STAT 689: Large-scale and high-dimensional statistical inference **Fall 2024**

Lecture 12

1 Lasso

In statistics and machine learning, Lasso (least absolute shrinkage and selection operator) is a regression analysis method that performs both variable selection and regularization to enhance the prediction accuracy and interpretability of the resulting statistical model. Here, we describe Lasso in linear regression. Consider i.i.d. samples (x_i, y_i), $i = 1, 2, \ldots, n$ from the linear model

$$y_i = x_i^\top \beta_0 + \epsilon_i,$$

where $\beta_0 \in \mathbb{R}^p$ is an unknown coefficient vector, and $\{\epsilon_i\}_{i=1}^n$ are random errors with mean zero. We can more succinctly express this data model as

$$Y = X\beta_0 + \epsilon,$$

where $Y = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$ is the vector of responses, X is the matrix of predictor variables, with ith row $x_i^\top$, and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^\top$ is the vector of errors.

1.1 Regularization

Regularization is the process of adding information in order to solve an ill-posed problem or to prevent overfitting. When $p \gg n$, least squares estimation is ill-posed and regularization is needed. Let's consider three canonical choices: the l_0, l_1, and l_2 norms:

$$\|\beta\|_0 = \sum_{j=1}^p \mathbf{1}\{\beta_j \neq 0\},$$
$$\|\beta\|_1 = \sum_{j=1}^p |\beta_j|,$$
$$\|\beta\|_2^2 = \sum_{j=1}^p \beta_j^2.$$

In constrained form, these norms give rise to the following problems:

$$\text{Best subset selection: } \min_{\beta \in \mathbb{R}^p} \|Y - X\beta\|^2 \text{ subject to } \|\beta\|_0 = \sum_{j=1}^p \mathbf{1}\{\beta_j \neq 0\} \leq t,$$

$$\text{Lasso: } \min_{\beta \in \mathbb{R}^p} \|Y - X\beta\|^2 \text{ subject to } \|\beta\|_1 = \sum_{j=1}^p |\beta_j| \leq t,$$

$$\text{Ridge regression: } \min_{\beta \in \mathbb{R}^p} \|Y - X\beta\|^2 \text{ subject to } \|\beta\|_2^2 = \sum_{j=1}^p \beta_j^2 \leq t.$$

In penalized form, Lasso is defined as

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{n}\|Y - X\beta\|^2 + \lambda\|\beta\|_1.$$

The best subset selection provides a sparse solution, but the corresponding optimization problem is nonconvex and difficult to solve. Ridge regression, while convex, does not perform variable selection effectively as the solution is non-sparse. On the other hand, Lasso addresses this by providing a sparse solution through the solution of a convex optimization problem.

1.2 Consistency of Lasso

We provide some consistent results about Lasso without giving proof. For more details, please refer to STAT 620 lecture notes.

Denote by s_0 the number of nonzero components in the true regression coefficient β_0.

Compatibility condition: Let $\widehat{\Sigma} = X^\top X/n \in \mathbb{R}^{p\times p}$. If for some $\phi_0 > 0$, and for all β satisfying $\|\beta_{S_0^c}\|_1 \le 3\|\beta_{S_0}\|_1$, it holds that

$$\|\beta_{S_0}\|_1^2 \le s_0(\beta^\top \widehat{\Sigma}\beta)/\phi_0^2.$$

Main result: Under the compatibility condition, we have

$$\|X(\hat{\beta}-\beta_0)\|^2/n + \lambda\|\hat{\beta}-\beta_0\|_1 \le 4\lambda^2 s_0/\phi_0^2.$$

As a result, we have

$$\|X(\hat{\beta}-\beta_0)\|^2/n \le 4\lambda^2 s_0/\phi_0^2,$$
$$\|\hat{\beta}-\beta_0\|_1 \le 4\lambda s_0/\phi_0^2.$$

