Bounds on prediction error and variable selection

HDS 7.4-7.5

Yangjianchen Xu

Department of Biostatistics
University of North Carolina at Chapel Hill

03/19/2021

Overview

1. 7.4 Bounds on prediction error

2. 7.5 Variable or subset selection

Prediction error

In the previous analysis, we have focused exclusively on the problem of parameter recovery in noiseless and noisy settings. In other applications, we might be interested in finding a good predictor, meaning a vector $\hat{\theta} \in \mathbb{R}^d$ such that the mean-squared prediction error

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

is small. In general, the problem of finding a good predictor should be easier than estimating θ^* well in ℓ_2-norm because the prediction problem does not require that θ^* even be identifiable.

Prediction error bounds

Theorem 7.20 (Prediction error bounds)

Consider the Lagrangian Lasso with a strictly positive regularization parameter $\lambda_n \geq 2\|\frac{\mathrm{X}^{\mathrm{T}} w}{n}\|_\infty$

(a) Any optimal solution $\widehat{\theta}$ satisfies the bound

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

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

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

Remarks on 7.20(a)

Theorem 7.20 (Prediction error bounds)

Consider the Lagrangian Lasso with a strictly positive regularization parameter $\lambda_n \geq 2\|\frac{\mathrm{X}^{\mathrm{T}} w}{n}\|_\infty$

(a) Any optimal solution $\widehat{\theta}$ satisfies the bound

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

As previously discussed in Example 7.14, when the noise vector w has i.i.d. zero-mean σ-sub-Gaussian entries and the design matrix is C-column normalized, the choice $\lambda_n = 2C\sigma(\sqrt{\frac{2\log d}{n}} + \delta)$ is valid with probability at least $1 - 2e^{-\frac{n\delta^2}{2}}$. In this case, Theorem 7.20(a) implies the upper bound $\frac{\|\mathrm{X}(\widehat{\theta}-\theta^*)\|_2^2}{n} \leq 24 \|\theta^*\|_1 \, C\sigma \left(\sqrt{\frac{2\log d}{n}} + \delta \right)$ (slow rates) with the same high probability.

Remarks on 7.20(b)

Theorem 7.20 (Prediction error bounds)

Consider the Lagrangian Lasso with a strictly positive regularization parameter $\lambda_n \geq 2\|\frac{\mathrm{X}^{\mathrm{T}} w}{n}\|_\infty$

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

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

On the other hand, when θ^* is s-sparse and in addition, the design matrix satisfies an RE condition, then Theorem 7.20(b) guarantees the bound $\frac{\|\mathrm{X}(\widehat{\theta} - \theta^*)\|_2^2}{n} \leq \frac{72}{\kappa} C^2 \sigma^2 \left(\frac{2s \log d}{n} + s\delta^2\right)$ (fast rates) with the same high probability.

Proof of Theorem 7.20(a)

We first show that $\|\widehat{\Delta}\|_1 \leq 4\,\|\theta^*\|_1$ under the stated conditions, where $\hat{\Delta} = \hat{\theta} - \theta^*$. From the Lagrangian basic inequality $L(\hat{\theta}; \lambda_n) \leq L(\theta^*; \lambda_n)$, we have

$$0 \leq \frac{1}{2n}\|X\widehat{\Delta}\|_2^2 \leq \frac{w^{\mathrm{T}}X\widehat{\Delta}}{n} + \lambda_n \left\{ \|\theta^*\|_1 - \|\widehat{\theta}\|_1 \right\}$$

By Hölder's inequality and our choice of λ_n, we have

$$\left| \frac{w^{\mathrm{T}}X\widehat{\Delta}}{n} \right| \leq \left\| \frac{X^{\mathrm{T}}w}{n} \right\|_\infty \|\widehat{\Delta}\|_1 \leq \frac{\lambda_n}{2}\|\widehat{\Delta}\|_1 \leq \frac{\lambda_n}{2}\left\{ \|\theta^*\|_1 + \|\widehat{\theta}\|_1 \right\},$$

where the final step also uses the triangle inequality. Putting together the pieces yields

