High Dimensional Data and Statistical Learning

Lecture 6: Theory of Lasso

Weixing Song

Department of Statistics
Kansas State University

Outline

Outline

Model:

$$y_i = x_{i1}\beta_1 + \cdots + x_{ip}\beta_p + \varepsilon_i, \quad 1 \leq i \leq n$$

Notations:

- Response: $\mathbf{y} = (y_1, \ldots, y_n)^T$.
- Predictors: $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^T$, $j = 1, 2, \ldots, p$.
- Design Matrix: $\mathbf{X}_{n\times p} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$.
- Residuals: $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n)^T$.
- Regression Coefficients: $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$.
- True Regression Coefficients: $\boldsymbol{\beta}^o = (\beta_1^o, \ldots, \beta_p^o)^T$.
- Oracle Set: $\mathcal{O} = \{j : \beta_j^o \neq 0\}$.
- Underlying Model Dimension: $d^0 = \|\mathcal{O}\| = \#\{j : \beta_j^o \neq 0\}$.

Centering and Standardization

WLOG, we assume that the response and predictors are centered and the predictors are standardized as follows

$$\sum_{i=1}^{n} y_i = 0, \quad \sum_{i=1}^{n} x_{ij} = 0, \quad \sum_{i=1}^{n} x_{ij}^2 = n, \quad 1 \leq j \leq p.$$

Then there is no intercept in the model.

Each predictor is standardized to have the same magnitude in L_2. So the corresponding regression coefficients are "comparable".

After model fitting, the results can be readily transformed back to the original scale.

Recall that

$$\hat{\beta}_{\text{Lasso}} = \text{argmin}_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\},$$

where $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$ is the L_1-norm of $\boldsymbol{\beta}$.

What theoretical properties does the Lasso estimator have?

- Estimation consistency and asymptotic distribution.
- Prediction properties.
- Selection consistency.

More General Questions

There are some fundamental theoretical questions in the area of high-dimensional problems:

- What is the limit of dimensionality that the regularization methods can handle?
- What is the role of penalty function?
- What is the optimality of the regularized estimator?

Optimality of the regularization methods (Bickel et al,. 2006; Candès, 2006):

- Consistency.
- For a function class $\mathcal{F}$, whether the estimate $\hat{f}$ attains the minimax risk

$$\inf_{\hat{f}} \max_{f \in \mathcal{F}} MSE(f, \hat{f}).$$

- For a function class $\mathcal{F}$ with a prior π, whether the estimator $\hat{f}$ achieves the minimum average MSE or Bayes risk $E_\pi MSE(f, \hat{f})$.
- Can the estimator match with the "oracle" choice?

Oracle Properties

One important concept is the so called *oracle properties*.

Ideally, we want our model to exactly select the set of true covariates as $n \to \infty$. This property is called *consistency* and it is the first requirement of an oracle procedure.

Definition (Oracle Procedure)

Denote $\hat{\boldsymbol{\beta}}(\delta)$ the coefficient estimator for fitting procedure δ. We call δ an oracle procedure if $\hat{\boldsymbol{\beta}}(\delta)$ (asymptotically) has the following properties:

- Identifies right subset model (consistency): $\{j : \hat{\beta}_j(\delta) \neq 0\} = \mathcal{O}$;
- Has optimal estimation rate:

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{O}} - \boldsymbol{\beta}^o_{\mathcal{O}}) \xrightarrow{\mathscr{L}} N(0, \Sigma_0),$$

where Σ_0 is the covariance matrix knowing the true subset model.

In general, we want to establish some "oracle inequalities", which relates the performance of a real estimator with that of an ideal estimator which relies on perfect information supplied by an oracle, and which is not available in practice.

Outline

Unveiling the Mystery of "Oracle Inequality"

The importance of this section relies upon the fact that

- it shows the strength of shrinkage estimation.
- it introduces the idea of an oracle inequality.

Consider the problem of estimating a (possibly infinite) vector $\boldsymbol{\theta} \in \mathbb{R}^p$ from observations $\mathbf{y} \sim N(\boldsymbol{\theta}, I)$, and focus on the statistical underpinnings of this problem.

What is the maximum likelihood estimator of $\boldsymbol{\theta}$?

What is the MSE of the MLE?

We would like to estimate $\boldsymbol{\theta} \in \mathbb{R}^p$ from observations $\mathbf{y} \sim N(\boldsymbol{\theta}, I)$, and use $MSE(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = E\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2$ to measure the performance.

The MLE is given by $\mathbf{y}$ and $MSE(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = p$.

Everybody would agree that the MLE is a good estimator. After all, what other estimators could we use in the absence of any additional information about $\boldsymbol{\theta}$?

The surprising discovery of James and Stein (1961) is that when $p > 2$, the MLE is not admissible!

That is, there exist estimators which are more accurate than the MLE (or better than the sample mean in the case where one gets independent copies of $\mathbf{y}$).

The James-Stein Estimator is defined as

$$\hat{\boldsymbol{\theta}}^{\mathrm{JS}} = \left(1 - \frac{p-2}{\|\mathbf{y}\|^2}\right)\mathbf{y},$$

which shrinks the data $\mathbf{y}$ towards the origin.

MSE Comparison between MLE and James-Stein Estimator

$$MSE(\hat{\boldsymbol{\theta}}^{\mathrm{JS}}, \boldsymbol{\theta}) < MSE(\hat{\boldsymbol{\theta}}^{\mathrm{MLE}}, \boldsymbol{\theta}), \quad \text{for all } \boldsymbol{\theta} \in \mathbb{R}^p, p > 2. \tag{1}$$

To show the above result, we need the following Stein's Lemma.

Stein's Lemma

If $Y \sim N(\theta, 1)$ and $h(y)$ is any differentiable function such that $Eh'(Y) < \infty$. Then

$$E[h(Y)(Y - \theta)] = Eh'(Y).$$

More general, we have

Stein's Unbiased Risk Estimation

If $\mathbf{y} \sim N(\boldsymbol{\theta}, 1)$ and $\hat{\boldsymbol{\mu}} = \mathbf{y} + g(\mathbf{y})$, where $g : \mathbb{R}^p \to \mathbb{R}^p$ is a differentiable function. Then under mild integrability assumptions:

$$E\|\mathbf{y} + g(\mathbf{y}) - \boldsymbol{\theta}\|^2 = E[p + 2\nabla g(\mathbf{y}) + \|g(\mathbf{y})\|^2],$$

where $\nabla g(\mathbf{y})$ is the divergence of g, $\nabla g(\mathbf{y}) = \sum_{i=1}^{p} \partial g_i(\mathbf{y})/\partial y_i$

Proof of (1): We shall show a more general result. For any fixed $\boldsymbol{\mu} \in \mathbb{R}^p$, let

$$\delta(\mathbf{y}) = \boldsymbol{\mu} + \left(1 - \frac{p-2}{S^2}\right)(\mathbf{y} - \boldsymbol{\mu}), \quad S^2 = \|\mathbf{y} - \boldsymbol{\mu}\|^2 = \sum_{j=1}^{p} (Y_j - \mu_j)^2.$$

