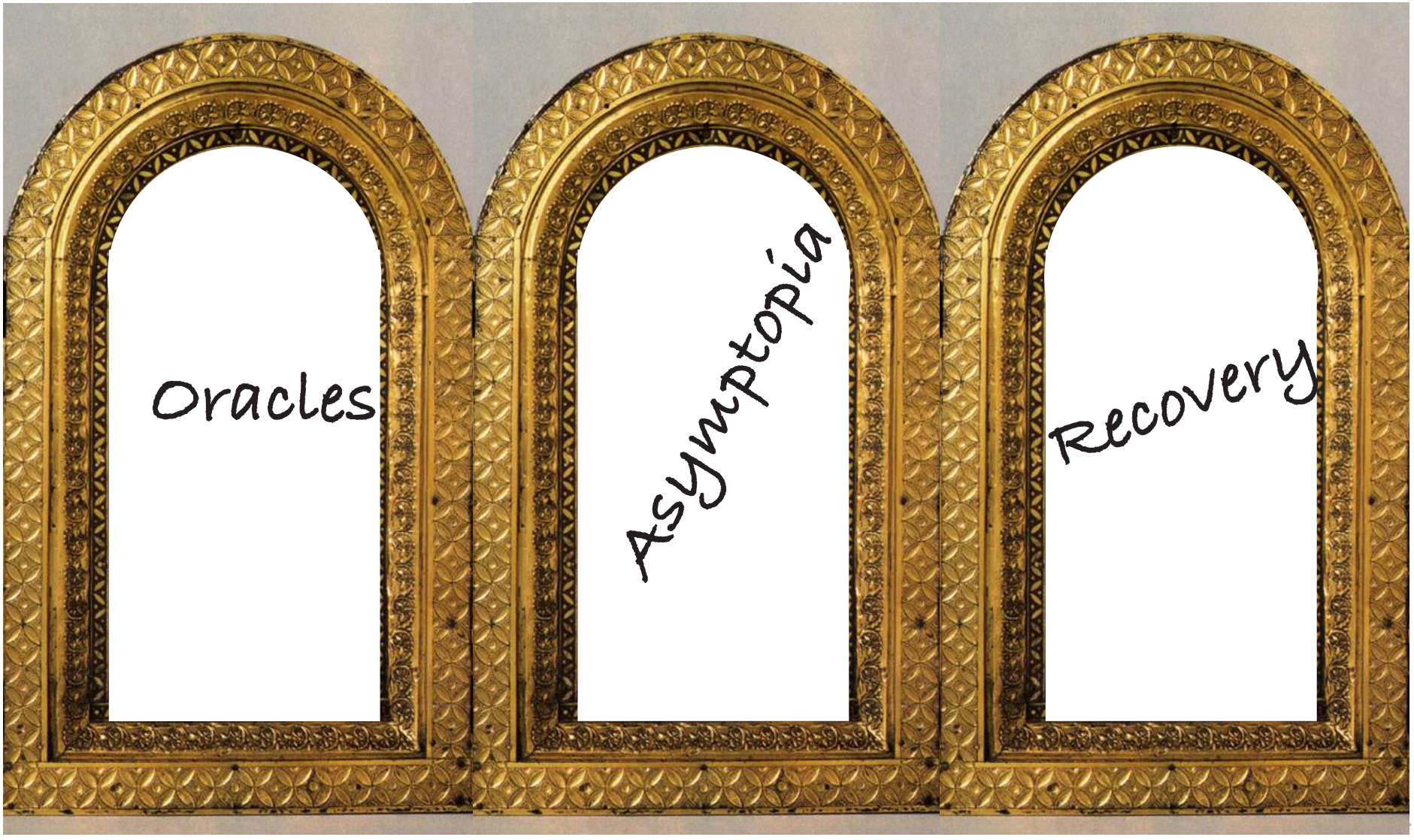
High-dimensional statistics: a triptych

Sara van de Geer



Panel I: Oracle inequalities

Panel II: Asymptotic normality and CRLB's

Panel III : Lower bounds for restricted eigenvalues

Panel I: Oracle inequalities

Panel II: Asymptotic normality and CRLB's

Panel III : Lower bounds for restricted eigenvalues

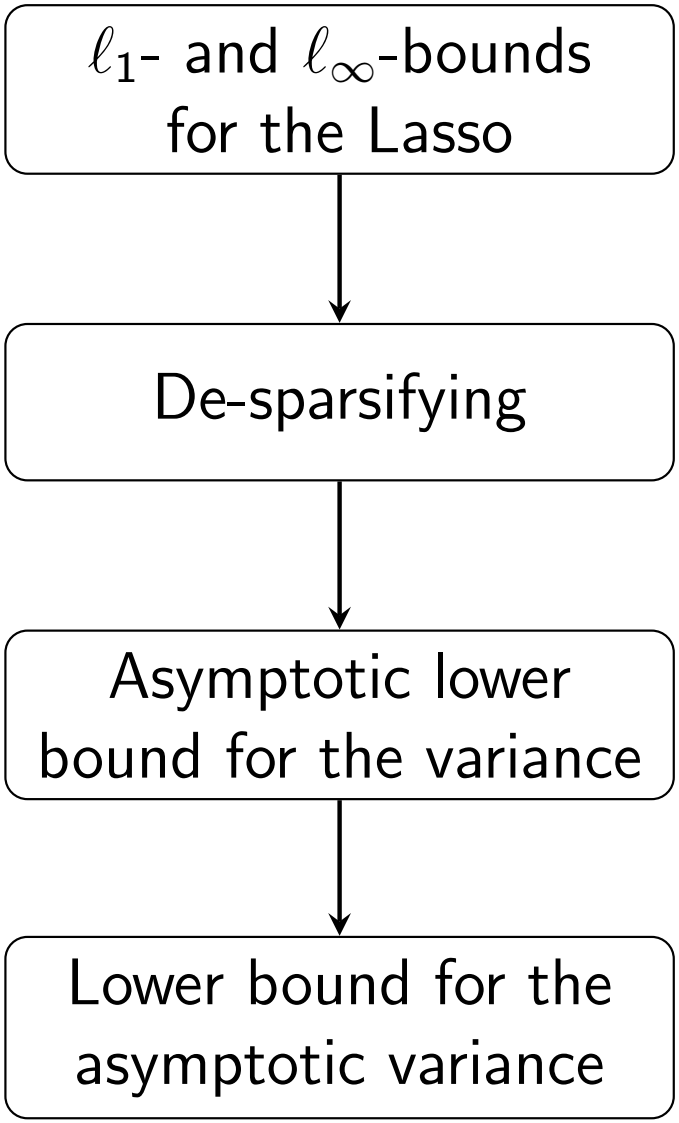
High-dimensional statistics: a triptych

Sara van de Geer

July 13, 2016

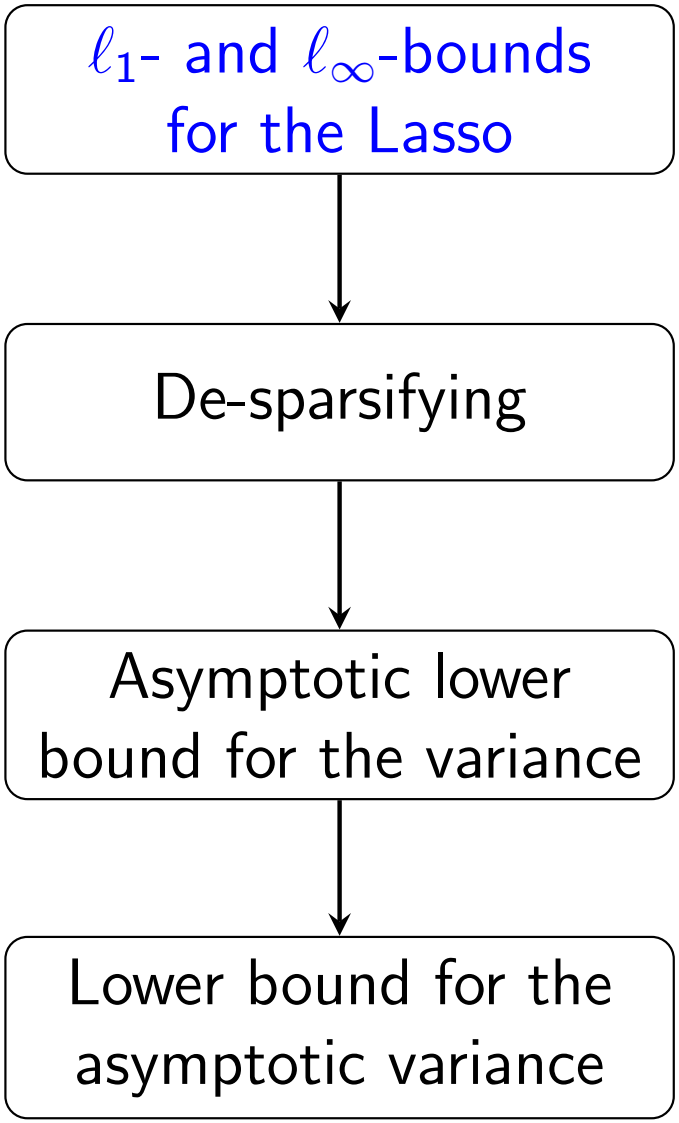
Panel II:
Asymptotic normality
and Cramèr-Rao lower bounds

Joint work with:
Andreas Elsener, Alan Muro, Jana Janková, Benjamin Stucky



Model:

Linear model
Graphical model
Other



Model:

Linear model
Graphical model
Other

Narrator

we will first study confidence intervals
for the high-dimensional linear model

for that we need bounds for the ℓ_1-estimation error
of the Lasso

The Lasso

$Y \in \mathbb{R}^n$ response
$X = (X_1, \ldots, X_p) \in \mathbb{R}^{n \times p}$ co-variables
$\boxed{p > n}$

