SimSel –
a Method for Variablen Selection

Silvelyn Zwanzig

Uppsala University, zwanzig@math.uu.se

LinStat2014, Linköping, August 29, 2014

Outline

- The Problem
- Our Answer: SimSel
- The Procedure
- Generalization of SimSel
- Theoretical Background
- Outlook

The Problem

Given an $n \times (p+1)$ data matrix

$$(\mathbf{Y}, \mathbf{X}_1, \ldots, \mathbf{X}_p)$$

containing observations of the response $\mathbf{y}$ and of the variables $\mathbf{x}_1, \ldots, \mathbf{x}_p$.

Wanted

A model which explains $\mathbf{y}$ and only includes relevant variables $\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_m}$:

$$E(\mathbf{y} \mid \mathbf{x}_1, \ldots, \mathbf{x}_p) = F(\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_m})$$

- BIG AIM:

$$E(\mathbf{y} \mid \mathbf{x}_1, \ldots, \mathbf{x}_p) = F(\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_m}).$$

- Essential step for finding a model: Select the **relevant** variables $\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_m}$ from $\mathbf{x}_1, \ldots, \mathbf{x}_p$.
- First step: Take one variable $\mathbf{x}_1$ and decide: Is this variable $\mathbf{x}_1$ relevant (important)?

A variable $\mathbf{x}_1$ is **unimportant** iff for all Δ

$$E(\mathbf{y} \mid \mathbf{x}_1, \ldots, \mathbf{x}_p) = E(\mathbf{y} \mid \mathbf{x}_1 + \Delta, \ldots, \mathbf{x}_p).$$

Pertubation Methods

Observed data set:

$$(\mathbf{Y}, \mathbf{X}_1, \ldots, \mathbf{X}_p)$$

- Disturb the response by random deviations:

$$(\mathbf{Y} + \delta, \mathbf{X}_1, \ldots, \mathbf{X}_p)$$

- Disturb variables by random errors:

$$\left(\mathbf{Y}, \mathbf{X}_1 + \sqrt{\lambda}\varepsilon, \ldots, \mathbf{X}_p\right), \ \lambda \in \{\lambda_1, \ldots, \lambda_K\}$$

- Extend the data by a pseudo variable **Z**, generated independently of $(\mathbf{Y}, \mathbf{X}_1, \ldots, \mathbf{X}_p)$:

$$(\mathbf{Y}, \mathbf{X}_1, \ldots, \mathbf{X}_p, \mathbf{Z})$$

Pertubation of Variables

General in iterature:

- STABILIZATION: "well known" method, $\mathbf{X}^T\mathbf{X}$ has no inverse, but $(\mathbf{X}+\delta\mathbf{I})^T(\mathbf{X}+\delta\mathbf{I})$ has.
- SIMEX
- PERTURBATION: huge literature in data engineering, data mining
  - additive data perturbation, each data element is randomized by adding random noise
  - multiplicative data perturbation, multiplicative noise

Aim: keep the statistical properties under preserving the privacy

Add Variables

Dissertation of Wu (2004), Wu et al, JASA (2007),102,235-243

Dissertation of Qi Tang (2010), Dec 2010 (Bayesian approach)

Add a set of independent pseudo variables to the data set.

"Intuitively, a good selection criterion should not include too many of the pseudo variables. If a procedure never selects pseudo variables, then the selection is too "ruthless" ".

Our Method SimSel

SimSel stands for simulation and selection.

no extrapolation step

no splitting of the data set

First Step: Study each variable x_i seperately.

published in

M. Eklund and S. Zwanzig (2012). SimSel - a new simulation method for variable selection, Journal of Statistical Computation & Simulation, 82,515-527.

Martin Eklund Department of Medical Epidemiology and Biostatistics, Karolinska Institute, Stockholm.

Embedding

Let $\mathbf{x}_1$ the feature of interest. We embed the original data set

$$(\mathbf{Y}, \mathbf{X}_1, \ldots, \mathbf{X}_p)$$

in

$$(\mathbf{Y}, \mathbf{X}_1 + \sqrt{\lambda}\varepsilon^*, \ldots, \mathbf{X_p}, \mathbf{Z}),\ \lambda \in \{\lambda_1, \ldots, \lambda_K\},$$

where

- $\mathbf{Z} = (z_1, \ldots, z_n)^T$ is an independent **pseudo variable**, independently generated of $\mathbf{Y}, \mathbf{X}_1, \ldots, \mathbf{X}_p$
- **pseudo errors** $\varepsilon^* = (\varepsilon_1^*, \ldots, \varepsilon_n^*)^T$, ε_i^* are i.i.d. P^*, with $E\varepsilon_i^* = 0$, $Var(\varepsilon_i^*) = 1$, $E(\varepsilon_i^*)^4 = \mu$.