Then

$$E\|\delta(\mathbf{y}) - \boldsymbol{\theta}\|^2 = E\left\|\mathbf{y} - \frac{p-2}{S^2}(\mathbf{y} - \boldsymbol{\mu}) - \boldsymbol{\theta}\right\|^2.$$

Let

$$g(\mathbf{y}) = -\frac{p-2}{S^2}(\mathbf{y} - \boldsymbol{\mu}), \quad \text{then} \quad \nabla g(\mathbf{y}) = -\frac{(p-2)^2}{S^2}.$$

So, from the result of Stein's unbiased risk estimation,

$$E\|\delta(\mathbf{y}) - \boldsymbol{\theta}\|^2 = p - (p-2)^2 E\left(\frac{1}{S^2}\right).$$

Note that when $p \geq 3$, the expectation is finite, so

$$E\|\delta(\mathbf{y}) - \boldsymbol{\theta}\|^2 = p - (p-2)^2 E\left(\frac{1}{S^2}\right) < p = E\|\mathbf{y} - \boldsymbol{\theta}\|^2.$$

Remark: See Bock, Judge and Yancey (1984) for the finiteness of $E(1/S^2)$.

A more general definition of James-Stein estimator is

$$\hat{\boldsymbol{\theta}}^{\mathrm{JS}} = \boldsymbol{\mu} + \left(1 - \frac{p-2}{\|\mathbf{y} - \boldsymbol{\mu}\|^2}\right)(\mathbf{y} - \boldsymbol{\mu}),$$

which shrink $\mathbf{y}$ towards an arbitrary $\boldsymbol{\mu}$.

Note that for small value of $\|\mathbf{y} - \boldsymbol{\mu}\|$, the shrinkage factor can be negative. A nonlinear shrinkage version is defined as

$$\hat{\boldsymbol{\theta}}^{\mathrm{JS}} = \boldsymbol{\mu} + \left(1 - \frac{p-2}{\|\mathbf{y} - \boldsymbol{\mu}\|^2}\right)_+ (\mathbf{y} - \boldsymbol{\mu}).$$

The above estimation problem can be stated more generally, i.e., we may assume $\mathbf{x}_1, \mathbf{x}_2, \ldots$ are independent Gaussian observation with mean $\boldsymbol{\theta}$ such that $\mathbf{y} = \bar{\mathbf{x}} \sim N(\boldsymbol{\theta}, I)$.

The approach can also be extended to a more general known covariate matrix Σ.

Implications of The James-Stein Estimator

The performance of the shrinkage estimator is superior to that of the sample mean for all values of the parameter $\boldsymbol{\theta}$.

This is surprising, because $\mathbf{y}$ may measure seemingly unrelated quantities such as the taste of clams and the age of the universe (Le Cam, 2000).

It is therefore surprising that by mixing information about completely disconnected problems, one can obtain an estimator with a total MSE that is smaller than that one would obtain by considering each problem separately.

This strange phenomenon is difficult to comprehend and has had an enormous influence on the theory of statistics.

James-Stein estimator can be motivated and interpreted from an empirical Bayes approach (Efron and Morris, 1975).

We will not attempt to dive into this literature and, instead, merely note that *nonlinear shrinkage improves performance.*

Ideal Linear Shrinkage Estimator and Oracle Inequalities

Unfortunately, James-Stein estimator is still not admissible.

It is time to revisit the main issue discussed so far: how much should we smooth or, rather, how much should we shrink?

To estimate $\boldsymbol{\theta} \in \mathbb{R}^p$ from observation $\mathbf{y} \sim N(\boldsymbol{\theta}, I)$, consider the family of estimators

$$\hat{\boldsymbol{\theta}}^c = c\mathbf{y}$$

where c is a scalar. The MSE of $\hat{\boldsymbol{\theta}}^c$ is

$$MSE(\hat{\boldsymbol{\theta}}^c, \boldsymbol{\theta}) = (1-c)^2\|\boldsymbol{\theta}\|^2 + c^2 p.$$

By differentiation, the ideal c-value to minimize $MSE(\hat{\boldsymbol{\theta}}^c, \boldsymbol{\theta})$ is

$$c^* = \frac{\|\boldsymbol{\theta}\|^2}{\|\boldsymbol{\theta}\|^2 + p}.$$

Accordingly,

$$MSE(\hat{\boldsymbol{\theta}}^{c^*}, \boldsymbol{\theta}) = \frac{p\|\boldsymbol{\theta}\|^2}{\|\boldsymbol{\theta}\|^2 + p}.$$

The estimator $\hat{\boldsymbol{\theta}}^{c^*}$ is ideal because we would of course not know which estimator c is the best.

To achieve the ideal MSE, one would need an oracle that would tell us which shrinkage factor to choose.

The difference from the James-Stein estimator is that the shrinkage factor in James-Stein estimator is estimated from data, while in the ideal scenario, the ideal shrinkage factor depends on $\|\boldsymbol{\theta}\|$.

Obviously, $\inf_c MSE(\hat{\boldsymbol{\theta}}^{c^*}, \boldsymbol{\theta}) \leq MSE(\hat{\boldsymbol{\theta}}^{\text{JS}}, \boldsymbol{\theta})$.

But a more interesting fact is that there is an inequality in the other direction.

An Oracle Inequality between $\hat{\boldsymbol{\theta}}^{\text{JS}}$ and $\hat{\boldsymbol{\theta}}^{c^*}$

For $\hat{\boldsymbol{\theta}}^{\text{JS}} = (1 - (p-2)/\|\mathbf{y}\|^2)\mathbf{y}$ and $\hat{\boldsymbol{\theta}}^{c} = c\mathbf{y}$, we have

$$MSE(\hat{\boldsymbol{\theta}}^{\text{JS}}, \boldsymbol{\theta}) \leq 4 + \inf_c MSE(\hat{\boldsymbol{\theta}}^{c}, \boldsymbol{\theta}).$$

Proof: It is known that

$$E\|\hat{\boldsymbol{\theta}}^{\mathrm{JS}} - \boldsymbol{\theta}\|^2 = p - (p-2)^2 E\left(\frac{1}{\|\mathbf{y}\|^2}\right) \leq p - \frac{(p-2)^2}{E\|\mathbf{y}\|^2},$$

and

$$E\|\mathbf{y}\|^2 = \|\theta\|^2 + p.$$

Therefore,

$$E\|\hat{\boldsymbol{\theta}}^{\mathrm{JS}} - \boldsymbol{\theta}\|^2 \leq p - \frac{(p-2)^2}{\|\boldsymbol{\theta}\|^2 + p} = \frac{\|\boldsymbol{\theta}\|^2 p}{\|\boldsymbol{\theta}\|^2 + p} + \frac{4(p-1)}{\|\boldsymbol{\theta}\|^2 + p}.$$

This completes the proof by noting that the last term is less than 4. □

The inequalities say that the James-Stein estimator is almost as good as the ideal estimator in a mean-squared error sense.

