ETH zürich

Statistical Inference for the Lasso

Yilei Zhang, Yici Yan

Content

Why should we do inference?

Why should we do inference?

We are trying to answer questions:

- What can we learn about the underlying distribution from the observed set of outcomes of a given sample size?
- How confident can we be of the inference based on a certain number of observations and a certain experiment design?

  $\Rightarrow$ The confidence affects the decision-making process

Why should we do inference?

We are trying to answer questions:

- What can we learn about the underlying distribution from the observed set of outcomes of a given sample size?
- How confident can we be of the inference based on a certain number of observations and a certain experiment design?

  $\Rightarrow$ The confidence affects the decision-making process

For lasso:

An attractive feature of l_1-regularized procedures is their ability to combine variable selection with parameter fitting. **It is sometimes of interest to determine the statistical strength of the included variables, as in "p-value" in traditional models.**

The Bayesian lasso

The Bayesian paradigm **treats the parameters as random quantities**, along with a **prior distribution** that characterizes our belief in what their values might be:

$$y|\beta, \lambda, \sigma \sim N(\mathbf{X}\beta, \sigma^2 \mathbf{I}_{N\times N}), \tag{6.1a}$$

$$\beta|\lambda, \sigma \sim \prod_{j=1}^{p} \frac{\lambda}{2\sigma} e^{-\frac{\lambda}{\sigma}|\beta_j|}, \tag{6.1b}$$

using the i.i.d Laplacian prior (6.1b). Under this model, the negative log posterior density for $\beta|y, \lambda, \sigma$ is given by:

$$\frac{1}{2\sigma^2}\|y - \mathbf{X}\beta\|_2^2 + \frac{\lambda}{\sigma}\|\beta\|_1, \tag{6.2}$$

where we have dropped an additive constant independent of β, and we assume that the columns of X are mean-centered, as is y.

The Bayesian lasso

The negative log posterior density for $\beta|y, \lambda, \sigma$ is given by:

$$\frac{1}{2\sigma^2}\|y - \mathbf{X}\beta\|_2^2 + \frac{\lambda}{\sigma}\|\beta\|_1. \tag{6.2}$$

Consequently, for any fixed values of σ and λ, **the posterior mode coincides with the lasso estimate(with regularization parameter $\sigma\lambda$)**.

In classical lasso, we don't assume any distribution of y, thus the distribution of $\hat{\beta}(y, \lambda)$ is unknown, but under the assumptions of Bayesian lasso, we are now able to get the distribution of $\hat{\beta}(y, \lambda)$, and then making inference is possible.

Main Ideas

Given distribution $f(\beta|\lambda,\sigma)$ and $f_X(y|\beta,\lambda,\sigma)$

Main Ideas

Given distribution $f(\beta|\lambda,\sigma)$ and $f_X(y|\beta,\lambda,\sigma)$

$\Downarrow$ *Bayes' theorem*

Get posterior distribution $h_X(\beta|y,\lambda,\sigma)$,
and $\hat{\beta}(y,\lambda,\sigma) = argmin_\beta - log(h_X)$

Main Ideas

Given distribution $f(\beta|\lambda,\sigma)$ and $f_X(y|\beta,\lambda,\sigma)$

$\Downarrow$ *Bayes' theorem*

Get posterior distribution $h_X(\beta|y,\lambda,\sigma)$,
and $\hat{\beta}(y,\lambda,\sigma) = argmin_\beta - log(h_X)$

$\Downarrow$

Get posterior distribution $g_x(\hat{\beta}|\lambda,\sigma)$
($\hat{\beta}$ is a function of y, and distribution of $y|\lambda,\sigma$ is known.)

Main Ideas

Given distribution $f(\beta|\lambda,\sigma)$ and $f_X(y|\beta,\lambda,\sigma)$

$\Downarrow$ *Bayes' theorem*

Get posterior distribution $h_X(\beta|y,\lambda,\sigma)$,
and $\hat{\beta}(y,\lambda,\sigma) = argmin_\beta - log(h_X)$

$\Downarrow$

Get posterior distribution $g_x(\hat{\beta}|\lambda,\sigma)$
($\hat{\beta}$ is a function of y, and distribution of $y|\lambda,\sigma$ is known.)

PROBLEM: in practice, exact Bayesian calculations are typically intractable, except for the simplest models.

Bayesian Lasso Simulations

PROBLEM: In practice, exact Bayesian calculations are typically intractable.
SOLUTION: simulations!

1. Given 100 λs: $\lambda_1, \lambda_2, \ldots, \lambda_{100}$
2. For each λ_i:
   a. sample 10,000 βs from $f(\beta|\lambda_i, \sigma)$ (6.1b);
   b. for each β, we sample y from $f_X(y|\beta, \lambda, \sigma)$ (6.1a);
   c. for each y we get, we use MCMC to sample realizations from distribution $h_X(\beta|y, \lambda, \sigma)$, then we can empirically estimate $\hat{\beta}(y)$.
3. For each λ_i, we obtain 10,000 $\hat{\beta}$s, thus we get the empirical distribution of $\hat{\beta}$ and confidence intervals follow.

Bayesian lasso results

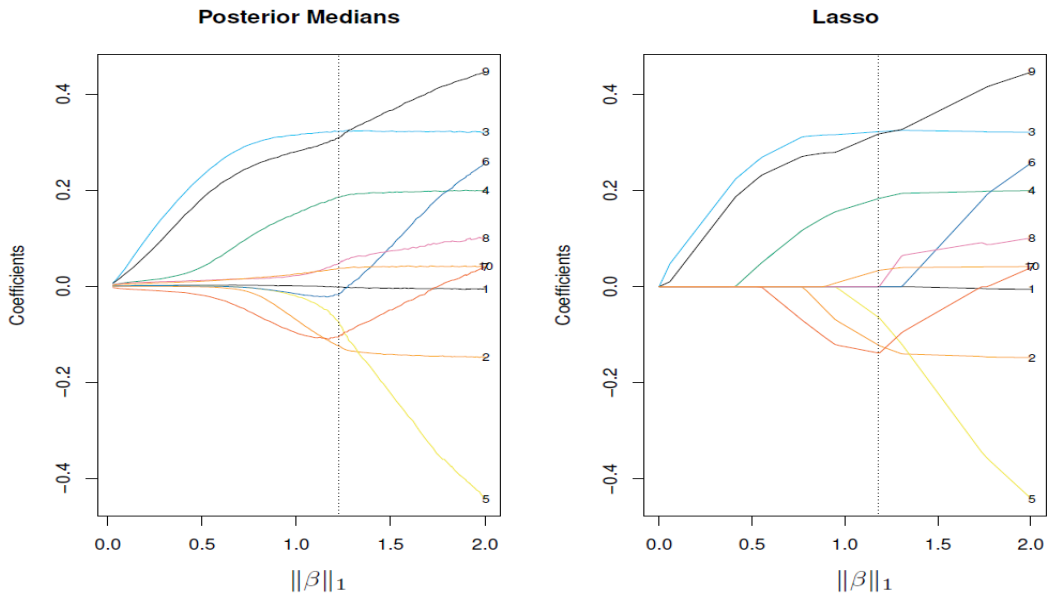


Figure 6.2 *Bayesian lasso on the diabetes data. The left plot shows the posterior medians from MCMC runs (conditional on λ). The right plot shows the lasso profile. In the left plot, the vertical line is at the posterior median of $\|\beta\|_1$ (from an unconditional model), while for the right plot the vertical line was found by N-fold cross-validation.*

Source: Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press, 2015, page 141.

