Thresholded Lasso for High Dimensional Variable Selection

Shuheng Zhou

Seminar for Statistics, ETH Zürich, Switzerland

Department of Statistics, University of California, Berkeley

Universit Paris-Est Marne-la-Vallée

May 20, 2010

Want to estimate a parameter $\beta \in \mathbf{R}^p$

Want to estimate a parameter $\beta \in \mathbf{R}^p$

- Example: How is a response $y \in \mathbf{R}$ related to the Parkinson's disease affected by a set of genes among the Chinese population?

Want to estimate a parameter $\beta \in \mathbf{R}^p$

- Example: How is a response $y \in \mathbf{R}$ related to the Parkinson's disease affected by a set of genes among the Chinese population?
- Construct a linear model: $y = \beta^T \vec{x} + \epsilon$, where $\mathbb{E}\left(y|\vec{x}\right) = \beta^T \vec{x}$.

Want to estimate a parameter $\beta \in \mathbf{R}^p$

- Example: How is a response $y \in \mathbf{R}$ related to the Parkinson's disease affected by a set of genes among the Chinese population?
- Construct a linear model: $y = \beta^T \vec{x} + \epsilon$, where $\mathbb{E}\left(y|\vec{x}\right) = \beta^T \vec{x}$.
  - Parameter: Non-zero entries in β (sparsity of β) identify a subset of genes and indicate how much they influence y.

Want to estimate a parameter $\beta \in \mathbf{R}^p$

- Example: How is a response $y \in \mathbf{R}$ related to the Parkinson's disease affected by a set of genes among the Chinese population?
- Construct a linear model: $y = \beta^T \vec{x} + \epsilon$, where $\mathbb{E}\left(y | \vec{x}\right) = \beta^T \vec{x}$.
  - Parameter: Non-zero entries in β (sparsity of β) identify a subset of genes and indicate how much they influence y.
- Take a random sample of (X, Y), and use the sample to estimate β; that is, we have $Y = X\beta + \epsilon$.

High dimensional linear model

Consider recovering $\beta \in \mathbf{R}^p$ in the following noisy linear model:

$$\begin{bmatrix} \\ Y \\ \\ \end{bmatrix}_n = \begin{bmatrix} & & \\ & X & \\ & & \end{bmatrix}_{n \times p} \begin{bmatrix} \\ \beta \\ \\ \end{bmatrix}_p + \begin{bmatrix} \\ \epsilon \\ \\ \end{bmatrix}_n$$

where we assume $p \gg n$ (i.e. given high-dimensional data).

High dimensional linear model

Consider recovering $\beta \in \mathbf{R}^p$ in the following noisy linear model:

$$\begin{bmatrix} \\ Y \\ \\ \end{bmatrix}_n = \begin{bmatrix} & & \\ & X & \\ & & \end{bmatrix}_{n \times p} \begin{bmatrix} \\ \beta \\ \\ \\ \end{bmatrix}_p + \begin{bmatrix} \\ \epsilon \\ \\ \end{bmatrix}_n$$

where we assume $p \gg n$ (i.e. given high-dimensional data).

- The paradigm has shifted to the setting where the dimensionality is much larger than the number of observations. Think of n, p as moderately large, e.g., between 10^3 to 10^6.

High dimensional linear model

Goal: to recover the unknown $\beta \in \mathbf{R}^p$ approximately from noisy data using computational feasible strategies,

$$\begin{bmatrix} \\ Y \\ \\ \end{bmatrix}_n = \begin{bmatrix} & & \\ & X & \\ & & \end{bmatrix}_{n \times p} \begin{bmatrix} \\ \beta \\ \\ \end{bmatrix}_p + \begin{bmatrix} \\ \epsilon \\ \\ \end{bmatrix}_n$$

where we assume $p \geq n$ (i.e., given high-dimensional data X).

High dimensional linear model

Goal: to recover the unknown $\beta \in \mathbf{R}^p$ approximately from noisy data using computational feasible strategies,

$$\begin{bmatrix} Y \end{bmatrix}_n = \begin{bmatrix} \quad X \quad \end{bmatrix}_{n \times p} \begin{bmatrix} \beta \end{bmatrix}_p + \begin{bmatrix} \epsilon \end{bmatrix}_n$$

where we assume $p \geq n$ (i.e., given high-dimensional data X).

- X has columns normalized to have ℓ_2 norm $\sqrt{n}$, and ϵ is the Gaussian noise: $\epsilon \sim N(0, \sigma^2 I_n)$.

Model selection and parameter estimation

When can we approximately recover β from n noisy observations Y?

- **Questions:** How many measurements n do we need in order to recover the non-zero positions in β?

Model selection and parameter estimation

When can we approximately recover β from n noisy observations Y?

- **Questions:** How many measurements n do we need in order to recover the non-zero positions in β?
- How does n scale with p or s, where s is the number of non-zero entries of β?

Model selection and parameter estimation

When can we approximately recover β from n noisy observations Y?