The Idea

$$(\mathbf{Y}, \mathbf{X}_1 + \sqrt{\lambda}\varepsilon^*, \ldots, \mathbf{X}_p, \mathbf{Z})$$

- The pseudo variable **Z** serves as an untreated control group in a biological experiment.
- The influence of the pseudo errors is controlled by stepwise increasing λ.

MAIN IDEA (due to Martin!)
If λ "does not matter" — then $\mathbf{x_1}$ is unimportant.

"does not matter"

Consider the data $(\mathbf{Y}, \mathbf{X}_1)$. Compare!

Model fit for the extended data: $(\mathbf{Y}, \mathbf{X}_1 + \sqrt{\lambda}\varepsilon^*, \mathbf{Z})$

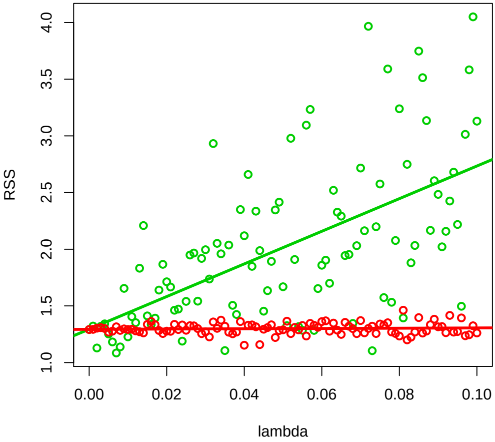
$$RSS_1(\lambda) = \min_{\beta_1, \beta_2} \left\| \mathbf{Y} - \beta_1 \left(\mathbf{X}_1 + \sqrt{\lambda}\varepsilon^* \right) - \beta_2 \mathbf{Z} \right\|^2.$$

Model fit for the extended data: $(\mathbf{Y}, \mathbf{X}_1, \mathbf{Z} + \sqrt{\lambda}\varepsilon^*)$

$$RSS_2(\lambda) = \min_{\beta_1, \beta_2} \left\| \mathbf{Y} - \beta_1 \mathbf{X}_1 - \beta_2 \left(\mathbf{Z} + \sqrt{\lambda}\varepsilon^* \right) \right\|^2.$$

Intuitively "does not matter" respects to a constant trend of $RSS(.)$.

Regression Step



- It looks like simple heteroscedastic linear regression.
- "does not matter" — the slope of RSS(.) is zero.

Testing Step

- Determine the distribution of the F-statistics by simulation.
- We repeat the regression and generate two samples of F-statistics of arbitrary size M.

  One sample is related to the variable under control $\mathbf{x}_i$

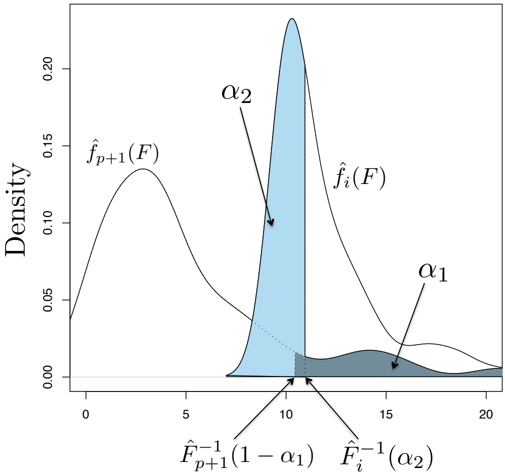
  $$F_{i,1}, \ldots, F_{i,M}.$$

  The other sample is related to the pseudo variable $\mathbf{z} = \mathbf{x}_{p+1}$ ("untreated control")