Bayesian lasso results

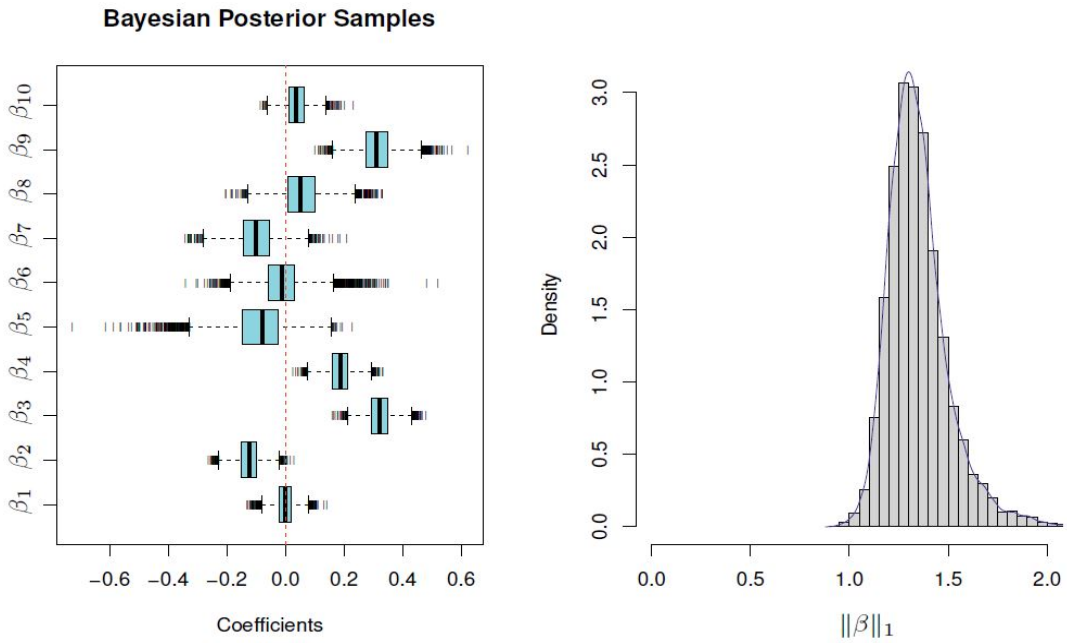


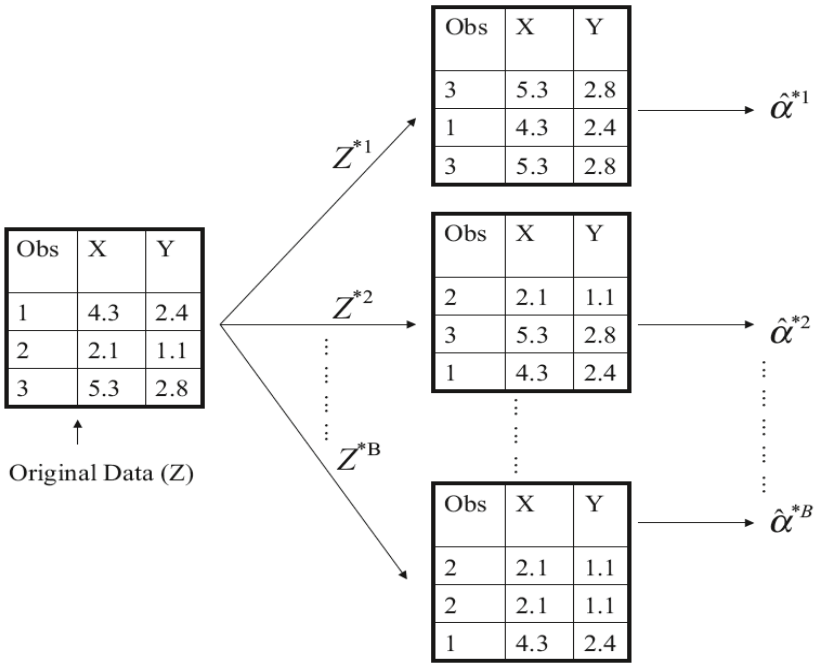
Figure 6.3 *Posterior distributions for the β_j and $\|\beta\|_1$ for the diabetes data. Summary of $10,000$ MCMC samples, with the first 1000 "burn-in" samples discarded.*

Source: Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press, 2015, page 142.

The Bootstrap

- Set-up:
  - $Z_1, \ldots, Z_n \sim_{i.i.d} P$.
  - We are interested in some parameter θ of P.
  - We have an estimator $\hat{\theta}_n = g(Z_1, \ldots, Z_n)$ and we want to know the distribution of $\hat{\theta}_n$, so that we can make inference about θ
- Main idea:
  - If we knew P, we could simulate many data sets to obtain the distribution of $\hat{\theta}_n$.
  - Since we don't know P, we simulate from an estimated version.
  - In **non-parametric bootstrap** we simulate from the empirical distribution $\hat{P}_n$ which places mass 1/n on each data point. This amounts to resampling the data with replacement.
  - In **parametric bootstrap** we first estimate θ by some estimate $\hat{\theta}$, then simulate bootstrap samples from $P_{\hat{\theta}_n}$

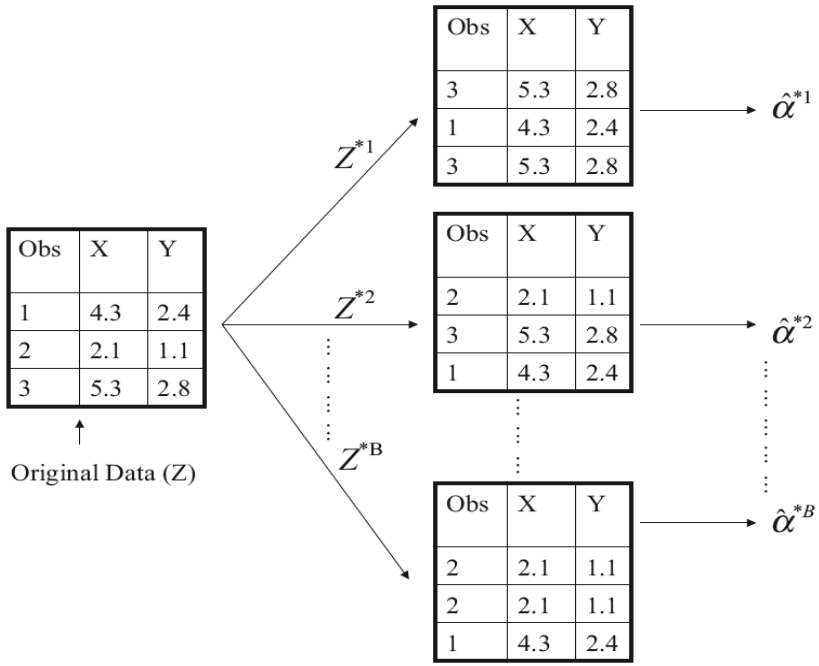
The Bootstrap



source: Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshiran, An Introduction to Statistical Learning with Applications in R, Springer, 2013 ,page 190.

- The distribution P^* of the bootstrapped estimates is a conditional distribution given the data.
- It is difficult to analyze P^* analytically, but easy to approximate it by simulation.

The Bootstrap



source: Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshiran, An Introduction to Statistical Learning with Applications in R, Springer, 2013 ,page 190.

- The distribution P^* of the bootstrapped estimates is a conditional distribution given the data.
- It is difficult to analyze P^* analytically, but easy to approximate it by simulation.

Bootstrap consistency for $\hat{\theta}_n$ with rate a_n:

$$P(a_n(\hat{\theta}_n - \theta)) \leq x) - P^*(a_n(\hat{\theta}_n^* - \hat{\theta}_n) \leq x) \xrightarrow{p} 0, \text{as n} \to \infty$$

Bootstrap for lasso inference

We first obtain an estimate $\hat{\beta}(\hat{\lambda}_{CV})$ for a lasso problem according to the following procedures:

1. Fit a lasso path to (X, y) over a dense grid of values $\Lambda = \{\lambda_l\}_{l=1}^{L}$.
2. Divide the training samples into 10 groups at random.
3. With the k^{th} group left out, fit a lasso path to the remaining 9/10ths, using the same grid Λ.
4. For each $\lambda \in \Lambda$ compute the mean-squared prediction error for the left-out group.
5. Average these errors to obtain a prediction error curve over the grid Λ.
6. Find the value $\hat{\lambda}_{CV}$ that minimizes this curve, and then return the coefficient vector from our original fit in step (1) at that value of λ.