- **Questions:** How many measurements n do we need in order to recover the non-zero positions in β?
- How does n scale with p or s, where s is the number of non-zero entries of β?
- What if some non-zero entries are really small, say within noise level?

Model selection and parameter estimation

When can we approximately recover β from n noisy observations Y?

- **Questions:** How many measurements n do we need in order to recover the non-zero positions in β?
- How does n scale with p or s, where s is the number of non-zero entries of β?
- What if some non-zero entries are really small, say within noise level?
- What assumptions about the data matrix X are reasonable?

Sparse recovery

When β is known to be s-sparse for some $1 \leq s \leq n$, which means that at most s of the coefficients of β can be non-zero:

- Assume every $2s$ columns of X are linearly independent: Identifiability condition (reasonable once $n \geq 2s$)

$$\Lambda_{\min}(2s) \triangleq \min_{\upsilon \neq 0, 2s\text{-sparse}} \frac{\|X\upsilon\|^2}{n\|\upsilon\|^2} > 0.$$

Sparse recovery

When β is known to be s-sparse for some $1 \le s \le n$, which means that at most s of the coefficients of β can be non-zero:

- Assume every $2s$ columns of X are linearly independent: Identifiability condition (reasonable once $n \ge 2s$)

$$\Lambda_{\min}(2s) \triangleq \min_{\upsilon \neq 0, 2s\text{-sparse}} \frac{\|X\upsilon\|^2}{n\|\upsilon\|^2} > 0.$$

- **Proposition:** (Candès-Tao 05). Suppose that any $2s$ columns of the $n \times p$ matrix X are linearly independent. Then, any s-sparse signal $\beta \in \mathbf{R}^p$ can be reconstructed uniquely from $X\beta$.

ℓ_0-minimization

How to reconstruct an s-sparse signal $\beta \in \mathbf{R}^p$ from the measurements $Y = X\beta$ given $\Lambda_{\min}(2s) > 0$?

- Let β be the unique sparsest solution to $X\beta = Y$:

ℓ_0-minimization

How to reconstruct an s-sparse signal $\beta \in \mathbf{R}^p$ from the measurements $Y = X\beta$ given $\Lambda_{\min}(2s) > 0$?

- Let β be the unique sparsest solution to $X\beta = Y$:

$$\beta = \arg\min_{\beta : X\beta = Y} \|\beta\|_0$$

where $\|\beta\|_0 := \#\{1 \leq i \leq p : \beta_i \neq 0\}$ is the sparsity of β.

ℓ_0-minimization

How to reconstruct an s-sparse signal $\beta \in \mathbf{R}^p$ from the measurements $Y = X\beta$ given $\Lambda_{\min}(2s) > 0$?

- Let β be the unique sparsest solution to $X\beta = Y$:

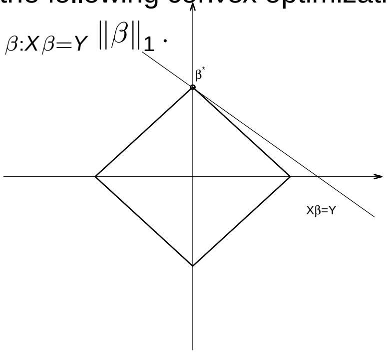
$$\beta = \arg\min_{\beta : X\beta = Y} \|\beta\|_0$$

  where $\|\beta\|_0 := \#\{1 \leq i \leq p : \beta_i \neq 0\}$ is the sparsity of β.

- Unfortunately, ℓ_0-minimization is computationally intractable; (in fact, it is an NP-complete problem).

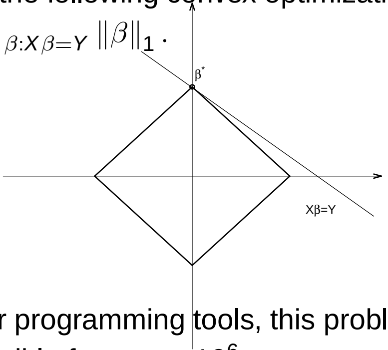
Basis pursuit

- We consider the following convex optimization problem
  $\beta^* := \arg\min_{\beta : X\beta = Y} \|\beta\|_1 .$



Basis pursuit

- We consider the following convex optimization problem $\beta^* := \arg\min_{\beta : X\beta = Y} \|\beta\|_1$.



By standard linear programming tools, this problem is computational feasible for $n, p \sim 10^6$. (This is studied by Chen, Donoho, Huo, Logan, Saunders, Candes, Romberg, Tao and others.)

To acquire the sparse signal β

- Basis pursuit works whenever the $n \times p$ measurement matrix X is sufficiently incoherent:

To acquire the sparse signal β

- Basis pursuit works whenever the $n \times p$ measurement matrix X is sufficiently incoherent:
- RIP (Candès-Tao 05) requires that for all $T \subset \{1, \ldots, p\}$ with $|T| \leq s$ and for all coefficients sequences $(c_j)_{j \in T}$, $(1 - \delta_s) \|c\|^2 \leq \|X_T c/n\|^2 \leq (1 + \delta_s) \|c\|^2$ holds for some $0 < \delta_s < 1$ (s-restricted isometry constant).