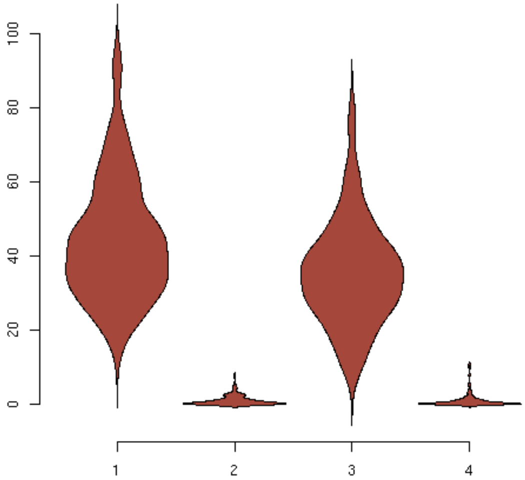
  $$F_{p+1,1}, \ldots, F_{p+1,M}.$$

- Calculate kernel estimates $\widehat{f_i}$, $\widehat{f_{p+1}}$.
- Compare $\widehat{f_i}$ and $\widehat{f_{p+1}}$.

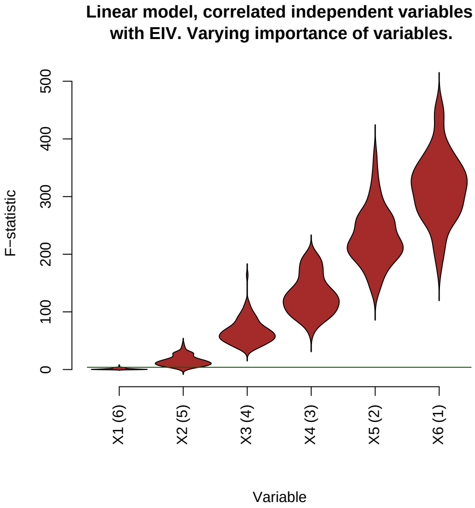
Significance - small overlapping



Graphic output - violin plots



Graphic output - violin plots



The SimSel - Algorithm

(1) Choose $0 \leq \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_K, M, \alpha_1, \alpha_2$ for $(m$ in $1:M)$ {

(2) Generate a non relevant pseudo variable $\mathbf{z} = \mathbf{x}_{p+1}$ for $(i$ in $1:p+1)$ {
for $(k$ in $1:K)$ {

(3),(4) generate and add pseudo errors to $\mathbf{X}_i$

(5) Compute $RSS_i(\lambda_k)$ }

(6) *Regression step*. Calculate $F_{i,m}$ }

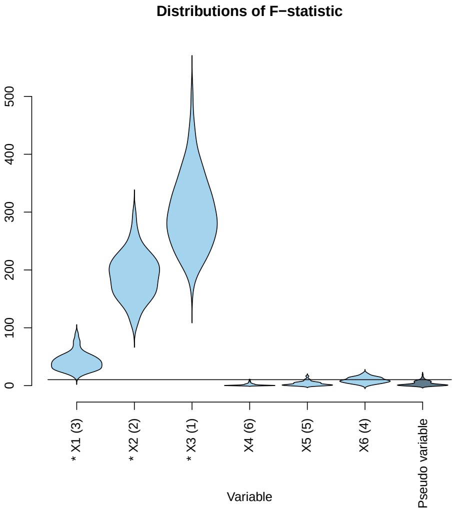
}

(7) *Plotting step*, violin plot of all $\widehat{f_i}$,

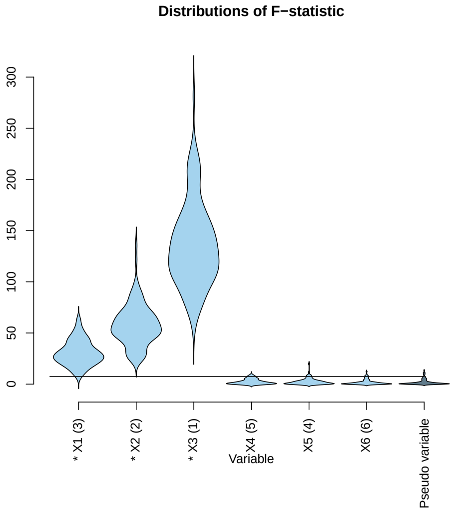
(8) *Ranking step*, according to the median of $\widehat{f_i}$