Bootstrap for lasso inference

Then we approximate the cumulative distribution F of the random pair (X, Y) by the empirical CDF $\hat{F}_N$ defined by the N samples:

- Draw N samples from $\hat{F}_N$ as a bootstrap sample, which amounts to drawing N samples with replacement from the given data set.
- Obtain $\hat{\beta}^*(\hat{\lambda}_{CV})$ by repeating steps 1-6 on each bootstrap sample.
- Draw 1000 bootstrap samples, and use the 1000 bootstrap realizations $\hat{\beta}^*(\hat{\lambda}_{CV})$ to make inference.

Results for non-parametric bootstrap

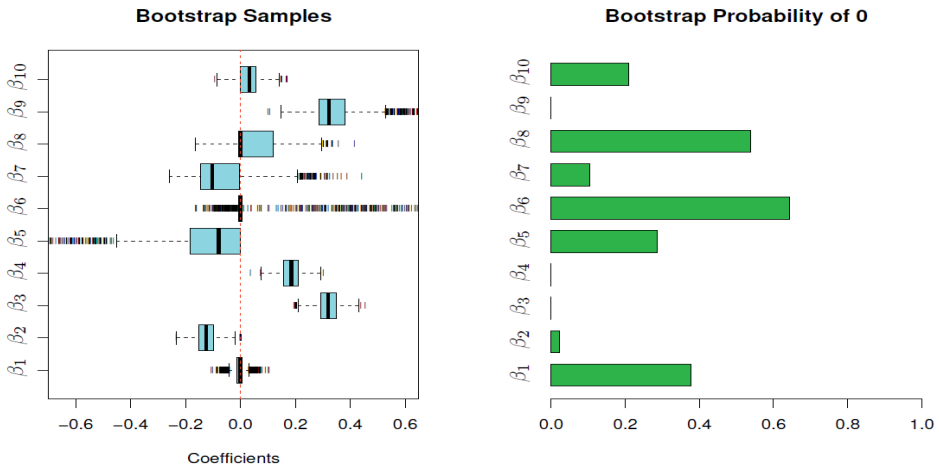


Figure 6.4 *[Left] Boxplots of 1000 bootstrap realizations of $\widehat{\beta}^*(\hat{\lambda}_{CV})$ obtained by the nonparametric bootstrap, which corresponds to re-sampling from the empirical CDF $\widehat{F}_N$. Comparing with the corresponding Bayesian posterior distribution in Figure 6.3, we see a close correspondence in this case. [Right] Proportion of times each coefficient is zero in the bootstrap distribution.*

There is a reasonable correspondence between this figure, and the corresponding Bayesian results in Figure 6.3.

Source: Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press, 2015, page 143

Parametric Bootstrap

In contrast to the non-parametric bootstrap, the parametric bootstrap samples from a parametric estimate of F:

- fix X and obtain estimates $\hat{\beta}$ and $\hat{\sigma}^2$ from the full least-squares fit $Y \sim X$.
- then we sample y^* from the Gaussian model (6.1a), that is, $y|\beta, \sigma \sim N(\mathbf{X}\beta, \sigma^2 \mathbf{I}_{N\times N})$.
- consider (X, y^*) as a new bootstrap sample, then repeat steps $1 \sim 6$ on this sample and get $\hat{\beta}^*(\hat{\lambda}_{CV})$.
- draw 1000 bootstrap samples, and use the 1000 bootstrap realizations $\hat{\beta}^*(\hat{\lambda}_{CV})$ to make inference.

Results for parametric bootstrap

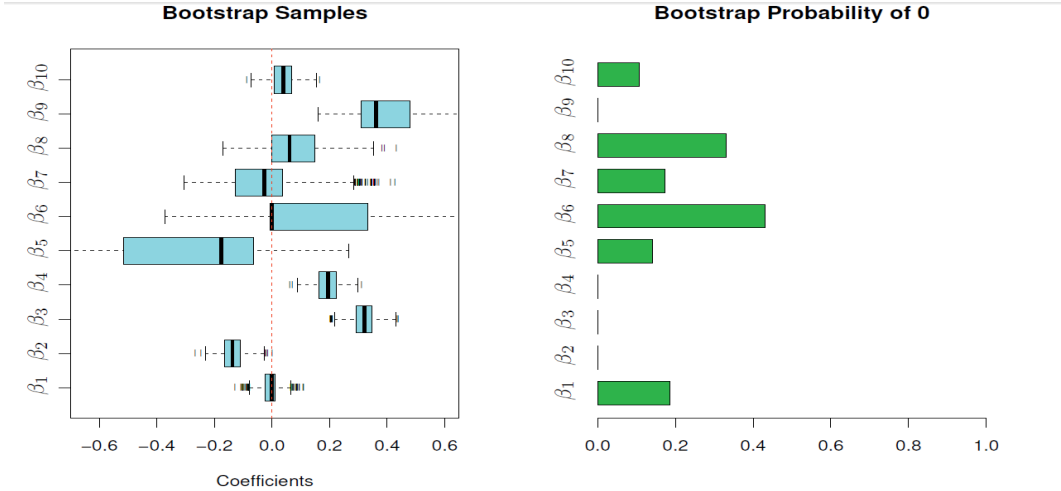


Figure 6.7 *[Left] Boxplots of 1000 parametric bootstrap realizations of $\widehat{\beta}^*(\hat{\lambda}_{CV})$. Comparing with the corresponding Bayesian posterior distribution in Figure 6.3, we again see a close correspondence. [Right] Proportion of times each coefficient is zero in the bootstrap distribution.*

The result is similar to both the non-parametric bootstrap results and those from the Bayesian lasso.

Source: Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press, 2015, page 146

Comparison of bootstrap and Bayesian lasso

Table 6.1 *Timings for Bayesian lasso and bootstrapped lasso, for four different problem sizes. The sample size is* $N = 400$.

p	Bayesian Lasso	Lasso/Bootstrap
10	3.3 secs	163.8 secs
50	184.8 secs	374.6 secs
100	28.6 mins	14.7 mins
200	4.5 hours	18.1 mins

- From table 6.1 we know Bayesian lasso is faster for small problems, but its complexity seems to scale as $O(p^2)$.
- In contrast, the scaling of the bootstrap seems to be closer to $O(p)$.
- As we move to GLMs, the Bayesian technical complexities grow, while bootstrap can be applied seamlessly in many situations.

source: Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press, 2015, page 145

Post-Selection Inference for the Lasso

In this section we present some relatively recent ideas on making inference after selection by adaptive methods such as lasso and forward-stepwise regression.

Assumptions:
The usual linear regression setup, with an outcome vector $y \in R^N$ and matrix of predictor variables $X \in R^{N \times p}$ related by:

$$y = \mathbf{X}\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I_{N \times N}),$$

where $\beta \in R^p$ are unknown coefficients to be estimated.

Problems with forward-stepwise regression

Traditional chi-square test:

Consider forward-stepwise regression. This procedure enters predictors one at a time, choosing the predictors that most decreases the residual sum of squares at each stage:

- Defining RSS_k to be the residual sum of squares for the model containing k predictors;
- we use this change in residual sum of squares to form a test statistic:

$$R_k = \frac{1}{\sigma^2}(RSS_{k-1} - RSS_k), \tag{6.4}$$

  with σ assumed to be known now.
- Compare R_k to a $\chi^2(1)$ distribution.

