Proof of prediction error bounds and sparsistency

Kaveh S. Nobari

Lectures in High-Dimensional Statistics

Department of Mathematics and Statistics
Lancaster University

Contents

Motivation

- At the end of the previous sessions, we had diverted our attention from the problem of parameter recovery (i.e. recovering the actual value of the regression vector θ^*) to finding a good predictor $\hat{\theta} \in \mathbb{R}^d$, such that the mean-squared prediction error

$$\frac{\|X(\hat{\theta} - \theta^*)\|_2^2}{n} = \frac{1}{n} \sum_{i=1}^{n} \Big(\langle x_i, \hat{\theta} - \theta^* \rangle \Big)^2$$

  is minimized.
- In this session, we provide proofs for the prediction error bounds outlined in the previous session.

Motivation

- Finally, we focus our attention on sparsistency or variable selection. As the Lasso behaves like a soft-thresholding operator, the solutions are sparse, which motivates the investigator to assess whether Lasso recovers the right support - i.e., estimate

$$S := \text{supp}\,(\theta^*)$$

exactly.

Lasso oracle inequality

Theorem (Prediction error bounds)

Once again consider the Lagrangian Lasso with a strictly positive $\lambda_n \geq 2\|X'w/n\|_\infty$.

(a) *Any optimal solution* $\hat{\theta}$ *satisfies*

$$\frac{\|X(\hat{\theta} - \theta^*)\|_2^2}{n} \leq 12\|\theta^*\|_1\lambda_n$$

(b) *If* θ^* *is supported on subset* S*, such that* $|S| = s$*, and the design matrix satisfies the* $(\kappa, 3)$*-RE condition over* S*, then any optimal solution satisfies the bound*

$$\frac{\|X(\hat{\theta} - \theta^*)\|_2^2}{n} \leq \frac{9}{\kappa}s\lambda_n^2$$

Proof

Proof of (a):

Let $\Delta = \hat{\theta} - \theta^*$. Recall from the Lagrangian basic inequality, we have

$$\frac{1}{2n}\|y - X\hat{\theta}\|_2^2 + \lambda_n\|\hat{\theta}\|_1 \leq \frac{1}{2n}\|y - X\theta^*\|_2^2 + \lambda_n\|\theta^*\|_1$$

$$0 \leq \frac{1}{2n}\|X\Delta\|_2^2 \leq \frac{\varepsilon' X\Delta}{n} + \lambda_n\{\|\theta^*\|_1 - \|\hat{\theta}\|_1\}.$$

Using Holder's inequality, we know that

$$\left|\frac{\varepsilon' X\Delta}{n}\right| \leq \left\|\frac{X'\varepsilon}{n}\right\|_\infty \|\Delta\|_1$$

Noting the choice of the regularization parameter - i.e., $\lambda_n \geq 2\|X'\varepsilon/n\|_\infty$, it can be claimed that

$$\left|\frac{\varepsilon' X\Delta}{n}\right| \leq \left\|\frac{X'\varepsilon}{n}\right\|_\infty \|\Delta\|_1 \leq \frac{\lambda_n}{2}\|\hat{\theta} - \theta^*\|_1 \leq \frac{\lambda_n}{2}\{\|\theta^*\|_1 + \|\hat{\theta}\|_1\}$$

Proof

Combining all the above information yields

$$\begin{aligned} 0 &\leq \frac{\lambda_n}{2}\{\|\theta^*\|_1 + \|\hat{\theta}\|_1\} + \lambda_n\{\|\theta^*\|_1 - \|\hat{\theta}\|_1\} \\ &\leq \lambda_n\{3\|\theta^*\|_1 - \|\hat{\theta}\|_1\} \end{aligned}$$

which implies $3\|\theta^*\|_1 \geq \|\hat{\theta}\|_1$. Using the triangle inequality on $\Delta = \hat{\theta} - \theta^*$, we obtain

$$\|\Delta\|_1 = \|\hat{\theta} - \theta^*\|_1 \leq \|\theta^*\|_1 + \|\hat{\theta}\|_1 \leq \|\theta^*\|_1 + 3\|\theta^*\|_1 = 4\|\theta^*\|_1$$

Noting that $\|\hat{\theta}\|_1 = \|\theta^* + \Delta\|_1$ and using the lower bound of the triangle inequality, we have

$$\|\theta^*\|_1 - \|\Delta\|_1 \leq \|\theta^* + \Delta\|_1$$

Returning to Lagrangian basic inequality, recall

Proof

$$\begin{aligned}\frac{\|X\Delta\|_2^2}{2n} &\leq \frac{\lambda_n}{2}\|\Delta\|_1 + \lambda_n\{\|\theta^*\|_1 - \|\theta^* + \Delta\|_1\} \\ &\leq \frac{\lambda_n}{2}\|\Delta\|_1 + \lambda_n\{\|\theta^*\|_1 - \|\theta^*\|_1 + \|\Delta\|_1\} \\ &\leq \frac{3\lambda_n}{2}\|\Delta\|_1.\end{aligned}$$

furthermore, we have established that $\|\Delta\|_1 \leq 4\|\theta^*\|_1$, which in turn implies

$$\frac{\|X\Delta\|_2^2}{2n} \leq \frac{12\lambda_n}{2}\|\Delta\|_1$$

or

$$\frac{\|X(\hat{\theta} - \theta^*)\|_2^2}{n} \leq 12\lambda_n\|\Delta\|_1$$

as (a) suggests.

