Variable selection with missing data for linear regression

Avner Bar-Hen,

CNAM

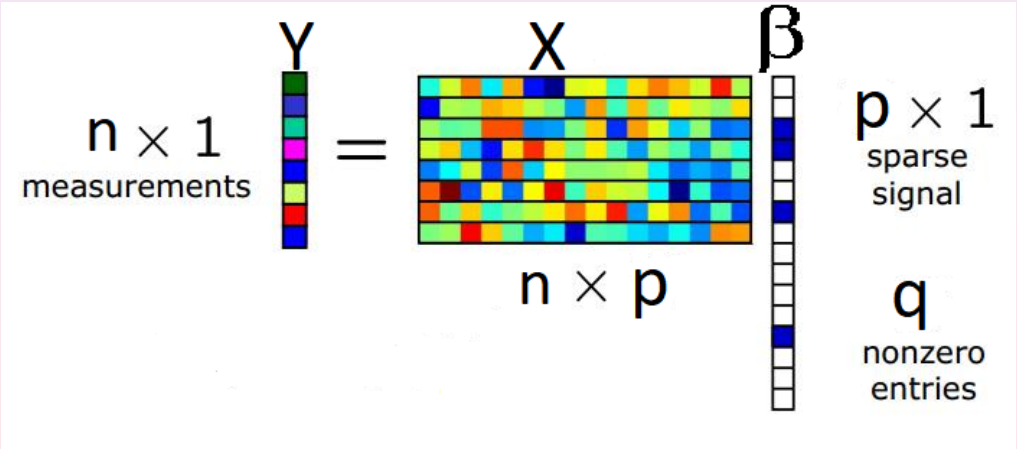
Journées GDR/SFB, ISPED-Bordeaux

5&6 octobre 2017

Sparse linear model

$$Y = X\beta + \varepsilon$$

- $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$
- n observations, p features
- β sparse (possibly $q < N << p$)
- possibly missing x_{ij} (Missing At Random)



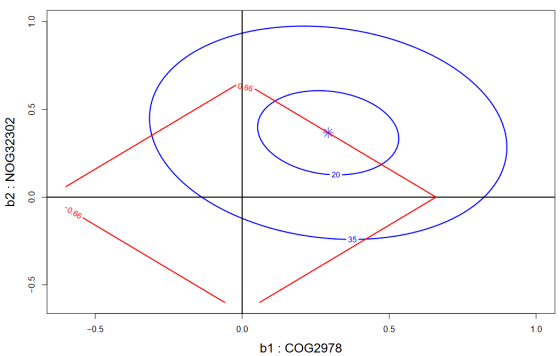
Goal: select a set of features X_j that are likely to be relevant to the response Y (without too many false positives)

Sparse regression : Lasso (1/2)

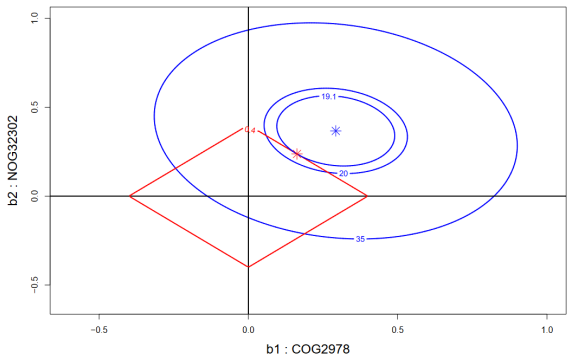
$$\beta_\lambda = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2} ||Y - X\beta||_2^2 + \lambda ||\beta||_1 \right\}$$

Asymptotically, Lasso will select the correct model (at a good λ)

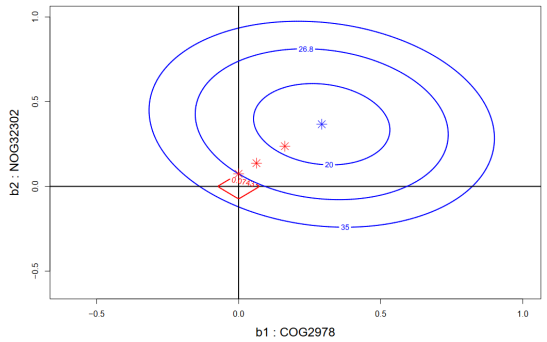
$\sum_{i=1}^{n}(y_i - x_i^1\beta_1 - x_i^2\beta_2)^2$, s.c. $|\beta_1| + |\beta_2| < 0.66$