$$0 \leq \frac{w^{\mathrm{T}}X\widehat{\Delta}}{n} + \left| \frac{w^{\mathrm{T}}X\widehat{\Delta}}{n} \right| \leq \frac{\lambda_n}{2}\left\{ \|\theta^*\|_1 + \|\widehat{\theta}\|_1 \right\} + \lambda_n \left\{ \|\theta^*\|_1 - \|\widehat{\theta}\|_1 \right\}$$

which implies $\|\widehat{\theta}\|_1 \leq 3\|\theta^*\|_1 \Rightarrow \|\widehat{\Delta}\|_1 \leq \|\theta^*\|_1 + \|\widehat{\theta}\|_1 \leq 4\,\|\theta^*\|_1$

Proof of 7.20(a)

We can now complete the proof. Returning to our earlier inequality

$$\frac{1}{2n}\|\mathrm{X}\widehat{\Delta}\|_2^2 \leq \frac{w^{\mathrm{T}}\mathrm{X}\widehat{\Delta}}{n} + \lambda_n \left\{ \|\theta^*\|_1 - \|\widehat{\theta}\|_1 \right\}$$

$$\textbf{First term: } \frac{w^{\mathrm{T}}\mathrm{X}\widehat{\Delta}}{n} \leq \left| \frac{w^{\mathrm{T}}\mathrm{X}\widehat{\Delta}}{n} \right| \leq \left\| \frac{\mathrm{X}^{\mathrm{T}} w}{n} \right\|_\infty \|\widehat{\Delta}\|_1 \leq \frac{\lambda_n}{2}\|\widehat{\Delta}\|_1$$

$$\textbf{Second term: } \lambda_n \left\{ \|\theta^*\|_1 - \|\widehat{\theta}\|_1 \right\} = \lambda_n\{\|\theta^*\|_1 - \|\theta^* + \widehat{\Delta}\|_1\} \leq \lambda_n\|\hat{\Delta}\|_1$$

These two inequalities imply

$$\frac{\|\mathrm{X}\widehat{\Delta}\|_2^2}{n} \leq 3\lambda_n\|\widehat{\Delta}\|_1 \leq 12\lambda_n\|\hat{\theta}^*\|_1$$

Proof of 7.20(b)

In this case, the same argument as in the proof of Theorem 7.13(a) leads to the basic inequality

$$\frac{\|X\widehat{\Delta}\|_2^2}{n} \leq 3\lambda_n \|\widehat{\Delta}_S\|_1 \leq 3\lambda_n \sqrt{s} \|\widehat{\Delta}\|_2$$

Similarly, the proof of Theorem 7.13(a) shows that the error vector $\widehat{\Delta}$ belongs to $\mathbb{C}_3(S)$, whence the $(\kappa; 3) - RE$ condition can be applied, this time to the right-hand side of the basic inequality. Doing so yields

$$\|\widehat{\Delta}\|_2 \leq \sqrt{\frac{1}{\kappa} \frac{\|X\widehat{\Delta}\|_2^2}{n}} = \frac{1}{\sqrt{\kappa}} \frac{\|X\widehat{\Delta}\|_2}{\sqrt{n}}$$

and hence that

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

Variable or subset selection

Thus far, we have focused on results that guarantee that either the ℓ_2-error or the prediction error of the Lasso is small. In other settings, we are interested in a somewhat more refined question, namely whether or not a Lasso estimate $\hat{\theta}$ has non-zero entries in the same positions as the true regression vector θ^*.

In terms of the Lasso, we ask the following question: given an optimal Lasso solution , when is its support set - denoted by $S(\hat{\theta})$ - exactly equal to the true support $S(\theta^*)$? We refer to this property as variable selection consistency.