Proof

Proof of (b):

Once again using Lagrangian basic inequality, we have

$$0 \leq \frac{1}{2n}\|X\Delta\|_2^2 \leq \frac{\varepsilon' X\Delta}{n} + \lambda_n\{\|\theta^*\|_1 - \|\hat{\theta}\|_1\}.$$

We know θ^* is S-sparse, so we may write

$$\begin{aligned}\|\theta^*\|_1 - \|\hat{\theta}\|_1 &= \|\theta^*_S + \theta^*_{S^c}\|_1 - \|\theta^*_S + \theta^*_{S^c} + \Delta_S\|_1 - \|\Delta_{S^c}\|_1 \\ &= \|\theta^*_S\|_1 - \|\theta^*_S + \Delta_S\|_1 - \|\Delta_{S^c}\|_1\end{aligned}$$

where substituting the above into the Lagrangian basic inequality yields,

$$0 \leq \frac{1}{2n}\|X\Delta\|_2^2 \leq \frac{\varepsilon' X\Delta}{n} + \lambda_n\{\|\theta^*_S\|_1 - \|\theta^*_S + \Delta_S\|_1 - \|\Delta_{S^c}\|_1\}$$

$$0 \leq \frac{1}{n}\|X\Delta\|_2^2 \leq 2\frac{\varepsilon' X\Delta}{n} + 2\lambda_n\{\|\theta^*_S\|_1 - \|\theta^*_S + \Delta_S\|_1 - \|\Delta_{S^c}\|_1\}$$

Proof

Once again using the Holder's inequality, we have

$$0 \leq \frac{1}{n}\|X\Delta\|_2^2 \leq 2\left\|\frac{X'\varepsilon}{n}\right\|_\infty \|\Delta\|_1 + 2\lambda_n\{\|\theta_S^*\|_1 - \|\theta_S^* + \Delta_S\|_1 - \|\Delta_{S^c}\|_1\}$$

Once again, substituting the choice of the regularization parameter yields,

$$0 \leq \frac{1}{n}\|X\Delta\|_2^2 \leq \lambda_n\|\Delta\|_1 + 2\lambda_n\{\|\theta_S^*\|_1 - \|\theta_S^* + \Delta_S\|_1 - \|\Delta_{S^c}\|_1\}$$

From the lower bound of the triangle inequality, we know that

$$\|\theta_S^*\|_1 - \|\Delta_S\|_1 \leq \|\theta_S^* + \Delta_S\|_1$$

which leads to

$$\begin{aligned} 0 \leq \frac{1}{n}\|X\Delta\|_2^2 &\leq \lambda_n\|\Delta\|_1 + 2\lambda_n\{\|\Delta_S\|_1 - \|\Delta_{S^c}\|_1\} \\ &\leq \lambda_n\{3\|\Delta_S\|_1 - \|\Delta_{S^c}\|_1\} \end{aligned}$$

Proof

For the above inequality to hold, $3\|\Delta_S\|_1 \geq \|\Delta_{S^c}\|_1$. Thus, using Cauchy-Schwarz inequality, we have

$$3\|\Delta_S\|_1 - \|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1 \leq 3\sqrt{s}\|\Delta\|_2$$

Hence,

$$\frac{\|X\Delta\|_2^2}{n} \leq 3\lambda_n\sqrt{s}\|\Delta\|_2.$$

Since, $\Delta \in \mathbb{C}_3(S)$, whence the $(\kappa, 3)$-RE condition can be applied.

$$\frac{\|X\Delta\|_2^2}{n} \leq 3\lambda_n\sqrt{s}\|\Delta\|_2.$$

we have

$$\|\Delta\|_2^2 \leq \frac{1}{\kappa}\frac{\|X\Delta\|_2^2}{n}.$$

Henceforth

$$\frac{\|X\Delta\|_2^2}{\sqrt{n}} \leq \frac{3}{\sqrt{\kappa}}\sqrt{s}\lambda_n.$$

Sparsistency

Following Wainwright (2019), let us begin by a deterministic design matrix X, and denote X_S as a sub-matrix of X that is composed of the columns of X belonging to an index set S. Variable selection requires some assumptions that are related, but distinct from the restricted eigenvalue condition:

- Lower eigenvalue:

$$\gamma_{\min}\left(\frac{X_S' X_S}{n}\right) \geq c_{\min} > 0$$

- Mutual incoherence: $\exists \alpha \in [0, 1)$ such that

$$\max_{j \in S^c} \|(X_S' X_S)^{-1} X_S' X_j\|_1 \leq \alpha$$

The intuition for these assumptions is as follows:

Lower eigenvalue:

1. The lower eigenvalue assumption on the sample covariance of the indexed sub-matrix is to ensure that that the model is identifiable, even if the support S is known a priori.
2. If the lower eigenvalue condition is violated, the sub-matrix X_S would have a non-trivial nullspace, leading to a non-identifiable model.

