High-Dimensional Hypothesis Testing with the Lasso

Sen Zhao
Department of Biostatistics
University of Washington
Joint work with Ali Shojaie and Daniela Witten

sen-zhao.com

High-Dimensional Data

- Is smoking associated with lung function, adjusting for age?

High-Dimensional Data

- Is smoking associated with lung function, adjusting for age?

$$\log(FEV1) = \beta_0 + \beta_1 \cdot 1(Smoking) + \beta_2 \cdot age + \epsilon.$$

High-Dimensional Data

- Is smoking associated with lung function, adjusting for age?

$$\log(FEV1) = \beta_0 + \beta_1 \cdot 1(Smoking) + \beta_2 \cdot age + \epsilon.$$

- What if we also want to adjust for the expression level of 10k genes?

High-Dimensional Data

- Is smoking associated with lung function, adjusting for age?

$$\log(FEV1) = \beta_0 + \beta_1 \cdot 1(Smoking) + \beta_2 \cdot age + \epsilon.$$

- What if we also want to adjust for the expression level of 10k genes?
- We rarely have more than 10k samples in the study...

High-Dimensional Data

- Is smoking associated with lung function, adjusting for age?

$$\log(FEV1) = \beta_0 + \beta_1 \cdot 1(Smoking) + \beta_2 \cdot age + \epsilon.$$

- What if we also want to adjust for the expression level of 10k genes?
- We rarely have more than 10k samples in the study...
  - We have more variables, p, than samples, n, i.e., high-dimensional data.
  - Ordinary least squares (OLS) do not fit.

High-Dimensional Data

- Is smoking associated with lung function, adjusting for age?

$$\log(FEV1) = \beta_0 + \beta_1 \cdot 1(Smoking) + \beta_2 \cdot age + \epsilon.$$

- What if we also want to adjust for the expression level of 10k genes?
- We rarely have more than 10k samples in the study...
  - We have more variables, p, than samples, n, i.e., high-dimensional data.
  - Ordinary least squares (OLS) do not fit.
- How do we examine the conditional association then?

High-Dimensional Estimation

- Formally, how to draw inference on $\boldsymbol{\beta}^*$ in the linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon},$$

when we have more variables than samples?

High-Dimensional Estimation

- Formally, how to draw inference on $\boldsymbol{\beta}^*$ in the linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon},$$

  when we have more variables than samples?
- One method to estimate $\boldsymbol{\beta}^*$ is to use regularization, e.g., the lasso:

$$\hat{\boldsymbol{\beta}}_\lambda = \underset{\boldsymbol{b}\in\mathbb{R}^p}{\arg\min}\left\{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|_2^2 + \lambda\|\boldsymbol{b}\|_1\right\}.$$

  - $\lambda > 0$ is the regularization tuning parameter.

High-Dimensional Inference with the Lasso

- Lasso estimates alone are not sufficient to draw scientific conclusion.
  - We also need p-values and confidence intervals.

High-Dimensional Inference with the Lasso

- Lasso estimates alone are not sufficient to draw scientific conclusion.
  - We also need *p*-values and confidence intervals.
- Can we compute *p*-values and confidence intervals directly from $\hat{\beta}_\lambda$, just as what we do with OLS estimates?

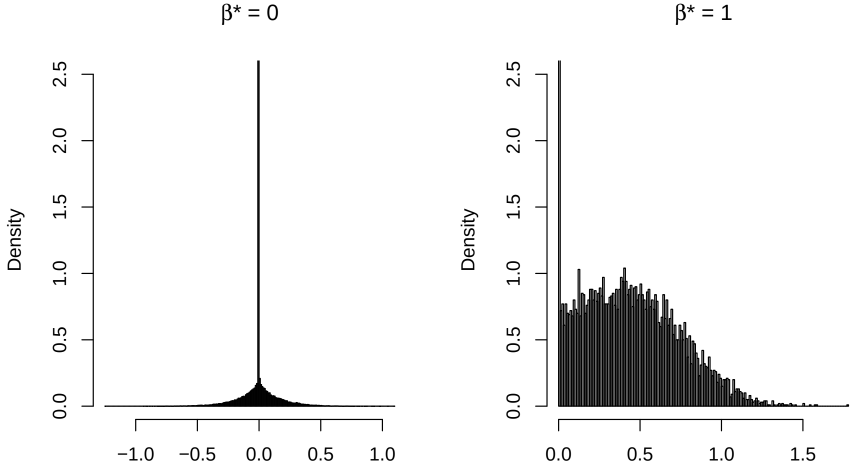
High-Dimensional Inference with the Lasso

- Lasso estimates alone are not sufficient to draw scientific conclusion.
  - We also need *p*-values and confidence intervals.
- Can we compute *p*-values and confidence intervals directly from $\hat{\beta}_\lambda$, just as what we do with OLS estimates?
- Very hard...

High-Dimensional Inference with the Lasso

- Lasso estimates alone are not sufficient to draw scientific conclusion.
  - We also need p-values and confidence intervals.
- Can we compute p-values and confidence intervals directly from $\hat{\boldsymbol{\beta}}_\lambda$, just as what we do with OLS estimates?
- Very hard...
  - $\hat{\boldsymbol{\beta}}_\lambda$ is biased, and the bias is unknown.
  - The distribution of $\hat{\boldsymbol{\beta}}_\lambda$ is also unknown – likely not Gaussian.

Distribution of Lasso Estimates



- **Solution 1**: add an one-step adjustment in $\hat{\beta}_\lambda$ to adjust for its bias.

High-Dimensional Inference with the Lasso

- **Solution 1**: add an one-step adjustment in $\hat{\beta}_\lambda$ to adjust for its bias.
- With some tricks, the resulting *debiased estimator* is
  - asymptotically unbiased.
  - asymptotically Gaussian.

High-Dimensional Inference with the Lasso

- **Solution 1**: add an one-step adjustment in $\hat{\beta}_\lambda$ to adjust for its bias.
- With some tricks, the resulting *debiased estimator* is
  - asymptotically unbiased.
  - asymptotically Gaussian.
- Great: obtain *p*-values and confidence intervals from debiased estimates just as from OLS estimates – *lasso debiased tests*.

High-Dimensional Inference with the Lasso

- **Solution 1**: add an one-step adjustment in $\hat{\beta}_\lambda$ to adjust for its bias.
- With some tricks, the resulting *debiased estimator* is
  - asymptotically unbiased.
  - asymptotically Gaussian.
- Great: obtain p-values and confidence intervals from debiased estimates just as from OLS estimates – *lasso debiased tests*.
- Not so great: computationally intensive.