Linear model:

$$Y = X\beta^0 + \sigma_0 \epsilon, \ \epsilon \sim \mathcal{N}_n(0, I)$$

Lasso

$\ell_1-\text{norm} \sum_{j=1}^p |\beta_j|$
$\downarrow$

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|_2^2/n + 2 \underbrace{\lambda\sigma_0}_{\text{tuning parameter}} \|\beta\|_1 \right\}$$

Notation

- $S \subset \{1, \ldots, p\}$, $s := |S|$
- $\|\beta_S\|_1 = \sum_{j \in S} |\beta_j|$
- $\|\beta_{-S}\|_1 = \sum_{j \notin S} |\beta_j|$

$$\beta = \begin{pmatrix} * \\ \vdots \\ * \\ * \\ * \\ \vdots \\ * \end{pmatrix} \qquad \beta_S := \begin{pmatrix} * & \leftarrow \in S \\ \vdots & \vdots \\ 0 & \leftarrow \notin S \\ 0 & \leftarrow \notin S \\ * & \leftarrow \in S \\ \vdots & \vdots \\ 0 & \leftarrow \notin S \end{pmatrix}$$

Definition

The compatibility constant is

$$\hat{\phi}^2(L, S) := \min\{s\|X\beta_S - X\beta_{-S}\|_2^2/n : \ \|\beta_S\|_1 = 1, \ \|\beta_{-S}\|_1 \leq L\}$$

- $L \geq 1$ is a "stretching factor"

Notation

- $S \subset \{1, \ldots, p\}$, $s := |S|$
- $\|\beta_S\|_1 = \sum_{j \in S} |\beta_j|$
- $\|\beta_{-S}\|_1 = \sum_{j \notin S} |\beta_j|$

$$\beta = \begin{pmatrix} * \\ \vdots \\ * \\ * \\ * \\ \vdots \\ * \end{pmatrix} \quad \beta_S := \begin{pmatrix} * \\ \vdots \\ 0 \\ 0 \\ * \\ \vdots \\ 0 \end{pmatrix} \quad \beta_{-S} := \begin{pmatrix} 0 \\ \vdots \\ * \\ * \\ 0 \\ \vdots \\ * \end{pmatrix}$$

Definition

The compatibility constant is

$$\hat{\phi}^2(L, S) := \min\{s\|X\beta_S - X\beta_{-S}\|_2^2/n : \ \|\beta_S\|_1 = 1, \ \|\beta_{-S}\|_1 \leq L\}$$

- $L \geq 1$ is a "stretching factor"

Example: $S := \{1\}$

$$\hat{\phi}^2(L, \{1\}) = \min\left\{ \|X_1 - X_{-1}\beta_{-1}\|_2^2/n : \ \|\beta_{-1}\|_1 \leq L \right\}$$

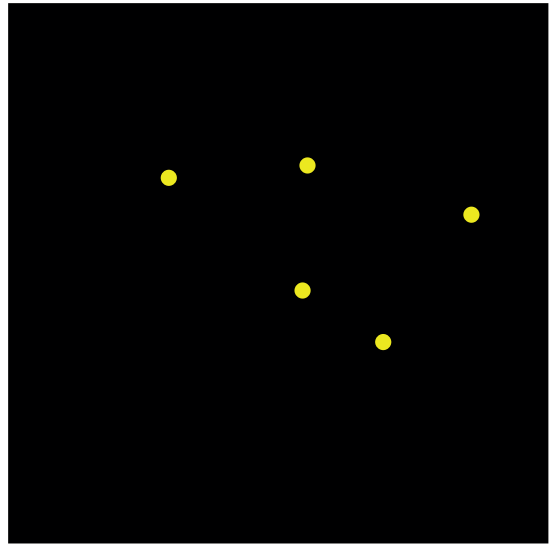
Wald lecture 3 (Friday):

Discussion of bounds for $\hat{\phi}^2(L, S)$

Strong sparsity:

$$\sum_{j=1}^{p} 1\{\beta_j^0 \neq 0\} \text{ is "small"}$$

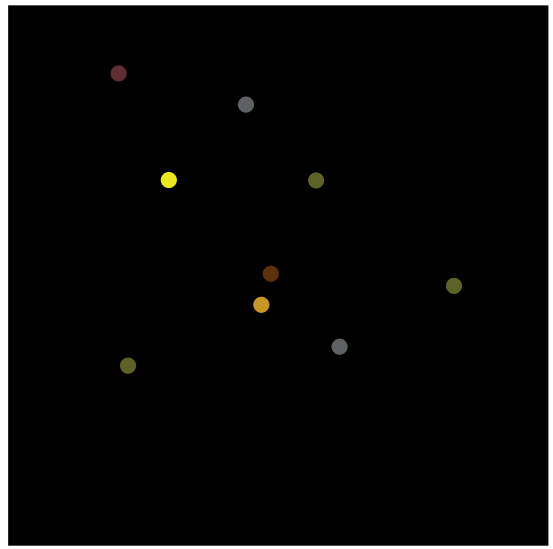
$$(r = 0)$$



Weak sparsity:

For some $0 < r < 1$,

$$\sum_{j=1}^{p} |\beta_j^0|^r \text{ is "small"}$$



ℓ_1-estimation error of the Lasso

Lemma (See Tuesday's lecture)

Let $\sim$"noise level" $\downarrow$

- $\lambda_0 := \sqrt{2\log(2p/\alpha)}$ *and* $\lambda > \lambda_0$
- $L := 3 \times \frac{\lambda+\lambda_0}{\lambda-\lambda_0}$
- $S_* := \{j : \ |\beta_j^0|/\sigma_0 > L(\lambda+\lambda_0)\}$

and

$$
\begin{aligned}
s_0 &:= |S_0|, \ S_0 := \{j : \ \beta_j^0 \neq 0\}, \\
s_r &:= \sum_{j=1}^{p} |\beta_j^0|^r/\sigma_0^r, \ 0 < r < 1,
\end{aligned}
$$

With probability at least $1-\alpha$

$$
\boxed{\|\hat{\beta} - \beta^0\|_1} \leq C \times \boxed{(\lambda+\lambda_0)^{1-r} s_r} \ \frac{\sigma_0}{\hat{\phi}^2(L,S_*)}
$$

□

Asymptopia

$$\boxed{(\lambda + \lambda_0)^{1-r} s_r} = \mathcal{O}\left(\sqrt{\frac{\log p}{n}}\right)^{1-r} s_r.$$

So modulo compatibility

$$s_r = o\left(\sqrt{\frac{n}{\log p}}\right)^{1-r} \Rightarrow \|\hat{\beta} - \beta_0\|_1 = o_{\mathbb{P}}(1).$$

Special case: strong sparsity ($r = 0$)

Modulo compatibility

$$s_0 = o\left(\sqrt{\frac{n}{\log p}}\right) \Rightarrow \|\hat{\beta} - \beta_0\|_1 = o_{\mathbb{P}}(1).$$

recall Lasso

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|_2^2/n + 2 \underbrace{\lambda\sigma_0}_{\text{tuning parameter}} \|\beta\|_1 \right\}$$

the tuning parameter of the Lasso depends on (an estimate) of the unknown variance σ_0^2

$\rightsquigarrow$ square-root Lasso

for the construction of confidence intervals
we will perform *many* Lasso's

the square-root Lasso can do this
using only *one* tuning parameter

The square-root Lasso

$\sqrt{\text{Lasso}}$

$$\hat{\beta} := \arg\min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|_2/\sqrt{n} + \lambda \|\beta\|_1 \right\}$$

[Belloni et al. 2011]

ℓ_1-error of the $\sqrt{\text{Lasso}}$

Lemma

- Let $\lambda_0 := \sqrt{\frac{2\log(2p/\alpha)}{n-1}}$, $\quad \bar{\alpha} := \mathbb{P}\left(\left|\|\epsilon\|_2^2/n - \sigma_0^2\right| > \eta\right)$

  $n-1$ instead of n ↑ ; $\|\epsilon\|_2^2 = \sum_{i=1}^n \epsilon_i^2$ ↑

- Assume

$$\frac{1}{1-\eta}\boxed{\lambda_0} < \boxed{\sigma_0/\|\beta^0\|_1}\,\eta^2.$$

- Take

$$\frac{1}{1-\eta}\boxed{\lambda_0} < \lambda < \boxed{\sigma_0/\|\beta^0\|_1}\,\eta^2.$$

Then with probability at least $1-\alpha-\bar{\alpha}$

$$\boxed{\|\hat{\beta}-\beta^0\|_1} \leq C \times \boxed{(\lambda(1+\eta)+\lambda_0)^{1-r} s_r}\,\frac{\sigma_0}{\hat{\phi}^2(L,S_*)}$$

Asymptopia

If $\|\beta^0\|_1 = o(\sqrt{\frac{n}{\log p}})$

then also for the $\sqrt{\text{Lasso}}$, modulo compatibility

$$s_r = o\left(\sqrt{\frac{n}{\log p}}\right)^{1-r} \quad \Rightarrow \quad \|\hat{\beta} - \beta^0\|_1 = o_{\mathbb{P}}(1).$$

Narrator

we now consider bounds for ℓ_∞-error of the Lasso

and bounds for the expectation

this reveals the bias

which is then removed

ℓ_∞-bounds for the Lasso

Consider the Lasso[1]

$$\hat{\beta} := \arg\min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|_2^2/n + 2\lambda\sigma_0\|\beta\|_1 \right\}.$$

Notation

- $\hat{\Sigma} := X^T X/n$ Gram matrix
- $\Theta \in \mathbb{R}^{p \times p}$ some matrix
- $\Theta := (\Theta_1, \cdots, \Theta_p)$
- $|||\Theta|||_1 := \max_k \|\Theta_k\|_1$ "ℓ_1-*operator norm*"