7.5.1 Variable selection consistency for the Lasso

For variable selection, we consider the following conditions:

Conditions

(A3) Lower eigenvalue: The smallest eigenvalue of the sample covariance submatrix indexed by S is bounded below:

$$\gamma_{\min}\left(\frac{\mathrm{X}_S^{\mathrm{T}}\mathrm{X}_S}{n}\right) \geq c_{\min} > 0$$

(A4) Mutual incoherence: There exists some $\alpha \in [0, 1)$ such that

$$\max_{j \in S^c} \|(\mathrm{X}_S^{\mathrm{T}}\mathrm{X}_S)^{-1}\mathrm{X}_S^{\mathrm{T}}X_j\|_1 \leq \alpha$$

Remarks on Condition (A3)

Conditions

(A3) Lower eigenvalue: The smallest eigenvalue of the sample covariance submatrix indexed by S is bounded below:

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

Condition (A3) is very mild: it would be required in order to ensure that the model is identifiable, even if the support set S were known a priori. If assumption (A3) were violated, then the submatrix X_S would have a non-trivial nullspace, leading to a non-identifiable model.

Remarks on Condition (A4)

Conditions

(A4) Mutual incoherence: There exists some $\alpha \in [0, 1)$ such that

$$\max_{j \in S^c} \|(\mathrm{X}_S^{\mathrm{T}} \mathrm{X}_S)^{-1} \mathrm{X}_S^{\mathrm{T}} X_j\|_1 \leq \alpha$$

Suppose that we tried to predict the column vector X_j using a linear combination of the columns of X_s. The best weight vector $\widehat{\omega} \in \mathbb{R}^{|S|}$ is given by

$$\widehat{\omega} = \arg \min_{\omega \in \mathbb{R}^{|S|}} \|X_j - \mathrm{X}_S \omega\|_2^2 = (\mathrm{X}_S^{\mathrm{T}} \mathrm{X}_S)^{-1} \mathrm{X}_S^{\mathrm{T}} X_j$$

and the mutual incoherence condition is a bound on $\|\widehat{\omega}\|_1$. In the ideal case, if the column space of X_S were orthogonal to X_j, then the optimal weight vector $\widehat{\omega}$ would be identically zero. In general, we cannot expect this orthogonality to hold, but the mutual incoherence condition (A4) imposes a type of approximate orthogonality.

Main theorem

Theorem 7.21

Consider an S-sparse linear regression model for which the design matrix satisfies conditions (A3) and (A4). Then for any choice of regularization parameter such that

$$\lambda_n \geq \frac{2}{1-\alpha} \|X_{S^c}^{\mathrm{T}} \Pi_{S^\perp}(X) \frac{w}{n}\|_\infty \quad (7.44)$$

where $\Pi_{S^\perp}(X) = I_n - X_S(X_S^{\mathrm{T}} X_S)^{-1} X_S^{\mathrm{T}}$, the Lagrangian Lasso has the following properties:

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

(b) No false inclusion: This solution has its support set $\hat{S}$ contained within the true support set S.

Main theorem

Theorem 7.21

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

$$\|\widehat{\theta}_S - \theta_S^*\|_\infty \le \underbrace{\left\| \left(\frac{\mathrm{X}_S^{\mathrm{T}} \mathrm{X}_S}{n} \right)^{-1} \mathrm{X}_S^{\mathrm{T}} \frac{w}{n} \right\|_\infty + \left\| \left(\frac{\mathrm{X}_S^{\mathrm{T}} \mathrm{X}_S}{n} \right)^{-1} \right\|_\infty^{-1} \lambda_n}_{B(\lambda_n;\mathrm{X})} \quad (7.45)$$

where $\|A\|_\infty = \max_{i=1} \dots s_i \sum_j |A_{ij}|$ is the matrix ℓ_∞-norm.

(d) No false exclusion: The Lasso includes all indices $i \in S$ such that $|\theta_i^*| > B\left(\lambda_n; \mathrm{X}\right)$, and hence is variable selection consistent if $\min_{i \in S} |\theta_i^*| > B\left(\lambda_n; \mathrm{X}\right)$.