When p is large, the additive factors are small compared to the MSE of the MLE, which is p.

The oracle inequality relates the performance of a real estimator with that of an ideal estimator which relies on perfect information supplied by an oracle, and which is not available in practice.

Oracle inequality is a powerful concept that is used extensively in high dimensional analysis.

The linear estimators can be highly ineffective. The James-Stein estimator, which is essentially a linear estimator, albeit with a nonlinear data-dependent shrinkage factor, can also be very ineffective.

The thresholding rules which are true nonlinear estimation procedures may perform well in much more complicated settings.

A foundational result in modern estimation is that correctly tuned thresholding rules nearly achieve the risk of ideal projections. For details, see Donoho and Johnstone (1994).

Outline

Some Inequalities

We will use some matrix inequalities in the theoretical derivation. Here are two important ones.

- Cauchy-Schwarz Inequality

  Suppose $\mathbf{x} \in \mathbb{R}^p, \mathbf{y} \in \mathbb{R}^p$, A is a $p \times p$ positive definite matrix. Then

  $$\mathbf{x}^T\mathbf{y} \leq \|\mathbf{x}\| \cdot \|\mathbf{y}\|, \quad \mathbf{x}^T\mathbf{y} \leq \sqrt{\mathbf{x}^T A\mathbf{y} \cdot \mathbf{x}^T A^{-1}\mathbf{y}}.$$

- Maximization Lemma

  Suppose $\mathbf{x} \in \mathbb{R}^p$ and A is $p \times p$ positive matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$ and eigenvectors $e_1, e_2, \ldots, e_p$. Then

  $$\lambda_p \mathbf{x}^T\mathbf{x} \leq \mathbf{x}^T A\mathbf{x} \leq \lambda_1 \mathbf{x}^T\mathbf{x},$$

  $$\max_{\mathbf{x}\neq 0} \frac{\mathbf{x}^T A\mathbf{x}}{\mathbf{x}^T\mathbf{x}} = \lambda_1 \text{ which is attained at } \mathbf{x} = e_1,$$

  $$\min_{\mathbf{x}\neq 0} \frac{\mathbf{x}^T A\mathbf{x}}{\mathbf{x}^T\mathbf{x}} = \lambda_p \text{ which is attained at } \mathbf{x} = e_p.$$

p Fixed: An MSE Inequality

Let $c_{\min}$ be the smallest eigenvalue of $\Sigma = \mathbf{X}^T\mathbf{X}/n$.

In the following, we shall use $\hat{\boldsymbol{\beta}}_n$ to denote the Lasso estimator of $\boldsymbol{\beta}$.

Theorem 1

Suppose that $c_{\min} > 0$ and that $\varepsilon_1, \ldots, \varepsilon_n$ are independent random variables with $E\varepsilon_i = 0$ and $E\varepsilon_i^2 = \sigma^2$. Then

$$E\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|^2 \leq \frac{8\sigma^2 p}{nc_{\min}} + \frac{16\lambda^2 p}{c_{\min}^2}.$$

Proof: Let $\lambda_n = n\lambda$. By the definition of $\hat{\boldsymbol{\beta}}_n$,

$$\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_n\|^2 + 2\lambda_n\|\hat{\boldsymbol{\beta}}_n\|_1 \leq \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^o\|^2 + 2\lambda_n\|\boldsymbol{\beta}^o\|_1.$$

Thus

$$\|\mathbf{X}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^o)\|^2 - 2\boldsymbol{\varepsilon}^T\mathbf{X}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^o) \leq 2\lambda_n\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^o\|_1. \tag{2}$$

Denote $\eta_n = \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o)$. Then

$$\|\eta_n\|^2 + 2\boldsymbol{\varepsilon}^T\eta_n \leq 2\lambda_n\sqrt{p}\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^o\|.$$

Let $\boldsymbol{\varepsilon}^* = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}$. By Cauchy-Schwarz inequality,

$$2|\boldsymbol{\varepsilon}^T\eta_n| \leq 2\|\boldsymbol{\varepsilon}^*\| \cdot \|\eta_n\| \leq 2\|\boldsymbol{\varepsilon}^*\|^2 + \frac{1}{2}\|\eta_n\|^2.$$

It follows that $\|\eta_n\|^2 \leq 4\|\boldsymbol{\varepsilon}^*\|^2 + 4\lambda_n\sqrt{p}\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^o\|$. Furthermore, we have

$$nc_{\min}\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^o\|^2 \leq 4\|\boldsymbol{\varepsilon}^*\|^2 + \frac{(4\lambda_n\sqrt{p})^2}{2nc_{\min}} + \frac{1}{2}nc_{\min}\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^o\|^2.$$

Simple arrangement leads to

$$\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^o\|^2 \leq \frac{8\|\boldsymbol{\varepsilon}^*\|^2}{nc_{\min}} + \frac{16\lambda_n^2 p}{n^2 c_{\min}^2}.$$

The result follows by noting that $\lambda_n = n\lambda$.

The keys in the proof of Theorem 1:

- Start with the so called "Basic inequality (2)" easily derived from the Lasso objective function.
- Try to bound the stochastic part of the problem.
- Use elementary inequalities to get everything to the L_2 world.

Theorem 1 implies that if $\lambda = o(1)$, then $\hat{\boldsymbol{\beta}}_n$ is consistent.

Note that the consistency in estimation is not the same as consistency in variable selection.

The latter apparently may require stronger assumption.

Fixed p: Asymptotic Distribution

Let $\boldsymbol{\beta} = \boldsymbol{\beta}^o + n^{-1/2}\mathbf{t}$, and define

$$L_n(\mathbf{t}) = \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1 = \frac{1}{2n}\|\mathbf{y} - \mathbf{X}(\boldsymbol{\beta}^o + n^{-1/2}\mathbf{t})\|^2 + \lambda\|\boldsymbol{\beta}^o + n^{-1/2}\mathbf{t}\|_1.$$

Let

$$\hat{\mathbf{t}}_n = \text{argmin}_{\mathbf{t}}\, V_n(\mathbf{t})$$

where

$$\begin{aligned} V_n(\mathbf{t}) &= n[L_n(\mathbf{t}) - L_n(0)] \\ &= \frac{1}{2}\left[\|\boldsymbol{\varepsilon} - n^{-1/2}\mathbf{X}\mathbf{t}\|^2 - \|\boldsymbol{\varepsilon}\|^2\right] + n\lambda(\|\boldsymbol{\beta}^o + n^{-1/2}\mathbf{t}\|_1 - \|\boldsymbol{\beta}^o\|_1). \end{aligned}$$

Then $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^o) = \hat{\mathbf{t}}_n$.

The idea is to show that $V_n(\mathbf{t}) \Longrightarrow V(\mathbf{t})$ weakly. By the argmin continuous mapping theorem of Kim and Pollard (1990), we have

$$\hat{\mathbf{t}}_n \Longrightarrow_d \text{argmin}_{\mathbf{t}}\, V(\mathbf{t}).$$

We shall assume $\mathbf{X}^T\mathbf{X}/n \to \Sigma$ and $\sqrt{n}\lambda \to \lambda_0$ as $n \to \infty$.