A common choice of λ is

$$\lambda = C_0\sqrt{\frac{\log(p)}{n}}$$

for some constant C_0, which leads to

$$\|\hat{\beta}-\beta_0\|_1 \le Cs_0\sqrt{\frac{\log(p)}{n}}, \quad C = \frac{4C_0}{\phi_0^2}.$$

1.3 Asymptotic distribution of Lasso estimator

In the low-dimensional setup (i.e., $n \gg p$), the Lasso estimator has a nonstandard limiting distribution; see Fu and Knight (2000, AOS). Suppose $\sqrt{n}\lambda \to \lambda_0$ and $\hat{\Sigma} \to^p \Sigma_0$. Fu and Knight showed that

$$\sqrt{n}(\hat{\beta}-\beta_0) \xrightarrow{d} \text{argmin}_{u\in\mathbb{R}^p} V(u),$$

where

$$V(u) = -2u^\top W + u^\top\Sigma_0 u + \lambda_0\sum_{j=1}^p [u_j \text{sign}(\beta_j)\mathbf{1}\{\beta_j \neq 0\} + |u_j|\mathbf{1}\{\beta_j = 0\}]$$

with $W \sim N(0, \sigma^2\Sigma_0)$.

2 Debiased Lasso

In the high-dimensional setup (i.e., $n \approx p$ or $n \ll p$), the (asymptotic) distribution of the Lasso estimator $\hat{\beta}$ is not tractable due to its bias. One way to address this issue is through dibiasing, which aims to remove/reduce the bias in the Lasso estimator so that the resulting estimator has a tractable asymptotic distribution.

Suppose the design matrix X has i.i.d rows with mean zero and covariance matrix Σ. The construction of the debiased Lasso estimator involves a suitable approximation (say $\widehat{\Theta}$) for the inverse of Σ (Σ^{-1} is also called the precision matrix). We will briefly discuss the construction of $\widehat{\Theta}$ below. The debiased Lasso estimator is defined as

$$\widetilde{\beta} = \widehat{\beta} + \widehat{\Theta}X^\top(Y - X\widehat{\beta})/n.$$

An expansion for $\widetilde{\beta}$: Recall that $Y = X\beta_0 + \epsilon$. The debiased Lasso estimator can be decomposed as

$$\begin{aligned}\sqrt{n}(\widetilde{\beta} - \beta_0) =& \sqrt{n}(\widehat{\beta} - \beta_0) + \widehat{\Theta}X^\top(Y - X\widehat{\beta})/\sqrt{n} \\ =& \widehat{\Theta}X^\top\epsilon/\sqrt{n} + \sqrt{n}(\widehat{\beta} - \beta_0) + \widehat{\Theta}X^\top(X\beta_0 - X\widehat{\beta})/\sqrt{n} \\ =& \widehat{\Theta}X^\top\epsilon/\sqrt{n} + \sqrt{n}(\widehat{\beta} - \beta_0) + \sqrt{n}\widehat{\Theta}\widehat{\Sigma}(\beta_0 - \widehat{\beta}) \\ =& \widehat{\Theta}X^\top\epsilon/\sqrt{n} + \sqrt{n}(I - \widehat{\Theta}\widehat{\Sigma})(\widehat{\beta} - \beta_0),\end{aligned}$$

where $\widehat{\Theta}X^\top\epsilon/\sqrt{n}$ is the leading term and $\Delta := \sqrt{n}(I - \widehat{\Theta}\widehat{\Sigma})(\widehat{\beta} - \beta_0)$ is a remainder term that is of smaller order.

Leading term. If the errors are normally distributed, i.e., $\epsilon \sim N(0, \sigma^2 I)$, then conditional on X, we have

$$\widehat{\Theta}X^\top\epsilon/\sqrt{n} \sim N(0, \sigma^2\widehat{\Theta}\widehat{\Sigma}\widehat{\Theta}^\top).$$

Note that $\widehat{\Theta}$ only depends on X. When ϵ's are not normally distributed, one can still establish the asymptotic normally using the CLT under suitable assumptions.