Main theorem

Corollary 7.22

Consider the S-sparse linear model based on a noise vector w with zero-mean i.i.d. σ-sub-Gaussian entries, and a deterministic design matrix X that satisfies assumptions $(A3)$ and $(A4)$, as well as the C -column normalization condition $\left(\max_{j=1,\dots,d} \|X_j\|_2 / \sqrt{n} \leq C\right)$. Suppose that we solve the Lagrangian Lasso with regularization parameter $\lambda_n = \frac{2C\sigma}{1-\alpha}\{\sqrt{\frac{2\log(d-s)}{n}} + \delta\}$ for some $\delta > 0$. Then the optimal solution θ is unique with its support contained within S, and satisfies the ℓ_∞-error bound

$$\|\widehat{\theta}_S - \theta^*_S\|_\infty \leq \frac{\sigma}{\sqrt{C_{\min}}}\left\{\sqrt{\frac{2\log s}{n}} + \delta\right\} + \left\|\left(\frac{\mathrm{X}_S^{\mathrm{T}}\mathrm{X}_S}{n}\right)^{-1}\right\|_\infty \lambda_n$$

all with probability at least $1 - 4e^{-\frac{n\delta^2}{2}}$.

Proof of Corollary 7.22

We first verify that the given choice of regularization parameter satisfies the bound (7.44) with high probability. It suffices to bound the maximum absolute value of the random variables

$$Z_j := X_j^{\mathrm{T}} \underbrace{\left[\mathrm{I}_n - \mathrm{X}_S \left(\mathrm{X}_S^{\mathrm{T}} \mathrm{X}_S\right)^{-1} \mathrm{X}_S^{\mathrm{T}}\right]}_{\Pi_s \perp(\mathrm{X})} \left(\frac{w}{n}\right) \qquad \text{for } j \in S^c$$

Since $\Pi_{S^\perp}(\mathrm{X})$ is an orthogonal projection matrix, we have

$$\|\Pi_{S^\perp}(\mathrm{X}) X_j\|_2 \leq \|X_j\|_2 \leq C\sqrt{n}$$

Therefore, each variable Z_j is sub-Gaussian with parameter at most $C^2\sigma^2/n$. From standard sub-Gaussian tail bounds (Chapter 2), we have

$$\mathbb{P}\left[\max_{j \in S^c} |Z_j| \geq t\right] \leq 2(d-s)e^{-\frac{nt^2}{2C^2\sigma^2}}$$

from which we see that our choice of λ_n ensures that the bound (7.44) holds with the claimed probability.

Proof of Corollary 7.22

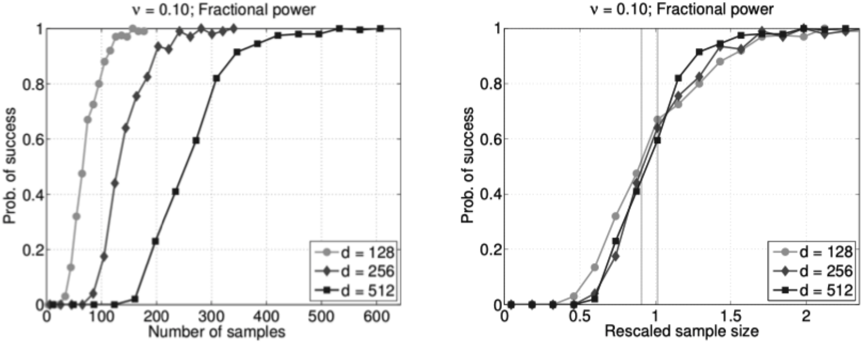
The only remaining step is to simplify the ℓ_∞-bound (7.45). The second term in this bound is a deterministic quantity, so we focus on bounding the first term. For each $i = 1, \ldots, s$, consider the random variable $\widetilde{Z}_i := e_i^{\mathrm{T}}(\frac{1}{n}\mathsf{X}_S^{\mathrm{T}}\mathsf{X}_S)^{-1}\mathsf{X}_S^{\mathrm{T}} w/n$. Since the elements of the vector w are i.i.d. σ-sub-Gaussian, the variable $\widetilde{Z}_i$ is zero-mean and sub-Gaussian with parameter at most

$$\frac{\sigma^2}{n}\left\|\left(\frac{1}{n}\mathsf{X}_S^{\mathrm{T}}\mathsf{X}_S\right)^{-1}\right\|_2 \leq \frac{\sigma^2}{c_{\min} n}$$

where we have used the eigenvalue condition (A3). Consequently, for any $\delta > 0$, we have $\mathbb{P}\left[\max_{i=1,\ldots,s}\left|\widetilde{Z}_i\right| > \frac{\sigma}{\sqrt{c_{\min}}}\left\{\sqrt{\frac{2\log s}{n}} + \delta\right\}\right] \leq 2e^{-\frac{n\delta^2}{2}}$, from which the claim follows.