First, we have

$$\frac{1}{2}\left(\|\boldsymbol{\varepsilon} - n^{-1/2}\mathbf{X}\mathbf{t}\|^2 - \|\boldsymbol{\varepsilon}\|^2\right) \Longrightarrow -\sigma\mathbf{t}^T\Sigma^{1/2}\mathbf{z} + \frac{1}{2}\mathbf{t}^T\Sigma\mathbf{t},$$

where $\mathbf{z} \sim N(0, I_p)$.

Second, we have

$$n\lambda(\|\boldsymbol{\beta}^o + n^{-1/2}\mathbf{t}\|_1 - \|\boldsymbol{\beta}^o\|_1) \to \lambda_0 \sum_{j=1}^{p} \left[t_j \mathrm{sgn}(\beta_{0j}) I(\beta_{0j} \neq 0) + |t_j| I(\beta_{0j} = 0)\right].$$

Therefore,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^o) \Longrightarrow \mathrm{argmin}_{\mathbf{t}}\, V(\mathbf{t}),$$

where

$$V(\mathbf{t}) = -\sigma\mathbf{t}^T\Sigma^{1/2}\mathbf{z} + \frac{1}{2}\mathbf{t}^T\Sigma\mathbf{t} + \lambda_0 \sum_{j=1}^{p} \left[t_j \mathrm{sgn}(\beta_{0j}) I(\beta_{0j} \neq 0) + |t_j| I(\beta_{0j} = 0)\right].$$

The keys in developing the asymptotic distribution:

- Transform the objective function so that it is optimized at the point of our target, i.e., $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^o)$.
- Rescale or adjust the objective function and find its limit.
- Invoke the Argmax/Argmin continuous mapping theorem, which roughly says that: under certain conditions, the convergence of the objective function implies the convergence of its optimizer.

When $\lambda_0 = 0$, i.e.,$\sqrt{n}\lambda \to 0$, then

$$V(\mathbf{t}) = -\sigma \mathbf{t}^T \Sigma^{1/2} \mathbf{z} + \frac{1}{2} \mathbf{t}^T \Sigma \mathbf{t}.$$

It is minimized at $\sigma \Sigma^{-1/2} \mathbf{z} \sim N(0, \sigma^2 \Sigma^{-1})$, which is the limit distribution of the LSE. However, this result is uninteresting. For such λ, the Lasso essentially behaves like the LS estimator, which does not do variable selection.

The right order of growth for λ is $\sqrt{n}\lambda \to \lambda_0 > 0$.

The asymptotic distribution here is complicated and it is not clear how to make use of it.

An interesting question is how to make statistical inference based on the Lasso estimator, e.g., constructing confidence intervals and conducting statistical tests etc..

It is tempting to bootstrap the Lasso estimator for statistical inference. However, it appears that the bootstrap does not work here: the bootstrap is not consistent for Lasso.

As recently shown by Chatterjee and Lahiri (2011), a modified bootstrap method provides valid approximation to the distribution of a Lasso estimator; moreover, they have shown that the standard residual bootstrap can consistently estimate the distribution of an adaptive Lasso estimator due to its oracle properties.

Note that we have used a very important tool called Argmax/Argmin continuous mapping theorem. See more details in Knight and Fu (1998).

$p \gg n$ Case: Theorem 2

The following theorem provides a prediction bound for Lasso in terms of L_1 sparsity.

Theorem 2

In the event $\left\{\|\mathbf{X}^T\boldsymbol{\varepsilon}\|_\infty \leq n\lambda\right\} = \left\{\max_{1\leq j\leq p}\left|\sum_{i=1}^n x_{ij}\varepsilon_i\right| \leq n\lambda\right\}$, we have $\|\mathbf{X}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^o)\|_2 \leq 4n\lambda\|\boldsymbol{\beta}^o\|_1$.

Proof: By the definition of $\hat{\boldsymbol{\beta}}_n$,

$$\frac{1}{2n}\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_n\|^2 + \lambda\|\hat{\boldsymbol{\beta}}_n\|_1 \leq \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^o\|^2 + \lambda\|\boldsymbol{\beta}^o\|_1.$$

This implies

$$\frac{1}{2n}\|\mathbf{X}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^o)\|^2 \leq \frac{1}{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^o)^T\mathbf{X}^T\varepsilon + \lambda(\|\boldsymbol{\beta}^o\|_1 - \|\hat{\boldsymbol{\beta}}_n\|_1).$$

Thus, in the event $\left\{\|\mathbf{X}^T\boldsymbol{\varepsilon}\|_\infty \leq n\lambda\right\}$,

$$\frac{1}{2n}\|\mathbf{X}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^o)\|^2 \leq \lambda(\|\hat{\boldsymbol{\beta}}_n\|_1 + \|\boldsymbol{\beta}^o\|_1) + \lambda(\|\boldsymbol{\beta}^o\|_1 - \|\hat{\boldsymbol{\beta}}_n\|_1) = 2\lambda\|\boldsymbol{\beta}^o\|_1.$$

The inequality follows.

A Guideline of Choosing Tuning Sequences

Suppose $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are i.i.d. as $N(0, \sigma^2)$. Let $\tau_j = \sum_{i=1}^n x_{ij}\varepsilon_i/(\sqrt{n}\sigma)$, we have $\tau_j \sim N(0,1)$. Then

$$\begin{aligned} P\left(\max_{1\le j\le p}\left|\sum_{i=1}^n x_{ij}\varepsilon_i\right| \ge n\lambda\right) &= P\left(\max_{1\le j\le p}|\tau_j| \ge \lambda\sqrt{n}/\sigma\right) \\ &\le \sum_{j=1}^p P\left(|\tau_j| \ge \lambda\sqrt{n}/\sigma\right) \le 2p(1-\Phi(\lambda\sqrt{n}/\sigma)). \end{aligned}$$

Since $1-\Phi(x) \le (\sqrt{2\pi}x)^{-1}\exp(-x^2/2)$, we have

$$P\left(\max_{1\le j\le p}\left|\sum_{i=1}^n x_{ij}\varepsilon_i\right| \ge n\lambda\right) \le \frac{2p\sigma}{\sqrt{2n\pi}\lambda}\exp\left(-\frac{n\lambda^2}{2\sigma^2}\right).$$

When $\lambda = a_0\sigma\sqrt{\log(p)/n}$ with $a_0 \ge \sqrt{2}$,

$$P\left(\max_{1\le j\le p}\left|\sum_{i=1}^n x_{ij}\varepsilon_i\right| \ge n\lambda\right) \le \frac{2p^{1-a_0^2/2}}{a_0\sqrt{2\pi\log(p)}} \to 0, \quad \text{as } p\to\infty.$$

Corollary to Theorem 2

Suppose $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent and identically distributed as $N(0, \sigma^2)$. Let $\lambda = a_0 \sigma \sqrt{\log(p)/n}$ with $a_0 \geq \sqrt{2}$. Then with probability at least

$$1 - \frac{2p^{1-a_0^2/2}}{a_0\sqrt{2\pi \log(p)}},$$

we have

$$\|\mathbf{X}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^o)\| \leq 4n\lambda \|\boldsymbol{\beta}^o\|_1.$$

In particular, when $a_0 = \sqrt{2}$, then with probability at least $1 - 1/\sqrt{\pi \log p}$, the above inequality holds.