Problems with forward-stepwise regression

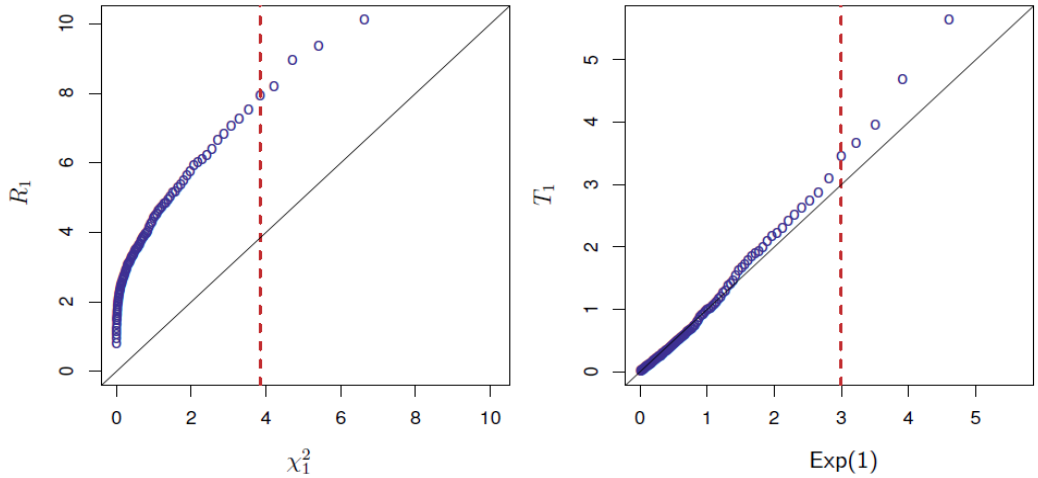
Suppose at step k-1, there are n candidates $x_1, \dots, x_n$ that can be added. For each candidate i, we can calculate:

$$R_k(i) = \frac{1}{\sigma^2}(RSS_{k-1} - RSS_k(i)),$$

where $RSS_k(i)$ is the RSS after adding the ith candidate. Then $R_k(i) \sim \chi^2(1)$ under null hypothesis($\beta_i = 0$), $\forall i \in 1, \dots, n$.

However, in forward-stepwise regression, at each step we intentionally choose the candidate with largest $R_k(i)$, that is, $R_k = max_i R_k(i)$, which is actually not $\chi^2(1)$ distribution anymore!

Problems with forward-stepwise regression



(a) Forward stepwise (b) Lasso

Figure 6.8 *A simulation example with $N = 100$ observations and $p = 10$ orthogonal predictors and $\beta = 0$. (a) a quantile-quantile plot, constructed over 1000 simulations, of the standard chi-squared statistic R_1 in (6.4), measuring the drop in residual sum-of-squares for the first predictor to enter in forward stepwise regression, versus the χ^2_1 distribution. The dashed vertical line marks the 95% quantile of the χ^2_1 distribution. (b) a quantile-quantile plot of the covariance test statistic T_1 in (6.5) for the first predictor to enter in the lasso path, versus its asymptotic null distribution Exp(1). The covariance test explicitly accounts for the adaptive nature of lasso modeling, whereas the chi-squared test is not appropriate for adaptively selected models as in forward-stepwise regression.*

Source: Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press, 2015, page 148.

The covariance test for lasso

Surprisingly, it turns out that for the lasso, a simple test can be derived that properly accounts for the adaptivity:

- Denote the knots returned by the LAR algorithm by $\lambda_1 > \lambda_2 \cdots > \lambda_K$. These are the values of regularization parameter λ where there is a change in the set of active predictors.
- Suppose that we wish to test significance of the predictor entered by LAR at λ_k.
- Let $\mathcal{A}_{k-1}$ be the active set (the predictors with nonzero coefficients) before this predictor was added;
- Let the estimate at the end of this step be $\hat{\beta}(\lambda_{k+1})$
- We refit the lasso, keeping $\lambda = \lambda_{k+1}$ but using just the variables in $\mathcal{A}_{k-1}$. This yields the estimate $\hat{\beta}_{\mathcal{A}_{k-1}}(\lambda_{k+1})$.

The covariance test for lasso

The covariance test statistic is defined by:

$$T_k = \frac{1}{\sigma^2}(\langle y, \mathbf{X}\hat{\beta}(\lambda_{k+1})\rangle - \langle y, \mathbf{X}\hat{\beta}_{\mathcal{A}_{k-1}}(\lambda_{k+1})\rangle). \tag{6.5}$$

Remark:

- The statistic measures how much of the covariance between the outcome and the fitted model can be attributed to the predictor that has just entered the model.
- Interestingly, for forward-stepwise regression, the corresponding covariance statistic is equal to R_k (6.4), however, for the lasso this is not the case.

The covariance test for lasso

The covariance test statistic is defined by:

$$T_k = \frac{1}{\sigma^2}(\langle y, \mathbf{X}\hat{\beta}(\lambda_{k+1})\rangle - \langle y, \mathbf{X}\hat{\beta}_{\mathcal{A}_{k-1}}(\lambda_{k+1})\rangle). \tag{6.5}$$

Remark:

- Remarkably, under the null hypothesis that all k-1 signal variables are in the model, and under general conditions on the model matrix X, for the predictor entered at the next step we have:

$$T_k \xrightarrow{d} Exp(1), \text{as N,p} \longrightarrow \infty$$

- When σ^2 is unknown, we estimate it using the full model: $\hat{\sigma}^2 = \frac{1}{N-p} RSS_p$. We plug this into (6.5), and the exponential test becomes an $F_{2,N-p}$ test.

The covariance test for lasso

The covariance test statistic is defined by:

$$T_k = \frac{1}{\sigma^2}(\langle y, \mathbf{X}\hat{\beta}(\lambda_{k+1})\rangle - \langle y, \mathbf{X}\hat{\beta}_{\mathcal{A}_{k-1}}(\lambda_{k+1})\rangle). \tag{6.5}$$

Remark:

- Fig 6.8(b) shows the quantile-quantile plot for T_1 versus Exp(1).
- Why is the mean of the forward-stepwise statistic R_1 much larger than one, while the mean of T_1 is approximately equal to one?
- The reason is shrinkage: the lasso picks the best predictor available at each stage, but does not fit it fully by least squares. It uses shrunken estimates of the coefficients, and this shrinkage compensates exactly for the inflation due to the selection.

The covariance test for lasso

Table 6.2 *Results of forward stepwise regression and LAR/lasso applied to the diabetes data introduced in Chapter 2. Only the first ten steps are shown in each case. The p-values are based on (6.4), (6.5), and (6.11), respectively. Values marked as 0 are* < 0.01.

Forward Stepwise			LAR/lasso		
Step	Term	p-value	Term	p-value	
				Covariance	Spacing
1	bmi	0	bmi	0	0
2	ltg	0	ltg	0	0
3	map	0	map	0	0.01
4	age:sex	0	hdl	0.02	0.02
5	bmi:map	0	bmi:map	0.27	0.26
6	hdl	0	age:sex	0.72	0.67
7	sex	0	glu^2	0.48	0.13
8	glu^2	0.02	bmi^2	0.97	0.86
9	age^2	0.11	age:map	0.88	0.27
10	tc:tch	0.21	age:glu	0.95	0.44

Remark:

- The forward-stepwise regression enters 8 terms at level 0.05, while the covariance test enters only 4.

Source: Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press, 2015, page 150.

General Scheme

We now propose the general scheme for post-selection inference yields exact p-values and confidence intervals in the Gaussian case.

General Scheme

We now propose the general scheme for post-selection inference yields exact p-values and confidence intervals in the Gaussian case.

- Claim: Selection events such as LAR, forward-stepwise regression, and the Lasso for fixed λ can be written as $\{\boldsymbol{A}y \leq b\}$ for some matrix $\boldsymbol{A}$ and vector b.

General Scheme

We now propose the general scheme for post-selection inference yields exact p-values and confidence intervals in the Gaussian case.

- Claim: Selection events such as LAR, forward-stepwise regression, and the Lasso for fixed λ can be written as $\{\boldsymbol{A}y \leq b\}$ for some matrix $\boldsymbol{A}$ and vector b.

Now suppose that $y \sim \boldsymbol{N}(\mu, \sigma^2 \boldsymbol{I})$, and that we want to make inferences conditional on the event $\{\boldsymbol{A}y \leq b\}$. In particular, we wish to make inferences about $\eta^T \mu$, where η might depend on the selection event.

General Scheme

We now propose the general scheme for post-selection inference yields exact p-values and confidence intervals in the Gaussian case.

- Claim: Selection events such as LAR, forward-stepwise regression, and the Lasso for fixed λ can be written as $\{\boldsymbol{A}y \leq b\}$ for some matrix $\boldsymbol{A}$ and vector b.