Restricted Isometry Property (RIP)

- The “good” matrices for compressed sensing should satisfy the inequalities for the largest possible s:

Restricted Isometry Property (RIP)

- The “good” matrices for compressed sensing should satisfy the inequalities for the largest possible s:
- For example, for Gaussian random matrix, or any sub-Gaussian ensemble, for $0 < \delta_s < 1$, it holds with $s \asymp n / \log(p/n)$.

Restricted Isometry Property (RIP)

- The “good” matrices for compressed sensing should satisfy the inequalities for the largest possible s:
- For example, for Gaussian random matrix, or any sub-Gaussian ensemble, for $0 < \delta_s < 1$, it holds with $s \asymp n/\log(p/n)$.
- These algorithms are also robust with regards to noise, and RIP will be replaced by more relaxed conditions.

Sparse recovery for $Y = X\beta + \epsilon$

- Lasso (Tibshirani 96), a.k.a. Basis Pursuit (Chen, Donoho and Saunders 98, and others):

$$\widetilde{\beta} = \arg\min_{\beta} \|Y - X\beta\|^2/2n + \lambda_n \|\beta\|_1,$$

where the scaling factor $1/(2n)$ is chosen by convenience.

- Dantzig selector (Candès-Tao 07):

$$(DS) \quad \arg\min_{\widetilde{\beta} \in \mathbf{R}^p} \|\widetilde{\beta}\|_1 \quad \text{subject to} \quad \|X^T(Y - X\widetilde{\beta})/n\|_\infty \leq \lambda_n.$$

References: Greenshtein-Ritov 04, Meinshausen-Bühlmann 06, Zhao-Yu 06, Candès-Tao 07, van de Geer 08, Wainwright 09, Koltchinskii 09, Meinshausen-Yu 09, Bickel-Ritov-Tsybakov 09, and others.

When X is a Gaussian random matrix

- Numerical experiments suggest that in practice, most s-sparse signals are in fact recovered exactly once $n \geq 4s$ or so for noiseless model $Y = X\beta$;

When X is a Gaussian random matrix

- Numerical experiments suggest that in practice, most s-sparse signals are in fact recovered exactly once $n \geq 4s$ or so for noiseless model $Y = X\beta$;
- This shows a strong contrast with the ordinary Lasso's behavior in the noisy case:

When X is a Gaussian random matrix

- Numerical experiments suggest that in practice, most s-sparse signals are in fact recovered exactly once $n \geq 4s$ or so for noiseless model $Y = X\beta$;
- This shows a strong contrast with the ordinary Lasso's behavior in the noisy case:
  The lower bound for the Lasso: (Wainwright 09). For the noisy linear model $Y = X\beta + \epsilon$, where $\epsilon \sim N(0, I_p)$. Then the probability of success in terms of exact recovery of the sparsity pattern tends to zero when $n < 2s \log(p - s)$, for any s-sparse vector.

Prelude

Is there a way to bridge the difference?

Prelude

Is there a way to bridge the difference?

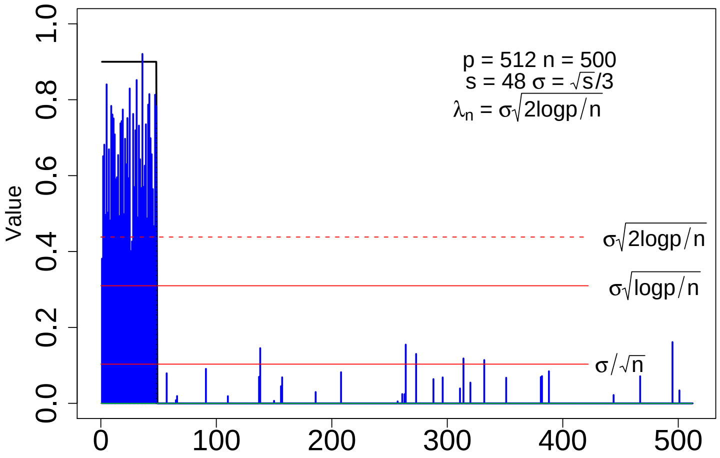
- **Linear sparsity:** How can we design an estimator to can recover a sparse model using nearly a constant number of measurements per non-zero element despite noise?

Prelude

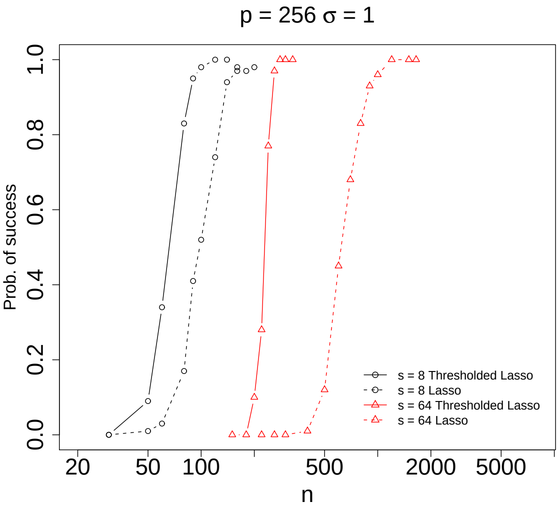
Is there a way to bridge the difference?