[1] Similar results for $\sqrt{\text{Lasso}}$

Lemma

Let

- $\lambda_0 := \sqrt{2\log(2p/\alpha)/n}$ *(as before)*
- $\tilde{\lambda} := \|I - \Theta^T\hat{\Sigma}\|_\infty$ *(Θ is a "surrogate inverse" of $\hat{\Sigma}$)*

Then with probability at least $1-\alpha$

$$\|\hat{\beta} - \beta^0\|_\infty \leq (\lambda + \lambda_0)\sigma_0 |||\Theta|||_1 + \tilde{\lambda} \overbrace{\|\hat{\beta} - \beta^0\|_1}^{=o_{\mathbb{P}}(1) \text{ under sparsity (see before)}}$$

□

Asymptopia[2]

- $\tilde{\lambda} \asymp \lambda \asymp \lambda_0$
- $\|\hat{\beta} - \beta^0\|_1 = o_{\mathbb{P}}(1)$

$$\Rightarrow \|\hat{\beta} - \beta^0\|_\infty = \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log p}{n}}\right) |||\Theta|||_1$$

[2] When $\mathbb{E}\hat{\Sigma} = \Sigma_0$ and $\Theta = \Sigma_0^{-1}$
then $\tilde{\lambda}$ is the maximum of $\sim p^2$ averages-minus-expectations

The bias of the Lasso

Lemma *We have*

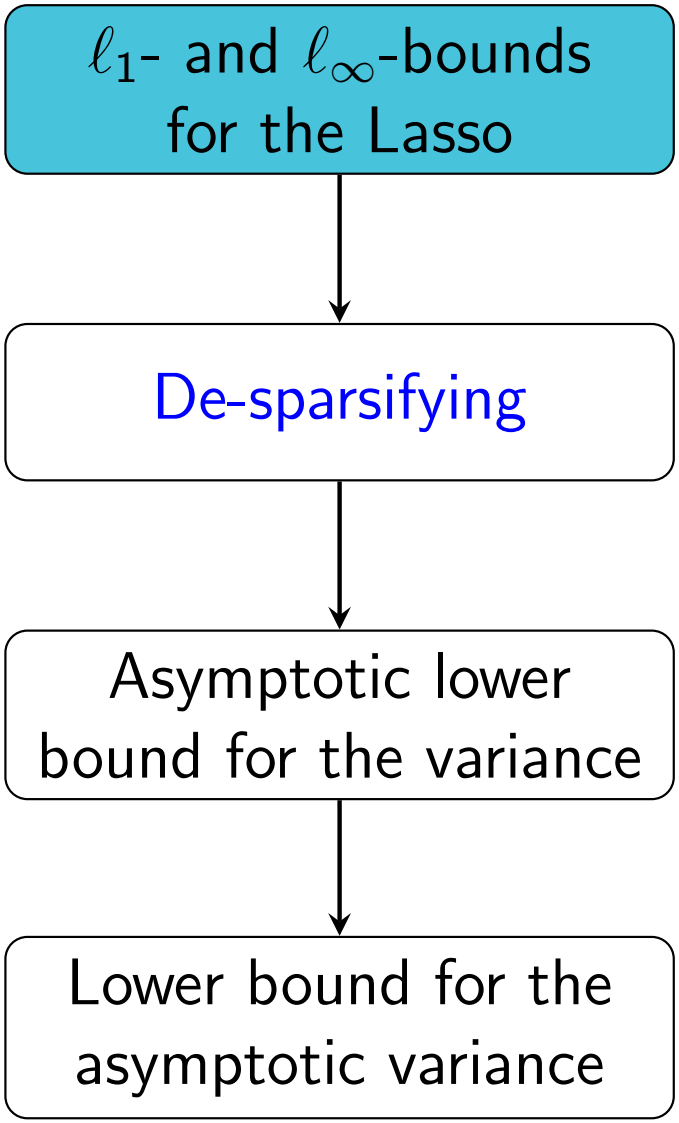
$$\left\| \mathbb{E}(\hat{\beta} - \beta_0) \right\|_\infty \leq \underbrace{\lambda \sigma_0 |||\Theta|||_1}_{\substack{\text{bias} \\ (\text{no } \lambda_0)}} + \tilde{\lambda} \underbrace{\mathbb{E}\|\hat{\beta} - \beta^0\|_1}_{\substack{=o_{\mathbb{P}}(1) \text{ under sparsity} \\ (\text{see later})}} .$$

$\uparrow$
note the norm is
outside the expectation

□

Conclusion

The bias of the Lasso is mainly the term $\boxed{\lambda \sigma_0 |||\Theta|||_1}$



Model:

Linear model
Graphical model
Other

Partly removing the bias:

Definition

The de-sparsified Lasso is

$$\hat{b} := \hat{\beta} + \Theta^T X^T (Y - X\hat{\beta})/n.$$

[Zhang and Zhang, 2014]

Lemma *We have*

$$\left\| \mathbb{E}(\hat{b} - \beta_0) \right\|_\infty \leq \quad \cdots \quad \tilde{\lambda} \mathbb{E} \|\hat{\beta} - \beta^0\|_1.$$

Partly removing the bias:

Definition

The de-sparsified Lasso is

$$\hat{b} := \hat{\beta} + \Theta^T X^T (Y - X\hat{\beta})/n.$$

[Zhang and Zhang, 2014]

Lemma *We have*

$$\left\| \mathbb{E}(\hat{\beta} - \beta_0) \right\|_\infty \leq \lambda \sigma_0 |||\Theta|||_1 + \tilde{\lambda} \mathbb{E} \|\hat{\beta} - \beta^0\|_1.$$

Partly removing the bias:

Definition

The de-sparsified Lasso is

$$\hat{b} := \hat{\beta} + \Theta^T X^T (Y - X\hat{\beta})/n.$$

[Zhang and Zhang, 2014]

Lemma *We have*

$$\left\| \mathbb{E}(\hat{b} - \beta_0) \right\|_\infty \leq \quad \cdots \quad \tilde{\lambda} \mathbb{E} \|\hat{\beta} - \beta^0\|_1.$$

Asymptopia

$$\left.\begin{array}{c}\tilde{\lambda} = \mathcal{O}(\sqrt{\log p/n}) \\ \& \\ \mathbb{E}\|\hat{\beta} - \beta_0\|_1 = o(1/\sqrt{\log p})\end{array}\right. \Rightarrow \left\|\mathbb{E}(\hat{b} - \beta_0)\right\|_\infty = o(1/\sqrt{n}).$$

[3] This will come back...

Asymptopia

$$\tilde{\lambda} = \mathcal{O}(\sqrt{\log p/n}) \quad \& \quad \mathbb{E}\|\hat{\beta} - \beta_0\|_1 = o(1/\sqrt{\log p}) \quad \Rightarrow \quad \left\|\mathbb{E}(\hat{b} - \beta_0)\right\|_\infty = o(1/\sqrt{n}).$$

Recall :

compatibility conditions on X & (weak) sparsity conditions e.g. $s_0 = o(\sqrt{n}/\log p)$ $\Rightarrow$ $\|\hat{\beta} - \beta_0\|_1 = o_{\mathbb{P}}(1/\sqrt{\log p})$ and actually indeed[3] $\mathbb{E}\|\hat{\beta} - \beta_0\|_1 = o(1/\sqrt{\log p})$

[3] This will come back...

De-sparsifying using the node-wise $\sqrt{\text{Lasso}}$

Let for $j = 1, \ldots, p$

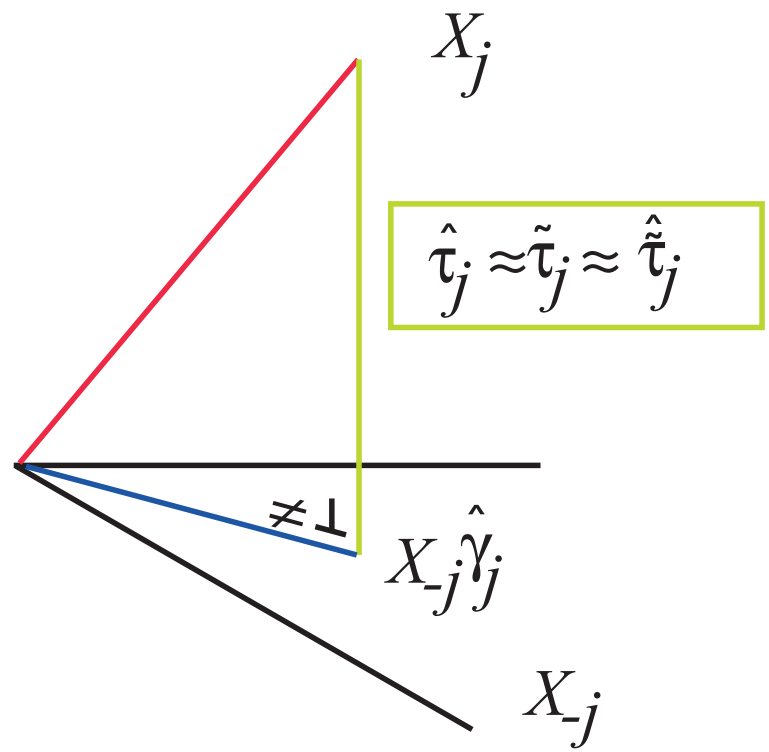
$$\hat{\gamma}_j := \arg \min_{\gamma \in \mathbb{R}^{p-1}} \left\{ \|X_j - X_{-j}\gamma\|_2/\sqrt{n} + \lambda_\sharp \|\gamma\|_1 \right\},$$