Remarks on Corollary 7.22

Corollary 7.22 applies to linear models with a fixed matrix X of covariates. An analogous result - albeit with a more involved proof - can be proved for Gaussian random covariate matrices. Exercise 7.19 shows that the incoherence condition holds with high probability with $n \gtrsim s \log(d - s)$.



(a)

(b)

Proof of Theorem 7.21

For the Lagrangian Lasso program, we say that a pair $(\widehat{\theta}, \widehat{z}) \in \mathbb{R}^d \times \mathbb{R}^d$ is primal-dual optimal if $\widehat{\theta}$ is a minimizer and $\widehat{z} \in \partial \|\widehat{\theta}\|_1$. Any such pair must satisfy the zero-subgradient condition

$$\frac{1}{n} \mathsf{X}^{\mathrm{T}}(\mathsf{X}\widehat{\theta} - y) + \lambda_n \widehat{z} = 0 \quad (7.48)$$

Our proof of Theorem 7.21 is based on a constructive procedure, known as a primal-dual witness method, which constructs a pair $(\widehat{\theta}, \widehat{z})$ satisfying the zero-subgradient condition (7.48) and such that $\widehat{\theta}$ has the correct signed support. When this procedure succeeds, the constructed pair is primal-dual optimal, and acts as a witness for the fact that the Lasso has a unique optimal solution with the correct signed support.

Proof of Theorem 7.21

Primal–dual witness (PDW) construction

1. Set $\widehat{\theta}_{S^c} = 0$.
2. Determine $(\widehat{\theta}_S, \widehat{z}_S) \in \mathbb{R}^s \times \mathbb{R}^s$ by solving the oracle subproblem

$$\widehat{\theta}_S \in \arg\min_{\theta_s \in \mathbb{R}^s} \{ \underbrace{\frac{1}{2n} \|y - \mathrm{X}_S \theta_S\|_2^2}_{=:f(\theta_s)} + \lambda_n \|\theta_S\|_1 \} \quad (7.49)$$

   and then choosing $\widehat{z}_S \in \partial \|\widehat{\theta}_S\|_1$ such that $\nabla f\,(\theta_S)\,|_{\theta_{\mathrm{S}}=\widehat{\theta}_{\mathrm{S}}} + \lambda_n \widehat{z}_S = 0$.
3. Solve for $\widehat{z}_{S^c} \in \mathbb{R}^{d-s}$ via the zero-subgradient equation (7.48), and check whether or not the strict dual feasibility condition $\|\widehat{z}_{S^c}\|_\infty < 1$ holds.

Remarks on PDW

By construction, the subvectors $\widehat{\theta}_S, \widehat{z}_S$ and $\widehat{z}_{S^c}$ satisfy the zero-subgradient condition (7.48). By using the fact that $\widehat{\theta}_{S^c} = \theta^*_{S^c} = 0$ and writing out this condition in block matrix form, we obtain

$$\frac{1}{n}\begin{bmatrix} \mathrm{X}_S^{\mathrm{T}}\mathrm{X}_S & \mathrm{X}_S^{\mathrm{T}}\mathrm{X}_{S^c} \\ \mathrm{X}_{S^c}^{\mathrm{T}}\mathrm{X}_S & \mathrm{X}_{S^c}^{\mathrm{T}}\mathrm{X}_{S^c} \end{bmatrix}\begin{bmatrix} \hat{\theta}_S - \theta^*_S \\ 0 \end{bmatrix} - \frac{1}{n}\begin{bmatrix} \mathrm{X}_S^{\mathrm{T}} w \\ \mathrm{X}_{S^c}^{\mathrm{T}} w \end{bmatrix} + \lambda_n \begin{bmatrix} \widehat{z}_S \\ \widehat{z}_{S^c} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad (7.50)$$

We say that the PDW construction succeeds if the vector $\widehat{z}_S$ constructed in step 3 satisfies he strict dual feasibility condition.