- **Linear sparsity:** How can we design an estimator to can recover a sparse model using nearly a constant number of measurements per non-zero element despite noise?
- **More generally:** How to design a sparse estimator whose accuracy depends upon the information content of the object we wish to recover?

Linear sparsity



Compare probability of success for $s = 8$ and 64



The Thresholded Lasso estimator

Define $S = \mathrm{supp}(\beta) := \{j : \beta_j \neq 0\}$, Let $s = |S|$. For some $s_0 \leq s$ to be defined.

- First we obtain an initial estimator β_{init} using the Lasso with $\lambda_n = c\sigma\sqrt{2\log p/n}$ for some constant c.

The Thresholded Lasso estimator

Define $S = \mathrm{supp}(\beta) := \{j : \beta_j \neq 0\}$, Let $s = |S|$. For some $s_0 \leq s$ to be defined.

- First we obtain an initial estimator β_{init} using the Lasso with $\lambda_n = c\sigma\sqrt{2\log p/n}$ for some constant c.
- Threshold the estimator β_{init} with t_0, and set $\mathcal{I} = \{j \in \{1, \ldots, p\} : \beta_{j,\mathrm{init}} \geq t_0\}$ with the general goal such that, we get an set $\mathcal{I}$ with cardinality at most $2s_0$.

The Thresholded Lasso estimator

Define $S = \mathrm{supp}(\beta) := \{j : \beta_j \neq 0\}$, Let $s = |S|$. For some $s_0 \leq s$ to be defined.

- First we obtain an initial estimator β_{init} using the Lasso with $\lambda_n = c\sigma\sqrt{2\log p/n}$ for some constant c.
- Threshold the estimator β_{init} with t_0, and set $\mathcal{I} = \{j \in \{1, \ldots, p\} : \beta_{j,\text{init}} \geq t_0\}$ with the general goal such that, we get an set $\mathcal{I}$ with cardinality at most $2s_0$.
- Feed $(Y, X_{\mathcal{I}})$ to the ordinary least squares (OLS) estimator: $\widehat{\beta}_{\mathcal{I}} = (X_{\mathcal{I}}^T X_{\mathcal{I}})^{-1} X_{\mathcal{I}}^T Y$ to obtain $\widehat{\beta}$, where $\widehat{\beta}_{\mathcal{I}^c} = 0$.

Variable selection under the RE condition

- Restricted eigenvalue assumption $RE(s, k_0, X)$: (Bickel-Ritov-Tsybakov 09). For some integer $1 \leq s \leq p$ and a positive number k_0, the following holds for all $\upsilon \neq 0$

$$\frac{1}{K(s, k_0)} \triangleq \min_{\substack{J_0 \subseteq \{1,\ldots,p\}, |J_0| \leq s \\ \|\upsilon_{J_0^c}\|_1 \leq k_0 \|\upsilon_{J_0}\|_1}} \frac{\|X\upsilon\|_2}{\sqrt{n}\left\|\upsilon_{J_0}\right\|_2} > 0.$$

Variable selection under the RE condition

- Restricted eigenvalue assumption $RE(s, k_0, X)$: (Bickel-Ritov-Tsybakov 09). For some integer $1 \leq s \leq p$ and a positive number k_0, the following holds for all $\upsilon \neq 0$

$$\frac{1}{K(s, k_0)} \triangleq \min_{\substack{J_0 \subseteq \{1,\dots,p\}, |J_0| \leq s \\ \|\upsilon_{J_0^c}\|_1 \leq k_0 \|\upsilon_{J_0}\|_1}} \frac{\|X\upsilon\|_2}{\sqrt{n}\left\|\upsilon_{J_0}\right\|_2} > 0.$$

- **Theorem (BRT 09).** It is sufficient for the Lasso and the Dantzig selector to achieve squared ℓ_2 loss $\|\beta_{\text{init}} - \beta\|^2$ of $O(s\sigma^2 \log p/n)$ with high probability.

Theorem (Z 09): Suppose that $RE(s, k_0, X)$ condition holds. Suppose $\beta_{\min} := \min_{j \in S} |\beta_j| \geq C\lambda_n\sqrt{s}$ for λ_n chosen below. Then with $\mathbb{P}(\mathcal{T}_a) \geq 1 - (\sqrt{\pi \log p} p^a)^{-1}$, the multi-step procedure returns $\widehat{\beta}$ with $\mathrm{supp}(\widehat{\beta}) := \mathcal{I}$ such that

$S \subseteq \mathcal{I}$ and $|\mathcal{I} \setminus S| < c_1$ and $\|\widehat{\beta} - \beta\|^2 \leq O(s\sigma^2 \log p/n)$,