Remainder term. Let $\|\cdot\|_p$ be the l_p norm of a vector. We note that

$$\begin{aligned}\|\Delta\|_\infty =& \sqrt{n}\|(\widehat{\Theta}\widehat{\Sigma} - I)(\widehat{\beta} - \beta_0)\|_\infty \\ \le& \sqrt{n}\|\widehat{\Theta}\widehat{\Sigma} - I\|_\infty\|\widehat{\beta} - \beta_0\|_1,\end{aligned}$$

where $\|A\|_\infty = \max_{i,j}|a_{ij}|$. For a carefully constructed estimator $\widehat{\Theta}$, one can show that

$$\|\widehat{\Theta}\widehat{\Sigma} - I\|_\infty \le C'\sqrt{\frac{\log(p)}{n}}$$

for some constant C'. For example, the above rate is satisfied for the precision matrix estimators as described below. In this case, we have

$$\begin{aligned}\|\Delta\|_\infty \le& \sqrt{n}\|\widehat{\Theta}\widehat{\Sigma} - I\|_\infty\|\widehat{\beta} - \beta_0\|_1 \\ =& O_p\left(\frac{s_0\log(p)}{\sqrt{n}}\right) = o_p(1)\end{aligned}$$

if $s_0\log(p) \ll \sqrt{n}$. Note that this is a stronger requirement than achieving the consistency for the Lasso estimator, which requires $s_0^2\log(p) \ll n$.

Combing the results, we have

$$\sqrt{n}(\widetilde{\beta}_j - \beta_{0,j}) \overset{d}{\approx} N(0, \sigma^2(\widehat{\Theta}\widehat{\Sigma}\widehat{\Theta}^\top)_{j,j}).$$

Therefore, a $1-\alpha$ confidence interval for $\beta_{0,j}$ can be constructed as

$$\left[\widetilde{\beta}_j - z_{1-\alpha/2}\widehat{\sigma}\sqrt{(\widehat{\Theta}\widehat{\Sigma}\widehat{\Theta}^\top)_{j,j}/n}, \widetilde{\beta}_j + z_{1-\alpha/2}\sqrt{(\widehat{\Theta}\widehat{\Sigma}\widehat{\Theta}^\top)_{j,j}/n}\right]$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution and $\widehat{\sigma}^2$ is some estimate of the noise level σ^2. The so-called scaled Lasso is one way to estimate σ^2; see Sun and Zhang (2012, Biometrika).

3 Precision matrix estimation

3.1 Gaussian graphical models

Suppose $Z = (Z_1, \ldots, Z_p)^\top \sim N(0, \Sigma)$ and $\Theta = \Sigma^{-1}$ is the precision matrix. The conditional distribution of Z_j given Z_{-j} is equal to $N(\mu_j(Z_{-j}), \tau_j^2)$

$$\begin{aligned}\mu_j(Z_{-j}) &= \Sigma_{j,-j}\Sigma_{-j,-j}Z_{-j}, \\ \tau_j^2 &= \Sigma_{jj} - \Sigma_{j,-j}\Sigma_{-j,-j}^{-1}\Sigma_{-j,j}.\end{aligned}$$

Therefore, we can write

$$Z_j = Z_{-j}^\top \beta_j^* + \epsilon_j$$

where $\beta_j^* = \Sigma_{-j,-j}^{-1}\Sigma_{-j,j}$ and $\epsilon_j \sim N(0, \tau_j^2)$.