&

$$\hat{\tau}_j^2 := \|X_j - X_{-j}\hat{\gamma}_j\|_2/\sqrt{n}$$

$$\tilde{\tau}_j^2 := \hat{\tau}_j^2 + \lambda_\sharp \hat{\tau}_j \|\hat{\gamma}_j\|_1$$

$$\hat{\tilde{\tau}}_j := \tilde{\tau}_j^2/\hat{\tau}_j$$



Let

$$\hat{\Theta} := \begin{pmatrix} 1/\tilde{\tau}_1^2 & -\hat{\gamma}_{1,2}/\tilde{\tau}_2^2 & \cdots & -\hat{\gamma}_{1,p}/\tilde{\tau}_p^2 \\ -\hat{\gamma}_{2,1}/\tilde{\tau}_1^2 & 1/\tilde{\tau}_2^2 & \cdots & -\hat{\gamma}_{2,p}/\tilde{\tau}_p^2 \\ \vdots & \vdots & \ddots & \vdots \\ -\hat{\gamma}_{p,1}/\tilde{\tau}_1^2 & -\hat{\gamma}_{p,2}/\tilde{\tau}_2^2 & \cdots & 1/\tilde{\tau}_p^2 \end{pmatrix}$$

Then

$$\left| I - \hat{\Theta}^T \hat{\Sigma} \right| \underbrace{\leq}_{\text{entry wise}} \lambda_\sharp \begin{pmatrix} 0 & \hat{\tilde{\tau}}_2^{-1} & \cdots & \hat{\tilde{\tau}}_p^{-1} \\ \hat{\tilde{\tau}}_1^{-1} & 0 & \cdots & \hat{\tilde{\tau}}_p^{-1} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\tilde{\tau}}_1^{-1} & \hat{\tilde{\tau}}_2^{-1} & \cdots & 0 \end{pmatrix}$$

Definition *The de-sparsified Lasso using* $\hat{\Theta}$ *is*

$$\hat{\hat{b}} := \hat{\beta} + \hat{\Theta}^T X^T (Y - X\hat{\beta})/n.$$

[Zhang and Zhang, 2014]

Lemma *For all* j

$$\hat{\hat{\tau}}_j(\hat{\hat{b}}_j - \beta_j^0) = \mathcal{N}(0, \sigma_0^2/n) + \Delta_j,$$

where $\|\Delta\|_\infty \leq \lambda_\sharp \|\hat{\beta} - \beta^0\|_1$.

Asymptopia

$$\left.\begin{array}{c} \lambda_\sharp = \mathcal{O}_{\mathbb{P}}(\sqrt{\log p/n}) \\ \& \\ \|\hat{\beta} - \beta^0\|_1 = o_{\mathbb{P}}(1/\sqrt{\log p}) \end{array}\right. \quad \Rightarrow \quad \hat{b}_j - \beta_j^0 \approx \mathcal{N}\left(0, \frac{\sigma_0^2}{n\hat{\tilde{\tau}}_j^2}\right)$$

Remark

No sparsity conditions on the design X are imposed[4]

[4] but they do occur when looking at asymptotic efficiency!

Asymptopia

$$\left.\begin{array}{c} \lambda_\sharp = \mathcal{O}_{\mathbb{P}}(\sqrt{\log p/n}) \\ \& \\ \|\hat{\beta} - \beta^0\|_1 = o_{\mathbb{P}}(1/\sqrt{\log p}) \end{array}\right. \quad \Rightarrow \quad \hat{b}_j - \beta_j^0 \approx \mathcal{N}\left(0, \frac{\sigma_0^2}{n\hat{\tilde{\tau}}_j^2}\right)$$

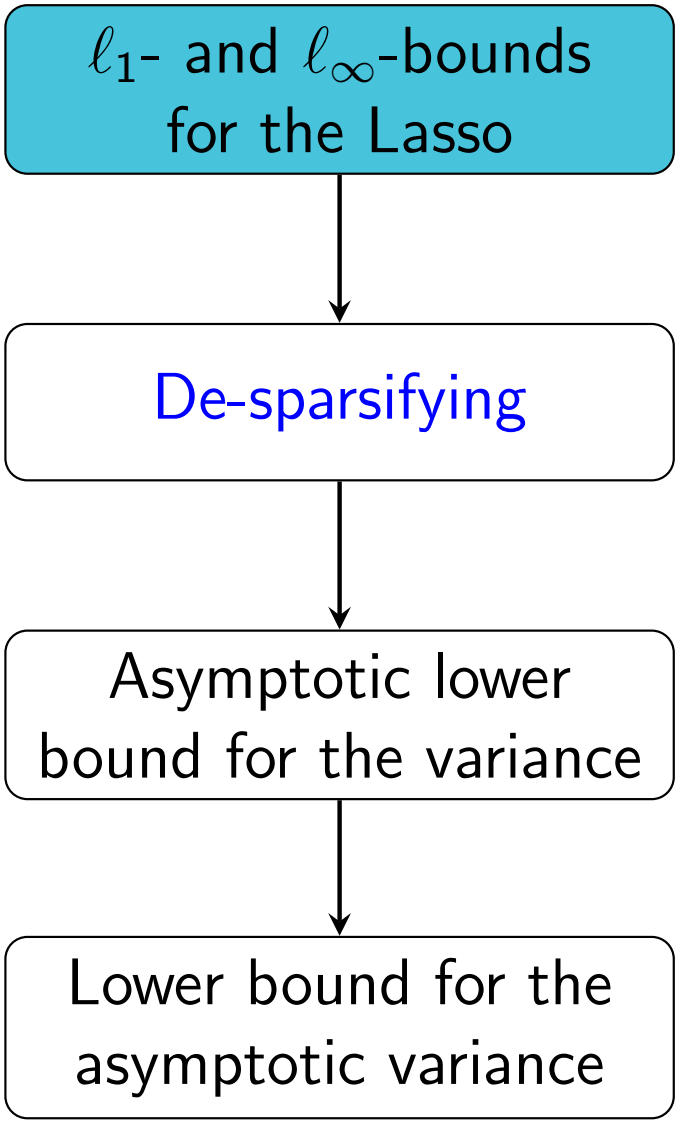
Remark

No sparsity conditions on the design X are imposed[4]

Recall :

$$\left.\begin{array}{c} \text{compatibility conditions on } X \\ \& \\ \text{(weak) sparsity conditions} \\ \text{e.g. } s_0 = o(\sqrt{n}/\log p) \end{array}\right. \quad \Rightarrow \quad \|\hat{\beta} - \beta_0\|_1 = o_{\mathbb{P}}(1/\sqrt{\log p})$$

[4] but they do occur when looking at asymptotic efficiency!



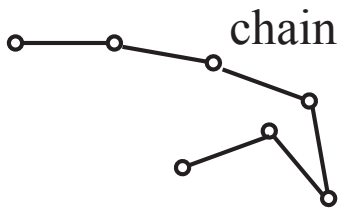
Model:

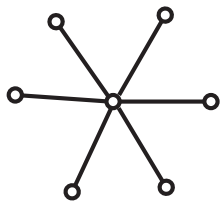
Linear model
Graphical model
Other

Extensions

- χ^2-confidence sets for groups [vdG and Stucky, 2015]
- Confidence intervals for the precision matrix
  - using the graphical Lasso [Janková and vdG, 2015]
  -
-
  -
  -
  -

- Some simulations with graphical Lasso[5]





Chain graph	S_0 Avgcov	S_0 Avglength	S_0^c Avgcov	S_0^c Avglength
graphical Lasso	0.934	0.247	0.972	0.215
MLE with specified S_0	0.963	0.293	-	-
Sample covariance	0.459	0.428	0.897	0.367

Star graph, $d = 8$	S_0 Avgcov	S_0 Avglength	S_0^c Avgcov	S_0^c Avglength
graphical Lasso	0.948	0.328	0.951	0.247
MLE with specified S_0	0.956	0.337	-	-
Sample covariance	0.124	0.499	0.897	0.367

[5] all simulations shown carried out by Jana Janková

Extensions

- χ^2-confidence sets for groups [vdG and Stucky, 2015]
- Confidence intervals for the precision matrix
  - using the graphical Lasso [Janková and vdG, 2015]
  -

Extensions

- χ^2-confidence sets for groups [vdG and Stucky, 2015]
- Confidence intervals for the precision matrix
  - using the graphical Lasso [Janková and vdG, 2015]
  - using the node-wise Lasso [Janková and vdG, 2016]

Extensions

- χ^2-confidence sets for groups [vdG and Stucky, 2015]
- Confidence intervals for the precision matrix
  - using the graphical Lasso [Janková and vdG, 2015]
  - using the node-wise Lasso [Janková and vdG, 2016]
- Confidence intervals in GLM's [Janková and vdG, 2016]

Extensions

- χ^2-confidence sets for groups [vdG and Stucky, 2015]
- Confidence intervals for the precision matrix
  - using the graphical Lasso [Janková and vdG, 2015]
  - using the node-wise Lasso [Janková and vdG, 2016]
- Confidence intervals in GLM's [Janková and vdG, 2016]
  - LAD

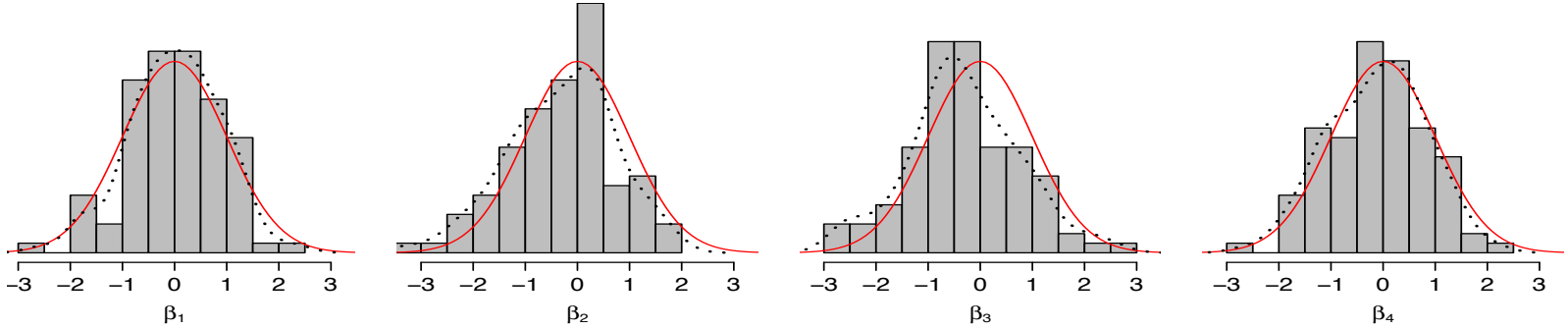
Extensions

- χ^2-confidence sets for groups [vdG and Stucky, 2015]
- Confidence intervals for the precision matrix
  - using the graphical Lasso [Janková and vdG, 2015]
  - using the node-wise Lasso [Janková and vdG, 2016]
- Confidence intervals in GLM's [Janková and vdG, 2016]
  - LAD
  - Huber