Theorem (Z 09): Suppose that $RE(s, k_0, X)$ condition holds. Suppose $\beta_{\min} := \min_{j \in S} |\beta_j| \geq C \lambda_n \sqrt{s}$ for λ_n chosen below. Then with $\mathbb{P}(\mathcal{T}_a) \geq 1 - (\sqrt{\pi \log p} p^a)^{-1}$, the multi-step procedure returns $\widehat{\beta}$ with $\mathrm{supp}(\widehat{\beta}) := \mathcal{I}$ such that $S \subseteq \mathcal{I}$ and $|\mathcal{I} \setminus S| < c_1$ and $\|\widehat{\beta} - \beta\|^2 \leq O(s\sigma^2 \log p/n)$,

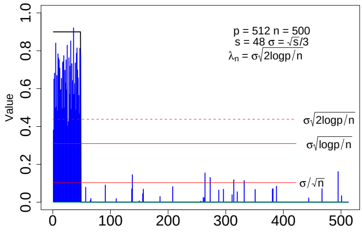
- where $\lambda_n \geq 2\sigma\sqrt{1+a}\sqrt{2\log p/n}$, where $a \geq 0$; and

Theorem (Z 09): Suppose that $RE(s, k_0, X)$ condition holds. Suppose $\beta_{\min} := \min_{j \in S} |\beta_j| \geq C\lambda_n\sqrt{s}$ for λ_n chosen below. Then with $\mathbb{P}(\mathcal{T}_a) \geq 1 - (\sqrt{\pi \log p} p^a)^{-1}$, the multi-step procedure returns $\widehat{\beta}$ with $\mathrm{supp}(\widehat{\beta}) := \mathcal{I}$ such that $S \subseteq \mathcal{I}$ and $|\mathcal{I} \setminus S| < c_1$ and $\|\widehat{\beta} - \beta\|^2 \leq O(s\sigma^2 \log p/n)$,

- where $\lambda_n \geq 2\sigma\sqrt{1+a}\sqrt{2\log p/n}$, where $a \geq 0$; and
- $\mathcal{T}_a := \left\{\epsilon : \left\|X^T\epsilon/n\right\|_\infty \leq \sigma\sqrt{1+a}\sqrt{2\log p/n}\right\}$.

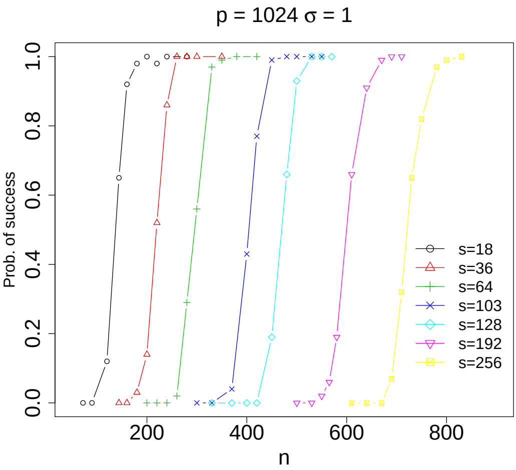
Theorem (Z 09): Suppose that $RE(s, k_0, X)$ condition holds. Suppose $\beta_{\min} := \min_{j \in S} |\beta_j| \geq C\lambda_n \sqrt{s}$ for λ_n chosen below. Then with $\mathbb{P}(\mathcal{T}_a) \geq 1 - (\sqrt{\pi \log p} p^a)^{-1}$, the multi-step procedure returns $\widehat{\beta}$ with $\mathrm{supp}(\widehat{\beta}) := \mathcal{I}$ such that $S \subseteq \mathcal{I}$ and $|\mathcal{I} \setminus S| < c_1$ and $\|\widehat{\beta} - \beta\|^2 \leq O(s\sigma^2 \log p/n)$,

- where $\lambda_n \geq 2\sigma\sqrt{1+a}\sqrt{2\log p/n}$, where $a \geq 0$; and
- $\mathcal{T}_a := \left\{\epsilon : \left\|X^T\epsilon/n\right\|_\infty \leq \sigma\sqrt{1+a}\sqrt{2\log p/n}\right\}$.

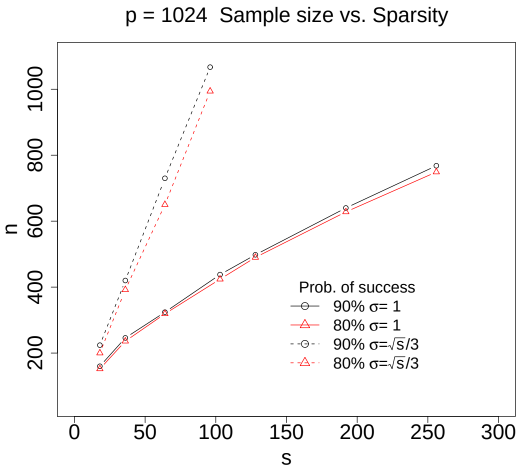


- $k_0 = 1$ for the Dantzig selector and $= 3$ for the Lasso; $c_1 = 1/64\Lambda^2_{\min}(2s)$; Proof imposes $s \geq K^4(s, k_0)$.