Now suppose that $y \sim \boldsymbol{N}(\mu, \sigma^2 \boldsymbol{I})$, and that we want to make inferences conditional on the event $\{\boldsymbol{A}y \leq b\}$. In particular, we wish to make inferences about $\eta^T \mu$, where η might depend on the selection event.
Simple OLS example: $\theta = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \mu$ is the regression coefficient vector upon which we want to make inference.

Polyhedral Lemma

The previous selection event can be expressed as

$$\{\boldsymbol{A}y \le b\} = \{V^-(y) \le \eta^T y \le V^+(y), V^0(y) \ge 0\},$$

where, denoting $\alpha = \boldsymbol{A}\eta / \| \eta \|_2^2$,

$$V^-(y) = \max_{j:\alpha_j<0} \frac{b_j - (\boldsymbol{A}y)_j + \alpha_j \eta^T y}{\alpha_j},$$

$$V^+(y) = \min_{j:\alpha_j>0} \frac{b_j - (\boldsymbol{A}y)_j + \alpha_j \eta^T y}{\alpha_j},$$

$$V^0(y) = \min_{j:\alpha_j=0} (b_j - (\boldsymbol{A}y)_j).$$

Furthermore, $\eta^T y$ and $(V^-(y), V^+(y), V^0(y))$ are statistically independent.

Intuitive Graphical Explanation

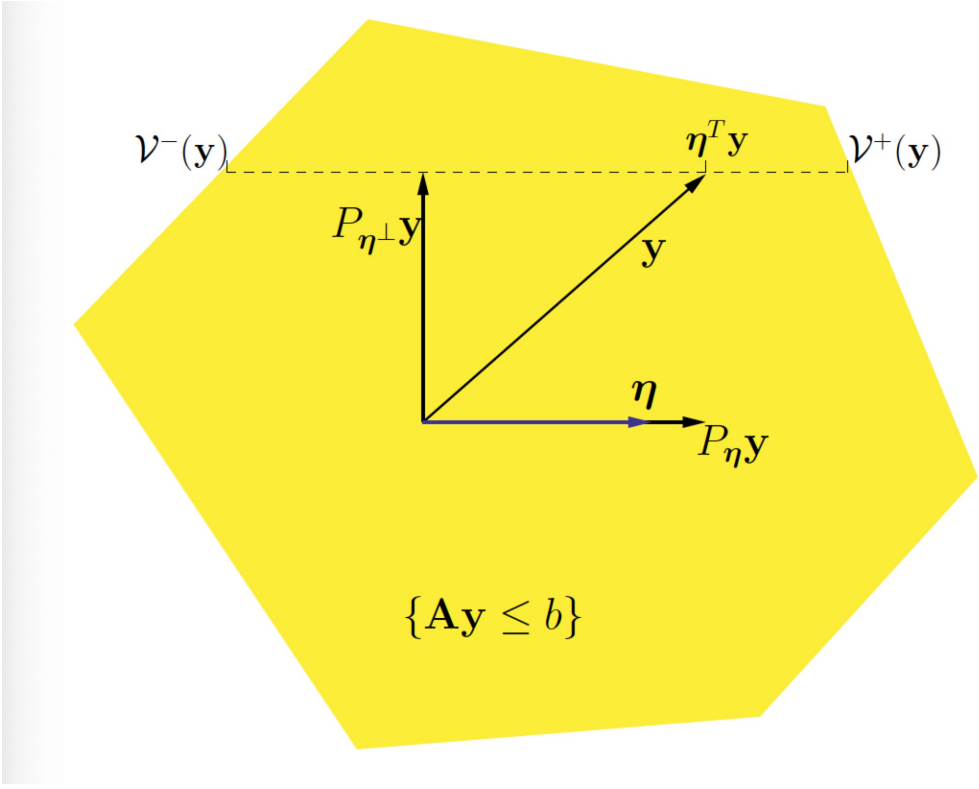


Figure: Source: Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press, 2015, page 152.

Finding the Pivot

Since y is Gaussian, the above lemma suggests that the conditional inference on $\eta^T \mu$ can be made using the truncated distribution of $\eta^T y$, which is a truncated normal distribution.

Finding the Pivot

Since y is Gaussian, the above lemma suggests that the conditional inference on $\eta^T \mu$ can be made using the truncated distribution of $\eta^T y$, which is a truncated normal distribution.
Concretely, denote

$$F^{c,d}_{\mu,\sigma^2}(x) = \frac{\mathbf{\Phi}((x-\mu)/\sigma) - \mathbf{\Phi}((c-\mu)/\sigma)}{\mathbf{\Phi}((d-\mu)/\sigma) - \mathbf{\Phi}((c-\mu)/\sigma)},$$

where $\mathbf{\Phi}$ is the CDF of the standard Gaussian. This is just how the truncated normal distribution with support on $[c, d]$ is defined.
It follows that

$$F^{V^-,V^+}_{\eta^T\mu,\sigma^2\|\eta\|_2^2}(\eta^T y) \mid \{\boldsymbol{A}y \le b\} \sim \boldsymbol{U}(0,1).$$

Spacing Test

We now apply the above inference procedure to successive steps of the LAR algorithm. For testing the global null hypothesis, we set $\eta^T y = \lambda_1 = \max_j |\langle x_j, y\rangle|$. We observe that $V^- = \lambda_2$, $V^+ = +\infty$, and hence

$$R_1 = 1 - F_{0,\sigma^2}^{\lambda_2,+\infty}(\lambda_1) \mid \{\boldsymbol{A}y \le b\} = \frac{1 - \boldsymbol{\Phi}(\lambda_1/\sigma)}{1 - \boldsymbol{\Phi}(\lambda_2/\sigma)} \sim \boldsymbol{U}(0,1).$$

Spacing Test

We now apply the above inference procedure to successive steps of the LAR algorithm. For testing the global null hypothesis, we set $\eta^T y = \lambda_1 = \max_j \mid \langle x_j, y \rangle \mid$. We observe that $V^- = \lambda_2$, $V^+ = +\infty$, and hence

$$R_1 = 1 - F_{0,\sigma^2}^{\lambda_2,+\infty}(\lambda_1) \mid \{\boldsymbol{A}y \leq b\} = \frac{1 - \boldsymbol{\Phi}(\lambda_1/\sigma)}{1 - \boldsymbol{\Phi}(\lambda_2/\sigma)} \sim \boldsymbol{U}(0,1).$$

Remark: this uniform distribution above holds exactly for finite N and p, and for any $\boldsymbol{X}$; and that it is a nonasymptotic version of the covariance test, and is asymptotically equivalent to it (Taylor et al. 2014).

Spacing Test

We now apply the above inference procedure to successive steps of the LAR algorithm. For testing the global null hypothesis, we set $\eta^T y = \lambda_1 = \max_j \mid \langle x_j, y \rangle \mid$. We observe that $V^- = \lambda_2$, $V^+ = +\infty$, and hence

$$R_1 = 1 - F_{0,\sigma^2}^{\lambda_2,+\infty}(\lambda_1) \mid \{\boldsymbol{A}y \leq b\} = \frac{1 - \boldsymbol{\Phi}(\lambda_1/\sigma)}{1 - \boldsymbol{\Phi}(\lambda_2/\sigma)} \sim \boldsymbol{U}(0,1).$$

Remark: this uniform distribution above holds exactly for finite N and p, and for any $\boldsymbol{X}$; and that it is a nonasymptotic version of the covariance test, and is asymptotically equivalent to it (Taylor et al. 2014).

In what follows, we use the same method to test other variables.

Test Results

LAR/lasso

Term	p-value	
	Covariance	Spacing
bmi	0	0
ltg	0	0
map	0	0.01
hdl	0.02	0.02
bmi:map	0.27	0.26
age:sex	0.72	0.67
glu^2	0.48	0.13
bmi^2	0.97	0.86
age:map	0.88	0.27
age:glu	0.95	0.44

Figure: Source: Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press, 2015, page 150.

What Hypothesis Is Being Tested?

What Hypothesis Is Being Tested?

Covariance Test (complete null hypothesis):

At each stage of LAR, we are testing whether the coefficients of all other predictors not yet in the model are zero.

What Hypothesis Is Being Tested?