The value $\lambda = \sigma\sqrt{2\log(p)/n}$ is often called the universal threshold value.

Implications of Theorem 2:

- "Sparsity" is a general concept, and it can be measured in various ways.
- Theorem 2 is about L_1 sparsity.
- The tuning parameter is of order $\sqrt{\log p/n}$, to ensure that the probability with which the statement holds converges to 1.
- Theorem 2 implies that $\|\mathbf{X}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^o)\|_2/n \to 0$ in probability requires a sparsity assumption of the from

$$\|\boldsymbol{\beta}^o\|_1 = o\left(\sqrt{\frac{n}{\log p}}\right).$$

We will introduce a more refined oracle inequality.

$p \gg n$ Case: A Basic Lemma

In the following, we will use $\hat{\boldsymbol{\beta}}$ to denote the Lasso solution.

Basic Lemma

On the set $\{\|\mathbf{X}^T\boldsymbol{\varepsilon}\|_\infty \leq \lambda/2\}$,

$$\frac{1}{2n}\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o)\|^2 + \frac{\lambda}{2}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 \leq 2\lambda\|\hat{\boldsymbol{\beta}}_{\mathcal{O}} - \boldsymbol{\beta}^o_{\mathcal{O}}\|_1,$$

In particular,

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 \leq 4\|\hat{\boldsymbol{\beta}}_{\mathcal{O}} - \boldsymbol{\beta}^o_{\mathcal{O}}\|_1.$$

Proof: By the definition of the Lasso,

$$\frac{1}{2n}\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 + \lambda\|\hat{\boldsymbol{\beta}}\|_1 \leq \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^o\|^2 + \lambda\|\boldsymbol{\beta}^o\|_1.$$

This implies

$$\frac{1}{2n}\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o)\|^2 \leq \frac{1}{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o)^T\mathbf{X}^T\boldsymbol{\varepsilon} + \lambda(\|\boldsymbol{\beta}^o\|_1 - \|\hat{\boldsymbol{\beta}}\|_1).$$

Thus in the event $\{\|\mathbf{X}^T\boldsymbol{\varepsilon}\|_\infty \le \lambda/2\}$,

$$\frac{1}{2n}\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o)\|^2 \le \frac{\lambda}{2}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 + \lambda(\|\boldsymbol{\beta}^o\|_1 - \|\hat{\boldsymbol{\beta}}\|_1),$$

and

$$\frac{1}{2n}\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o)\|^2 + \frac{\lambda}{2}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 \le \lambda\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 + \lambda(\|\boldsymbol{\beta}^o\|_1 - \|\hat{\boldsymbol{\beta}}\|_1).$$

Note that for all $j \notin \mathcal{O} = \|\boldsymbol{\beta}^o_{\mathcal{O}}\|_0 = \|\boldsymbol{\beta}^o\|_0$,

$$|\hat{\beta}_j - \beta^o_j| + |\beta^o_j| - |\hat{\beta}_j| = 0.$$

So we have

$$\frac{1}{2n}\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o)\|^2 + \frac{\lambda}{2}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 \le 2\lambda\|\hat{\boldsymbol{\beta}}_{\mathcal{O}} - \boldsymbol{\beta}^o_{\mathcal{O}}\|_1.$$

In particular, this implies the desired result.

Restricted Eigenvalue Condition

Denote

$$\mathcal{B} = \{\mathbf{b} \in \mathbb{R}^p : \|\mathbf{b}\|_1 \leq 4\|\mathbf{b}_{\mathcal{O}}\|_1\}.$$

The design matrix $\mathbf{X}$ satisfies the restricted eigenvalue (RE) condition if there exists a constant $c_* > 0$, such that, on the set $\mathcal{B}$,

$$\min_{\|\mathbf{b}\| \neq 0, \mathbf{b} \in \mathcal{B}} \frac{\mathbf{b}^T \mathbf{X}^T \mathbf{X} \mathbf{b}/n}{\|\mathbf{b}\|^2} \geq c_*,$$

where $\|\cdot\|$ is the L_2-norm.

Now we introduce an oracle inequality for Lasso in terms of L_0 sparsity.

Theorem 3

Suppose the RE condition holds. In the event $\{\|\mathbf{X}^T \boldsymbol{\varepsilon}\|_\infty / n \leq \lambda/2\}$, we have

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|^2 \leq 16 \frac{\lambda^2}{c_*^2} \|\boldsymbol{\beta}^o\|_0.$$

Proof: By the basic Lemma and the RE condition, on the set $\{\|\mathbf{X}^T\boldsymbol{\varepsilon}\|_\infty/n \leq \lambda/2\}$,

$$\frac{c_*}{2}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|^2 \leq \frac{1}{2n}\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o)\|^2 \leq 2\lambda\|\hat{\boldsymbol{\beta}}_{\mathcal{O}} - \boldsymbol{\beta}^o_{\mathcal{O}}\|_1.$$

By Cauchy-Schwarz inequality, it follows that

$$\frac{c_*}{2}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|^2 \leq 2\lambda\sqrt{|\mathcal{O}|}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|.$$

Using $2ab \leq 4a^2 + b^2/4$,

$$\frac{c_*}{2}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|^2 \leq \frac{4\lambda^2|\mathcal{O}|}{c_*} + \frac{c_*}{4}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|^2.$$

This implies the desired result.

A Remark on Restricted Eigenvalue Assumption

If we relax the RE condition to: There exists a constant $c_* > 0$ such that, for any $\mathbf{b} \in \mathbb{R}^p$ that satisfies

$$\sum_{j \neq \mathcal{O}} |b_j| \leq 4\sqrt{|\mathcal{O}|}\sqrt{\sum_{j \in \mathcal{O}} b_j^2},$$

it holds that

$$\frac{1}{n}\|\mathbf{X}\mathbf{b}\|^2 \geq c_* \sum_{j \in \mathcal{O}} b_j^2.$$

Then, in the event $\{\|\mathbf{X}^T\boldsymbol{\varepsilon}\|_\infty/n \leq \lambda/2\}$, we have

$$\frac{1}{n}\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o)\|^2 \leq 16\frac{\lambda^2}{c_*^2}|\mathcal{O}|.$$

This implies

$$\|\hat{\boldsymbol{\beta}}_{\mathcal{O}} - \boldsymbol{\beta}^o_{\mathcal{O}}\|^2 \leq 16\frac{\lambda^2}{c_*^2}|\mathcal{O}|,$$

and

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 \leq 4\sqrt{|\mathcal{O}|}\|\hat{\boldsymbol{\beta}}_{\mathcal{O}} - \boldsymbol{\beta}^o_{\mathcal{O}}\| \leq 4\sqrt{|\mathcal{O}|}4\frac{\lambda}{c_*}\sqrt{|\mathcal{O}|} \leq 16\frac{\lambda}{c_*}|\mathcal{O}|.$$

Implication of Theorem 3

- Theorem 3 presents us an oracle inequality based on L_0 sparsity.
- Note that λ is of order $\sqrt{\log(p)/n}$, which means $n^{-1}\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o)\|^2$ is of order $\log(p)|O|/n$.
- It implies that, up to the $\log p$ term (and the compatibility constant c_*^2), the mean squared prediction error is of the same order as of one know a priori which covariates are relevant and using ordinary LSE based on the true relevant $|\mathcal{O}|$ variables only.
- The $\log p$ factor can be viewed as the price we have to pay for not knowing the oracle set.