High-Dimensional Inference with the Lasso

- **Solution 1**: add an one-step adjustment in $\hat{\beta}_\lambda$ to adjust for its bias.
- With some tricks, the resulting *debiased estimator* is
  - asymptotically unbiased.
  - asymptotically Gaussian.
- Great: obtain p-values and confidence intervals from debiased estimates just as from OLS estimates – *lasso debiased tests*.
- Not so great: computationally intensive.
- References: Javanmard and Montanari (2013, 2014a,b); Zhang and Zhang (2014); van de Geer et al. (2014); Dezeure et al. (2015); Zhao and Shojaie (2016); Ning and Liu (2016).

- **Solution 2**: only use a small subset of variables $\mathcal{M}$ to draw inference:

$$\boldsymbol{y} = \boldsymbol{X}_{\mathcal{M}}\boldsymbol{\beta}^{\mathcal{M}} + \boldsymbol{\epsilon}.$$

High-Dimensional Inference with the Lasso

- **Solution 2**: only use a small subset of variables $\mathcal{M}$ to draw inference:

$$\boldsymbol{y} = \boldsymbol{X}_{\mathcal{M}}\boldsymbol{\beta}^{\mathcal{M}} + \boldsymbol{\epsilon}.$$

- Great: if $|\mathcal{M}| < n$, we can use OLS for the above model!

- **Solution 2**: only use a small subset of variables $\mathcal{M}$ to draw inference:

$$\boldsymbol{y} = \boldsymbol{X}_{\mathcal{M}}\boldsymbol{\beta}^{\mathcal{M}} + \boldsymbol{\epsilon}.$$

- Great: if $|\mathcal{M}| < n$, we can use OLS for the above model!
- How to choose $\mathcal{M}$?

- **Solution 2**: only use a small subset of variables $\mathcal{M}$ to draw inference:

$$\boldsymbol{y} = \boldsymbol{X}_{\mathcal{M}}\boldsymbol{\beta}^{\mathcal{M}} + \boldsymbol{\epsilon}.$$

- Great: if $|\mathcal{M}| < n$, we can use OLS for the above model!
- How to choose $\mathcal{M}$?
  - $\mathcal{M}$ should be small (to reduce variance).

- **Solution 2**: only use a small subset of variables $\mathcal{M}$ to draw inference:

$$\boldsymbol{y} = \boldsymbol{X}_{\mathcal{M}}\boldsymbol{\beta}^{\mathcal{M}} + \boldsymbol{\epsilon}.$$

- Great: if $|\mathcal{M}| < n$, we can use OLS for the above model!
- How to choose $\mathcal{M}$?
  - $\mathcal{M}$ should be small (to reduce variance).
  - $\mathcal{M}$ should contain all confounding variables (otherwise $\boldsymbol{\beta}^{\mathcal{M}} \neq \boldsymbol{\beta}^*_{\mathcal{M}}$).

High-Dimensional Inference with the Lasso

- <u>**Solution 2**</u>: only use a small subset of variables $\mathcal{M}$ to draw inference:

$$\boldsymbol{y} = \boldsymbol{X}_{\mathcal{M}}\boldsymbol{\beta}^{\mathcal{M}} + \boldsymbol{\epsilon}.$$

- Great: if $|\mathcal{M}| < n$, we can use OLS for the above model!
- How to choose $\mathcal{M}$?
  - $\mathcal{M}$ should be small (to reduce variance).
  - $\mathcal{M}$ should contain all confounding variables (otherwise $\boldsymbol{\beta}^{\mathcal{M}} \neq \boldsymbol{\beta}^{*}_{\mathcal{M}}$).
- How about choosing $\mathcal{M}$ using lasso? – <u>*Lasso post-selection inference*</u>.
  - With the lasso (and some mild conditions), $|\mathcal{M}| < n$.

Problems in Lasso Post-Selection Inference

1. Lasso is not guaranteed to pick all $\beta_j^* \neq 0$.

Problems in Lasso Post-Selection Inference

1. Lasso is not guaranteed to pick all $\beta_j^* \neq 0$.
    - We may answer a different question from our question of interest due to confounding!

Problems in Lasso Post-Selection Inference

1. Lasso is not guaranteed to pick all $\beta_j^* \neq 0$.
   - We may answer a different question from our question of interest due to confounding!
   - Argument for the usefulness of post-selection inference: in practice, we almost never get a complete list of relevant variables in their correct forms. So we essentially never answer our question of interest.

Problems in Lasso Post-Selection Inference

1. Lasso is not guaranteed to pick all $\beta_j^* \neq 0$.
   - We may answer a different question from our question of interest due to confounding!
   - Argument for the usefulness of post-selection inference: in practice, we almost never get a complete list of relevant variables in their correct forms. So we essentially never answer our question of interest.
   - "All models are wrong, but some are useful." – George E. P. Box

Problems in Lasso Post-Selection Inference

2. We peek the data twice!
   - Once in lasso variable selection and once in OLS sub-model fit.

[1] Pötscher (1991); Kabaila (1998); Leeb and Pötscher (2003, 2005, 2006a,b, 2008); Kabaila (2009); Berk et al. (2013)

Problems in Lasso Post-Selection Inference

2 We peek the data twice!
   - Once in lasso variable selection and once in OLS sub-model fit.
   - Estimates from the OLS sub-model fit are no longer Gaussian[1].

[1] Pötscher (1991); Kabaila (1998); Leeb and Pötscher (2003, 2005, 2006a,b, 2008); Kabaila (2009); Berk et al. (2013)

Problems in Lasso Post-Selection Inference

2 We peek the data twice!

- Once in lasso variable selection and once in OLS sub-model fit.
- Estimates from the OLS sub-model fit are no longer Gaussian[1].
- Solutions:
  - Sample-splitting: e.g., Cox (1975); Wasserman and Roeder (2009).
  - Exact inference: e.g., Lee et al. (2016); Tibshirani et al. (2016).

[1] Pötscher (1991); Kabaila (1998); Leeb and Pötscher (2003, 2005, 2006a,b, 2008); Kabaila (2009); Berk et al. (2013)

How Bad is Double-Peeking?

- Leeb et al. (2015) performed simulations to examine the coverage probability of *naïve confidence intervals* – CIs directly from OLS.

How Bad is Double-Peeking?

- Leeb et al. (2015) performed simulations to examine the coverage probability of *naïve confidence intervals* – CIs directly from OLS.
- They found that ignoring the issue of double-peeking is
  - bad if we use best subset selection to choose variables.
    - Coverage probability on $\boldsymbol{\beta}^{\mathcal{M}} \approx 0.87$ for 95% CIs.