Compare probability of success for $p = 1024$



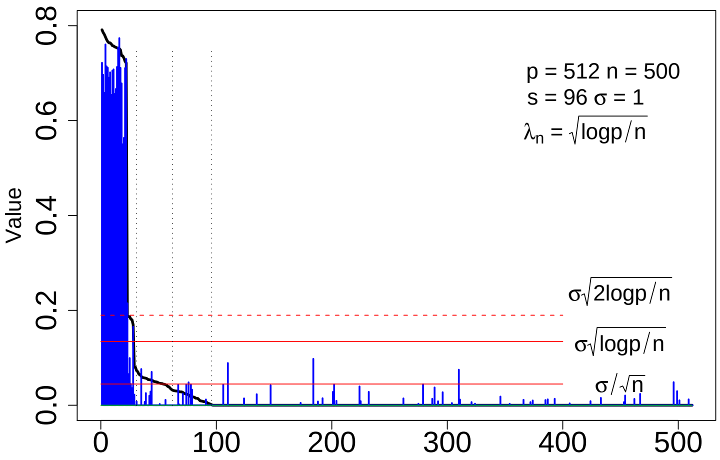
Sample size increases almost linearly with *s*



Linear sparsity result: summary

- The thresholded Lasso requires that $n \asymp s \log(p/n)$, in order to achieve (almost) exact recovery of the sparsity pattern for (sub)Gaussian random matrix when $\beta_{\min}$ is sufficiently large.
- This shows a strong contrast with the ordinary Lasso: to reach the same goal, the required sample size is much larger.

Detection limit



Ideal model selection: sparse oracle inequalities

Contributions: Define a meaningful criterion for variable selection when some non-zero elements are well below $\sigma / \sqrt{n}$;

Ideal model selection: sparse oracle inequalities

Contributions: Define a meaningful criterion for variable selection when some non-zero elements are well below $\sigma/\sqrt{n}$;

- Identify the relevant set of variables that are significant;

Ideal model selection: sparse oracle inequalities

Contributions: Define a meaningful criterion for variable selection when some non-zero elements are well below $\sigma/\sqrt{n}$;

- Identify the relevant set of variables that are significant;
- Estimation accuracy: recovers a good approximation $\widehat{\beta}$ to β, with ℓ_2 loss tightly bounded – in an "oracle" sense.

Ideal model selection: sparse oracle inequalities

Contributions: Define a meaningful criterion for variable selection when some non-zero elements are well below $\sigma/\sqrt{n}$;

- Identify the relevant set of variables that are significant;
- Estimation accuracy: recovers a good approximation $\widehat{\beta}$ to β, with ℓ_2 loss tightly bounded – in an "oracle" sense.

In addition to RE, we assume

$$\Lambda_{\max}(2s) \triangleq \max_{\upsilon \neq 0, 2s\text{-sparse}} \frac{\|X\upsilon\|^2}{n\|\upsilon\|^2} < \infty.$$

Nearly ideal model selection

Consider subset least squares estimators $\widehat{\beta}_I = (X_I^T X_I)^{-1} X_I^T Y$:

Nearly ideal model selection

Consider subset least squares estimators $\widehat{\beta}_I = (X_I^T X_I)^{-1} X_I^T Y$:

- **Question:** How to find a sparse subset $\mathcal{I}$ such that $|\mathcal{I}| \leq 2s_0$ and $\mathbb{E}\|\widehat{\beta}_{\mathcal{I}} - \beta\|^2 = O(\log p)\mathbb{E}\left\|\beta^\star - \beta\right\|^2$,

Nearly ideal model selection

Consider subset least squares estimators $\widehat{\beta}_I = (X_I^T X_I)^{-1} X_I^T Y$:

- **Question:** How to find a sparse subset $\mathcal{I}$ such that $|\mathcal{I}| \leq 2s_0$ and $\mathbb{E}\|\widehat{\beta}_{\mathcal{I}} - \beta\|^2 = O(\log p)\mathbb{E}\,\|\beta^\star - \beta\|^2$, where $\beta^\star$ is the ideal least-squares estimator which minimizes the expected mean squared error (MSE) $\mathbb{E}\,\|\beta^\star - \beta\|^2 = \arg\min_{I \subset \{1,\ldots,p\}} \mathbb{E}\|\widehat{\beta}_I - \beta\|^2$.