Outline

Sign Consistency

For Lasso method, another important question is: under what conditions, the Lasso estimator $\hat{\boldsymbol{\beta}}$ is selection consistent, in the sense that, with high probability, if $\beta_j^o = 0$, then $\hat{\beta}_j = 0$, and if $\beta_j^o \neq 0$, then $\hat{\beta}_j \neq 0$.

Actually, we want to ask for a stronger result, that is, under what conditions, $\hat{\boldsymbol{\beta}}$ is sign consistent.

What is **Sign Consistency**?

For any vectors $\mathbf{x}$ and $\mathbf{y}$, we say $\mathbf{x} =_s \mathbf{y}$ if $\text{sgn}(\mathbf{x}) = \text{sgn}(\mathbf{y})$ componentwise, where $\text{sgn}(x) = -1, 0$, or 1 if $x < 0$, $= 0$ or > 0.

Our goal is to find reasonable sufficient conditions under which

$$P(\hat{\boldsymbol{\beta}}(\lambda) =_s \boldsymbol{\beta}^o) \to 1.$$

We will consider the fixed design case.

Denote $\lambda_n = n\lambda$.

Theorem: A Characterization of Lasso Solution

Denote the gradient of $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$ by $G(\boldsymbol{\beta}) = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$. Then a necessary and sufficient condition for $\hat{\boldsymbol{\beta}}$ to be a Lasso solution is

$$\begin{aligned} G_j(\hat{\boldsymbol{\beta}}) = \mathbf{x}_j^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) &= \lambda_n \text{sgn}(\hat{\beta}_j), \quad & \hat{\beta}_j \neq 0, \\ |G_j(\hat{\boldsymbol{\beta}})| = |\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})| &\leq \lambda_n, & \hat{\beta}_j = 0. \end{aligned} \tag{3}$$

Moreover, if the Lasso solution is not unique (e.g., if $p > n$) and $G_j(\hat{\boldsymbol{\beta}}) < \lambda$ for some solution $\hat{\boldsymbol{\beta}}$, then $\hat{\beta}_j = 0$ for all Lasso solutions.

Proof: For the first statements regarding a necessary and sufficient characterization of the solution, we invoke sub-differential calculus. Denote the criterion function by

$$Q_\lambda(\boldsymbol{\beta}) = \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1.$$

For a minimizer $\hat{\boldsymbol{\beta}}(\lambda)$ of $Q_\lambda(\cdot)$ it is necessary and sufficient that the subdifferential at $\hat{\boldsymbol{\beta}}(\lambda)$ is zero. If the j-th component $\hat{\beta}_j(\lambda) \neq 0$, this means that the ordinary first derivative at $\hat{\boldsymbol{\beta}}(\lambda)$ has to be zero:

$$\frac{\partial Q_\lambda(\boldsymbol{\beta})}{\partial \beta_j}\bigg|_{\beta=\hat{\boldsymbol{\beta}}(\lambda)} = -\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})/n + \lambda \text{sgn}(\beta_j)\bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}(\lambda)} = 0.$$

This is equivalent to the first condition stated in the theorem.

On the other hand, if $\hat{\beta}_j(\lambda) = 0$, the subdifferential at $\hat{\boldsymbol{\beta}}(\lambda)$ has to include the zero element. That is

$$G_j(\hat{\boldsymbol{\beta}}(\lambda)) + \lambda e = 0, \quad \text{for some } e \in [-1, 1].$$

But this is equivalent to

$$|G_j(\hat{\boldsymbol{\beta}}(\lambda))| \leq \lambda \quad \text{if } \hat{\beta}_j(\lambda) = 0,$$

and this is equivalent to the second condition stated in the theorem.

Regarding to the uniqueness of the zeros among different solutions we argue as follows. Assume that there exist two solutions $\hat{\boldsymbol{\beta}}^{(1)}$ and $\hat{\boldsymbol{\beta}}^{(2)}$ such that for a component j we have $\hat{\beta}_j^{(1)} = 0$ with $|G_j(\hat{\boldsymbol{\beta}}^{(1)})| < \lambda$, but $\hat{\beta}_j^{(2)} \neq 0$. Because the set of all solutions is convex,

$$\hat{\boldsymbol{\beta}}_\rho = (1-\rho)\hat{\boldsymbol{\beta}}^{(1)} + \rho\hat{\boldsymbol{\beta}}^{(2)}$$

is also a minimizer for all $\rho \in [0,1]$. By assumption and for $0 < \rho < 1$, $\hat{\beta}_{\rho,j} \neq 0$ and hence, by the first statement from the KKT conditions, $G_j(\hat{\beta}_\rho) = \lambda_n$ for all $\rho \in (0,1)$. Hence, it holds for $g(\rho) = |G_j(\hat{\boldsymbol{\beta}}_\rho)|$ that $g(0) < \lambda$ and $g(\rho) = \lambda$ for all $\rho \in [0,1]$. But this contradicts to the fact that $g(\cdot)$ is continuous. Hence, a non-active (i.e., zero) component j with $|G_j(\hat{\boldsymbol{\beta}})| < \lambda$ cannot be active (i.e. non-zero) in any other solution.

Remark:

(i). The necessary and sufficient conditions stated in the above theorem is known as Karush-Kunh-Tucker conditions (KKT Conditions).

(ii). Question: When is the Lasso solution unique?

Derivation of Irrepresentable Condition

Let $\mathbf{s}_1 = (\text{sgn}(\beta_{0j}), j \in \mathcal{O})'$, and $\mathbf{X}_1 = (\mathbf{x}_j : j \in \mathcal{O})$ and

$$\hat{\boldsymbol{\beta}}_{n1} = (\mathbf{X}_1^T \mathbf{X}_1)^{-1}(\mathbf{X}_1^T \mathbf{y} - \lambda_n \mathbf{s}_1) = \boldsymbol{\beta}_{01} + \Sigma_{11}^{-1}(\mathbf{X}_1^T \boldsymbol{\varepsilon} - \lambda_n \mathbf{s}_1)/n, \tag{4}$$

where $\Sigma_{11} = (\mathbf{X}_1^T \mathbf{X}_1)/n$.