Covariance Test (complete null hypothesis):
At each stage of LAR, we are testing whether the coefficients of all other predictors not yet in the model are zero.

Spacing test and Fixed λ test (incremental null hypothesis):
At the first step, it tests the global null hypothesis, as does the covariance test. But at subsequent steps, it tests whether the partial correlation of the given predictor entered at that step is zero, adjusting for other variables that are currently in the model.

Debiased Lasso

Aim: directly estimates confidence intervals for the full set of population regression parameters under an assumed linear model, instead of attempting to make inferences about the partial regression coefficients in models derived by LAR or the lasso.

Debiased Lasso

Aim: directly estimates confidence intervals for the full set of population regression parameters under an assumed linear model, instead of attempting to make inferences about the partial regression coefficients in models derived by LAR or the lasso.

Method: Use debias operation to "invert" the KKT conditions.

Debiased Lasso

Aim: directly estimates confidence intervals for the full set of population regression parameters under an assumed linear model, instead of attempting to make inferences about the partial regression coefficients in models derived by LAR or the lasso.

Method: Use debias operation to "invert" the KKT conditions.

Reference:Van de Geer, Sara, et al. "On asymptotically optimal confidence regions and tests for high-dimensional models." The Annals of Statistics 42.3 (2014): 1166-1202.

Debiased Lasso Setup

Consider as before the high dimensional linear model with Gaussian error $\epsilon \sim \boldsymbol{N}(0, \sigma_\epsilon \boldsymbol{I})$:

$$Y = \boldsymbol{X}\beta^0 + \epsilon.$$

Define the lasso as:

$$\hat{\beta} = \hat{\beta}(\lambda) := \arg\min_{\beta \in R^p}(\| Y - \boldsymbol{X}\beta \|_2^2 / n + 2\lambda \| \beta \|_1).$$

By the KKT Theorem,

$$\lambda\hat{\kappa} = \boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\hat{\beta})/n.$$

where $\| \hat{\kappa} \|_\infty \leq 1$ and $\hat{\kappa}_j = sign(\hat{\beta}_j)$ if $\hat{\beta}_j \neq 0$.

Inverting the KKT Conditions

Denote $\hat{\mathbf{\Sigma}} = \boldsymbol{X}^T \boldsymbol{X}/n$, thus

$$\hat{\mathbf{\Sigma}}(\hat{\beta} - \beta^0) + \lambda \hat{\kappa} = \boldsymbol{X}^T \epsilon / n.$$

We estimate the population quantity $\hat{\mathbf{\Theta}} := \hat{\mathbf{\Sigma}}^{-1}$ by nodewise regression on $\boldsymbol{X}$.

Inverting the KKT Conditions

Denote $\hat{\mathbf{\Sigma}} = \boldsymbol{X}^T \boldsymbol{X}/n$, thus

$$\hat{\mathbf{\Sigma}}(\hat{\beta} - \beta^0) + \lambda\hat{\kappa} = \boldsymbol{X}^T\epsilon/n.$$

We estimate the population quantity $\hat{\mathbf{\Theta}} := \hat{\mathbf{\Sigma}}^{-1}$ by nodewise regression on $\boldsymbol{X}$.

For each $j = 1, \ldots, p$, define

$$\hat{\gamma}_j := \underset{\gamma \in \mathbb{R}^{p-1}}{\arg\min}(\| X_j - \boldsymbol{X}_{-j}\gamma \|_2^2 /n + 2\lambda_j \| \gamma \|_1).$$

where X_j is the jth column of the design matrix, and $\boldsymbol{X}_{-j}$ denotes the design matrix without the jth column.

Inverting the KKT Conditions

Slightly abusing the notation, we see components of

$$\hat{\gamma}_j = \{\hat{\gamma}_{j,k}; k = 1, \dots, p, k \neq j\}.$$

Inverting the KKT Conditions

Slightly abusing the notation, we see components of

$$\hat{\gamma}_j = \{\hat{\gamma}_{j,k}; k = 1, \dots, p, k \neq j\}.$$

Denote a $p \times p$ matrix $\hat{C}$ as:

$$\hat{C} = \begin{bmatrix} 1 & -\hat{\gamma}_{1,2} & \dots & -\hat{\gamma}_{1,p} \\ -\hat{\gamma}_{2,1} & 1 & \dots & -\hat{\gamma}_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ -\hat{\gamma}_{p,1} & -\hat{\gamma}_{p,2} & \dots & 1 \end{bmatrix}$$

Inverting the KKT Conditions

We then define for all $j = 1, \dots, p$,

$$\hat{\tau}_j^2 := \| X_j - \boldsymbol{X}_{-j}\hat{\gamma}_j \|_2^2 / n + \lambda_j \| \hat{\gamma}_j \|_1 .$$

We further define $\hat{\boldsymbol{T}}^2 := \mathit{diag}(\hat{\tau}_1^2, \dots, \hat{\tau}_p^2)$.

Inverting the KKT Conditions

We then define for all $j = 1, \dots, p$,

$$\hat{\tau}_j^2 := \| X_j - \boldsymbol{X}_{-j}\hat{\gamma}_j \|_2^2 / n + \lambda_j \| \hat{\gamma}_j \|_1 .$$

We further define $\hat{\boldsymbol{T}}^2 := diag(\hat{\tau}_1^2, \dots, \hat{\tau}_p^2)$.

Now we are finally ready to define an approximation of $\hat{\boldsymbol{\Sigma}}^{-1}$:

$$\hat{\boldsymbol{\Theta}}_{Lasso} := \hat{\boldsymbol{T}}^{-2}\hat{\boldsymbol{C}}.$$

We want to use the estimate to construct asymptotic pivots of the Lasso.

Debiasing the Lasso

By previous arguments:

$$\hat{\boldsymbol{\Sigma}}(\hat{\beta} - \beta^0) + \lambda\hat{\kappa} = \boldsymbol{X}^T\epsilon/n.$$

Simple calculation yields

$$\hat{\beta} - \beta^0 + \hat{\boldsymbol{\Theta}}_{Lasso}\lambda\hat{\kappa} = \hat{\boldsymbol{\Theta}}_{Lasso}\boldsymbol{X}^T\epsilon/n - \boldsymbol{\Delta}/\sqrt{n}.$$

where

$$\boldsymbol{\Delta} := \sqrt{n}(\hat{\boldsymbol{\Theta}}_{Lasso}\hat{\boldsymbol{\Sigma}} - \boldsymbol{I})(\hat{\beta} - \beta^0).$$

Define estimator

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

Assumptions and Notations

We assume for now that the rows of $\boldsymbol{X}$ are i.i.d. realizations from a Gaussian distribution whose p-dimensional inner product matrix $\boldsymbol{\Sigma}$ has strictly positive smallest eigenvalue Λ_{min}^2 satisfying $1/\Lambda_{min}^2 = \mathcal{O}(1)$. Furthermore, $max_j \boldsymbol{\Sigma}_{j,j} = \mathcal{O}(1)$.

Assumptions and Notations

We assume for now that the rows of $\boldsymbol{X}$ are i.i.d. realizations from a Gaussian distribution whose p-dimensional inner product matrix $\boldsymbol{\Sigma}$ has strictly positive smallest eigenvalue Λ_{min}^2 satisfying $1/\Lambda_{min}^2 = \mathcal{O}(1)$. Furthermore, $max_j \boldsymbol{\Sigma}_{j,j} = \mathcal{O}(1)$.

We will use the following notation to further assume sparsity w.r.t. rows of $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$:

Define for all $j = 1, \dots, p$

$$s_j := \mid \{k \neq j : \boldsymbol{\Theta}_{j,k} \neq 0\} \mid .$$

Define the active set of variables $S_0 := \{j; \beta_j^0 \neq 0\}$, and its cardinality $s_0 = \mid S_0 \mid$.

We finally denote $\hat{\boldsymbol{\Omega}} := \hat{\boldsymbol{\Theta}}_{Lasso} \hat{\boldsymbol{\Sigma}} \hat{\boldsymbol{\Theta}}_{Lasso}^T$.