Nearly ideal model selection

Consider subset least squares estimators $\widehat{\beta}_I = (X_I^T X_I)^{-1} X_I^T Y$:

- **Question:** How to find a sparse subset $\mathcal{I}$ such that $|\mathcal{I}| \leq 2s_0$ and $\mathbb{E}\|\widehat{\beta}_{\mathcal{I}} - \beta\|^2 = O(\log p)\mathbb{E}\,\|\beta^\star - \beta\|^2$, where $\beta^\star$ is the ideal least-squares estimator which minimizes the expected mean squared error (MSE) $\mathbb{E}\,\|\beta^\star - \beta\|^2 = \arg\min_{I \subset \{1,\ldots,p\}} \mathbb{E}\|\widehat{\beta}_I - \beta\|^2$.
- We show $\|\widehat{\beta}_{\mathcal{I}} - \beta\|^2 = O(\log p)\sum_{i=1}^p \min(\beta_i^2, \sigma^2/n)$,

Nearly ideal model selection

Consider subset least squares estimators $\widehat{\beta}_I = (X_I^T X_I)^{-1} X_I^T Y$:

- **Question:** How to find a sparse subset $\mathcal{I}$ such that $|\mathcal{I}| \leq 2s_0$ and $\mathbb{E}\|\widehat{\beta}_{\mathcal{I}} - \beta\|^2 = O(\log p)\mathbb{E}\,\|\beta^\star - \beta\|^2$, where $\beta^\star$ is the ideal least-squares estimator which minimizes the expected mean squared error (MSE) $\mathbb{E}\,\|\beta^\star - \beta\|^2 = \arg\min_{I \subset \{1,\ldots,p\}} \mathbb{E}\|\widehat{\beta}_I - \beta\|^2$.
- We show $\|\widehat{\beta}_{\mathcal{I}} - \beta\|^2 = O(\log p)\sum_{i=1}^p \min(\beta_i^2, \sigma^2/n)$, given **Proposition:** (Candès-Tao 07). For $\Lambda_{\max}(s) < \infty$, then $\mathbb{E}\,\|\beta - \beta^\star\|^2 \geq \min\left(1, 1/\Lambda_{\max}(s)\right) \sum_{i=1}^p \min(\beta_i^2, \sigma^2/n)$.

Nearly ideal model selection

Consider subset least squares estimators $\widehat{\beta}_I = (X_I^T X_I)^{-1} X_I^T Y$:

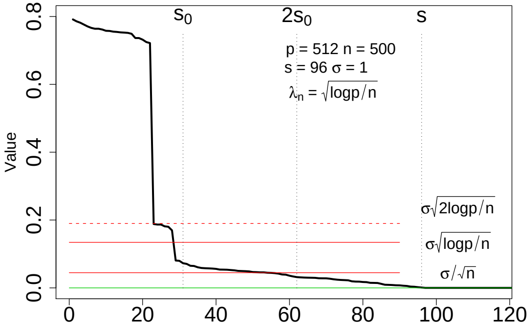
- **Question:** How to find a sparse subset $\mathcal{I}$ such that $|\mathcal{I}| \leq 2s_0$ and $\mathbb{E}\|\widehat{\beta}_{\mathcal{I}} - \beta\|^2 = O(\log p)\mathbb{E}\|\beta^\star - \beta\|^2$, where $\beta^\star$ is the ideal least-squares estimator which minimizes the expected mean squared error (MSE) $\mathbb{E}\|\beta^\star - \beta\|^2 = \arg\min_{I \subset \{1,\ldots,p\}} \mathbb{E}\|\widehat{\beta}_I - \beta\|^2$.
- We show $\|\widehat{\beta}_{\mathcal{I}} - \beta\|^2 = O(\log p)\sum_{i=1}^p \min(\beta_i^2, \sigma^2/n)$, given **Proposition:** (Candès-Tao 07). For $\Lambda_{\max}(s) < \infty$, then $\mathbb{E}\|\beta - \beta^\star\|^2 \geq \min(1, 1/\Lambda_{\max}(s)) \sum_{i=1}^p \min(\beta_i^2, \sigma^2/n)$.
- Note $\sum_{i=1}^p \min(\beta_i^2, \sigma^2/n) = \min_{I \subset \{1,\ldots,p\}} \|\beta - \beta_I\|^2 + |I|\sigma^2/n$ represents the ideal squared bias and variance tradeoff.

Defining $2s_0$

- Let $0 \leq s_0 \leq s$ be the smallest integer such that $\sum_{i=1}^{p} \min(\beta_i^2, \lambda^2\sigma^2) \leq s_0\lambda^2\sigma^2$, where $\lambda = \sqrt{2\log p/n}$.

Defining $2s_0$

- Let $0 \leq s_0 \leq s$ be the smallest integer such that $\sum_{i=1}^{p} \min(\beta_i^2, \lambda^2\sigma^2) \leq s_0 \lambda^2 \sigma^2$, where $\lambda = \sqrt{2 \log p / n}$.
- If we order the β_j's in decreasing order of magnitude $|\beta_1| \geq |\beta_2| ... \geq |\beta_p|$, then $|\beta_j| < \lambda\sigma \; \forall j > s_0$.