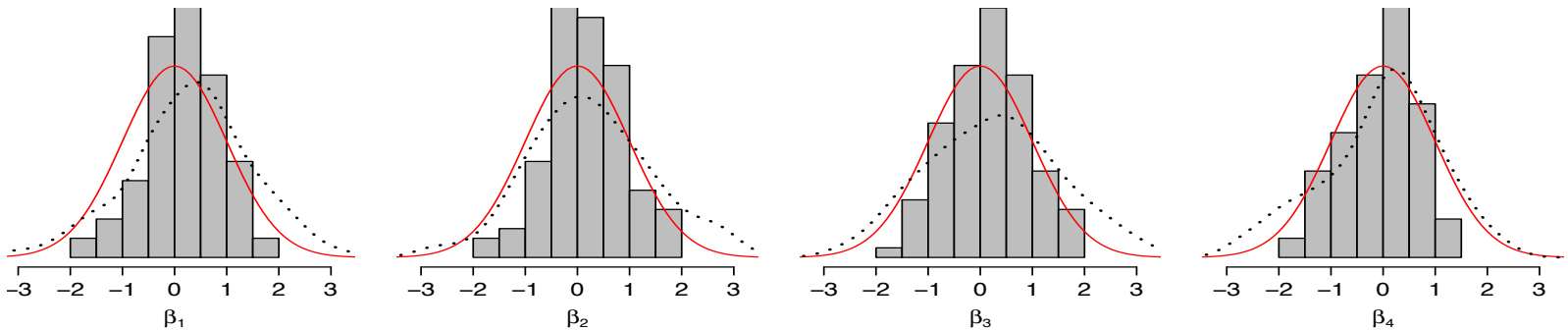
Extensions

- χ^2-confidence sets for groups [vdG and Stucky, 2015]
- Confidence intervals for the precision matrix
  - using the graphical Lasso [Janková and vdG, 2015]
  - using the node-wise Lasso [Janková and vdG, 2016]
- Confidence intervals in GLM's [Janková and vdG, 2016]
  - LAD
  - Huber
  - logistic regression

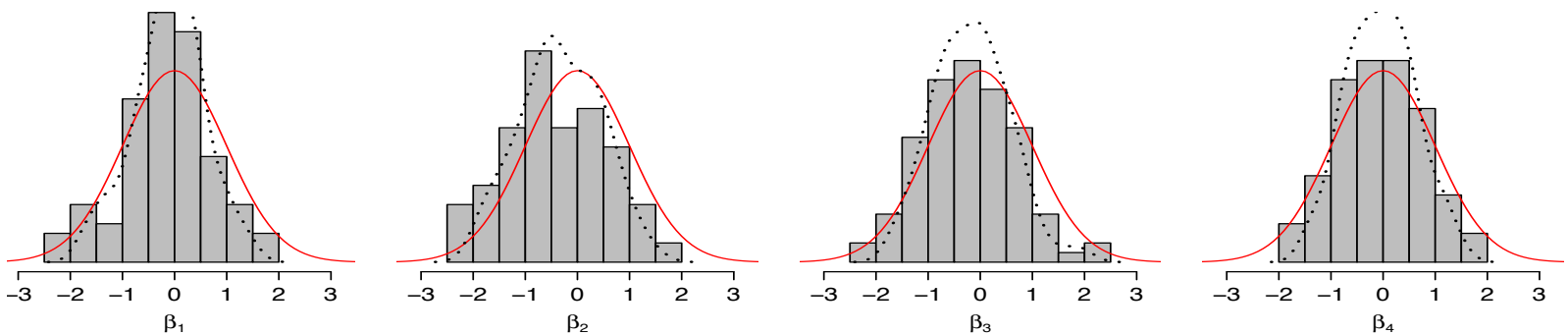
De-sparsified Lasso



De-sparsified LAD



De-sparsified Huber estimator

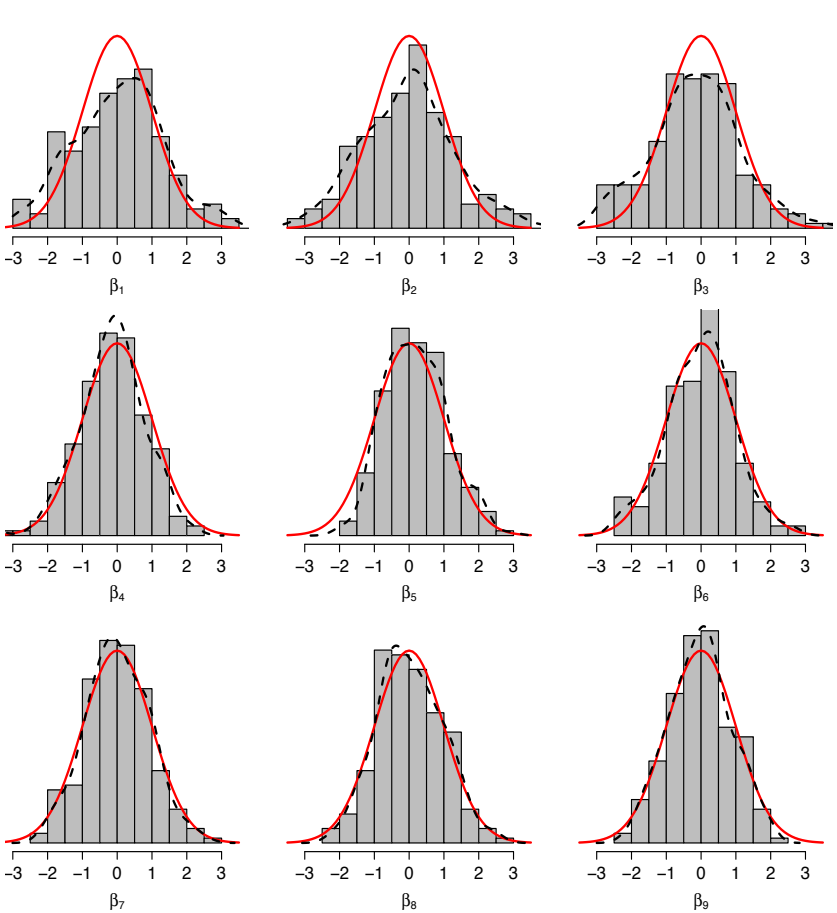


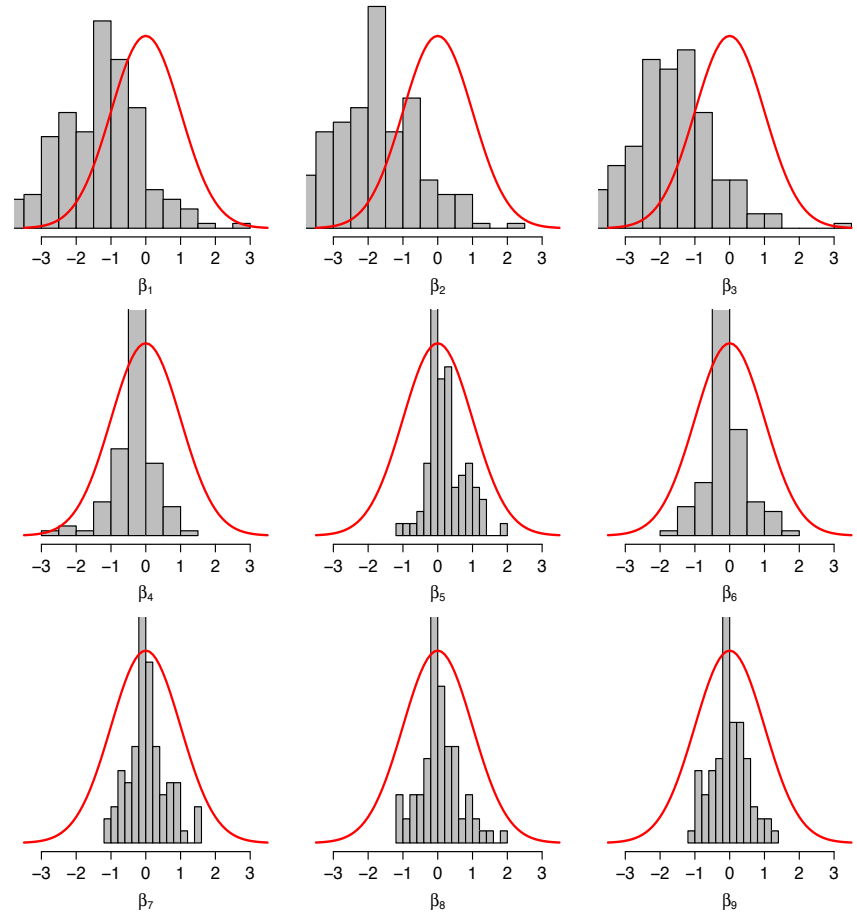
[6] [**Janková** and vdG, 2016]

Histograms for coefficients in logistic regression

De-sparsified logistic Lasso

Logistic Lasso





[7] [**Janková** and vdG, 2016]