Mutual incoherence:

1. The general idea is that the columns of subset matrix X_S and the sub-matrix X_{S^c} must possess low correlation. Suppose, we wish to predict the column vector X_j using a linear combination of columns of X_S, the best weight vector $\hat{\omega} \in \mathbb{R}^{|S|}$ is given by

$$\hat{\omega} = \arg \min_{\omega \in \mathbb{R}^{|S|}} \|X_j - X_S \omega\|_2^2 = (X_S' X_S)^{-1} X_S' X_j.$$

2. The mutual incoherence condition is a bound on $\|\omega\|_1$, with the ideal weight being zero, which suggests orthogonality between the columns of X_S and X_{S^c}.

With this setup and the assumptions outlined earlier, we provide the following Theorem that applies to the Lagrangian Lasso, when applied to an instance of the linear observational model, such that the true parameter θ^* is supported on a subset S with cardinality s.

Before we derive the results, we introduce the orthogonal projection matrix

$$\Pi_{S^\perp}(X) = I_n - X_S(X_S'X_S)^{-1}X_S'$$

Theorem (Part I)

Consider an S-sparse linear regression model, with a design matrix that satisfies the lower eigenvalue and mutual incoherence assumptions. Then, for any regularization parameter

$$\lambda_n \geq \frac{2}{1-\alpha} \left\| X'_{S^c} \Pi_{S^\perp}(X) \frac{\varepsilon}{n} \right\|_\infty$$

the Lagrangian Lasso has these properties:

(a) *Uniqueness: There exists a unique optimal solution $\hat{\theta}$.*

(b) *No erroneous inclusion: The optimal solution has its support set $\hat{S}$ contained within the support set S, or in other words*

$$supp(\hat{S}) \subseteq supp(S)$$

Theorem (Part II)

(c) ℓ_∞ *bounds:* *The error* $\hat{\theta} - \theta^*$ *satisfies*

$$\|\hat{\theta}_S - \theta^*_S\|_\infty \leq \underbrace{\left\| \left(\frac{X_S' X_S}{n} \right)^{-1} X_S' \frac{\varepsilon}{n} \right\|_\infty + \left\| \left(\frac{X_S' X_S}{n} \right)^{-1} \right\|_\infty \lambda_n}_{B(\lambda_n, X)},$$

where $|||A|||_\infty = \max_{i \in [S]} \sum |A_{ij}|$ *is the matrix* ℓ_∞*-norm.*

(d) *No erroneous exclusion:* *The Lasso includes all indices* $i \in S$*, such that* $|\theta_i^*| > B(\lambda_n; X)$*, and hence it is sparsistent if* $\min_{i \in S} |\theta_i^*| > B(\lambda_n; X)$.

The intuition for these results is as follows:

Uniqueness:

- This is not trivial, as since $d > n$, although the Lasso objective is convex, it cannot be strictly convex.
- Based on the uniquness claim, we may unambiguously talk about the support $\hat{\theta}$.

No erroneous inclusion (and exclusion):

- This claim guarantees that the Lasso estimate $\hat{\theta}$ does not erroneously include variables that are not in the true support of θ^* - i.e, $\hat{\theta}_{S^c} = 0$.
- Similarly (d) is a consequence of the sup-norm bound (c), which suggests so long as the minimum value of $|\theta_i^*|$ over $i \in S$ is not too small, then the Lasso is sparsistent.

Corollary

Consider the S-sparse linear model, with noise vector ε with zero-mean i.i.d. that is sub-Gaussian with the sub-Gaussianity parameter σ. Furthermore, suppose that the deterministic design matrix X satisfies the lower eigenvalue and mutual incoherence assumptions, as well as the C-column normalisation condition ($\max_{j=1,\cdots,d} \|X_j\|_2/\sqrt{n} \leq C$). Suppose, we solve the Lagranian Lasso with regularizaion parameter

$$\lambda_n = \frac{2C\sigma}{1-\alpha}\left\{\sqrt{\frac{2\log(d-s)}{n}} + \delta\right\}$$

for $\delta > 0$. Then the optimal solution is unique, with $supp(\hat{\theta}) \subseteq supp(\theta^*)$, and satisfied the ℓ_∞ bound

$$P\left[\|\hat{\theta}_S - \theta^*_S\|_\infty \leq \frac{\sigma}{\sqrt{c_{\min}}}\left\{\sqrt{\frac{2\log s}{n}} + \delta\right\} + \left|\left|\left|\left(\frac{X_S' X_S}{n}\right)^{-1}\right|\right|\right|_\infty \lambda_n\right]$$
$$\geq 1 - 4\exp\left(-\frac{n\delta^2}{2}\right).$$

References

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.