If $\hat{\boldsymbol{\beta}}_{n1} =_s \boldsymbol{\beta}_{01}$, then the equation (3) holds for $\hat{\boldsymbol{\beta}}_n = (\hat{\boldsymbol{\beta}}_{n1}^T, \mathbf{0}^T)^T$. Thus, since $\mathbf{X}\hat{\boldsymbol{\beta}}_n = \mathbf{X}_1 \hat{\boldsymbol{\beta}}_{n1}$ for this $\hat{\boldsymbol{\beta}}_n$ and $(\mathbf{x}_j : j \in \mathcal{O})$ are linearly independent,

$$\hat{\boldsymbol{\beta}}_n =_s \boldsymbol{\beta}_0 \quad \text{if} \quad \begin{cases} \hat{\boldsymbol{\beta}}_{n1} =_s \boldsymbol{\beta}_{01}, \\ |\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}_1 \hat{\boldsymbol{\beta}}_{n1})| \le \lambda_n, \quad \forall j \notin \mathcal{O}. \end{cases} \tag{5}$$

Let $H_n = I_n - \mathbf{X}_1 \Sigma_{11}^{-1} \mathbf{X}_1^T/n$. It follows from (4) that

$$\mathbf{y} - \mathbf{X}_1 \hat{\boldsymbol{\beta}}_{n1} = \boldsymbol{\varepsilon} - \mathbf{X}_1(\hat{\boldsymbol{\beta}}_{n1} - \boldsymbol{\beta}_{01}) = H_n \boldsymbol{\varepsilon} + \mathbf{X}_1 \Sigma_{11}^{-1} \mathbf{s}_1 \lambda_n / n$$

so that by (5),

$$\hat{\boldsymbol{\beta}}_n =_s \boldsymbol{\beta}_0 \quad \text{if} \quad \begin{cases} |\beta_j^o - \hat{\beta}_{nj}| \le |\beta_j^o|, \quad \forall j \in \mathcal{O} \\ |\mathbf{x}_j^T(H_n \boldsymbol{\varepsilon} + \mathbf{X}_1 \Sigma_{11}^{-1} \mathbf{s}_1 \lambda_n / n)| \le \lambda_n, \quad \forall j \notin \mathcal{O}. \end{cases} \tag{6}$$

By (4),

$$\hat{\boldsymbol{\beta}}_n =_s \boldsymbol{\beta}_0 \quad \text{if} \quad \begin{cases} |\mathbf{e}_j^T \Sigma_{11}^{-1}(\mathbf{X}_1^T \boldsymbol{\varepsilon} - \lambda_n \mathbf{s}_1)/n| \le |\beta_j^o|, & \forall j \in \mathcal{O} \\ |\mathbf{x}_j^T (H_n \boldsymbol{\varepsilon} + \mathbf{X}_1 \Sigma_{11}^{-1} \mathbf{s}_1 \lambda_n / n)| \le \lambda_n, & \forall j \notin \mathcal{O}, \end{cases} \tag{7}$$

where $\mathbf{e}_j$ is the unit vector in the direction of the j-th coordinate. Therefore,

$$\hat{\boldsymbol{\beta}}_n =_s \boldsymbol{\beta}_0 \quad \text{if} \quad \begin{cases} |\mathbf{e}_j^T \Sigma_{11}^{-1} \mathbf{X}_1^T \boldsymbol{\varepsilon}|/n + \lambda_n |\mathbf{e}_j^T \Sigma_{11}^{-1} \mathbf{s}_1 / n| \le |\beta_j^o|, & \forall j \in \mathcal{O} \\ |\mathbf{x}_j^T H_n \boldsymbol{\varepsilon}| \le \lambda_n (1 - |\mathbf{x}_j^T \mathbf{X}_1 \Sigma_{11}^{-1} \mathbf{s}_1|/n), & \forall j \notin \mathcal{O}. \end{cases} \tag{8}$$

If $1 - |\mathbf{X}_j^T \mathbf{X}_1 \Sigma_{11}^{-1} \mathbf{s}_1|/n \ge \eta$ for some $\eta > 0$, then a sufficient condition for (8) is

$$\hat{\boldsymbol{\beta}}_n =_s \boldsymbol{\beta}_0 \quad \text{if} \quad \begin{cases} |\mathbf{e}_j^T \Sigma_{11}^{-1} \mathbf{X}_1^T \boldsymbol{\varepsilon}|/n + \lambda_n |\mathbf{e}_j^T \Sigma_{11}^{-1} \mathbf{s}_1 / n| \le |\beta_j^o|, & \forall j \in \mathcal{O} \\ |\mathbf{x}_j^T H_n \boldsymbol{\varepsilon}| \le \lambda_n \eta, & \forall j \notin \mathcal{O}. \end{cases} \tag{9}$$

Let $\beta_* = \min\{|\beta_j^o| : j \in \mathcal{O}\}$. Define

$$\Omega_1 = \left\{ \max_{j \in \mathcal{O}} \left(|\mathbf{e}_j^T \Sigma_{11}^{-1} \mathbf{X}_1^T \boldsymbol{\varepsilon}| + \lambda_n |\mathbf{e}_j \Sigma_{11}^{-1} \mathbf{s}|/n \right) \geq \beta_* \right\},$$

$$\Omega_2 = \left\{ \max_{j \notin \mathcal{O}} |\mathbf{x}_j^T H_n \boldsymbol{\varepsilon}| \geq \lambda_n \eta \right\}.$$

Theorem

Suppose that $|\mathbf{X}_j^T \mathbf{X}_1 \Sigma_{11}^{-1} \mathbf{s}_1|/n \leq 1 - \eta$ for some $\eta > 0$. Then

$$P(\hat{\boldsymbol{\beta}}_n \neq_s \boldsymbol{\beta}^o) \leq P(\Omega_1) + P(\Omega_2).$$

Therefore, $P(\Omega_1) + P(\Omega_2) \to 0 \Longrightarrow P(\hat{\boldsymbol{\beta}}_n =_s \boldsymbol{\beta}^o) \to 0$. If this holds, we say that $\hat{\boldsymbol{\beta}}_n$ is sign consistent. Note that sign consistency implies selection consistency.

We will

(1). Find an upper bound for $P(\Omega_1) + P(\Omega_2)$,

(2). Find conditions to ensure that the upper bound converges to zero.

First consider $P(\Omega_2) = P(\max_{j\notin\mathcal{O}} |\mathbf{x}_j^T H_n \boldsymbol{\varepsilon}| \geq n\lambda\eta)$.