(9) *Testing step*

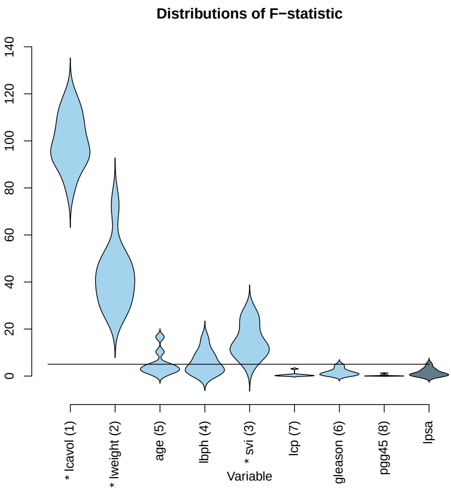
Simulations linear model



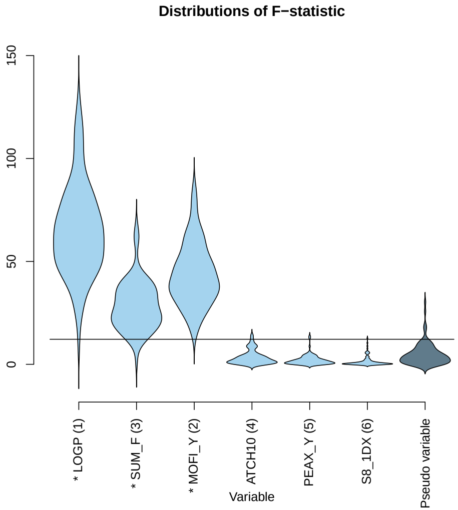
Simulations nonlinear model with errors in variables



Prostate Data Set



Selwood



Distributions of F−statistic
150
100
50
0
* LOGP (1)
* SUM_F (3)
* MOFI_Y (2)
ATCH10 (4)
PEAX_Y (5)
S8_1DX (6)
Pseudo variable
Variable

Theoretical Background

Under the assumption that $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$ exists, it holds

$$\frac{1}{n}RSS(\lambda) = \frac{1}{n}RSS + \frac{\lambda}{1 + h_{11}\lambda}\left(\widehat{\beta}_1\right)^2 + o_{P^*}(1)$$

where h_{11} is the $(1,1)-$element of $\left(\frac{1}{n}\mathbf{X}^T\mathbf{X}\right)^{-1}$and $\widehat{\beta}_1$ is the first component of the LSE estimator $\widehat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$.

Thus in case $\widehat{\beta}_1 = 0$, it holds $\frac{1}{n}RSS(\lambda) \approx const$.

Idea of the proof

It holds

$$\frac{1}{n} RSS(\lambda) = \frac{1}{n} \mathbf{Y}^T \mathbf{Y} - \frac{1}{n} \mathbf{Y}^T P(\lambda) \mathbf{Y} \tag{1}$$

with

$$P(\lambda) = \mathbf{X}(\lambda) \left(\mathbf{X}(\lambda)^T \mathbf{X}(\lambda) \right)^{-1} \mathbf{X}(\lambda)^T. \tag{2}$$

Idea of the proof cont.

$$\frac{1}{n}\mathbf{X}(\lambda)^T\mathbf{Y} = \left(\frac{1}{n}\mathbf{X} + \frac{1}{n}\sqrt{\lambda}\Delta\right)^T \mathbf{Y},$$

where Δ is the $(n \times p)-$ matrix

$$\Delta = \begin{pmatrix} \varepsilon_1^* & 0 & \cdots & 0 \\ \varepsilon_2^* & 0 & \cdots & 0 \\ \vdots & \vdots & 0 & \vdots \\ \varepsilon_{n-1}^* & 0 & \cdots & \vdots \\ \varepsilon_n^* & 0 & \cdots & 0 \end{pmatrix}.$$

and by the LLN applied to the pseudo errors only

$$\frac{1}{n}\mathbf{X}(\lambda)^T\mathbf{Y} = \frac{1}{n}\mathbf{X}^T\mathbf{Y} + o_{P^*}(1). \tag{3}$$

Idea of the proof cont.