Estimated coverage probabilities and lengths

p	n		avgcov	avglength	avgcov	avglength
100	400	D-S Logistic Lasso	0.816	0.423	0.927	0.402
		Maximum likelihood	0.320	0.730	0.891	0.638

Conclusion from simulations:

(for $p < n$)

the de-sparsified estimator seems to be better than MLE

Narrator

based on the asymptotic results
one constructs (asymptotic) $(1-\alpha)$-confidence intervals

the question is now:

does the proposed method gives the shortest intervals?

or

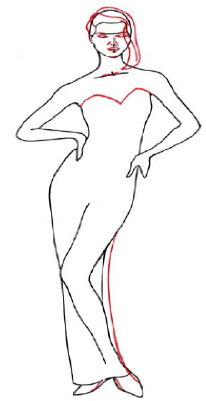
is the variance about the smallest possible?

as we are in high-dimensions, this requires some thought....
as the model changes with n!

let us set up the situation in general terms

Lower bounds and efficiency

- Let $\mathbf{X}_n := (X_{1,n}, \ldots, X_{n,n})$ have distribution $\mathbf{P}_n$, $n = 1, 2, \ldots$



Supermodel:

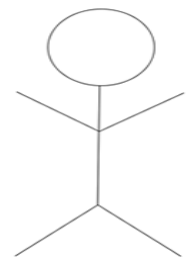
$\mathbf{P}_n \in \mathcal{P}_n := \{\mathbf{P}_{\beta_n,n} : \ \beta_n \in \mathcal{B}_n\}$

with $\mathcal{B}_n \subset \mathbb{R}^{p_n}$ *convex and open*

- Let $S_{\beta_n} := \{j : \beta_{j,n} \neq 0\}$, $s_{\beta_n} := |S_{\beta_n}|$.
- Let $m_n \in \mathbb{N}$ given, $m_n/n \to 0$
- Let $\mathcal{B}_n(m_n) := \{\beta_n \in \mathcal{B}_n : \ s_{\beta_n} \leq m_n\}$

Sparse model:[8]

$\mathbf{P}_n \in \mathcal{P}_n(m_n) := \{\mathbf{P}_{\beta_n,n} : \ \beta_n \in \mathcal{B}_n(m_n)\}$



[8] Possible extension to *weakly* sparse models

let $g_n(\beta_n) \in \mathbb{R}$ be the parameter of interest.

Preview:

we will eventually show a Le Cam type of result.

the *main condition* will be

$$\boxed{{\beta_n}^0 + \mathcal{I}_n^{-1}({\beta_n}^0)\dot{g}_n({\beta_n}^0) \in \mathcal{B}_n(m_n)}, \; m_n = o(\sqrt{n/\log p}).$$

here

- $\dot{g}_n$ is the derivative of g_n
- $\mathcal{I}_n$ be the Fisher-information matrix
- ${\beta_n}^0$ is a fixed sequence

then the asymptotic variance

of an asymptotically linear estimator

is at least

$$\dot{g}_n({\beta_n}^0)^T \mathcal{I}_n^{-1}({\beta_n}^0)\dot{g}_n({\beta_n}^0)$$

Narrator

but

we first consider
- "strong asymptotic unbaisedness"
- in the linear
- and the graphical model.

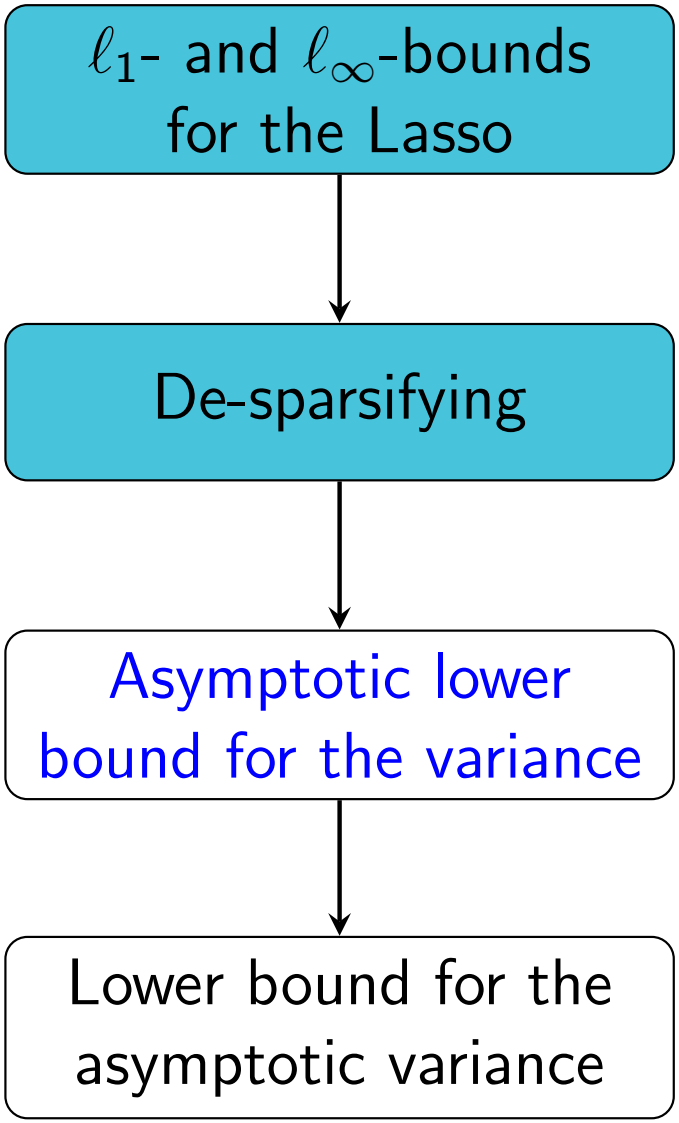
Let d_n be some norm on $\mathbb{R}^{p_n}$
Let $\sqrt{n}\delta_n \to 0$.
Let $\mathcal{B}_n(m_n, \delta_n) := \{\beta_n \in \mathcal{B}_n(m_n) : \ d_n(\beta_n - {\beta_n}^0) \leq \delta_n\}$
(neighbourhood of the fixed (sequence) ${\beta_n}^0$)

Definition

We call $T_n = T_n(\mathbf{X}_n)$ a strongly asymptotically unbiased estimator at ${\beta_n}^0$ if

$$\mathbb{E}_{\beta_n}(T_n) = g_n(\beta_n) + o(\delta_n) \ \forall \ \beta_n \in \mathcal{B}_n(m_n, \delta_n).$$

$\uparrow$
parameter of interest



Model:

Linear model
Graphical model
Other

Lower bounds for the linear model

Let $Y_n \in \mathbb{R}^n$ and $X_n \in \mathbb{R}^{n \times p_n}$, $\mathcal{B}_n := \mathbb{R}^{p_n}$.

Supermodel:

$\mathbf{P}_{\beta_n,n}$ *corresponds to the linear model*

$Y_n = X_n\beta_n + \epsilon_n,\ \epsilon_n \sim \mathcal{N}_n(0, I),\ \beta_n \in \mathbb{R}^{p_n}$

Sparse model:

$\beta_n \in \{\beta_n :\ s_{\beta_n} \leq m_n\}$

- Let $\hat{\beta}_n$ be the Lasso with tuning parameter $\lambda := 2\lambda_0$
- Let $\hat{\Theta}_n$ be the node-wise square root Lasso estimator with tuning parameter $\lambda_\sharp = \mathcal{O}(\sqrt{\log p_n / n})$.
- Let $\hat{\hat{b}}_n$ be the de-biased estimator

$$\hat{\hat{b}}_n := \hat{\beta}_n + \hat{\Theta}_n^T X_n^T (Y_n - X_n\hat{\beta}_n)/n$$

Consider estimating $g_n(\beta_n) := \beta_{j,n}$ (say).

Theorem

Suppose that

- $m_n = o(\sqrt{n}/\log p)$
- $\hat{\Theta}_{j,j,n} = \mathcal{O}(1)$
- $\max_{S \subset \{1,\ldots,p_n\}:\ |S| \leq m_n} \hat{\phi}_n^{-2}(3, S) = \mathcal{O}(1)$.

Then $\hat{\hat{b}}_{j,n}$ is a strongly asymptotically unbiased estimator. □

Theorem [Janková and vdG, 2016]

Suppose that T_n is a strongly asymptotically unbiased estimator at ${\beta_n}^0$.

Assume that $\boxed{{\beta_n}^0 + \hat{\Theta}_{j,n} \in \mathcal{B}_n(m_n)}$.

$\uparrow$

the j^{th} column of $\hat{\Theta}_n$

Then

$$\text{var}_{{\beta_n}^0}(T_n) \geq \frac{\hat{\Theta}_{j,j,n}}{n}(1 - o(1)).$$

Corollary

Under the above condition $\boxed{}$ *with*

- $m_n = o(\sqrt{n}/\log p)$,
- $\hat{\Theta}_{j,j,n} = \mathcal{O}(1)$,
- $\max_{S \subset \{1,\ldots,p_n\}:\ |S| \leq m_n} \hat{\phi}_n^{-2}(3, S) = \mathcal{O}(1)$.

the de-sparsified estimator $\hat{\hat{b}}_{j,n}$ has asymptotically

the smallest *variance*

among all strongly asymptotically unbiased estimators.

An example of a model for the design:

The rows of X_n are i.i.d. copies of $X_n^0 \sim \mathcal{N}_{p_n}(0, \Sigma_n^0)$

The largest eigenvalue of Σ_n^0 is bounded

The smallest eigenvalue of Σ_n^0 is bounded away from zero.

Let Θ_n^0 be the inverse of Σ_n^0.