Note that $\mathbf{x}_j^T H_n \boldsymbol{\varepsilon}$ is normal with

$$E(\mathbf{x}_j H_n \boldsymbol{\varepsilon}/\sqrt{n}) = 0, \quad \mathrm{Var}(\mathbf{x}_j H_n \boldsymbol{\varepsilon}/\sqrt{n}) = \mathbf{x}_j^T H_n \mathbf{x}_j \sigma^2/n \leq \sigma^2,$$

we have

$$\begin{aligned} P(\Omega_2) &= P(\max_{j\notin\mathcal{O}} |\mathbf{x}_j^T H_n \boldsymbol{\varepsilon}|/(\sigma\sqrt{n}) > \sqrt{n}\lambda\eta/\sigma) \\ &\leq 2(p-d^o)(1-\Phi(\sqrt{n}\lambda\eta/\sigma)) \\ &\leq 2(p-d^o)\frac{\sigma}{\sqrt{2\pi}\eta\lambda\sqrt{n}}\exp\left(-\frac{-\eta^2 n\lambda^2}{2\sigma^2}\right). \end{aligned}$$

When $\eta\sqrt{n}\lambda/\sigma \geq a_0\sqrt{\log(p-d^o)}$ with $a_0 \geq \sqrt{2}$, we have

$$P(\Omega_2) \leq \frac{2(p-d^o)^{1-a_0^2/2}}{a_0\sqrt{2\pi\log(p-d^o)}}.$$

Now consider $P(\Omega_1)$. We have

$$P(\Omega_1) \leq P\left(\max_{j\in\mathcal{O}} |\mathbf{e}_j^T \Sigma_{11}^{-1} \mathbf{X}_1^T \boldsymbol{\varepsilon}|/n \geq \beta_*/2\right) + P\left(\lambda_n \max_{j\in\mathcal{O}} |\mathbf{e}_j \Sigma_{11}^{-1} \mathbf{s}_1|/n \geq \beta_*/2\right). \quad (10)$$

Denote the smallest eigenvalue of Σ_{11} by c_1. We assume that $c_1 > 0$.

For the second term in (10), we have

$$|\mathbf{e}_j \Sigma_{11}^{-1} \mathbf{s}_1| \leq \|\mathbf{e}_j\| \cdot \|\Sigma_{11}^{-1}\| \cdot \|\mathbf{s}_1\| \leq c_1^{-1}\sqrt{d^o}.$$

Thus, we want

$$\beta_* > 2c_1^{-1}\lambda\sqrt{d^o}$$

which is needed for sign consistency.

Implication: The above inequality simply says that the original nonzero parameters cannot be too small, which indeed is a pretty natural requirement.

For the first term in (10), since $\mathbf{e}_j^T \Sigma_{11}^{-1} \mathbf{X}_1^T \boldsymbol{\varepsilon}/\sqrt{n} \sim N(0, \sigma_j^2)$, where $\sigma_j^2 = \mathbf{e}_j^T \Sigma_{11}^{-1} \mathbf{e}_j \sigma^2 \leq \sigma^2/c_1$.

Therefore, the first term in (10) is bounded above by

$$\begin{aligned} & \sum_{j \in \mathcal{O}} P\left(|\mathbf{e}_j^T \Sigma_{11}^{-1} \mathbf{X}_1^T \boldsymbol{\varepsilon}|/(\sigma_j \sqrt{n}) \geq \sqrt{n}\beta_*/(2\sigma_j) \right) \\ \leq \quad & 2 \sum_{j \in \mathcal{O}} (1 - \Phi(\beta_* \sqrt{n}/(2\sigma_j))) \leq 2d^o (1 - \Phi(\sqrt{c_1}\beta_* \sqrt{n}/(2\sigma))) \\ \leq \quad & \frac{4d^o \sigma}{\beta_* \sqrt{nc_1}} \exp\left(-\frac{\beta_*^2 n c_1}{8\sigma^2} \right). \end{aligned}$$

Note that when $\beta_* > 2\lambda\sqrt{d^o}/c_1$ and $\sqrt{n}\lambda \geq a_0 \sigma \sqrt{\log(p - d^o)}/\eta$ with $a_0 \geq \sqrt{2}$, we have

$$P\left(\max_{j \in \mathcal{O}} |\mathbf{e}_j^T \Sigma_{11}^{-1} \mathbf{X}_1^T \boldsymbol{\varepsilon}|/n \geq \beta_*/2 \right) \leq \frac{2\eta\sqrt{c_1 d^o}}{\sqrt{\log(p - d^o)}} \exp\left(-\frac{a_0^2 d^o \log(p - d^o)}{2c_1 \eta^2} \right).$$

We summarize the above calculation in the following theorem.

Theorem 4: Sign Consistency

Suppose

(a). $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ i.i.d. $\sim N(0, \sigma^2)$,

(b). (Irrepresentable Condition) $|\mathbf{x}_j^T \mathbf{X}_1 \Sigma_{11}^{-1} \mathbf{s}_1|/n < 1 - \eta$ for some $\eta > 0$,

(c). $c_1 > 0$, $\beta_* > 2\lambda\sqrt{d^o}/c_1$ and $\sqrt{n}\lambda \geq a_0 \sigma \sqrt{\log(p - d^o)}/\eta$ with $a_0 \geq \sqrt{2}$.

Then

$$P(\hat{\boldsymbol{\beta}}_n \neq_s \boldsymbol{\beta}^o) \leq P(\Omega_1) + P(\Omega_2),$$

where

$$P(\Omega_1) \leq \frac{2\eta\sqrt{c_1 d^o}}{\sqrt{\log(p - d^o)}} \exp\left(-\frac{a_0^2 d^o \log(p - d^o)}{2c_1\eta^2}\right)$$

and

$$P(\Omega_2) \leq \frac{2(p - d^o)^{1 - a_0^2/2}}{a_0\sqrt{2\pi \log(p - d^o)}}.$$

Consequently, $P(\hat{\boldsymbol{\beta}}_n \neq_s \boldsymbol{\beta}^o) \to 0$ as $p - d^o \to \infty$.

Outline

References

- Bickel, P., Li, B. (2006). Regularization in statistics. *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research* 15(2), 271-344.
- Bock, M.E., Judge, G.G. and Yancey, T.A. (1984). A simple form for the inverse moments of non-central χ^2 and F random variables and certain confluent hypergeometric functions. *Journal of Econometrics* 25(1-2), 217-234.
- Candès, E. and T. Tao (2007). The dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* 35(6), 2313-2351.
- Candès, E. J. (2006). Modern statistical estimation via oracle inequalities. *Acta Numerica.* 15, 257-325.
- Chatterjee, A. (2011). Bootstrapping Lasso Estimators. *Journal of the American Statistical Association* 106(494), 608-625.
- Donoho, D. L. (1995, May). De-noising by soft-thresholding. *IEEE Transactions on Information Theory* 41(3), 613-627.
- Efron, B. and C. Morris (1975). Data Analysis Using Stein Estimator and its Generalizations. *Journal of the American Statistical Association* 70(350), 311-319.

- James, W. and J. Stein (1961). Estimation with Quadratic Loss. In J. Neyman (Ed.), *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability,* 361-379. University of California Press.
- Knight, K. and W. Fu (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics* 28(5), 1356-1378.
- Stein, C. M. (1981). Estimation of the Mean of a Multivariate Normal Distribution. *The Annals of Statistics* 9(6), 1135-1151.
- Zhao, P. and B. Yu (2006). On model selection consistency of lasso. *J. Mach. Learn. Res.* 7, 2541-2563.