Lemma 7.23

If the lower eigenvalue condition (A3) holds, then success of the PDW construction implies that the vector $(\widehat{\theta}_S, 0) \in \mathbb{R}^d$ is the unique optimal solution of the Lasso.

Proof of Lemma 7.23

Let $\widetilde{\theta}$ be any other optimal solution. If we introduce the shorthand notation $F(\theta) = \frac{1}{2n}\|y - \mathrm{X}\theta\|_2^2$, then we are guaranteed that

$$F(\widehat{\theta}) + \lambda_n(\widehat{z}, \widehat{\theta}\rangle = F(\widetilde{\theta}) + \lambda_n\|\widetilde{\theta}\|_1$$

$$F(\widehat{\theta}) - \lambda_n\langle \widehat{z}, \widetilde{\theta} - \widehat{\theta}\rangle = F(\tilde{\theta}) + \lambda_n\left(\|\widetilde{\theta}\|_1 - \langle \widehat{z}, \tilde{\theta}\rangle\right)$$

But by the zero-subgradient conditions (7.48), we have $\lambda_n\widehat{z} = -\nabla F(\widehat{\theta})$, which implies that

$$F(\widehat{\theta}) + \langle \nabla F(\widehat{\theta}), \widetilde{\theta} - \widehat{\theta}\rangle - F(\widetilde{\theta}) = \lambda_n\left(\|\widetilde{\theta}\|_1 - \langle \widehat{z}, \widetilde{\theta}\rangle\right)$$

$$\Rightarrow \|\widetilde{\theta}\|_1 \leq \langle \widehat{z}, \widetilde{\theta}\rangle \leq \|\widetilde{z}\|_\infty\|\widetilde{\theta}\|_1 \leq \|\widetilde{\theta}\|_1$$

$$\Rightarrow \|\widetilde{\theta}\|_1 = \langle \widehat{z}, \widetilde{\theta}\rangle$$

Since $\|\widehat{z}_{S^c}\|_\infty < 1$, this equality can only occur if $\widetilde{\theta}_j = 0$ for all $j \in S^c$. Thus, all optimal solutions are supported only on S , and hence can be obtained by solving the oracle subproblem (7.49). Given the lower eigenvalue condition (A3), this subproblem is strictly convex, and so has a unique minimizer.

Proof of Theorem 7.21

To prove Theorem 7.21(a) and (b), it suffices to show that the vector $\widehat{z}_{S^c} \in \mathbb{R}^{d-s}$ constructed in step 3 satisfies the strict dual feasibility condition. From (7.50),

$$\widehat{z}_{S^c} = -\frac{1}{\lambda_n n} \mathrm{X}_{S^c}^{\mathrm{T}} \mathrm{X}_S (\widehat{\theta}_S - \theta_S^*) + \mathrm{X}_{S^c}^{\mathrm{T}} \left(\frac{w}{\lambda_n n} \right)$$

$$\widehat{\theta}_S - \theta_S^* = (\mathrm{X}_S^{\mathrm{T}} \mathrm{X}_S)^{-1} \mathrm{X}_S^{\mathrm{T}} w - \lambda_n n (\mathrm{X}_S^{\mathrm{T}} \mathrm{X}_S)^{-1} \widehat{z}_S$$

$$\Rightarrow \widehat{z}_{S^c} = \underbrace{\mathrm{X}_{S^c}^{\mathrm{T}} \mathrm{X}_S (\mathrm{X}_S^{\mathrm{T}} \mathrm{X}_S)^{-1} \widehat{z}_S}_{\mu} + \underbrace{\mathrm{X}_{S^c}^{\mathrm{T}} \left[I - \mathrm{X}_S (\mathrm{X}_S^{\mathrm{T}} \mathrm{X}_S)^{-1} \mathrm{X}_S^{\mathrm{T}} \right] \left(\frac{w}{\lambda_n n} \right)}_{V_{S^c}}$$

$$\Rightarrow \| \widehat{z}_{S^c} \|_\infty \leq \| \mu \|_\infty + \| V_{S^c} \|_\infty \leq \alpha + (1-\alpha)/2 = (1+\alpha)/2 < 1$$

by the mutual incoherence condition (A4) and our choice of regularization parameter λ_n.