Nearly ideal model selection under the RE

Theorem: (Z 10) Suppose $RE(s_0, 6, X)$ holds with $K(s_0, 6)$, and $2s$-sparse eigenvalue conditions hold. Then with probability at least $1 - (\sqrt{\pi \log p} p^a)^{-1}$, the *Thresholded Lasso* estimator achieves sparse oracle inequalities:

$$
\begin{aligned}
|\mathcal{I}| &\leq 2s_0 \text{ and } |\mathcal{I} \setminus S| \leq s_0 \leq s \text{ and} \\
\|\widehat{\beta} - \beta\|^2 &\leq O(\log p) \sum_{i=1}^{p} \min(\beta_i^2, \sigma^2/n).
\end{aligned}
$$

Nearly ideal model selection under the RE

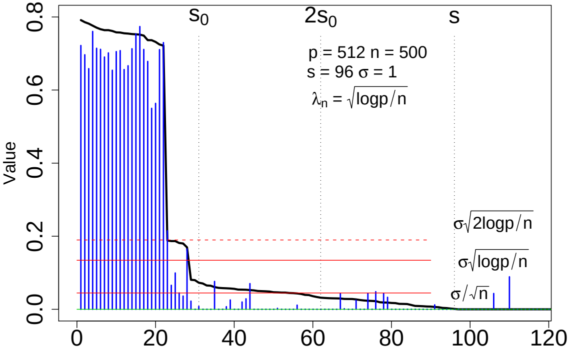
Theorem: (Z 10) Suppose $RE(s_0, 6, X)$ holds with $K(s_0, 6)$, and $2s$-sparse eigenvalue conditions hold. Then with probability at least $1 - (\sqrt{\pi \log p} p^a)^{-1}$, the *Thresholded Lasso* estimator achieves sparse oracle inequalities:

$$\begin{aligned} |\mathcal{I}| &\leq 2s_0 \text{ and } |\mathcal{I} \setminus S| \leq s_0 \leq s \text{ and} \\ \|\widehat{\beta} - \beta\|^2 &\leq O(\log p) \sum_{i=1}^{p} \min(\beta_i^2, \sigma^2/n). \end{aligned}$$

- Obtain β_{init} using the Lasso with $\lambda_n \geq 2\sigma\sqrt{1+a}\lambda$, where $\lambda = \sqrt{2\log p/n}$; Threshold β_{init} with t_0 chosen from $(D_1\lambda\sigma, C_4\lambda\sigma]$, where $D_1 = \Lambda_{\max}(s - s_0) + 9K^2(s_0, 6)/2$ and $C_4 \geq D_1$; and refit with model $\mathcal{I}$ using OLS.

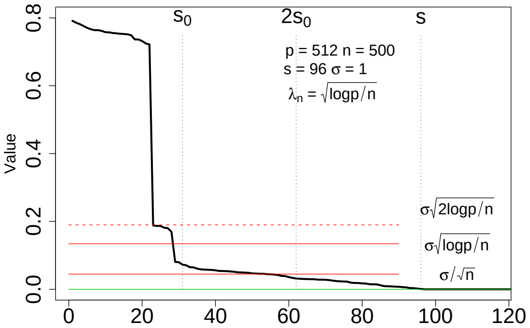
Oracle inequalities for the Lasso

- **Theorem (Z 10).** $RE(s_0, 6, X)$ is a sufficient condition for the Lasso to achieve squared ℓ_2 loss of $O(s_0\sigma^2 \log p/n)$ so long as $\Lambda_{\max}(2s) < \infty$ and $\Lambda_{\min}(2s_0) > 0$.



Decompose the ℓ_2 loss

- $\left\|\widehat{\beta}_{\mathcal{I}} - \beta\right\|^2 = \left\|\widehat{\beta}_{\mathcal{I}} - \beta_{\mathcal{I}}\right\|^2 + \|\beta_{\mathcal{I}} - \beta\|^2$



- Each term above is bounded by $O(s_0\lambda^2\sigma^2)$, where $s_0\lambda^2\sigma^2 \leq O(\log p)\mathbb{E}\|\beta - \beta^\star\|^2$.

- **Theorem (Z 09).** Under RIP type of condition, the Gauss-Dantzig selector proposed by Candès-Tao 07 achieves such sparse oracle inequalities.
- Analysis builds upon Candès-Tao's result for the initial Dantzig selector.

Summary on the general thresholding rules

When $\beta_{\min}$ is well below the noise level

- We show how to choose a sparse model $\mathcal{I}$, upon which the OLS estimator achieves the sparse oracle inequalities.
- We consider the bound on ℓ_2-loss as a natural criterion to evaluate a sparse model when it is not exactly S.
- Variables in model $\mathcal{I}$ are essential in predicting $X\beta$.

Subset selection: related work

- Oracle inequalities in ℓ_2 loss have been studied in Donoho-Johnstone 94 and Candès-Tao 07.
- Also relevant is the work of Meinshausen and Yu 09, Wasserman and Roeder 09, and Zhang 09.
- A final note: this method was called “selection/estimation (s/e) procedure” in Foster and George 94, and “subset least squares” by Mallows 73.

Conclusion

- In the high dimensional linear model, it is possible to estimate the parameter β and its significant set of variables accurately using the Thresholded Lasso.

Conclusion

- In the high dimensional linear model, it is possible to estimate the parameter β and its significant set of variables accurately using the Thresholded Lasso.
- In a joint work with Peter Buehlmann, Philipp Rutimann and Min Xu, we apply the thresholding/re-estimation idea to Gaussian graphical model selection and covariance estimation.

- **That is it! Thank you very much!**