Consider now $\mathbf{X}(\lambda)^T\mathbf{X}(\lambda)$:

$$= \frac{1}{n}\left(\mathbf{X}+\sqrt{\lambda}\Delta\right)^T\left(\mathbf{X}+\sqrt{\lambda}\Delta\right) \tag{4}$$

$$= \frac{1}{n}\mathbf{X}^T\mathbf{X}+\frac{1}{n}\sqrt{\lambda}\mathbf{X}^T\Delta+\frac{1}{n}\sqrt{\lambda}\Delta^T\mathbf{X}+\frac{1}{n}\lambda\Delta^T\Delta \tag{5}$$

Hence

$$\left(\frac{1}{n}\mathbf{X}(\lambda)^T\mathbf{X}(\lambda)\right)^{-1} = \left(\frac{1}{n}\mathbf{X}^T\mathbf{X}+\lambda\mathbf{e}_1\mathbf{e}_1^T\right)^{-1} + o_{P^*}(1).$$

Remarks

- We use in the procedure

$$\frac{\lambda}{1 + h_{11}\lambda} \approx \lambda.$$

- We have not required any model assumption for this result; only least squares fits are compared.
- In linear errors-in-variable models the naive LSE is inconsistent. But if β_1 is zero, then the naive LSE also converges to zero. This gives the motivation for successful application of SimSel to errors-in-variables models.

Approximative Model

Compare the fit of an approximative model.
We have chosen a quadratic model.
We organize the quadratic approximation such that the first terms include $\mathbf{x}_1$:

$$\begin{aligned} & H(\mathbf{x}_1, \ldots, \mathbf{x}_{p+1}) = \mathbf{H}\beta \\ = \quad & \beta_1\, \mathbf{x}_1 + \beta_2\, (\mathbf{x}_1 \mathbf{x}_2) + \ldots + \beta_{p+2}\, (\mathbf{x}_1 \mathbf{x}_{p+1}) + \beta_{p+3}\, \mathbf{x}_1^2 \\ & + \beta_{p+4}\, \mathbf{x}_2 + \ldots + \beta_m\, \mathbf{x}_{p+1}^2 \end{aligned}$$

$\beta \in \mathbb{R}^m$, where $m = \frac{1}{2}((p+1)^2 + 3(p+1))$

Theoretical Result

Under the assumption, that $(\frac{1}{n}\mathbf{H}^T\mathbf{H})^{-1}$ exists it holds

$$\frac{1}{n}RSS(\lambda) = \frac{1}{n}RSS + \lambda\widehat{\beta}^T\mathbf{D}(\lambda)\widehat{\beta} + o_{P^*}(1)$$

where $\widehat{\beta}^T\mathbf{D}(\lambda)\widehat{\beta}$ includes $\widehat{\beta}_1, \ldots, \widehat{\beta}_{p+3}$ only. $\mathbf{D}(\lambda) = \ldots$ is positive definite .

Generalization of SimSel

Wanted: to study the dependence structure between variables.

- Disturb q variables simultaneously.
- Add k simulated control variables $\mathbf{z_1}, \ldots, \mathbf{z_k}$ to the data.
- Allow $rank(\mathbf{X}) = r < p$.
- Use the ridge criterion instead of least squares.

Remind Ridge

Here we do not require that **X** has full rank.

$$\min_{\beta}(\|\mathbf{Y}-\mathbf{X}\beta\|^2 + k\|\beta\|^2) = \left\|\mathbf{Y}-\mathbf{X}\widehat{\beta}_{ridge}\right\|^2 + k\left\|\widehat{\beta}_{ridge}\right\|^2$$

delivers an unique parameter estimator

$$\widehat{\beta}_{ridge} = \left(\mathbf{X}^T\mathbf{X}+k\mathbf{I}_p\right)^{-1}\mathbf{X}^T\mathbf{Y}. \quad (6)$$

$$RIDGE(k) = \left\|\mathbf{Y}-\widehat{\mathbf{Y}}_{ridge}\right\|^2 + k\left\|\widehat{\beta}_{ridge}\right\|^2$$

$$RIDGE(k) = \mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}+k\mathbf{I}_p\right)^{-1}\mathbf{X}^T\mathbf{Y}.$$

No projection!

Approximation of the Criterion

Disturb the variables $X_{j_1}, ..., X_{j_q}$ simultaneously.

$$X_{j_1}(\lambda) = X_{j_l} + \sqrt{\lambda}\varepsilon^*_{j_l},\ l = 1, ..., q$$

Thus

$$\mathbf{X}(\lambda) = \mathbf{X} + \sqrt{\lambda}\mathbf{E}^{(*)}$$

.

$$\frac{1}{n}Ridge(\beta, \lambda, k) = \frac{1}{n}\|\mathbf{Y} - \mathbf{X}(\lambda)\beta\|^2 + k\|\beta\|^2$$

$$= \frac{1}{n}\|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda\beta^T\Delta\beta + k\|\beta\|^2 + o_{P^*}(1)$$

where $\Delta = diag(0, \ldots, 1, \ldots, 0, 1, 0, \ldots)$

with $\Delta_{j_l j_l} = 1$ for $l = 1, \ldots, q$ and zero otherwise.

