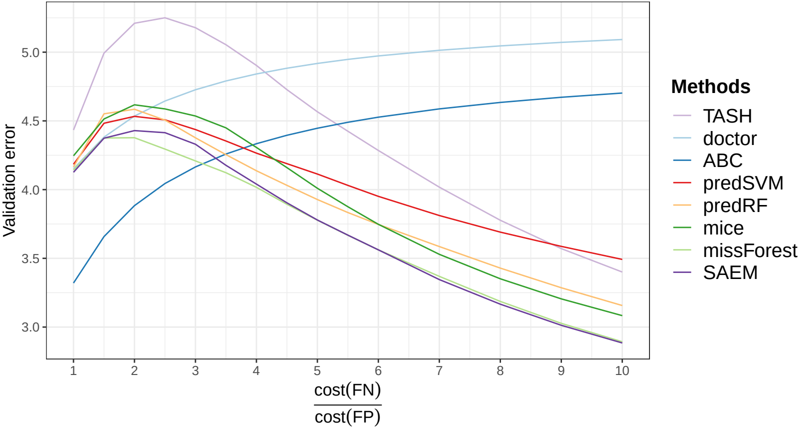
Exploration of dataset

Data preprocessing ⇒ **6384 patients** in the dataset.
Clinical experience ⇒ **14 influential quantitative measurements**
Based on **penalized observed log-likelihood**:
⇒ Observations resulting in a very small value of the log-likelihood.
⇒ wrong records



Predictive performance

Random split : training set (70%) + test set (30%) (repeated 15 times)



Appendix 2:
ABSLOPE

False discovery rate control

In an orthogonal design:

$$\tilde{y} = X^T y = X^T X\beta + X^T\varepsilon = \beta + X^T\varepsilon \sim \mathcal{N}(\beta, \sigma^2 I_p).$$

Selecting model $\Leftrightarrow$ multiple tests: $H_{0,j} : \beta_j = 0$. To control the FDR at level q, (Benjamini and Hochberg, 1995)

1. sort $|\tilde{y}|_{(1)} \geq \cdots \geq |\tilde{y}|_{(p)}$
2. corresponding hypotheses $H_{(1)}, \cdots, H_{(p)}$
3. rejects all $H_{(i)}$ for which

$$i \leq i_{BH} = \max\left\{ i : \frac{|\tilde{y}|_{(i)}}{\sigma} \geq \phi^{-1}(1 - q_i) \right\}, \quad q_i = \frac{iq}{2p},$$

False discovery rate control

In an orthogonal design:

$$\tilde{y} = X^T y = X^T X\beta + X^T\varepsilon = \beta + X^T\varepsilon \sim \mathcal{N}(\beta, \sigma^2 I_p).$$

Selecting model $\Leftrightarrow$ multiple tests: $H_{0,j} : \beta_j = 0$. To control the FDR at level q, (Benjamini and Hochberg, 1995)

1. sort $|\tilde{y}|_{(1)} \geq \cdots \geq |\tilde{y}|_{(p)}$
2. corresponding hypotheses $H_{(1)}, \cdots, H_{(p)}$
3. rejects all $H_{(i)}$ for which

$$i \leq i_{BH} = \max\left\{ i : \frac{|\tilde{y}|_{(i)}}{\sigma} \geq \phi^{-1}(1 - q_i) \right\}, \quad q_i = \frac{iq}{2p},$$

For SLOPE, if we set $\lambda_{BH}(j) = \phi^{-1}(1 - q_j), \quad q_j = \frac{jq}{2p}$, then

$$FDR = \mathbb{E}\left[\frac{\text{\#False rejections}}{\text{\#Rejections}}\right] \leq q\frac{p_0}{p}, \quad p_0 = \text{\# true null hypotheses}$$