On the other hand, by the inverse formula for block matrices, we have

$$\begin{aligned}\theta_{jj} =& (\Sigma_{jj} - \Sigma_{j,-j}\Sigma_{-j,-j}\Sigma_{-j,j})^{-1} \\ =& (\Sigma_{jj} - 2\Sigma_{j,-j}\beta_j^* + \beta_j^{*\top}\Sigma_{-j,-j}^{-1}\beta_j^*)^{-1}\end{aligned}$$

and

$$\theta_{-j,j} = -\theta_{jj}\beta_j^*.$$

It thus implies that

$$\theta_{jj} = \frac{1}{\tau_j^2}, \quad \theta_{-j,j} = -\frac{\beta_j^*}{\tau_j^2}.$$

These results inspire the nodewise Lasso method introduced below.

3.2 Nodewise Lasso

Nodewise Lasso was originally proposed by Meinshausen and Bühlmann (2006, AOS) as a way to estimate the precision matrix when $n \ll p$.

Let X_{-j} be the design matrix without the jth column X_j. For $j = 1, 2, \ldots, p$, consider

$$\widehat{\gamma}_j = \{\widehat{\gamma}_{j,k} : 1 \le k \le p : k \ne j\} = \text{argmin}_{\gamma \in \mathbb{R}^{p-1}} \left(\|X_j - X_{-j}\gamma\|_2^2/n + 2\lambda_j\|\gamma\|_1 \right).$$

Let $\widehat{C} = (\widehat{c}_{i,j})_{i,j=1}^p$ with $\widehat{c}_{i,i} = 1$ and $\widehat{c}_{i,j} = -\widehat{\gamma}_{i,j}$ for $i \ne j$. Further, let

$$\widehat{\tau}_j^2 = \|X_j - X_{-j}\widehat{\gamma}_j\|_2^2/n + 2\lambda_j\|\widehat{\gamma}_j\|_1.$$

Write $\widehat{T}^2 = \text{diag}(\widehat{\tau}_1^2, \ldots, \widehat{\tau}_p^2)$. Finally, the nodewise Lasso estimator for Θ is constructed as

$$\widehat{\Theta} = \widehat{T}^{-2}\widehat{C}.$$

It can be shown that

$$\|\widehat{\Theta}\widehat{\Sigma} - I\|_\infty = O_p\left(\sqrt{\frac{\log(p)}{n}}\right).$$

3.3 Optimal projection

In this section, we provide an interpretation of debiased Lasso based on optimal projection. Suppose we want to test the significance of the jth covariate, i.e., to test if $\beta_{0,j} = 0$ where $\beta_{0,j}$ is the jth component of the true regression coefficient β_0. We can rewrite the model $Y = X\beta_0 + \epsilon$ as

$$\eta_j := Y - X_{-j}\beta_{0,-j} = X_j\beta_{0,j} + \epsilon.$$

If the value of η_j is known, the problem will reduce to the inference about $\beta_{0,j}$ in a simple linear regression model. As η_j is not directly observable, the natural idea is to replace η_j with a suitable estimator defined as

$$\widehat{\eta}_j := Y - X_{-j}\widehat{\beta}_{-j} = X_j\beta_{0,j} + \epsilon + X_{-j}(\beta_{0,-j} - \widehat{\beta}_{-j}),$$

Measure	Method	Toeplitz		Equi corr	
		$U([0,2])$	$U([0,4])$	$U([0,2])$	$U([0,4])$
		Active set $S_0 = \{1,2,3\}$			
Avgcov S_0	Lasso-Pro	0.86	0.84	0.90	0.89
	Res-Boot	0.66	0.85	0.45	0.57
Avglength S_0	Lasso-Pro	0.786	0.787	0.762	0.760
	Res-Boot	0.698	0.918	0.498	0.670
Avgcov S_0^c	Lasso-Pro	0.95	0.95	0.95	0.95
	Res-Boot	1.00	1.00	1.00	1.00
Avglength S_0^c	Lasso-Pro	0.786	0.787	0.811	0.808
	Res-Boot	0.000	0.000	0.006	0.007

Figure 1: Lasso-Pro: Debiased Lasso; Res-Boot: Residual bootstrap. The nominal level is 95%.