Main Results and Corollaries

Theorem (Van de Geer et al.(2014))

Consider linear model as above with Gaussian error $\epsilon \sim \boldsymbol{N}(0, \sigma_\epsilon^2 \boldsymbol{I})$ *where* $\sigma_\epsilon^2 = \mathcal{O}(1)$. *Assume the previous assumption holds, and* $s_0 = o(\sqrt{n}/log(p))$ *and* $max_j s_j = o(n/log(p))$.

Main Results and Corollaries

Theorem (Van de Geer et al.(2014))

Consider linear model as above with Gaussian error $\epsilon \sim \boldsymbol{N}(0, \sigma_\epsilon^2 \boldsymbol{I})$ *where* $\sigma_\epsilon^2 = \mathcal{O}(1)$. *Assume the previous assumption holds, and* $s_0 = o(\sqrt{n}/log(p))$ *and* $max_j s_j = o(n/log(p))$.
If $\lambda \asymp \sqrt{log(p)/n}$ *for the lasso, and* $\lambda_j \asymp \sqrt{log(p)/n}$ *uniformly for the nodewise regression, then*

Main Results and Corollaries

Theorem (Van de Geer et al.(2014))

Consider linear model as above with Gaussian error $\epsilon \sim \boldsymbol{N}(0, \sigma_\epsilon^2 \boldsymbol{I})$ *where* $\sigma_\epsilon^2 = \mathcal{O}(1)$. *Assume the previous assumption holds, and* $s_0 = o(\sqrt{n}/log(p))$ *and* $max_j s_j = o(n/log(p))$.
If $\lambda \asymp \sqrt{log(p)/n}$ *for the lasso, and* $\lambda_j \asymp \sqrt{log(p)/n}$ *uniformly for the nodewise regression, then*

$$\sqrt{n}(\hat{b}_{lasso} - \beta^0) = \boldsymbol{W} + \boldsymbol{\Delta},$$

$$\boldsymbol{W} \mid \boldsymbol{X} \sim \boldsymbol{N}(0, \sigma_\epsilon^2 \hat{\boldsymbol{\Omega}}),$$

$$\| \boldsymbol{\Delta} \|_\infty = o_{\mathbb{P}}(1),$$

$$\| \hat{\boldsymbol{\Omega}} - \boldsymbol{\Sigma}^{-1} \|_\infty = o_{\mathbb{P}}(1).$$

Main Results and Corollaries

Corollary (1)

For any fixed subset $G \subset \{1, \dots, p\}$, *condition on* $\boldsymbol{X}$, *the asymptotic distribution of* $\max_{j \in G}(n \mid \hat{b}_{lasso;j} \mid^2 / \sigma_\epsilon^2 \hat{\boldsymbol{\Omega}}_{j,j})$, *under the null-hypothesis* $\boldsymbol{H}_{0,G} : \beta_j^0 = 0, \forall j \in G$, *is asymptotically equal to the maximum of dependent* $\chi^2(1)$ *variables* $\max_{j \in G} \mid \boldsymbol{W}_j \mid^2 / \sigma_\epsilon^2 \hat{\boldsymbol{\Omega}}_{j,j}$.

Main Results and Corollaries

Corollary (1)

For any fixed subset $G \subset \{1, \dots, p\}$, *condition on* $\boldsymbol{X}$, *the asymptotic distribution of* $\max_{j \in G}(n \mid \hat{b}_{lasso;j} \mid^2 / \sigma_\epsilon^2 \hat{\boldsymbol{\Omega}}_{j,j})$, *under the null-hypothesis* $\boldsymbol{H}_{0,G} : \beta_j^0 = 0, \forall j \in G$, *is asymptotically equal to the maximum of dependent* $\chi^2(1)$ *variables* $\max_{j \in G} \mid \boldsymbol{W}_j \mid^2 / \sigma_\epsilon^2 \hat{\boldsymbol{\Omega}}_{j,j}$.

Corollary (2)

Under the same assumptions as the above theorem, the limiting variance $\sigma_\epsilon^2 \hat{\boldsymbol{\Omega}}_{j,j}$ *reaches the Cramér-Rao lower bound, and thus* $\hat{b}_{lasso;j}$ *is asymptotically efficient in the sense of semi-parametric inference.*

Other Story of Post-Selection Inference

In practice, a model is more often "found" by a data-driven selection process. Consequently, the inferential guarantees derived from classical theory are invalidated, since a data-driven variable selection process produces a model that is itself stochastic, as is the hypothesis to be tested.

Other Story of Post-Selection Inference

In practice, a model is more often "found" by a data-driven selection process. Consequently, the inferential guarantees derived from classical theory are invalidated, since a data-driven variable selection process produces a model that is itself stochastic, as is the hypothesis to be tested.

Naturally, one would expect possible solutions of post-selection inference under a formal algorithm such as the Lasso or LAR. However in reality, the challenge is to devise inferences which are valid for all kinds of variable selections, be it formal, informal, post hoc, etc.

General Philosophy

We interpret the submodel coefficients as follows:
(1) The full model has no special status other than being the repository of available predictors.
(2) The coefficients of excluded predictors are not zero; they are not defined and therefore do not exist.
(3) The meaning of a predictor's coefficient depends on which other predictors are included in the selected model.
(4) No data generating or causal claims are implied in any submodels and its parameters.

Assumptions

We consider fixed design of full predictor matrix

$$\boldsymbol{X} = (X_1, X_2, \dots, X_p) \in \mathbb{R}^{n \times p},$$

and let n and p be arbitrary. We denote

$$d := \mathit{rank}(\boldsymbol{X}) = \mathit{dim}(\mathit{span}(\boldsymbol{X})).$$

The response vector $Y \in \mathbb{R}^n$ satisfies

$$Y \sim \boldsymbol{N}(\mu, \sigma^2 \boldsymbol{I}),$$

where $\mu = \mathbb{E}(Y)$, which does not necessarily reside in $\mathit{span}(\boldsymbol{X})$.

Notations and Other Assumptions

Denote the index set $M = \{j_1, j_2, \dots, j_m\} \subset M_F = \{1, \dots, p\}$, and let $\boldsymbol{X}_M = \{X_{j_1}, \dots, X_{j_m})$ denote the $n \times m$ submatrix of $\boldsymbol{X}$ indexed by M. We assume that $\boldsymbol{X}_M$ is of full rank.

Notations and Other Assumptions

Denote the index set $M = \{j_1, j_2, \dots, j_m\} \subset M_F = \{1, \dots, p\}$, and let $\boldsymbol{X}_M = \{X_{j_1}, \dots, X_{j_m})$ denote the $n \times m$ submatrix of $\boldsymbol{X}$ indexed by M.
We assume that $\boldsymbol{X}_M$ is of full rank.
Let $\hat{\beta}_M = (\boldsymbol{X}_M^T \boldsymbol{X}_M)^{-1} \boldsymbol{X}_M^T Y$ be the unique least square estimate in M, which is an estimator of

$$\beta_M := \mathbb{E}(\hat{\beta}_M) = \arg\min_{\beta' \in R^m} \| \mu - \boldsymbol{X}_M \beta' \|^2 .$$

Respectively, we denote multiple regression coefficients for all $j \in M$, $\hat{\beta}_{j \cdot M} = X_{j \cdot M}^T Y / \| X_{j \cdot M} \|^2$.

Notations and Other Assumptions

Denote the index set $M = \{j_1, j_2, \dots, j_m\} \subset M_F = \{1, \dots, p\}$, and let $\boldsymbol{X}_M = \{X_{j_1}, \dots, X_{j_m})$ denote the $n \times m$ submatrix of $\boldsymbol{X}$ indexed by M.
We assume that $\boldsymbol{X}_M$ is of full rank.
Let $\hat{\beta}_M = (\boldsymbol{X}_M^T \boldsymbol{X}_M)^{-1} \boldsymbol{X}_M^T Y$ be the unique least square estimate in M, which is an estimator of

$$\beta_M := \mathbb{E}(\hat{\beta}_M) = \arg\min_{\beta' \in R^m} \| \mu - \boldsymbol{X}_M \beta' \|^2 .$$

Respectively, we denote multiple regression coefficients for all $j \in M$, $\hat{\beta}_{j \cdot M} = X_{j \cdot M}^T Y / \| X_{j \cdot M} \|^2$.
We further assume the availability of a valid estimate $\hat{\sigma}^2$ of σ^2 which is independent of all estimates $\hat{\beta}_{j \cdot M}$, and $\hat{\sigma}^2 \sim \sigma^2 \chi_r^2 / r$ for r degrees of freedom.