SLOPE -> ABSLOPE

Proposition

Assume that a random variable $z = (z_1, z_2, \cdots, z_p)$ *has a SLOPE prior:*

$$p(z \mid \sigma^2; \lambda) \propto \prod_{j=1}^{p} \exp\left\{-\frac{1}{\sigma}\lambda_{r(z,j)}|z_j|\right\},$$

and then define $\beta = W^{-1}z = (\frac{z_1}{w_1}, \cdots, \frac{z_p}{w_p})$. *Finally the prior of* β *corresponds to ABSLOPE*

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

Details of Simulation step

$$\begin{aligned} X_{\text{mis}} &\sim \mathtt{p}(X_{\text{mis}} \mid \gamma, c, y, X_{\text{obs}}, \beta, \sigma, \theta, \mu, \Sigma) \\ &= \mathtt{p}(X_{\text{mis}} \mid y, X_{\text{obs}}, \beta, \sigma, \mu, \Sigma) \\ &\propto \mathtt{p}(y \mid X_{\text{obs}}, X_{\text{mis}}, \beta, \sigma)\, \mathtt{p}(X_{\text{mis}} \mid X_{\text{obs}}, \mu, \Sigma). \end{aligned}$$

Proposition

Let $\mathcal{M}$ be the set containing indexes for missing covariates and $\mathcal{O}$ for the observed ones. Assume that $p(x_{\text{obs}}, x_{\text{mis}}; \Sigma, \mu) \sim \mathcal{N}(\mu, \Sigma)$ and let $y = x\beta + \varepsilon$ where $\varepsilon \sim N(0, \sigma^2)$. For all the indexes of the missing covariates $i \in \mathcal{M}$, we denote:

$$m_i = \sum_{q=1}^{p} \mu_i s_{iq}, \quad u_i = \sum_{k \in \mathcal{O}} x_{\text{obs}}^k s_{ik}, \quad r = y - x_{\text{obs}} \beta_{\text{obs}}, \quad \tau_i = \sqrt{s_{ii} + \beta_i^2/\sigma^2},$$

with s_{ij} elements of Σ^{-1} and β_{obs} the observed elements of β.
Let $\tilde{\mu} = (\tilde{\mu}_i)_{i \in \mathcal{M}}$ be the solution of the following system of linear equations:

$$\frac{r\beta_i/\sigma^2 + m_i - u_i}{\tau_i} - \sum_{j \in \mathcal{M}, j \neq i} \frac{\beta_i \beta_j/\sigma^2 + s_{ij}}{\tau_i \tau_j} \tilde{\mu}_j = \tilde{\mu}_i \ , \quad \textit{for all } i \in \mathcal{M},$$

and let B be a matrix with elements: $B_{ij} = \begin{cases} \frac{\beta_i \beta_j/\sigma^2 + s_{ij}}{\tau_i \tau_j}, & \textit{if } i \neq j \\ 1, & \textit{if } i = j \end{cases}$ *, then for*

$z = (z_i)_{i \in \mathcal{M}}$ where $z_i = \tau_i x_{\text{mis}}^i$ we have

$$z \mid x_{\text{obs}}, y; \Sigma, \mu, \beta, \sigma^2 \sim N(\tilde{\mu}, B^{-1}) \ .$$

Stochastic Approximation step

When step-size $\eta_t = 1 \Leftrightarrow$ Stochastic EM (SEM) Estimation $\Leftrightarrow$ maximizing $\ell_{\text{comp}}\Big|_{X^t_{\text{mis}},\gamma^t,c^t}$

Update β for an example:

$$\beta^t = \arg\max_{\beta} -\frac{1}{2(\sigma^{t-1})^2}\|y - X^t\beta\|^2 - \frac{1}{\sigma^{t-1}}\sum_{j=1}^{p} w_j^t|\beta_j|\lambda_{r(W^t\beta,j)}$$

where $X^t = (X_{\text{obs}}, X^t_{\text{mis}})$.