How Bad is Double-Peeking?

- Leeb et al. (2015) performed simulations to examine the coverage probability of *naïve confidence intervals* – CIs directly from OLS.
- They found that ignoring the issue of double-peeking is
  - bad if we use best subset selection to choose variables.
    - Coverage probability on $\boldsymbol{\beta}^{\mathcal{M}} \approx 0.87$ for 95% CIs.
  - OK if we use lasso to choose variables!
    - Coverage probability on $\boldsymbol{\beta}^{\mathcal{M}} \approx 0.94$ for 95% CIs.

How Bad is Double-Peeking?

- Leeb et al. (2015) performed simulations to examine the coverage probability of *naïve confidence intervals* – CIs directly from OLS.
- They found that ignoring the issue of double-peeking is
  - bad if we use best subset selection to choose variables.
    - Coverage probability on $\boldsymbol{\beta}^{\mathcal{M}} \approx 0.87$ for 95% CIs.
  - OK if we use lasso to choose variables!
    - Coverage probability on $\boldsymbol{\beta}^{\mathcal{M}} \approx 0.94$ for 95% CIs.
- Why???

In This Talk

1. Show theoretical results that explain findings by Leeb et al. (2015).

In This Talk

1. Show theoretical results that explain findings by Leeb et al. (2015).
    - Yes, it is OK to use lasso to select variables and then perform naïve OLS to draw inference!

In This Talk

1. Show theoretical results that explain findings by Leeb et al. (2015).
    - Yes, it is OK to use lasso to select variables and then perform naïve OLS to draw inference!
    - (asymptotically, with appropriate assumptions...)

In This Talk

1. Show theoretical results that explain findings by Leeb et al. (2015).
   - Yes, it is OK to use lasso to select variables and then perform naïve OLS to draw inference!
   - (asymptotically, with appropriate assumptions...)
2. Based on the theoretical result, propose the *lasso score test*, which draws inference on $\boldsymbol{\beta}^*$.

In This Talk

1. Show theoretical results that explain findings by Leeb et al. (2015).
   - Yes, it is OK to use lasso to select variables and then perform naïve OLS to draw inference!
   - (asymptotically, with appropriate assumptions...)
2. Based on the theoretical result, propose the *lasso score test*, which draws inference on $\boldsymbol{\beta}^*$.
   - Recall that lasso post-selection inference examines hypotheses related to the sub-model.

Normality of Post-Selection OLS

- Let $\hat{\mathcal{A}}_\lambda$ be the set of selected variables by the lasso.
- Let $\tilde{\boldsymbol{\beta}}_\lambda$ be the OLS estimate on selected variables.

Normality of Post-Selection OLS

- Let $\hat{\mathcal{A}}_\lambda$ be the set of selected variables by the lasso.
- Let $\tilde{\boldsymbol{\beta}}_\lambda$ be the OLS estimate on selected variables.

Theorem

With appropriate assumptions (described later), for any $j \in \hat{\mathcal{A}}_\lambda$,

$$\frac{\tilde{\beta}_{\lambda,j} - \beta_j^{\hat{\mathcal{A}}_\lambda}}{\sigma_\epsilon \sqrt{\left[(\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda}^\top \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda})^{-1} \right]_{(j,j)}}} \to_d \mathcal{N}(0,1),$$

where σ_ϵ is the error standard deviation; $\boldsymbol{\beta}^{\hat{\mathcal{A}}_\lambda}$ is the sub-model coefficients.

Comments

- The same distribution as if $\hat{\mathcal{A}}_\lambda$ is selected without looking at the data.

Comments

- The same distribution as if $\hat{\mathcal{A}}_\lambda$ is selected without looking at the data.
- We can use naïve confidence intervals to achieve asymptotically correct coverage on $\boldsymbol{\beta}^{\hat{\mathcal{A}}_\lambda}$,

$$\boldsymbol{y} = \boldsymbol{X}_{\hat{\mathcal{A}}_\lambda} \boldsymbol{\beta}^{\hat{\mathcal{A}}_\lambda} + \boldsymbol{\epsilon}.$$

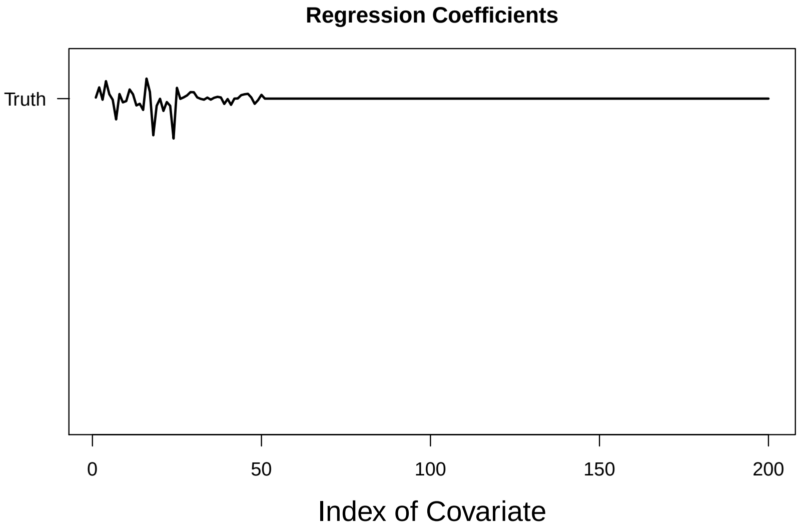
Idea Behind the Result

- Lasso selected active set is fixed with high probability.
  - No additional randomness introduced by lasso variable selection.

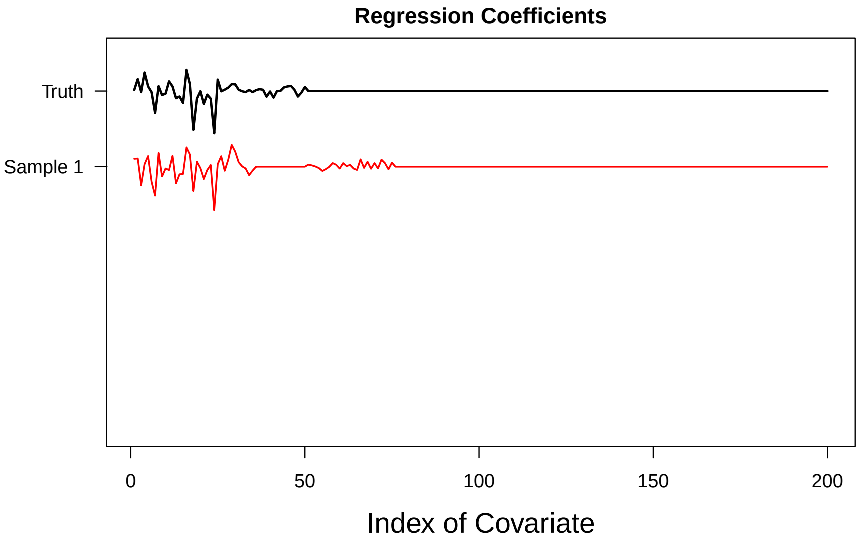
Idea Behind the Result

- Lasso selected active set is fixed with high probability.
  - No additional randomness introduced by lasso variable selection.
  - We effectively only look at the data once!