Theorem

Assume

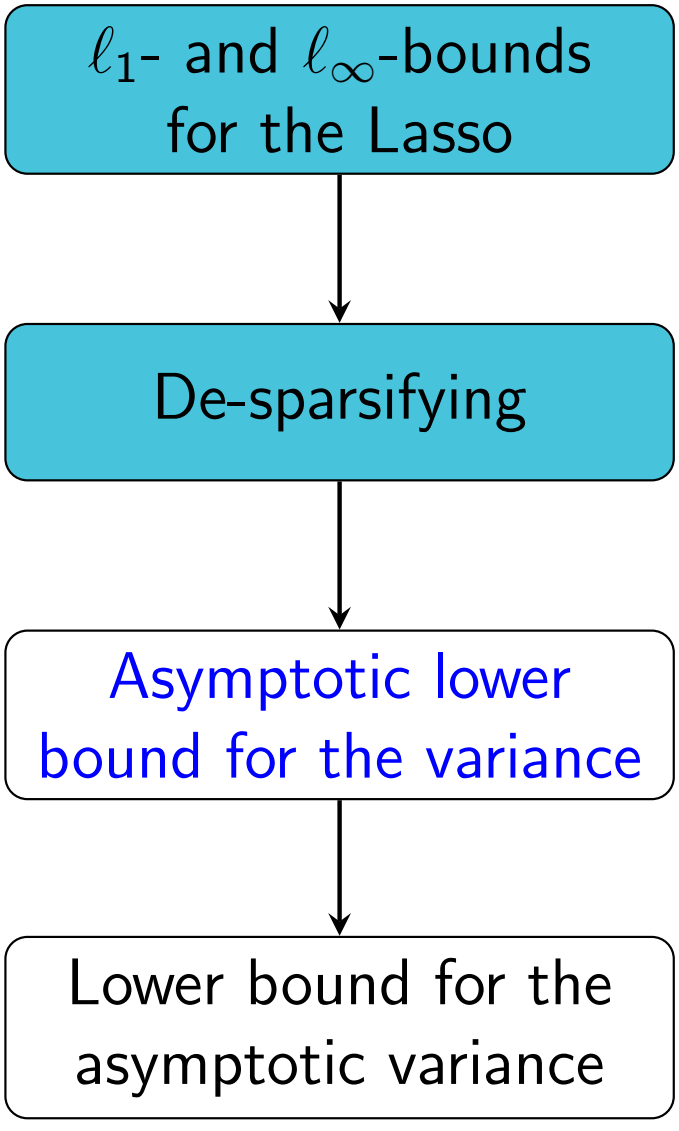
- *for some $m_n' = o(\sqrt{n}/\log n)$:*

$$\beta_n^0 \in \mathcal{B}_n(m_n') \text{ as wel as } \Theta_{j,n}^0 \in \mathcal{B}_n(m_n').$$

$$\uparrow$$

j^{th} column of the population precision matrix Θ_n^0

Then with high probability all the conditions of the previous corollary are met. □



Model:

Linear model
Graphical model
Other

Lower bounds for the graphical model

Supermodel:

The rows of X_n are i.i.d. copies of $X_n^0 \sim \mathcal{N}_{p_n}(0, \Sigma_n^0)$
The largest eigenvalue of Σ_n^0 is bounded
The smallest eigenvalue of Σ_n^0 is bounded away from zero.

Let Θ_n^0 be the inverse of Σ_n^0.
Let the maximal degree of a matrix $\Theta_n \geq 0$ in $\mathbb{R}^{p_n \times p_n}$ be

$$s_{\Theta_n} := \max_k |\{j \neq k : \ \Theta_{j,k,n} \neq 0\}|$$

Sparse model: $\mathcal{B}_n(m_n) := \{\Theta_n > 0 : \ s_{\Theta_n} \leq m_n\}$.

Let $g_n(\Theta_n) := \Theta_{j,k,n}$ for some fixed (sequences) j and k.

Theorem [Janková and vdG, 2016]
Suppose that T_n is a strongly unbiased estimator at $\Theta^0_{j,k,n}$. Assume that

$$\boxed{\Theta^0_n + \left[\Theta^0_{j,n}\Theta^{0T}_{k,n} + \Theta^0_{k,n}\Theta^{0T}_{j,n}\right] \in \mathcal{B}_n(m_n)}$$

Then

$$\mathrm{var}_{\Theta^0_n}(T_n) \geq \frac{\Theta^{02}_{j,k,n} + \Theta^0_{j,j,n}\Theta^0_{k,k,n}}{n}(1 - o(1)).$$

Under the additional condition

$$m_n = o(\sqrt{n}/\log p_n)$$

the de-sparsified node-wise estimator of Θ_n^0 is strongly asymptotically unbiased

and hence asymptotically efficient among such.

Narrator

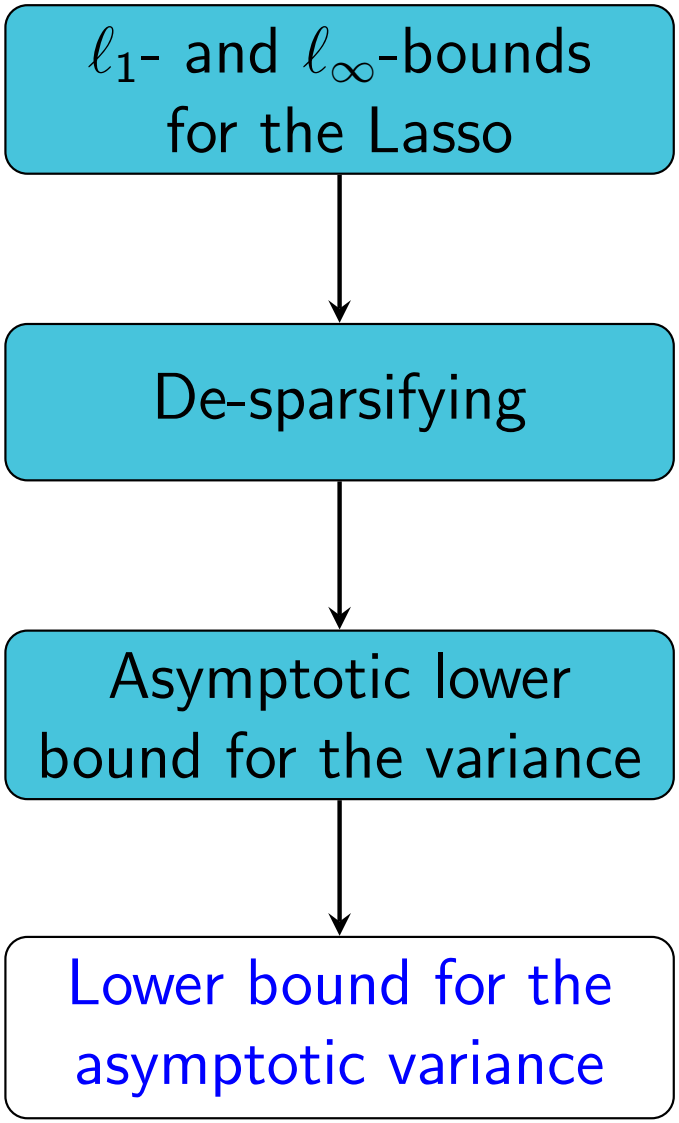
well...
the asymptotics for $\mathrm{var}(T_n)$ is something else than the asymptotic variance

let us now look at Le Cam's 3^{rd} Lemma

Narrator

well...
the asymptotics for $\mathrm{var}(T_n)$ is something else than the asymptotic variance

let us now look at Le Cam's 3$^{\mathrm{rd}}$ Lemma



Model:

Linear model
Graphical model
Other

Le Cam theory

We assume $X_{1,n}, \ldots, X_{n,n}$ are independent in $\mathcal{X}$

Notation

For functions $f_i : \mathcal{X} \to$ Euclidean space $(i = 1, \ldots, n)$:

$$\mathbf{P}_{\beta_n,n} f := \mathbb{E}_{\beta_n} f_i(X_{i,n})/n.$$

Definition *We call T_n asymptotically linear at ${\beta_n}^0$ if*

$$T_n - g_n({\beta_n}^0) = \frac{1}{n}\sum_{i=1}^{n} \ell_{i,n}(X_i) + o_{\mathbf{P}_{{\beta_n}^0,n}}\left(\frac{1}{\sqrt{n}}\right),$$

where

$\mathbb{E}_{{\beta_n}^0}\ \ell_n(X_{i,n}) = 0,\ i = 1, \ldots, n$

$v_n^2 := \mathbf{P}_{{\beta_n}^0,n}\ \ell_n^2 = \mathcal{O}(1)$

Remark

$\ell_n = (\ell_{1,n}, \ldots, \ell_{n,n})$ is called the influence function

v_n^2 is called the asymptotic variance

Let $\mathbf{p}_{i,\beta_n,n}$ be the density of $X_{i,n}$ under $\mathbf{P}_{\beta_n,n}$ w.r.t. some dominating measure μ_n $(i = 1, \ldots, n)$.