$\sum_{i=1}^{n}(y_i - x_i^1\beta_1 - x_i^2\beta_2)^2$, s.c. $|\beta_1| + |\beta_2| < 0.4$



$\sum_{i=1}^{n}(y_i - x_i^1\beta_1 - x_i^2\beta_2)^2$, s.c. $|\beta_1| + |\beta_2| < 0.0743$



Lasso is equivalent to soft thresholding

Sparse regression : Lasso (2/2)

One way to measure performance:

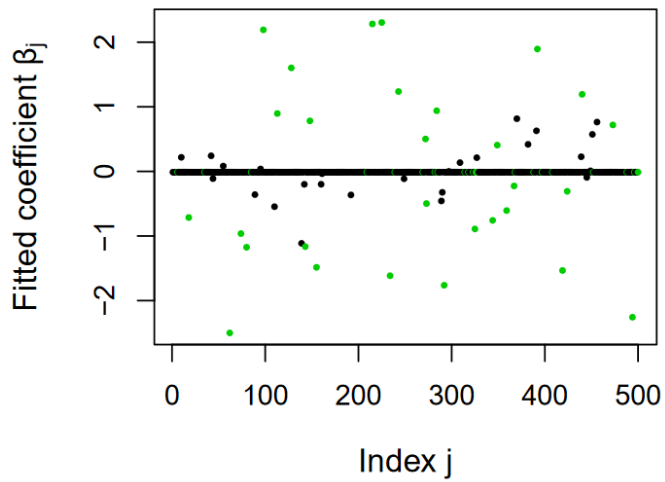
$$\underbrace{\text{FDR}}_{\text{False discovery rate}} = \mathbb{E}\left[\underbrace{\frac{\#\text{ false positives}}{\text{total }\#\text{ of features selected}}}_{\text{False discovery proportion}}\right] = \mathbb{E}\left[\frac{|S \cap \mathcal{H}_0|}{|S|}\right].$$

$S =$ set of selected features

$\mathcal{H}_0 =$ "null hypotheses" $= \{j : \beta_j^\star = 0\}$

Simulated data with $n = 1500$, $p = 500$.

Lasso fitted model for $\lambda = 1.75$:



FDP $= \frac{26}{55} = 47\%$

Sparse regression : before knockoff (1/2)

- For linear models, Miller (’84, ’02) creates "dummy" variables with entries drawn i.i.d. at random
- Forward selection procedure is applied to augmented list of variables
- Stop when selects a dummy variable for the first time
- Pseudovariables (permuted rows and variants): Wu, Boos and Stefanski (’07, ’09)

Sparse regression : before knockoff (2/2)

BUT there may be a relationship between X_1 and X_2 which makes the test statistic for a null variable larger than that of its permuted version

- Linear model with two variables X_1 and X_2

$$Y = X_1 + \varepsilon$$

- Standardized variables (mean zero and variance one) with

$$\text{Cor}(X_1, X_2) = 0.5$$

- Marginal correlations

$$\text{Cor}(Y, X_2) = \text{Cor}(X_1 + \varepsilon, X_2) = 0.5$$

- Marginal correlations with permuted features

$$\text{Cor}(Y, X_{2,\pi}) = 0$$

Statistics $X_2 y$ and $X_{2,\pi} y$ cannot be the same!

Knockoff

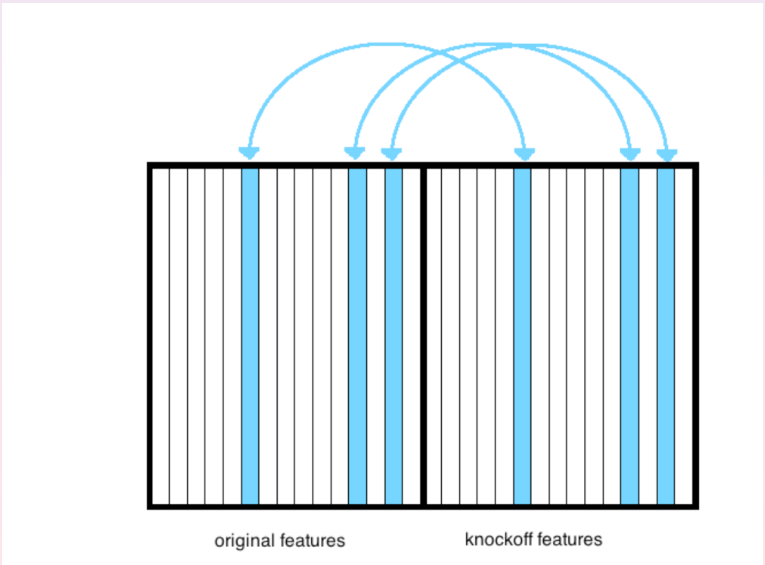
- The knockoffs replicate the correlation structure of X:

$$\tilde{X}_j{}^T \tilde{X}_j = X_j^T X_j \qquad \forall\, j,k$$

- Also preserve correlations between knockoffs and originals:

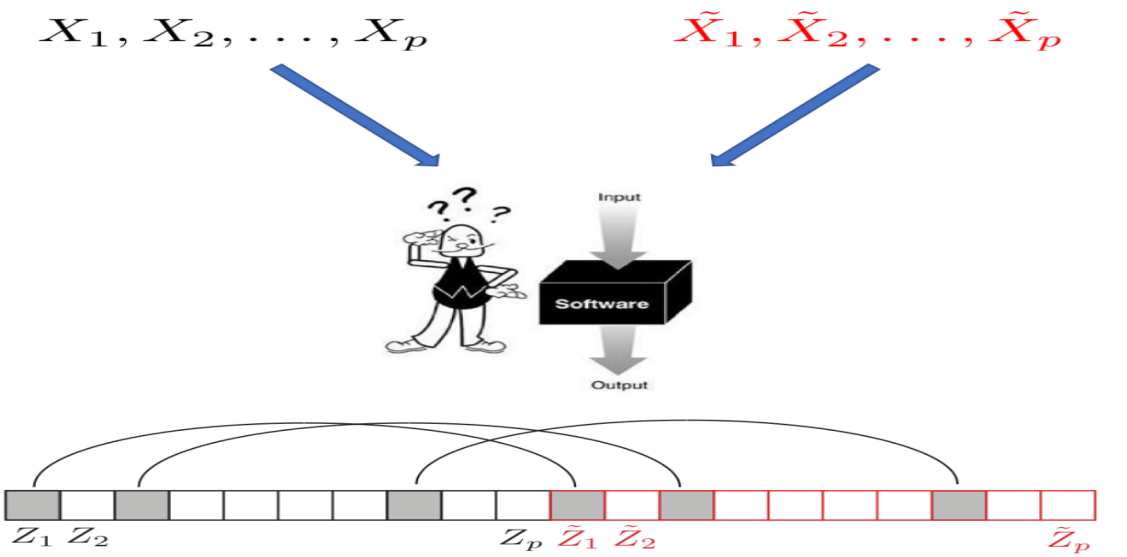
$$\tilde{X}_j{}^T X_j = X_j^T X_j \qquad \forall\, j \neq k$$

- Augmented design matrix $[X\tilde{X}]$

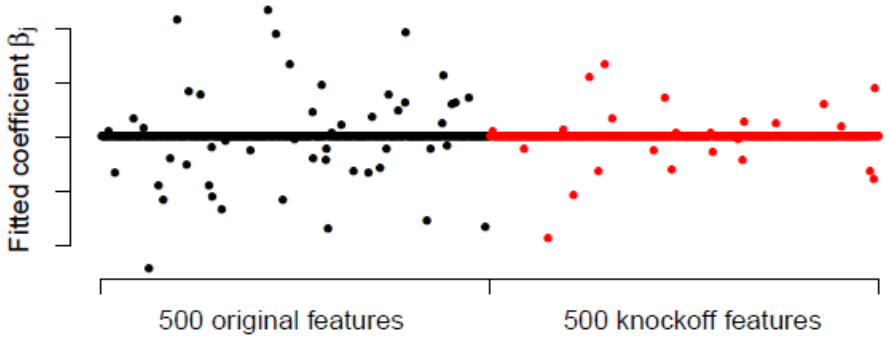


$$X_j^T Y = X_j^T X\beta + X_j^T \varepsilon \overset{\mathcal{D}}{=} \tilde{X}_j{}^T X\beta + \tilde{X}_j{}^T \varepsilon = \tilde{X}_j{}^T Y$$

Knockoff



Fitted model for $\lambda = 1.75$ on the simulated dataset:



- Lasso selects 49 original features & 24 knockoff features

$\Longrightarrow$ probably $\approx$ 24 false positives among the 49 original features

Knockoff

- Very nice properties to control FDR
- Allows to choose λ
- Extended to high dimensional setting

Not adapted to missing values

Missing data

Not all missing data are the same

$\Longrightarrow$ Missing completely at random (MCAR): Missing values are randomly distributed

$\Longrightarrow$ Missing at random (MAR): After accounting for one or more other variables, missing values are randomly distributed

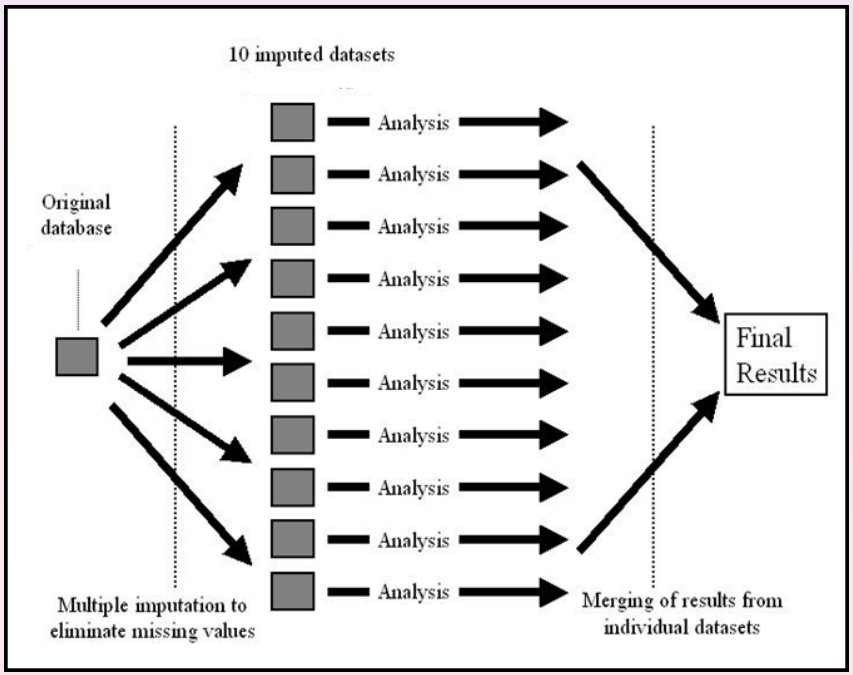
- Non-ignorable (NI): Missing values are functions of the variables themselves

Why care about missing data?

- MCAR data bias upward standard errors parameter estimates
- MAR or NI data bias BOTH parameter estimates and standard errors in unpredictable ways

Missing data

- Complete Case analysis: missing data problem is ignored, *ie.* observations with missing data are excluded
- Impute missing values and then carry out analysis with complete dataset: no subject excluded (many methods of estimating the missing values)
  - Full information maximum likelihood (FIML)
  - Multiple imputations



Missing data

How much missing data is too much?

- Hard to say
- Small amounts of missing data can sometimes greatly affect analysis if missing values are extreme
- Missing data are particularly problematic when the data are MAR or NI

Fewer variables implies less missing data

Proposed algorithm: ensemble regression

Create regression instances

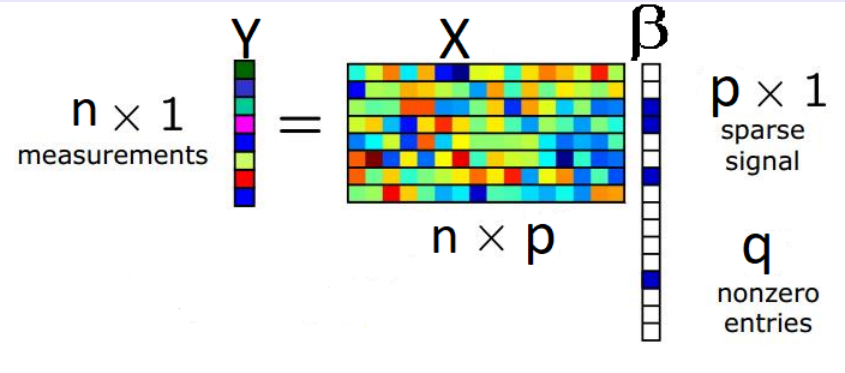
- Sample k variables among the p variables
- consider the largest subset of observations without missing observations for these k variables.
- For a given threshold, apply selection procedure (eg knockoff) to decide which of the k variables are significantly related to Y.
- Iterate the process B times ($\Rightarrow$ B regression instances).