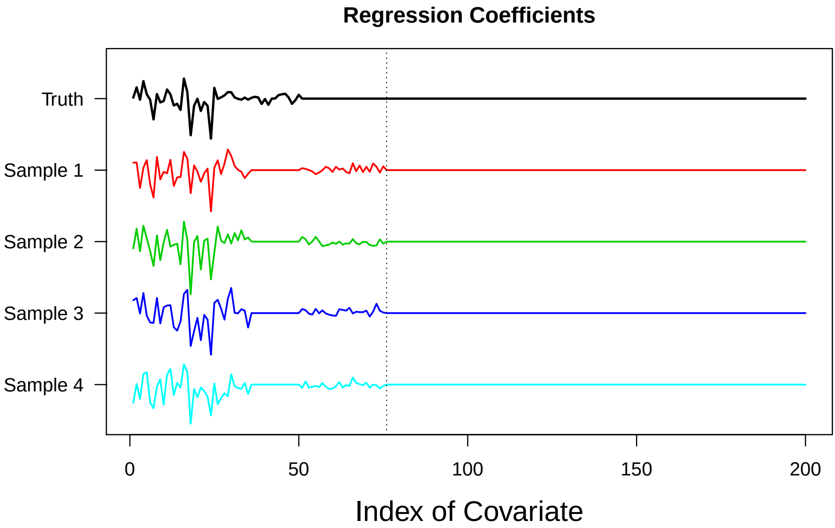
Illustration



Illustration



Illustration



Main Theorem

Theorem

Let $\hat{\mathcal{A}}_\lambda \equiv \mathrm{supp}(\hat{\boldsymbol{\beta}}_\lambda)$ *and* $\mathcal{A}_\lambda \equiv \mathrm{supp}(\boldsymbol{\beta}_\lambda)$, *where*

$$\text{(lasso):} \qquad \hat{\boldsymbol{\beta}}_\lambda \equiv \operatorname*{arg\,min}_{\boldsymbol{b}\in\mathbb{R}^p} \left\{ \frac{1}{2n} \left\| \boldsymbol{y} - \boldsymbol{X}\boldsymbol{b} \right\|_2^2 + \lambda \left\| \boldsymbol{b} \right\|_1 \right\},$$

$$\text{(noiseless lasso):} \qquad \boldsymbol{\beta}_\lambda \equiv \operatorname*{arg\,min}_{\boldsymbol{b}\in\mathbb{R}^p} \left\{ \frac{1}{2n} \mathrm{E} \left[\left\| \boldsymbol{y} - \boldsymbol{X}\boldsymbol{b} \right\|_2^2 \right] + \lambda \left\| \boldsymbol{b} \right\|_1 \right\}.$$

With appropriate conditions (described later), we have

$$\lim_{n\to\infty} \Pr \left[\hat{\mathcal{A}}_\lambda = \mathcal{A}_\lambda \right] = 1.$$

Main Theorem

Theorem

Let $\hat{\mathcal{A}}_\lambda \equiv \mathrm{supp}(\hat{\boldsymbol{\beta}}_\lambda)$ *and* $\mathcal{A}_\lambda \equiv \mathrm{supp}(\boldsymbol{\beta}_\lambda)$, *where*

(lasso): $$\hat{\boldsymbol{\beta}}_\lambda \equiv \arg\min_{\boldsymbol{b}\in\mathbb{R}^p} \left\{ \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|_2^2 + \lambda \|\boldsymbol{b}\|_1 \right\},$$

(noiseless lasso): $$\boldsymbol{\beta}_\lambda \equiv \arg\min_{\boldsymbol{b}\in\mathbb{R}^p} \left\{ \frac{1}{2n} \mathrm{E}\left[\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|_2^2 \right] + \lambda \|\boldsymbol{b}\|_1 \right\}.$$

With appropriate conditions (described later), we have

$$\lim_{n\to\infty} \Pr\left[\hat{\mathcal{A}}_\lambda = \mathcal{A}_\lambda\right] = 1.$$

- This is not the same as $\lim_{n\to\infty} \Pr[\hat{\mathcal{A}}_\lambda = \mathrm{supp}(\boldsymbol{\beta}^*)] = 1$!

Main Theorem

Theorem

Let $\hat{\mathcal{A}}_\lambda \equiv \mathrm{supp}(\hat{\boldsymbol{\beta}}_\lambda)$ *and* $\mathcal{A}_\lambda \equiv \mathrm{supp}(\boldsymbol{\beta}_\lambda)$, *where*

$$\text{(lasso):} \qquad \hat{\boldsymbol{\beta}}_\lambda \equiv \arg\min_{\boldsymbol{b}\in\mathbb{R}^p} \left\{ \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|_2^2 + \lambda \|\boldsymbol{b}\|_1 \right\},$$

$$\text{(noiseless lasso):} \qquad \boldsymbol{\beta}_\lambda \equiv \arg\min_{\boldsymbol{b}\in\mathbb{R}^p} \left\{ \frac{1}{2n} \mathrm{E}\left[\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|_2^2 \right] + \lambda \|\boldsymbol{b}\|_1 \right\}.$$

With appropriate conditions (described later), we have

$$\lim_{n\to\infty} \Pr\left[\hat{\mathcal{A}}_\lambda = \mathcal{A}_\lambda \right] = 1.$$

- This is not the same as $\lim_{n\to\infty} \Pr[\hat{\mathcal{A}}_\lambda = \mathrm{supp}(\boldsymbol{\beta}^*)] = 1$!
- To recover the true support requires stronger conditions (Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006; Wainwright, 2009).

Assumptions

- Fixed design matrix $\boldsymbol{X}$ – can be extended to random $\boldsymbol{X}$.