Definition

We say that the score condition is met at ${\beta_n}^0 \in \Theta$ *if for*

$$\mathbf{s}_{i,n} := \left. \frac{\partial \log \mathbf{p}_{i,\beta_n,n}}{\partial \beta_n} \right|_{\beta_n = {\beta_n}^0}, \quad i = 1, \ldots, n$$

it holds that

- $\mathbb{E}_{{\beta_n}^0} \mathbf{s}_{i,n}(X_{i,n}) = 0, \; i = 1, \ldots, n$
- $\mathcal{I}_n({\beta_n}^0) := \mathbf{P}_{{\beta_n}^0,n} \, \mathbf{s}_{,n}\mathbf{s}_n^T = \mathcal{O}(1)$
- $\max_{1 \le i \le n} \|\ddot{\mathbf{s}}_{i,n}\|_\infty = \mathcal{O}(1)$

and

- $$\left\| \frac{1}{n} \sum_{i=1}^{n} \dot{\mathbf{s}}_{i,n}(X_{i,n}) \dot{\mathbf{s}}_{i,n}^T(X_{i,n}) + \mathcal{I}_n({\beta_n}^0) \right\|_\infty = o_{\mathbf{P}_{{\beta_n}^0,n}}\left(\sqrt{\frac{\log p_n}{n}} \right)$$

Remark

$\mathbf{s}_n = (\mathbf{s}_{1,n}, \ldots, \mathbf{s}_{n,n})$ is called the score function

$\mathcal{I}(\beta_n^0) \in \mathbb{R}^{p_n \times p_n}$ is called the Fisher information matrix

Theorem *Suppose*

- ◇ *asymptotic linearity at* ${\beta_n}^0 \in \mathcal{B}_n(m_n)$
- ◇ *the score condition at* ${\beta_n}^0$
- ◇ *Lindeberg conditions*
- ◇ $m_n = \begin{cases} o(\sqrt{n/\log p}) & \text{if } \ddot{\mathbf{s}}_n \neq 0 \\ o(n^{1/3}) & \text{if } \ddot{\mathbf{s}}_n = 0 \end{cases}$

Let $\beta_n := {\beta_n}^0 + h_n \in \mathcal{B}_n(m_n)$ *where* $\|h_n\|_2 = \mathcal{O}(1)$.
Then

$$\frac{1}{v_n}\left\{\sqrt{n}(T_n - g_n(\beta_n)) - \underbrace{\left(h_n^T \mathbf{P}_{{\beta_n}^0,n}\mathbf{s}_n\ell_n - h_n^T \dot{g}_n({\beta_n}^0)\right)}_{\text{asymptotic bias}}\right\}$$

is asymptotically $\mathcal{N}(0,1)$ *under* $\mathbf{P}_{\beta_n,n}$.

Corollary

No asymptotic bias $\forall\ {\beta_n}^0 + h_n/\sqrt{n} \in \mathcal{B}_n(m_n)$

$$\Leftrightarrow$$

$$\left(h_n^T \mathbf{P}_{{\beta_n}^0,n} \mathbf{s}_n \ell_n - h_n^T \dot{g}_n({\beta_n}^0) \right) = 0 \ \forall\ {\beta_n}^0 + h_n/\sqrt{n} \in \mathcal{B}_n(m_n)$$

$$\Leftrightarrow$$

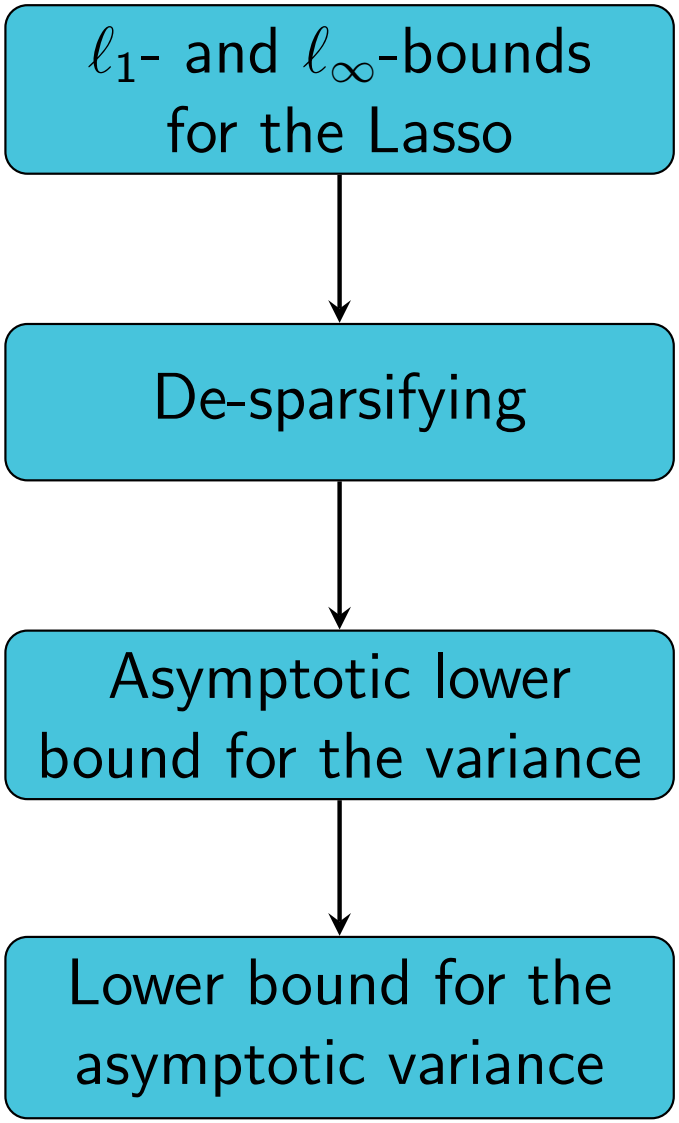
asymptotic variance $v_n^2 \geq \max_{h_n:\ {\beta_n}^0 + h_n/\sqrt{n} \in \mathcal{B}_n(m_n)} \frac{h_n^T \dot{g}_n({\beta_n}^0)}{h_n^T \mathcal{I}_n({\beta_n}^0) h_n} (1 - o(1))$

Hence, if

$$\boxed{\beta_n := {\beta_n}^0 + \mathcal{I}_n^{-1}({\beta_n}^0) \dot{g}_n({\beta_n}^0) \in \mathcal{B}_n(m_n)}$$

we have

$$v_n^2 \geq \dot{g}_n({\beta_n}^0) \mathcal{I}_n^{-1}({\beta_n}^0) \dot{g}_n({\beta_n}^0)(1 - o(1)).$$



Model:

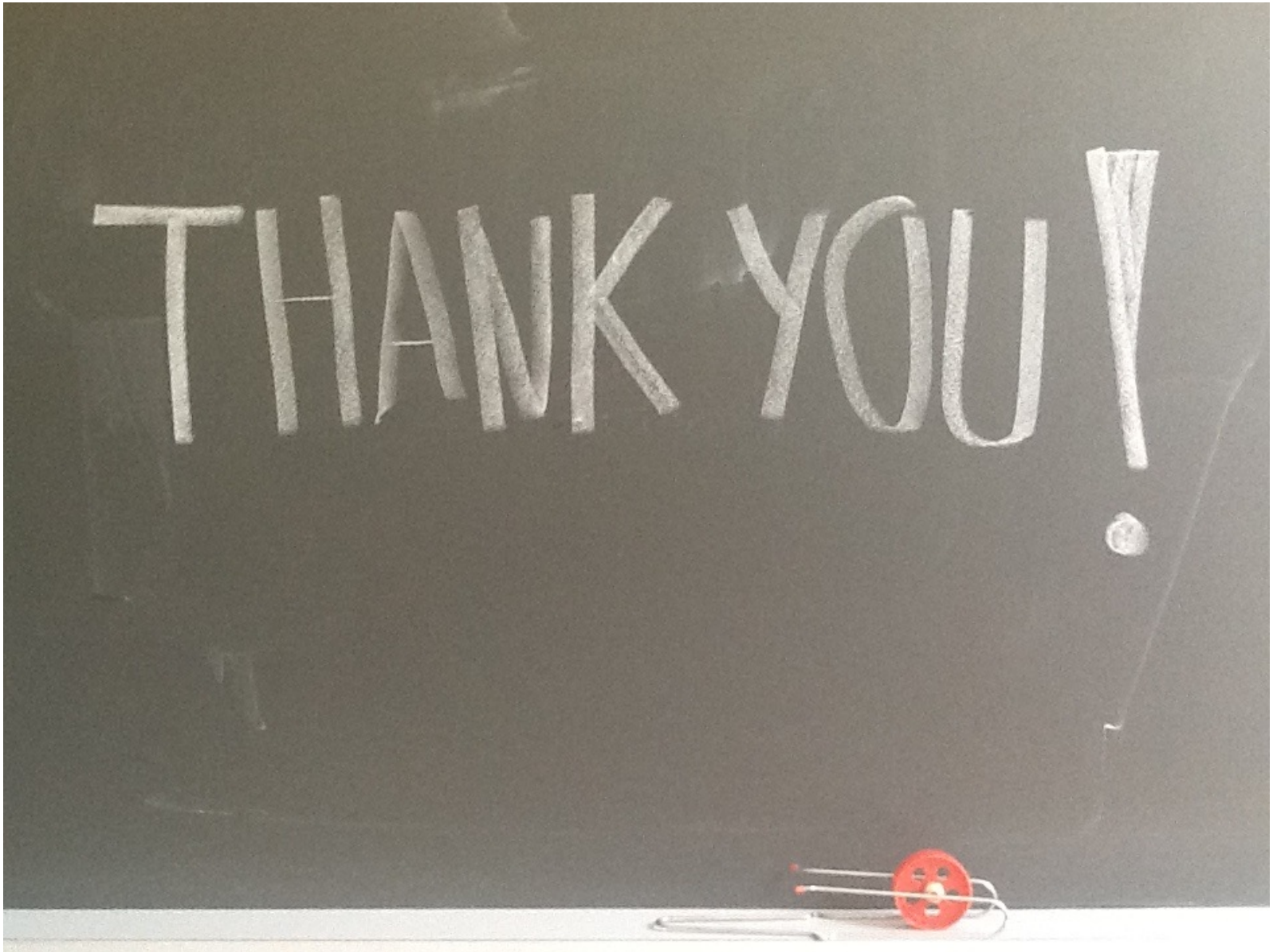
Linear model
Graphical model
Other

Some conclusions

de-sparsifying for obtaining
confidence sets in high dimensions
works theoretically when
$s_0 = o(\sqrt{n}/\log p)$
(or weak versions thereof)

it works in simulations for p not too large
may beat MLE for $p < n$

asymptotic efficiency may require sparseness of the inverse Fisher information


THANK YOU!