Aggregate the regression instances

- $r_i = \dfrac{\#\text{ times the variable } X_i \text{ is selected}}{\#\text{ time } X_i \text{ is present in the instances}}$
- Conclude that X_i is significantly related to Y if $r_i > r$.

Sample k variables among the p variables

Aim: cardinal of complete observations > number of variables



$$\delta_{ij} = \begin{cases} 1 & \text{if } x_{ij} \text{ is present} \\ 0 & \text{if } x_{ij} \text{ is missing} \end{cases}$$

$$n_{j_1 \cdots j_k} = \sum_{i=1}^{n} \delta_{i j_1} \cdots \delta_{i j_k}$$

$$\mathbb{P}(n_{j_1 \cdots j_k} < k) < \exp\left\{ -\left(1 - \frac{k}{n(1-\delta)^k}\right)^2 n(1-\delta)^k / 2 \right\} \quad \text{(Chernoff inequality)}$$

$$\mathbb{E}(n_{j_1 \cdots j_k}) \begin{cases} = n(1-\delta)^k & \text{if MCAR } (\delta = \mathbb{P}(\delta_{ij} = 0)) \\ < n & \text{if MAR} \end{cases}$$

B regression instances

- Z_i number of regression instances that contains X_i.
- $Z_i \sim \mathscr{B}(B, k/p)$
- $\mathbb{P}(\min_{i=1,\ldots,p} Z_i < \tilde{B}) < p \exp\left(-\left(1 - \frac{p\tilde{B}}{Bk}\right)^2 Bk/(2p)\right)$

Linear model

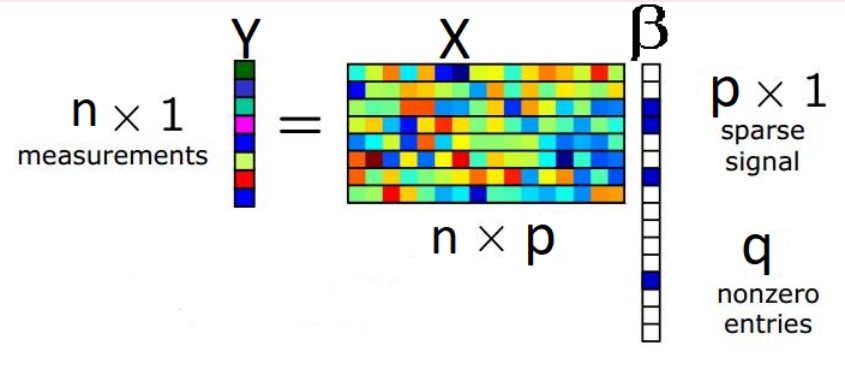
$$\delta_i = \begin{cases} 1 & \text{with probability } k/p \\ 0 & \text{otherwise} \end{cases}$$

$$\Delta = \text{diag}(\delta_1, \ldots, \delta_p)$$

$$\begin{aligned} Y &= X\beta + \varepsilon \\ &= X\Delta\beta + X(I - \Delta)\beta + \varepsilon \\ &= X\Delta\beta + \varepsilon' \end{aligned}$$

$$\hat{\beta} = (\Delta X' X \Delta)^{-1} \Delta X' Y$$

$$\begin{aligned} \mathbb{E}(\hat{\beta}) &= \Delta\beta + (\Delta X' X \Delta)^{-1} \Delta X' X (I - \Delta)\beta \\ \mathbb{V}(\hat{\beta}) &= (\Delta X' X \Delta)^{-1} \end{aligned}$$



Simulations

- $n = 200$ observations
- $p = 100$ or $p = 300$ variables
- Number of non zero β: 30
- $\beta = 1.5$, $\beta = 3$ or $\beta = 4.5$
- $\mathrm{Cor}(X_i, X_j) = 0$ or $\mathrm{Cor}(X_i, X_j) = 0.4$
- Missing values: $\delta = 0$, $\delta = 0.1$ or $\delta = 0.3$

$$\mathbb{P}(n_{j_1 \cdots j_k} < k) < \exp\left\{ -\left(1 - \frac{k}{n(1-\delta)^k}\right)^2 n(1-\delta)^k / 2 \right\}$$