Proof of Theorem 7.21

It remains to establish a bound on the ℓ_∞-norm of the error $\widehat{\theta}_S - \theta^*_S$. From equation

$$\widehat{\theta}_S - \theta^*_S = (\mathsf{X}_S^{\mathrm{T}}\mathsf{X}_S)^{-1}\mathsf{X}_S^{\mathrm{T}} w - \lambda_n n (\mathsf{X}_S^{\mathrm{T}}\mathsf{X}_S)^{-1}\widehat{z}_S$$

and the triangle inequality, we have

$$\left\|\widehat{\theta}_S - \theta^* s\right\|_\infty \leq \left\| \left(\frac{\mathsf{X}_S^{\mathrm{T}}\mathsf{X}_S}{n}\right)^{-1} \mathsf{X}_S^{\mathrm{T}} \frac{w}{n} \right\|_\infty + \left\| \left(\frac{\mathsf{X}_S^{\mathrm{T}}\mathsf{X}_S}{n}\right)^{-1} \right\|_\infty^{-1} \lambda_n$$

which completes the proof.

Extension to random Gaussian ensembles

The previous section treated the case of a deterministic X, which allowed for a relatively straightforward analysis. We now turn to the more complex case of random design matrix $X \in \mathbb{R}^{n \times p}$, in which each row $x_i, i = 1, \ldots, n$ is chosen as an i.i.d. Gaussian random vector with covariance matrix Σ.

Conditions

(A5) Lower eigenvalue: The smallest eigenvalue of the covariance submatrix indexed by S is bounded below:

$$\gamma_{\text{min}}\left(\Sigma_{SS}\right) \geq c_{\text{min}} > 0$$

(A6) Mutual incoherence: There exists some $\alpha \in [0, 1)$ such that

$$\max_{j \in S^c} \|\Sigma_{jS} \Sigma_{SS}^{-1}\|_1 \leq \alpha$$

Main theorem

Theorem

Consider the family of regularization parameters $\lambda_n(\phi_d) = \sqrt{\frac{\phi_d \rho_u(\Sigma_{S^c|S})}{(1-\alpha)^2} \frac{2\sigma^2 \log(d)}{n}}$ where $\Sigma_{S^c|S} := \Sigma_{S^cS^c} - \Sigma_{S^cS} \left(\Sigma_{SS}\right)^{-1} \Sigma_{SS^c}$, $\phi_d \geq 2$, $\rho_u(A) = \max_i A_{ii}$. If for some fixed $\delta > 0$, n, d, s and λ_n satisfy $\frac{n}{2s\log(d-s)} > (1+\delta)\frac{\rho_u(\Sigma_{S^c|S})}{c_{\min}(1-\alpha)^2}\left(1 + \frac{\sigma^2 c_{\min}}{\lambda_n^2 s}\right)$ then the following properties holds with probability greater than $1 - c_1 e^{-c_2 \min\{s, \log(d-s)\}}$

(a) The Lasso has a unique solution $\widehat{\theta}$ with support contained within S.

(b)

$$\|\widehat{\theta}_S - \theta^*_S\|_\infty \leq \underbrace{c_3 \lambda_n \left\|\Sigma_{SS}^{-1/2}\right\|_\infty^2 + 20\sqrt{\frac{\sigma^2 \log s}{c_{\min} n}}}_{B(\lambda_n)}$$

Proof

As with the proof of Theorem 7.21, the proof is based on the PDW method and we need to verify the strict dual feasibility condition

$$\|\widehat{z}_{S^c}\|_\infty < 1$$

For $j \in S^c$, define

$$W = X_{S^c} - X_S \Sigma_{SS}^{-1} \Sigma_{SS^c} \in \mathbb{R}^{n\times(d-s)}$$

where W is independent of X_S and $W_{ij} \sim N(0, [\Sigma_{S^c|S}]_{jj})$. Then