Further Notations

We denote the t-ratio for $\beta_{j\cdot M}$ that uses $\hat{\sigma}^2$ irrespective of M:

$$t_{j\cdot M} := \frac{\hat{\beta}_{j\cdot M} - \beta_{j\cdot M}}{((\boldsymbol{X}_M^T \boldsymbol{X}_M)^{-1})_{jj}^{1/2} \hat{\sigma}} = \frac{\hat{\beta}_{j\cdot M} - \beta_{j\cdot M}}{\hat{\sigma} \parallel X_{j\cdot M} \parallel} = \frac{(Y-\mu)^T X_{j\cdot M}}{\hat{\sigma} \parallel X_{j\cdot M} \parallel},$$

which has a central t-distribution with r degrees of freedom.

Further Notations

We denote the t-ratio for $\beta_{j\cdot M}$ that uses $\hat{\sigma}^2$ irrespective of M:

$$t_{j\cdot M} := \frac{\hat{\beta}_{j\cdot M} - \beta_{j\cdot M}}{((\boldsymbol{X}_M^T \boldsymbol{X}_M)^{-1})_{jj}^{1/2}\hat{\sigma}} = \frac{\hat{\beta}_{j\cdot M} - \beta_{j\cdot M}}{\hat{\sigma} \parallel X_{j\cdot M} \parallel} = \frac{(Y-\mu)^T X_{j\cdot M}}{\hat{\sigma} \parallel X_{j\cdot M} \parallel},$$

which has a central t-distribution with r degrees of freedom.

Remark: With such choice of $\hat{\sigma}^2$, the confidence intervals for $\beta_{j\cdot M}$ take the form

$$CI_{j\cdot M}(K) := [\hat{\beta}_{j\cdot M} \pm K((\boldsymbol{X}_M^T \boldsymbol{X}_M)^{-1}))_{jj}^{1/2}\hat{\sigma}] = [\hat{\beta}_{j\cdot M} \pm K\hat{\sigma}/ \parallel X_{j\cdot M} \parallel],$$

where $K = t_{r,1-\alpha/2}$ to be the $1-\alpha/2$ quantile of the t-distribution of r degrees of freedom, we have marginal $1-\alpha$ coverage guarantee

$$\mathbb{P}(\beta_{j\cdot M} \in CI_{j\cdot M}(K)) \geq 1-\alpha.$$

PoSI Selection Setup

As explained before, the index set M is the result of model selection that involves the stochastic component of data, namely Y. Thus we view a random variable selection procedure as

$$\hat{M} : \mathbb{R}^n \to \mathbb{M},$$

where $\mathbb{M} := \{M \mid M \subset \{1, \dots, p\}, rank(\boldsymbol{X}_M) = \mid M \mid\}$.

PoSI Selection Setup

As explained before, the index set M is the result of model selection that involves the stochastic component of data, namely Y. Thus we view a random variable selection procedure as

$$\hat{M} : \mathbb{R}^n \to \mathbb{M},$$

where $\mathbb{M} := \{M \mid M \subset \{1, \dots, p\}, rank(\boldsymbol{X}_M) = \mid M \mid\}$.
Remark: Now the selected model is random, and marginal guarantees for individual predictors requires some care, as $\beta_{j \cdot \hat{M}}$ might not exist for some $j \in \hat{M}$. Therefore marginal probabilities should be considered in a conditional sense, i.e.

$$\mathbb{P}(\beta_{j \cdot \hat{M}} \in CI_{j \cdot \hat{M}}(\cdot) \mid j \in \hat{M}).$$

PoSI Selection

Goal: Find a confidence interval with a constant K that provides universally valid post-selection inference for all model selection procedure $\hat{M}$:

$$\mathbb{P}(\beta_{j\cdot\hat{M}} \in CI_{j\cdot\hat{M}}(K), \forall j \in \hat{M}) \geq 1 - \alpha.$$

PoSI Selection

Goal: Find a confidence interval with a constant K that provides universally valid post-selection inference for all model selection procedure $\hat{M}$:

$$\mathbb{P}(\beta_{j\cdot\hat{M}} \in CI_{j\cdot\hat{M}}(K), \forall j \in \hat{M}) \geq 1 - \alpha.$$

Lemma

For any model selection procedure $\hat{M} : \mathbb{R}^n \to \mathbb{M}$, the following sharp inequality holds for all $Y \in \mathbb{R}^n$:

$$\max_{j \in \hat{M}(Y)} \mid t_{j\cdot\hat{M}(Y)}(Y) \mid \leq \max_{M \in \mathbb{M}} \max_{j \in M} \mid t_{j\cdot M}(Y) \mid .$$

Main Results and Corollaries

Theorem (Berk, et al.(2013))

Let K be the minimal value that satisfies

$$\mathbb{P}(\max_{M\in\mathbb{M}}\max_{j\in M} \mid t_{j\cdot M}(Y) \mid \leq K) \geq 1-\alpha,$$

Then for any model selection procedure $\hat{M}:\mathbb{R}^n \to \mathbb{M}$ we have

$$\mathbb{P}(\max_{j\in\hat{M}(Y)} \mid t_{j\cdot\hat{M}(Y)}(Y) \mid \leq K) \geq 1-\alpha,$$

where $K = K(\boldsymbol{X}, \mathbb{M}, \alpha, r)$ is referred to as "the PoSI constant".

Main Results and Corollaries

Corollary

Simultaneous post-selection confidence guarantees hold for any model selection procedure $\hat{M} : \mathbb{R}^n \to \mathbb{M}$,

$$\mathbb{P}(\beta_{j\cdot\hat{M}} \in CI_{j\cdot\hat{M}}(K), \forall j \in \hat{M}) \geq 1 - \alpha,$$

where K *is the PoSI constant.*

Main Results and Corollaries

Corollary

Simultaneous post-selection confidence guarantees hold for any model selection procedure $\hat{M} : \mathbb{R}^n \to \mathbb{M}$,

$$\mathbb{P}(\beta_{j\cdot\hat{M}} \in CI_{j\cdot\hat{M}}(K), \forall j \in \hat{M}) \geq 1 - \alpha,$$

where K *is the PoSI constant.*
Strong post-selection error control holds for any model selection procedure $\hat{M} : \mathbb{R}^n \to \mathbb{M}$,

$$\mathbb{P}(\exists j \in \hat{M} : \beta_{j\cdot\hat{M}} = 0, \mid t^{(0)}_{j\cdot\hat{M}} \mid > K) \leq \alpha,$$

where K *is again the PoSI constant, and* $t^{(0)}_{j\cdot M}$ *is the t-statistic for the null hypothesis* $\beta_{j\cdot M} = 0$.

Further Discussion

The above argument showed that simultaneity protection for all parameters $\beta_{j\cdot M}$ provides valid post-selection inference. In practice this means enlarging the constant $t_{1-\alpha/2,r}$ used in conventional inference to a constant $K(\boldsymbol{X}, \mathbb{M}, \alpha, r)$ that provides simultaneity protections.

Further Discussion

The above argument showed that simultaneity protection for all parameters $\beta_{j\cdot M}$ provides valid post-selection inference. In practice this means enlarging the constant $t_{1-\alpha/2,r}$ used in conventional inference to a constant $K(\boldsymbol{X}, \mathbb{M}, \alpha, r)$ that provides simultaneity protections.

It can be shown that the constant K depends strongly on the predictor matrix $\boldsymbol{X}$, and the asymptotic bound for K ranges between $\sqrt{2\log d}$ and $\sqrt{d}$. This wide asymptotic range suggests that computation is critical for problems with large numbers of predictors. However fast computational methods are still left to be discovered.

Questions?

Thanks For Listening!