Assumptions

- Fixed design matrix $\boldsymbol{X}$ – can be extended to random $\boldsymbol{X}$.
- Sub-Gaussian noise ϵ.

Assumptions

- Fixed design matrix $\boldsymbol{X}$ – can be extended to random $\boldsymbol{X}$.
- Sub-Gaussian noise ϵ.
- Restricted Eigenvalue Condition (Bickel et al., 2009; van de Geer and Bühlmann, 2009).

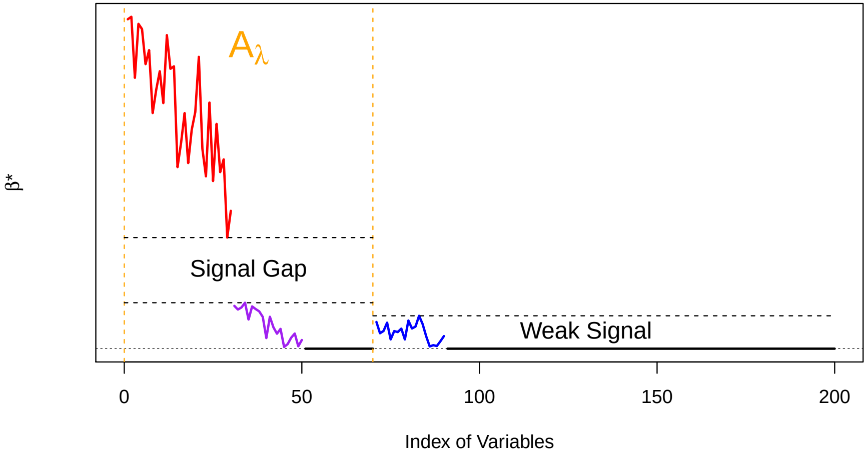
Assumptions

- Fixed design matrix $\boldsymbol{X}$ – can be extended to random $\boldsymbol{X}$.
- Sub-Gaussian noise ϵ.
- Restricted Eigenvalue Condition (Bickel et al., 2009; van de Geer and Bühlmann, 2009).
- $\lambda \succ_{asy} \sqrt{\log(p)/n}$.
    - The tuning parameter should converge to zero (slightly) slower than the usual estimation-optimal and prediction-optimal rate.

Assumptions

- Fixed design matrix $\boldsymbol{X}$ – can be extended to random $\boldsymbol{X}$.
- Sub-Gaussian noise ϵ.
- Restricted Eigenvalue Condition (Bickel et al., 2009; van de Geer and Bühlmann, 2009).
- $\lambda \succ_{asy} \sqrt{\log(p)/n}$.
  - The tuning parameter should converge to zero (slightly) slower than the usual estimation-optimal and prediction-optimal rate.
- A technical assumption on the subgradient of the noiseless lasso.

Assumption on β^*



Simulations 1 - Setting

- $p = 100, n \in \{300, 400, 500\}$.

Simulations 1 - Setting

- $p = 100, n \in \{300, 400, 500\}$.
- Power-law graph with edge density 0.05.

Simulations 1 - Setting

- $p = 100, n \in \{300, 400, 500\}$.
- Power-law graph with edge density 0.05.
- Let the graph adjacency matrix be the partial correlation matrix of $\boldsymbol{X}$, with partial correlation $\rho = 0.2$.

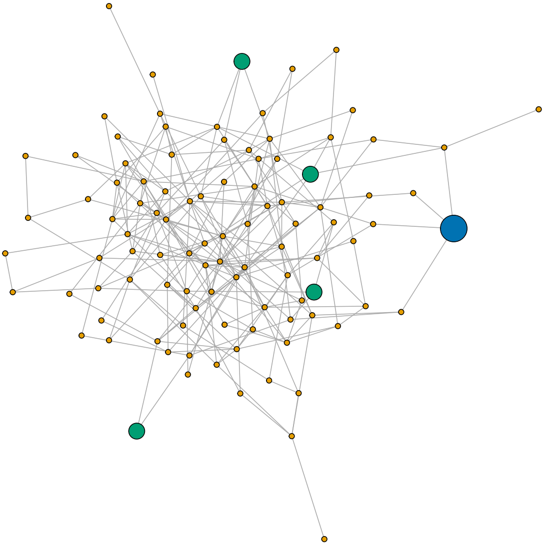
Simulations 1 - Setting

- $p = 100, n \in \{300, 400, 500\}$.
- Power-law graph with edge density 0.05.
- Let the graph adjacency matrix be the partial correlation matrix of $\boldsymbol{X}$, with partial correlation $\rho = 0.2$.
- Simulation $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$, with

$$\beta_j^* = \begin{cases} 1 & j = 10 \\ 0.1 & j \in \{20, 30, 40, 50\} \\ 0 & \text{otherwise} \end{cases},$$

and signal to noise ratio $SNR \in \{0.1, 0.3, 0.5\}$

Simulations 1 - Graph Structure



Simulations 1 - Coverage Probability

n	300			400			500		
SNR	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
exact	0.952	0.949	0.950	0.953	0.951	0.951	0.947	0.946	0.946
naïve	0.914	0.932	0.930	0.944	0.932	0.924	0.941	0.940	0.932

Simulations 2 - Setting

- $p = 100$, $n \in \{300, 400, 500\}$.

Simulations 2 - Setting

- $p = 100, n \in \{300, 400, 500\}$.
- Two graphs with 5 and 95 nodes, respectively, and edge density 0.3.

Simulations 2 - Setting

- $p = 100, n \in \{300, 400, 500\}$.
- Two graphs with 5 and 95 nodes, respectively, and edge density 0.3.
- Connect the two dense graphs using edges with density 0.05.

Simulations 2 - Setting

- $p = 100, n \in \{300, 400, 500\}$.
- Two graphs with 5 and 95 nodes, respectively, and edge density 0.3.
- Connect the two dense graphs using edges with density 0.05.
- Let the graph adjacency matrix be the partial correlation matrix of $\boldsymbol{X}$, with partial correlation $\rho = 0.2$.