$$\widehat{z}_{S^c} = \underbrace{\Sigma_{S^c S}\Sigma_{SS}^{-1}\widehat{z}_S}_{A} + \underbrace{W^{\mathrm{T}}\left\{X_S(X_S^T X_S)^{-1}\widehat{z}_S + \Pi_{X_S^\perp}\left(\frac{w}{\lambda_n n}\right)\right\}}_{B}$$

$$\widehat{z}_{S^c,j} = A_j + B_j = \Sigma_{jS}\Sigma_{SS}^{-1}\widehat{z}_S + W_j^{\mathrm{T}}\left\{X_S(X_S^T X_S)^{-1}\widehat{z}_S + \Pi_{X_S^\perp}\left(\frac{w}{\lambda_n n}\right)\right\}$$

Proof

By condition (A6),

$$\max_{j \in S^c} |A_j| \leq \alpha$$

Since $\operatorname{var}(W_{ij}) = \left[\Sigma_{S^c|S}\right]_{jj} \leq \rho_u\left(\Sigma_{S^c|S}\right)$, conditioned on X_S and w, the quantity B_j is zero-mean Gaussian with variance at most

$$\operatorname{var}(B_j) \leq \rho_u(\Sigma_{S^c|S}) \left\| X_S \left(X_S^T X_S\right)^{-1} \widehat{z}_S + \Pi_{X_{S^\perp}}\left(\tfrac{w}{\lambda_n n}\right)\right\|_2^2$$
$$= \rho_u \underbrace{\left\{\frac{1}{n}\widehat{z}_S^T \left(\frac{X_S^T X_S}{n}\right)^{-1} \widehat{z}_S + \left\|\Pi_{X_{S^\perp}}\left(\frac{w}{\lambda_n n}\right)\right\|_2^2\right\}}_{M_n}$$

Proof

Lemma

For any $\epsilon \in (0, 1/2)$, define the event $\bar{T}(\epsilon) = \left\{ M_n > \bar{M}_n(\epsilon) \right\}$, where

$$\bar{M}_n(\epsilon) := \left(1 + \max\left\{ \epsilon, \frac{8}{c_{\min}} \sqrt{\frac{s}{n}} \right\} \right) \left(\frac{s}{c_{\min} n} + \frac{\sigma^2}{\lambda_n^2 n} \right)$$

Then $\mathbb{P}(\bar{T}(\epsilon)) \le 4 \exp\left(-c_1 \min\left\{ n\epsilon^2, s \right\} \right)$ for some $c_1 > 0$.

$$\begin{aligned} \mathbb{P}\left[\max_{j \in S^c} |B_j| \ge (1-\alpha) \right] &= \mathbb{P}\left[\max_{j \in S^c} |B_j| \ge (1-\alpha) | \bar{T}^c(\epsilon) \right] \mathbb{P}[\bar{T}^c(\epsilon)] \\ &+ \mathbb{P}\left[\max_{j \in S^c} |B_j| \ge (1-\alpha) | \bar{T}(\epsilon) \right] \mathbb{P}[\bar{T}(\epsilon)] \\ &\le \mathbb{P}\left[\max_{j \in S^c} |B_j| \ge (1-\alpha) \mid \bar{T}^c(\epsilon) \right] + 4 \exp\left(-c_1 \min\left\{ n\epsilon^2, k \right\} \right) \end{aligned}$$

Proof

Conditioned on $\bar{T}^c(\epsilon)$, the variance of B_j is at most $\rho_u(\Sigma_{S^c|S})\bar{M}_n(\epsilon)$, so that by standard Gaussian tail bounds, we obtain the upper bound

$$\mathrm{P}\left[\max_{j \in S^c} |B_j| \geq (1-\alpha) \mid \bar{T}^c(\epsilon)\right] \leq 2(d-s)\exp\left(-\frac{(1-\alpha)^2}{2\rho_u \bar{M}_n(\epsilon)}\right)$$

Reference: M. J. Wainwright, "Sharp Thresholds for High-Dimensional and Noisy Sparsity Recovery Using ℓ_1 -Constrained Quadratic Programming (Lasso)," in IEEE Transactions on Information Theory, vol. 55, no. 5, pp. 2183-2202, May 2009, doi: 10.1109/TIT.2009.2016018.

The End