where $\widehat{\beta}$ is the Lasso estimator for β_0. Here, the extra term $X_{-j}(\beta_{0,-j} - \widehat{\beta}_{-j})$ quantifies the estimation effect. Now given a projection vector $v = (v_1, \ldots, v_n)^\top \in \mathbb{R}^n$ such that $v^\top X_j = n$, we define the projection-based estimator for $\beta_{0,j}$ as

$$\widetilde{\beta}_j = \frac{v^\top \widehat{\eta}_j}{n} = \beta_{0,j} + \frac{v^\top \epsilon}{n} + \frac{v^\top X_{-j}(\beta_{0,-j} - \widehat{\beta}_{-j})}{n},$$

which implies that

$$\sqrt{n}(\widetilde{\beta}_j - \beta_{0,j}) = \frac{v^\top \epsilon}{\sqrt{n}} + \frac{v^\top X_{-j}(\beta_{0,-j} - \widehat{\beta}_{-j})}{\sqrt{n}}.$$

Suppose $\epsilon \sim N(0, \sigma^2 I_p)$ and v is determined by the design matrix X. Then, conditional on the design

$$\frac{v^\top \epsilon}{\sqrt{n}} \sim N(0, \sigma^2 \|v\|^2/n).$$

On the other hand,

$$\frac{1}{\sqrt{n}} |v^\top X_{-j}(\beta_{0,-j} - \widehat{\beta}_{-j})| \le \sqrt{n} \|\beta_{0,-j} - \widehat{\beta}_{-j}\|_1 \|v^\top X_{-j}/n\|_\infty.$$

Recall that for the Lasso estimator,

$$\sqrt{n}\|\beta_{0,-j} - \widehat{\beta}_{-j}\|_1 = O_p(s_0 \sqrt{\log(p)}).$$

Suppose

(i) $\|v\|^2/n = O_p(1)$;

(ii) $s_0 \sqrt{\log(p)} \|v^\top X_{-j}/n\|_\infty = o_p(1)$.

Then, we have

$$\sqrt{n}(\widetilde{\beta}_j - \beta_{0,j}) \approx^d N(0, \sigma^2 \|v\|^2/n).$$

Essentially, we want to find a direction v which is "almost orthogonal" to X_k for all $k \neq j$ while satisfying that $v^\top X_j = n$. One way to achieve this goal is by finding v, which minimizes

$$\max_{k \neq j} |v^\top X_k| + \lambda \|v\|^2$$

subject to the constraint that $v^\top X_j = n$. See Yi and Zhang (2022) for a development based on this idea.

3.4 Javanmard and Montanari's approach

This approach (JM, hereafter) was proposed by Javanmard and Montanari (2014, JMLR). As seen before, the validity of the debiased Lasso estimator requires the control of the remainder term Δ, while its efficiency relies on the magnitude of $(\widehat{\Theta}\widehat{\Sigma}\widehat{\Theta}^\top)_{j,j}$. The JM method utilizes this insight and constructs $\widehat{\Theta}$ row by row. Specifically, the ith row of $\widehat{\Theta}$ is the solution of the convex optimization problem:

$$\min_{\theta} \theta^\top \widehat{\Sigma}\theta \quad \text{subject} \quad \|\widehat{\Sigma}\theta - e_j\|_\infty \le \eta,$$

where e_j is the vector with one at the jth position and zero everywhere else. Denote by $\widehat{\theta}_j$ the solution to the above problem. Then $\widehat{\Theta} = [\widehat{\theta}_1, \ldots, \widehat{\theta}_p]^\top$. When

$$\eta = C' \sqrt{\frac{\log(p)}{n}},$$

a feasible solution automatically satisfies that

$$\|\widehat{\Theta}\widehat{\Sigma} - I\|_\infty \le C' \sqrt{\frac{\log(p)}{n}}.$$

On the other hand, the minimization encourages the resulting debiased Lasso estimator to have a small asymptotic variance and thus a narrower confidence interval for $\beta_{0,j}$.