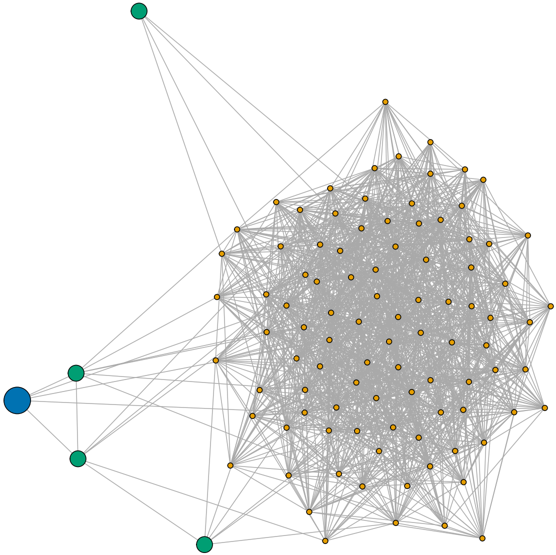
Simulations 2 - Setting

- $p = 100, n \in \{300, 400, 500\}$.
- Two graphs with 5 and 95 nodes, respectively, and edge density 0.3.
- Connect the two dense graphs using edges with density 0.05.
- Let the graph adjacency matrix be the partial correlation matrix of $\boldsymbol{X}$, with partial correlation $\rho = 0.2$.
- Simulation $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$, with

$$\beta_j^* = \begin{cases} 1 & j = 1 \\ 0.1 & j \in \{2, 3, 4, 5\} \\ 0 & \text{otherwise} \end{cases},$$

  and signal to noise ratio $SNR \in \{0.1, 0.3, 0.5\}$

Simulations 2 - Graph Structure



Simulations 2 - Coverage Probability

n	300			400			500		
SNR	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
exact	0.952	0.944	0.944	0.950	0.951	0.953	0.951	0.956	0.957
naïve	0.922	0.923	0.916	0.935	0.937	0.933	0.952	0.947	0.937

Inference on the Full Model Coefficients $\boldsymbol{\beta}^*$

- Recall that post-selection inference procedures make inference on $\boldsymbol{\beta}^{\mathcal{M}}$.

$$\boldsymbol{y} = \boldsymbol{X}_{\mathcal{M}}\boldsymbol{\beta}^{\mathcal{M}} + \boldsymbol{\epsilon}.$$

Inference on the Full Model Coefficients $\boldsymbol{\beta}^*$

- Recall that post-selection inference procedures make inference on $\boldsymbol{\beta}^{\mathcal{M}}$.

$$\boldsymbol{y} = \boldsymbol{X}_{\mathcal{M}}\boldsymbol{\beta}^{\mathcal{M}} + \boldsymbol{\epsilon}.$$

- $\boldsymbol{\beta}^{\mathcal{M}}$ in general is not the same as the fully adjusted regression coefficients, unless lasso does not miss any (strong) confounding variables.

Inference on the Full Model Coefficients $\boldsymbol{\beta}^*$

- Recall that post-selection inference procedures make inference on $\boldsymbol{\beta}^{\mathcal{M}}$.

$$\boldsymbol{y} = \boldsymbol{X}_{\mathcal{M}}\boldsymbol{\beta}^{\mathcal{M}} + \boldsymbol{\epsilon}.$$

- $\boldsymbol{\beta}^{\mathcal{M}}$ in general is not the same as the fully adjusted regression coefficients, unless lasso does not miss any (strong) confounding variables.
    - It is OK to miss some very weak confounding variables, as long as asymptotically their contributed bias is ignorable.

Inference on the Full Model Coefficients $\boldsymbol{\beta}^*$

- How to make sure lasso does not miss any strong confounders?

Inference on the Full Model Coefficients β^*

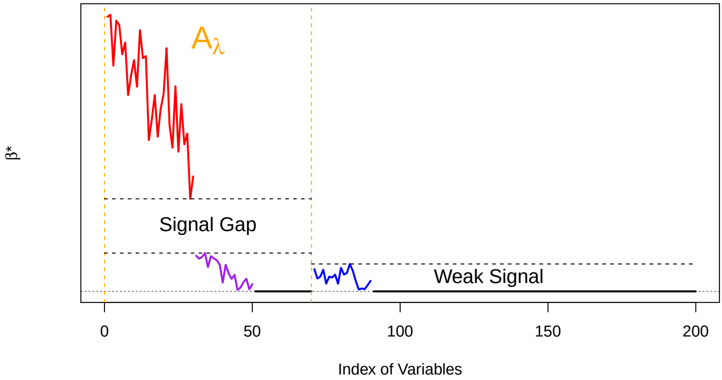
- How to make sure lasso does not miss any strong confounders?
- Recall the set of variable selected by the lasso is the same as the set of variable selected by the noiseless lasso with high probability.

Inference on the Full Model Coefficients β^*

- How to make sure lasso does not miss any strong confounders?
- Recall the set of variable selected by the lasso is the same as the set of variable selected by the noiseless lasso with high probability.
- We can assume the noiseless lasso does not miss any strong confounding variables.

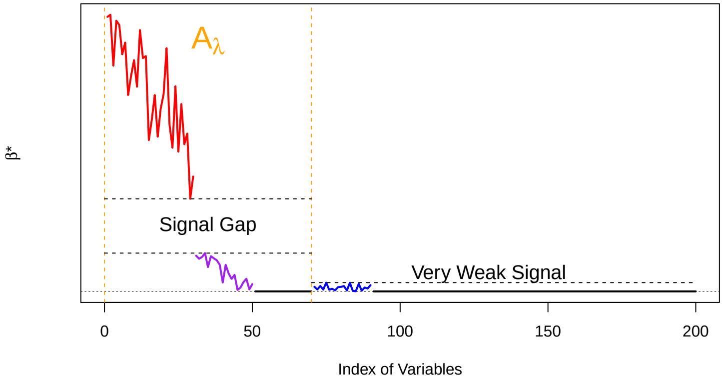
Assumption on $\boldsymbol{\beta}^*$

Inference on $\boldsymbol{\beta}^{\mathcal{A}}$



Assumption on β^*

Inference on β^*



Lasso Score Test

- Let
$$\tilde{S}_j \equiv \boldsymbol{x}_j^\top \left(\boldsymbol{y} - \boldsymbol{X} \tilde{\beta}_\lambda^0 \right),$$
where $\tilde{\beta}_\lambda^0$ is the OLS estimate with $\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda \setminus \{j\}}$.

Lasso Score Test

- Let
$$\tilde{S}_j \equiv \boldsymbol{x}_j^\top \left(\boldsymbol{y} - \boldsymbol{X} \tilde{\beta}_\lambda^0 \right),$$
where $\tilde{\beta}_\lambda^0$ is the OLS estimate with $\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda \setminus \{j\}}$.
- Under $H_0 : \beta_j^* = 0$, $\tilde{S}_j$ is asymptotically normally distributed with mean zero.