$\Leftrightarrow$ Solution of SLOPE, given W^t, X^t_{mis} and σ^{t-1}.

$\Rightarrow$ proximal gradient.

Basic Idea of proximal gradient

SLOPE is a convex optimization problem of the form

$$\min f(\beta) = g(\beta) + h(\beta)$$

g : smooth and convex $\quad h$: convex but not smooth
At each iteration, compute a local approximation to g:

$$g(\beta^t) + \langle \nabla g(\beta^t), x - \beta^t \rangle + \frac{1}{2r} \|x - \beta^t\|^2,$$

where r is a step size. Then update β^{t+1}

$$\begin{aligned}
\beta^{t+1} &= \arg\min_x g(\beta^t) + \langle \nabla g(\beta^t), x - \beta^t \rangle + \frac{1}{2r} \|x - \beta^t\|^2 + h(x) \\
&= \arg\min_x \frac{1}{2r} \|(\beta^t - t\nabla g(\beta^t)) - x\|^2 + h(x) \\
&= \mathrm{prox}_{t,h}(\beta^t - t\nabla g(\beta^t))
\end{aligned}$$

The prox of l_1 norm is given by entry-wise soft thresholding.

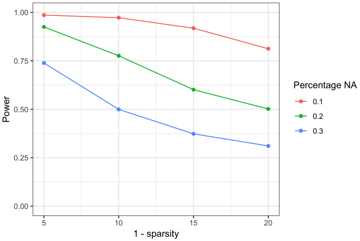
Criterion

True model		Model selection results: 1	Model selection results: 0
	1	True Positive (TP)	False Negative (FN)
	0	False Positive (FP)	True Negative (TN)

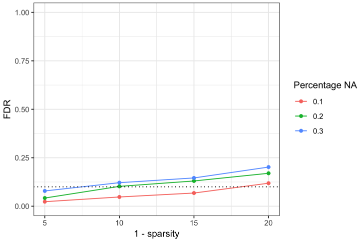
- FDR $= \frac{FN}{FN+TN}$;
- Power $= \frac{TP}{TP+FN}$;
- Relative MSE $= \frac{\|\hat{\beta}-\beta\|^2}{\|\beta\|^2}$.

Effect of missing percentage

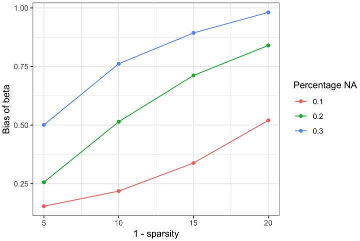
$n = p = 100$, with correlation and strong signal



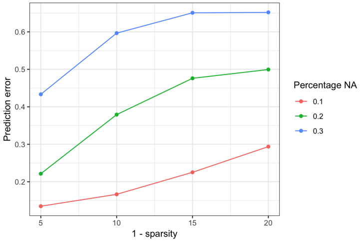
Power



FDR



Bias of β



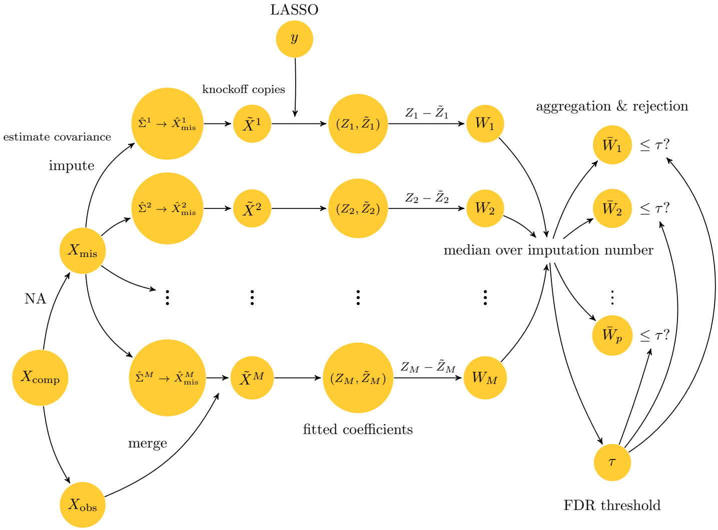
Prediction error

Appendix 3: missKnockoff

missKnockoff: handling missing values

Contributions:

- Combine multiple imputation ⇒ single knockoff on each imputed dataset values
- Suggest new aggregation rules (averaging the cases)



missKnockoff: handling missing values

Input: $X = (X_{\text{mis}}, X_{\text{obs}})$ (rows can have different pattern of missing values);

for $b = 1, 2, \cdots, B$ **do**

(Bootstrap: reflect sampling variability in covariance matrix estimate)

1. Bootstrap X with missing values.
2. On bootstrap samples, estimate the covariance (Schneider, 2001; Lounici et al., 2014):

$$\hat{\Sigma}^b = \left(\hat{\delta}^{-1} - \hat{\delta}^{-2}\right) \operatorname{diag}\left(\hat{\Sigma}_n\right) + \hat{\delta}^{-2}\hat{\Sigma}_n ,$$

with $\hat{\delta}$ the proportion of observed entries and $\hat{\Sigma}_n$ the linear shrinkage estimation on empirical covariance of initially imputed dataset by 0.

(Generate multiple knockoff and compute importance measures)

1. With $\hat{\Sigma}^b$, impute missing values $\hat{X}^b_{\text{mis}}$ from $\mathrm{p}(X_{\text{mis}} \mid X_{\text{obs}})$ and generate knockoff copies $\tilde{X}^b$ from $\mathrm{p}\left(\tilde{X} \mid X = \left(X_{\text{obs}}, \hat{X}^b_{\text{mis}}\right)\right)$.
2. On the set $(y, \hat{X}^{(b)}, \tilde{X}^{(b)})$, use LASSO to obtain fitted coefficient vectors and statistics: $Z_j^{(b)} = \left|\hat{\beta}_j^{(b)}\right|, \quad \tilde{Z}_j^{(b)} = \left|\hat{\beta}_{j+p}^{(b)}\right|$.
3. Calculate variable importance $W_j^{(b)} = Z_j^{(b)} - \tilde{Z}_j^{(b)}$, $j = 1, 2, \cdots, p$.

(Aggregation by averaging the cases)

1. Estimate the knockoff threshold: $\tau = \min \left\{ t : \frac{1}{B} \sum_{b=1}^{B} \frac{\#\left\{j : W_{bj} \leq -t\right\} + c}{\#\left\{j : W_{bj} \geq t\right\} \vee 1} \leq q \right\}$.
2. Calculate the median of $\{W_{mj}\}$ over $b = 1, 2, \cdots, B$ to obtain $\bar{W}_j$.

if $\bar{W}_j \leq \tau$ **then**

Reject j**-th variable**.

Output: Indexes for model selection $\{j : \bar{W}_j > \tau\}$.

Theoritical result

Theorem (FDR control for single missKnockoff)

missKnockoff procedure with single imputation from $\mathtt{p}(X_{mis}|X_{obs})$ controls FDR at the level q.

Sketch of proof:

If we generate values for missing covariates with:
$\hat{X}_{\text{mis}} \sim \mathtt{p}(X_{\text{mis}} \mid X_{\text{obs}})$,
$\Rightarrow (X_{\text{obs}}, \hat{X}_{\text{mis}}) \overset{d}{=} X$.
$\Rightarrow (X_{\text{obs}}, \hat{X}_{\text{mis}}, \tilde{X})_{\text{swap}(S)} \overset{d}{=} (X, \tilde{X})$.
$\Rightarrow$ Design matrix with imputed missing values satisfies the exchangeability condition.
$\Rightarrow$ it satisfies the definition of model-X knockoff.
$\Rightarrow$ FDR control.