- If $k = 5$ and $\delta = 0.3$: $\mathbb{P}(n_{j_1 \cdots j_k} < k) < 5.78126\ 10^{-08}$
- If $k = 5$ and $\delta = 0.1$: $\mathbb{P}(n_{j_1 \cdots j_k} < k) < 1.279093\ 10^{-25}$

Parameters				Knockoff			Our algorithm		
β	corr	miss	Nvar	OK	Minus	Extra	OK	Minus	Extra
1.5	0.0	0.0	100	30.00	0.00	2.26	30.00	0.00	0.00
3.0	0.0	0.0	100	30.00	0.00	1.34	30.00	0.00	0.00
4.5	0.0	0.0	100	30.00	0.00	0.00	30.00	0.00	0.00
1.5	0.4	0.0	100	30.00	0.00	1.98	30.00	0.00	0.00
3.0	0.4	0.0	100	30.00	0.00	1.28	30.00	0.00	0.00
4.5	0.4	0.0	100	30.00	0.00	0.00	30.00	0.00	0.00
1.5	0.0	0.1	100	30.00	0.00	1.99	30.00	0.00	0.00
3.0	0.0	0.1	100	30.00	0.00	1.45	30.00	0.00	0.00
4.5	0.0	0.1	100	30.00	0.00	0.00	30.00	0.00	0.00
1.5	0.4	0.1	100	30.00	0.00	2.15	30.00	0.00	0.00
3.0	0.4	0.1	100	30.00	0.00	1.04	30.00	0.00	0.00
4.5	0.4	0.1	100	30.00	0.00	0.00	30.00	0.00	0.00
1.5	0.0	0.3	100	30.00	0.00	1.99	28.76	1.24	0.00
3.0	0.0	0.3	100	30.00	0.00	1.41	29.58	0.42	0.00
4.5	0.0	0.3	100	30.00	0.00	0.00	29.63	0.37	0.00
1.5	0.4	0.3	100	30.00	0.00	2.35	28.13	1.87	0.00
3.0	0.4	0.3	100	30.00	0.00	0.97	29.24	0.76	0.00
4.5	0.4	0.3	100	30.00	0.00	0.00	29.47	0.53	0.00

Parameters				Knockoff			Our algorithm		
β	corr	miss	Nvar	OK	Minus	Extra	OK	Minus	Extra
1.5	0.0	0.0	300	27.76	2.24	2.37	30.00	0.00	0.22
3.0	0.0	0.0	300	28.06	1.94	1.92	30.00	0.00	0.03
4.5	0.0	0.0	300	27.77	2.23	2.30	30.00	0.00	0.02
1.5	0.4	0.0	300	27.44	2.56	2.26	30.00	0.00	0.21
3.0	0.4	0.0	300	27.65	2.35	2.62	30.00	0.00	0.03
4.5	0.4	0.0	300	27.45	2.55	2.27	30.00	0.00	0.00
1.5	0.0	0.1	300	27.90	2.10	1.75	30.00	0.00	0.01
3.0	0.0	0.1	300	27.86	2.14	1.81	30.00	0.00	0.00
4.5	0.0	0.1	300	27.88	2.12	1.77	30.00	0.00	0.00
1.5	0.4	0.1	300	27.33	2.67	2.16	30.00	0.00	0.02
3.0	0.4	0.1	300	27.62	2.38	2.42	30.00	0.00	0.00
4.5	0.4	0.1	300	27.60	2.40	1.80	30.00	0.00	0.00
1.5	0.0	0.3	300	28.08	1.92	2.33	29.25	0.75	0.00
3.0	0.0	0.3	300	27.99	2.01	1.85	29.55	0.45	0.00
4.5	0.0	0.3	300	27.75	2.25	2.24	29.75	0.25	0.00
1.5	0.4	0.3	300	27.57	2.43	2.10	29.08	0.92	0.00
3.0	0.4	0.3	300	27.60	2.40	2.25	29.66	0.34	0.00
4.5	0.4	0.3	300	27.70	2.30	2.37	29.60	0.40	0.00

Some possible extensions

- Generalized linear models
- Mixed models
- Outliers
- Estimates (and distribution) of the parameters

Thank you for your attention (and your questions)

Bibliography

Slides of Rina Barber

Slides of Emmanuel Candès

Slides of julien Chiquet