Lasso Score Test

- Let
$$\tilde{S}_j \equiv \boldsymbol{x}_j^\top \left(\boldsymbol{y} - \boldsymbol{X}\tilde{\beta}_\lambda^0 \right),$$
where $\tilde{\beta}_\lambda^0$ is the OLS estimate with $\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda \setminus \{j\}}$.
- Under $H_0 : \beta_j^* = 0$, $\tilde{S}_j$ is asymptotically normally distributed with mean zero.
  - $\tilde{S}_j$ follows the same distribution as if $\hat{\mathcal{A}}_\lambda$ is chosen without seeing data.

Lasso Score Test

- Let
$$\tilde{S}_j \equiv \boldsymbol{x}_j^\top \left(\boldsymbol{y} - \boldsymbol{X} \tilde{\beta}_\lambda^0 \right),$$
where $\tilde{\beta}_\lambda^0$ is the OLS estimate with $\boldsymbol{X}_{\hat{\mathcal{A}}_\lambda \setminus \{j\}}$.
- Under $H_0 : \beta_j^* = 0$, $\tilde{S}_j$ is asymptotically normally distributed with mean zero.
  - $\tilde{S}_j$ follows the same distribution as if $\hat{\mathcal{A}}_\lambda$ is chosen without seeing data.
  - Lasso score test derives p-values for all variables, including those with have zero lasso regression coefficients.

Simulations 1 - Setting

- $p = 500$, $n \in \{100, 200, 400\}$.

Simulations 1 - Setting

- $p = 500, n \in \{100, 200, 400\}$.
- Power-law graph with edge density 0.05.

Simulations 1 - Setting

- $p = 500, n \in \{100, 200, 400\}$.
- Power-law graph with edge density 0.05.
- Let the graph adjacency matrix be the partial correlation matrix of $\boldsymbol{X}$, with partial correlation $\rho = 0.2$.

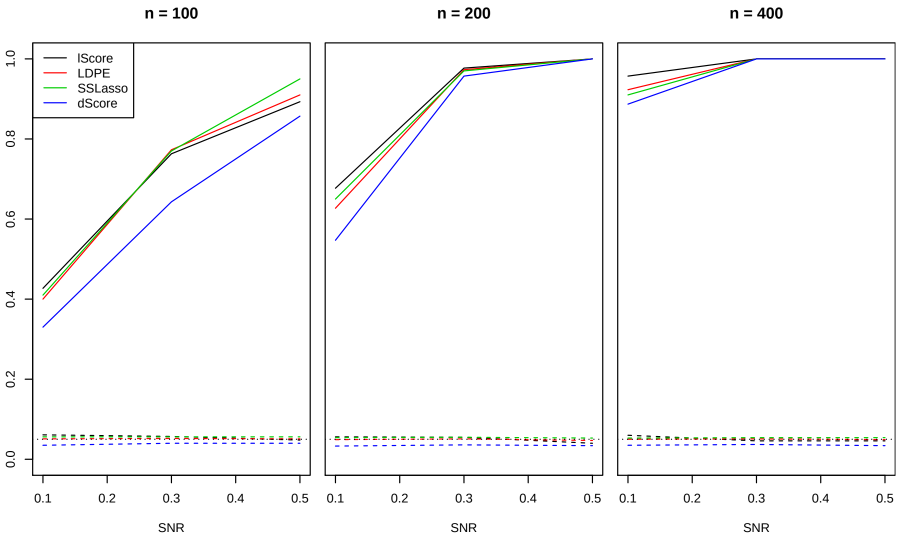
Simulations 1 - Setting

- $p = 500, n \in \{100, 200, 400\}$.
- Power-law graph with edge density 0.05.
- Let the graph adjacency matrix be the partial correlation matrix of $\boldsymbol{X}$, with partial correlation $\rho = 0.2$.
- Simulation $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$, with

$$\beta_j^* = \begin{cases} 1 & j \in \{30, 60, 90\} \\ 0.1 & j \in \{120, 150, 180, 210, 240, 270, 300\} \\ 0 & \text{otherwise} \end{cases},$$

  and signal to noise ratio $SNR \in \{0.1, 0.3, 0.5\}$

Simulations 1 - Power and Type-I Error Rate



Simulations 2 - Setting

- $p = 500$, $n \in \{100, 200, 400\}$.

Simulations 2 - Setting

- $p = 500, n \in \{100, 200, 400\}$.
- Two degree-free graphs with 10 and 490 nodes, respectively, and edge density 0.3.

Simulations 2 - Setting

- $p = 500, n \in \{100, 200, 400\}$.
- Two degree-free graphs with 10 and 490 nodes, respectively, and edge density 0.3.
- Connect the two dense graphs using edges with density 0.05.

Simulations 2 - Setting

- $p = 500, n \in \{100, 200, 400\}$.
- Two degree-free graphs with 10 and 490 nodes, respectively, and edge density 0.3.
- Connect the two dense graphs using edges with density 0.05.
- Let the graph adjacency matrix be the partial correlation matrix of $\boldsymbol{X}$, with partial correlation $\rho = 0.2$.

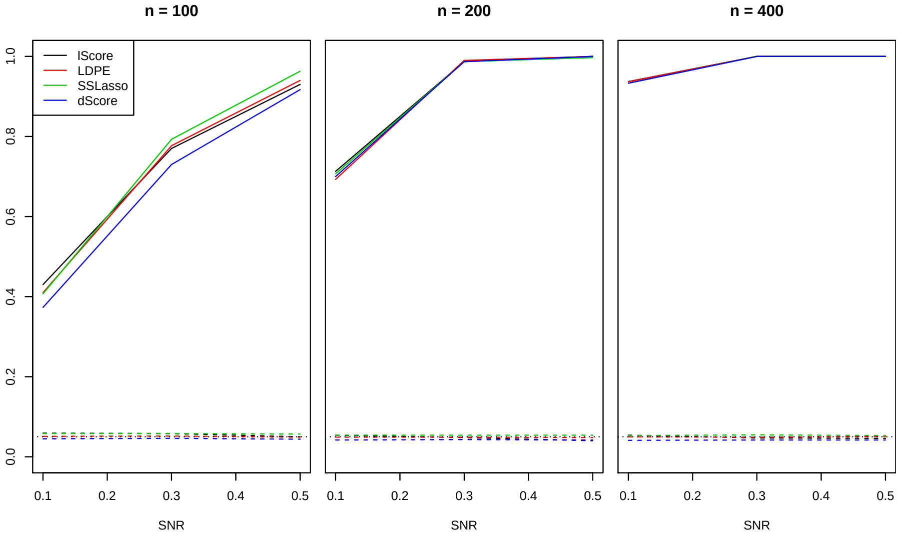
Simulations 2 - Setting

- $p = 500, n \in \{100, 200, 400\}$.
- Two degree-free graphs with 10 and 490 nodes, respectively, and edge density 0.3.
- Connect the two dense graphs using edges with density 0.05.
- Let the graph adjacency matrix be the partial correlation matrix of $\boldsymbol{X}$, with partial correlation $\rho = 0.2$.
- Simulation $\boldsymbol{y} = \boldsymbol{X}\beta^* + \epsilon$, with