Ridge Type Estimator

$$\min_{\beta \in \mathbb{R}^p} \left(\|\mathbf{Y} - \mathbf{X}\beta\|^2 + \beta^T B^T B \beta \right)$$

defined a least squares estimator in the "big" model

$$\begin{pmatrix} \mathbf{Y} \\ 0 \end{pmatrix} = \begin{pmatrix} \mathbf{X} \\ B \end{pmatrix} \beta + \begin{pmatrix} \varepsilon \\ 0 \end{pmatrix}$$

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$$

$$\min_{\beta \in \mathbb{R}^p} \|\mathbf{y} - \mathbf{X}\beta\|^2 = \|\mathbf{y} - \mathbf{P}\mathbf{y}\|^2$$

where

$$\mathbf{P} : \mathbb{R}^{n+p} \to \mathscr{L}(\mathbf{X}), \text{ projection}$$

OBS: The "big" model is misspecified!!!

$$\beta \neq 0, \ E\mathbf{y} \notin \mathscr{L}(\mathbf{X})$$

Bias Term

Set

$$B = A^T(X^TX) + A_2^T, \;\; A_2^T(X^TX) = 0$$

Then for $Ey = \mu_0, \;\; \mu_0 \in \mathscr{L}(X)$

$$BIAS = \mu_0^T XA(A^T(X^TX)A + I_p)^{-1} A^T X^T \mu_0$$

and for nonlinear relation, $\mu_0 \notin \mathscr{L}(X)$

$$BIAS = const - \mu_0^T XA(A^T(X^TX)A + I_p)^{-1} A^T X^T \mu_0$$

Note, it is not required that B or X have full rank!

The effect of the perturbation is included in A.

Special Cases

- orthogonal design and all variables are disturbed:

$$X^T X = I_p, \; B = \sqrt{\lambda} I_p, \; \mu_0 = X\beta_0$$

$$BIAS = \frac{\lambda}{1+\lambda} \|\beta_0\|^2$$

- singular design, only nonrelevant variables are disturbed:

$$B(X^T X) = 0 \text{ alternatively } B = A_2$$

$$BIAS = 0$$

Special Case

- Estimation procedure: $k = 0$, $\lambda_{\min}(X^T X) = \lambda_0 > 0$
- Perturbation: $B = \sqrt{\lambda}\, diag(1, \ldots, 1, 0, 0, \ldots 0)$ q variables simultaneously
- Model assumption: $Ey = (X_{i_1}, \ldots, X_{i_m})\beta_0$ all components of β_0 are not zero.
- Then
$$\frac{\lambda}{1 + \lambda\lambda_0^{-1}} \sum_{j \in J} \beta_{0,j}^2 \le Bias(\lambda) \le \lambda \sum_{j \in J} \beta_{0,j}^2,$$
where J set of variables which are in the model and which are disturbed.

Variance Term

$$tr(Cov(\mathbf{Y})(I - \mathbf{P})) = n - tr\left(\begin{pmatrix} I_n & 0 \\ 0 & 0 \end{pmatrix} \mathbf{P}\right)$$

$$\mathbf{P} : \mathbb{R}^{n+p} \to \mathscr{L}(\begin{pmatrix} \mathbf{X} \\ B \end{pmatrix}) \text{ projection}$$

stabilization effect

when $dim(\mathscr{L}(\begin{pmatrix} \mathbf{X} \\ B \end{pmatrix})) > dim(\mathscr{L}(\mathbf{X}))$

Lasso

Study

$$\frac{1}{n} Lasso(\beta, \lambda, k) = \frac{1}{n} \|\mathbf{Y} - \mathbf{X}(\lambda)\beta\|^2 + k|\beta|$$
$$= \frac{1}{n} \|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda \beta^T \Delta \beta + k|\beta| + o_{P^*}(1).$$

It is related to the elastic net procedure.

Simultaneous SimSel - Outlook

- Wanted: to study the dependence structure between variables.
- Need to study the behavior of bias term for singular design matrices.
- Algorithm for systematic simultaneously disturbtion.

Tack för uppmärksamheten!