$$\beta_j^* = \begin{cases} 1 & j = \{1, 2, 3\} \\ 0.1 & j \in \{4, 5, 6, 7, 8, 9, 10\} \\ 0 & \text{otherwise} \end{cases},$$

and signal to noise ratio $SNR \in \{0.1, 0.3, 0.5\}$

Simulations 2 - Power and Type-I Error Rate



Discussion

- Empirical evidence suggests the largest λ whose MSE is within one standard error of the minimum MSE (Hastie et al., 2009) works well.
  - λ should be larger than the prediction optimal ones.

Discussion

- Empirical evidence suggests the largest λ whose MSE is within one standard error of the minimum MSE (Hastie et al., 2009) works well.
  - λ should be larger than the prediction optimal ones.
- Lasso score test is closely related to both post-selection inference and debiased tests.
  - The two classes of inference procedures are in fact very similar.

A Bigger Picture

Classical statistical inference:

1. Formulate the model: what variables to adjust for? in what form, e.g., linear, polynomial, log-transformed, interactions?
2. Collect data.
3. Test hypothesis.

A Bigger Picture

Classical statistical inference:

1. Formulate the model: what variables to adjust for? in what form, e.g., linear, polynomial, log-transformed, interactions?
2. Collect data.
3. Test hypothesis.

Post-selection inference:

1. Collect data.
2. Find the "best" model based on the data.
3. Test hypothesis.

A Bigger Picture

Classical statistical inference:

1. Formulate the model: what variables to adjust for? in what form, e.g., linear, polynomial, log-transformed, interactions?
2. Collect data.
3. Test hypothesis.

Post-selection inference:

1. Collect data.
2. Find the "best" model based on the data.
3. Test hypothesis.

More efficient use of data, faster iterations...

A Bigger Picture

Classical statistical inference:

1. Formulate the model: what variables to adjust for? in what form, e.g., linear, polynomial, log-transformed, interactions?
2. Collect data.
3. Test hypothesis.

Post-selection inference:

1. Collect data.
2. Find the "best" model based on the data.
3. Test hypothesis.

More efficient use of data, faster iterations...

Bottomline: post-selection inference is a promising area of research; there are still lots of questions to be solved...

References I

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837.

Bickel, P., Ritov, Y., and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732.

Cox, D. R. (1975). A note on data-splitting for the evaluation of significance levels. *Biometrika*, 62(2):441–444.

Dezeure, R., Bühlmann, P., Meier, L., and van de Geer, S. (2015). High-dimensional inference: Confidence intervals, *p*-values and R-software `hdi`. *Statistical Science*, 30(4):533–558.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer-Verlag, New York.

Javanmard, A. and Montanari, A. (2013). Confidence intervals and hypothesis testing for high-dimensional statistical models. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 26*, pages 1187–1195. Curran Associates, Inc.

Javanmard, A. and Montanari, A. (2014a). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15(Oct):2869–2909.

References II

Javanmard, A. and Montanari, A. (2014b). Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *IEEE Transaction on Information Theory*, 60(10):6522 – 6554.

Kabaila, P. (1998). Valid confidence intervals in regression after variable selection. *Econometric Theory*, 14(4):463–482.

Kabaila, P. (2009). The coverage properties of confidence regions after model selection. *International Statistical Review*, 77(3):405–414.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927.

Leeb, H. and Pötscher, B. M. (2003). The finite-sample distribution of post-model-selection estimators and uniform versus nonuniform approximations. *Econometric Theory*, 19(1):100–142.

Leeb, H. and Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, 21(1):21–59.

Leeb, H. and Pötscher, B. M. (2006a). Can one estimate the conditional distribution of post-model-selection estimators? *The Annals of Statistics*, 34(5):2554–2591.

References III

Leeb, H. and Pötscher, B. M. (2006b). Performance limits for estimators of the risk or distribution of shrinkage-type estimators, and some general lower risk-bound results. *Econometric Theory*, 22(1):69–97.

Leeb, H. and Pötscher, B. M. (2008). Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory*, 24(2):338–376.

Leeb, H., Pötscher, B. M., and Ewald, K. (2015). On various confidence intervals post-model-selection. *Statistical Science*, 30(2):216–227.

Meinshausen, N. and Bühlmann, P. (2006). High dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462.

Ning, Y. and Liu, H. (2016). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics*, to appear.

Pötscher, B. M. (1991). Effects of model selection on inference. *Econometric Theory*, 7(2):163–185.

Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514):600–620.

van de Geer, S. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electronic Journals of Statistics*, 3:1360–1392.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programmming (lasso). *IEEE Transaction on Information Theory*, 55(5):2183 – 2202.

Wasserman, L. and Roeder, K. (2009). High-dimensional variable selection. *The Annals of Statistics*, 37(5A):2178–2201.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B*, 76(1):217–242.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, 7(Nov):2541–2563.

Zhao, S. and Shojaie, A. (2016). A significance test for graph-constrained estimation. *Biometrics*, 72(2):484–493.

Thanks!

Questions?

Assumption on the Subgradient

- Denote $\hat{\boldsymbol{\Sigma}} = \boldsymbol{X}^\top \boldsymbol{X}/n$. Define the subgradient of the noiseless lasso

$$\boldsymbol{\tau}_\lambda = \frac{1}{\lambda} \hat{\boldsymbol{\Sigma}} \left(\boldsymbol{\beta}^* - \boldsymbol{\beta}_\lambda \right).$$

Then,

$$\limsup_{n\to\infty} \|\boldsymbol{\tau}_{\lambda,\mathcal{A}_\lambda^c}\|_\infty < 1,$$

$$\frac{1}{\min_{j\in\mathcal{A}_\lambda\setminus\mathcal{S}^*} \left| \left(\hat{\boldsymbol{\Sigma}}_{(\mathcal{A}_\lambda,\mathcal{A}_\lambda)} \right)^{-1} \boldsymbol{\tau}_{\lambda,\mathcal{A}_\lambda} \right|_j} = \mathcal{O}\left(\lambda \sqrt{\frac{n}{\log(p)}} \right),$$

where $\mathcal{S}^*$ is the set of strong signal variables.