Statistics 210B Lecture 18 Notes

Daniel Raban

March 29, 2022

1 Sufficient Conditions for Exact Recovery in Sparse Linear Regression and Introduction to Noisy, Sparse Linear Regression

1.1 Recap: sparse linear regression via the restricted nullspace condition

Our model is a the high dimensional sparse linear model, $y = X\theta^* \in \mathbb{R}^n$, where $X \in \mathbb{R}^{n \times d}$, $\theta^* \in \mathbb{R}^d$ and the support of θ^* has cardinality $|S(\theta^*)| \leq s$. Given (y, X), we want to recover θ^*. When $d > n$, we want

$$\widehat{\theta} := \arg\min_{y = X\theta} \|\theta\|_1.$$

When can we have exact recovery? Last time, we had the following condition.

Definition 1.1 (Restricted nullspace)**.** Let $S \subseteq [d]$. $X \in \mathbb{R}^{n \times d}$ satisfies $\mathbf{RN}(S)$ if $\mathbb{C}(S) \cap \text{Null}(X) = \{0\}$, where

$$\mathbb{C}(S) := \{\Delta \in \mathbb{R}^d : \|\Delta_{S^c}\|_1 \leq \|\Delta_S\|_1\}.$$

Theorem 1.1. *The following are equivalent:*

1. *For all $\theta^* \in \mathbb{R}^d$ with $S(\theta^*) = S$,*

$$\arg\min_{\theta} \{\|\theta\|_1 : X\theta^* = X\theta\} = \theta^*\}.$$

2. *X satisfies* $\text{RN}(S)$*, i.e.* $\text{Null}(X)$ *cap*$\mathbb{C}(S) = \{0\}$.

However, it is hard to verify the restricted nullspace property for a matrix, since we need to check all subsets of $[d]$ of cardinality s. How can we find examples of matrices satisfying this property?

1.2 Two sufficient conditions for the restricted nullspace property

The intuition is that if $d < n$ (which is not the case we want to solve), X is full-rank, so we can take $X^\top X/n = I_d$. This implies that $\text{Null}(X) = \{0\}$ because

$$\|Xv\|_2^2/n = v^\top (X^\top X/n)v = v^\top I_d v = \|v\|_2^2.$$

Since $\mathbb{C}(S)$ is basically $\{\theta : S(\theta) = S\}$, we can restrict to S. So as long as we have $(X^\top X)_{s,s}/n = I_S$, if $v \in \{\theta : S(\theta) = s\} \cap \text{Null}(X)$, we can say

$$v^\top (X^\top X/n)v = v_s^\top (X^\top X/n)_{s,s} v_s = \|v_s\|_2^2.$$

This equals 0, so we get $v_S = 0$; i.e. $v = 0$.

This motivates the following definitions.

Definition 1.2. Let $\Gamma = X^\top X/n - I_d$. The **pairwise incorherence**[1] is

$$\delta_{\text{PW}}(X) = \max_{i,j} |\Gamma_{i,j}| = \max_{i,j} |(X^\top X/n - I_d)_{i,j}|.$$

The **restricted isometry constant**[2] is

$$\delta_s(X) = \max_{|S| \le s} \|\Gamma_{S,S}\|_{\text{op}} = \max_{|S| \le s} \|X_S^\top X_S/n - I_S\|_{\text{op}},$$

where $X_S \in \mathbb{R}^{n \times s}$ is the matrix where we only keep the columns in S.

Note that $\delta_d = \|\Gamma\|_{\text{op}}$.

1.2.1 The pairwise incoherence condition

Proposition 1.1 (Incoherence implies RN(S))**.** *If $\delta_{\text{PW}}(X) \le \frac{1}{3s}$, then X satisfies* RN(S) *for any $|S| \le s$.*

Proof. Assume that $\delta_{\text{PW}}(X) \le \frac{1}{3s}$, and take any $\theta \in \text{Null}(X) \setminus \{0\}$; we want to show that $\theta \notin \mathbb{C}(S)$. Let $S \subseteq [d]$ with $|S| \le s$. That is, our goal is to show that $\|\theta_{S^c}\|_1 > \|\theta_S\|_1$. The nullspace condition gives

$$0 = \|X\theta\|_2^2$$

We now want to decompose this into θ_S and θ_{S^c} so these two quantities appear. Writing $\theta_S \in \mathbb{R}^d$,

$$= \|X(\theta_{S^c} + \theta_S)\|_2^2$$

[1] The pairwise incorherence was introduced in 2001 by Donoho and Huo.

[2] The restricted isometry constant was introduced by Candès and Tao in 2005

$$= \theta_S^\top X_S^\top X_S \theta_S + 2\theta_{S^c} X_{S^c}^\top X_S \theta_S + \underbrace{\|X_{S^c}\theta_{S^c}\|_2^2}_{\geq 0}.$$

This implies that

$$\theta_S^\top X_S^\top X_S \theta_S \leq 2|\theta_{S^c}^\top X_{S^c}^\top X_S \theta_S|.$$

We can normalize by n to get

$$\theta_S^\top (X_S^\top X_S/n)\theta_S \leq 2|\theta_{S^c}^\top (X_{S^c}^\top X_S/n)\theta_S|.$$

The left hand side is

$$\theta_S^\top (X_S^\top X_S/n)\theta_S \geq \lambda_{\min}(X_S^\top X_S/n)\|\theta_S\|_2^2$$

Using the fact that $\|\theta\|_1^2 \leq \|\theta\|_0\|\theta_2^2$, we get the lower bound

$$\geq \lambda_{\min}(X_S^\top X_S/n)\|\theta_S\|_1^2/s.$$

To upper bound the right hand side, we use the fact that $a^\top Ab \leq \|a\|_1\|Ab\|_\infty \leq \|a\|_1\|A\|_{\max}\|b\|_1$. Then

$$2|\theta_{S^c}^\top (X_{S^c}^\top X_S/n)\theta_S| \leq \|\theta_S\|_1\|\theta_{S^c}\|_1\|X_{S^c}^\top X_S/n\|_{\max}/$$

Putting these inequalities together gives

$$\frac{\|\theta_{S^c}\|_1}{\|\theta_S\|_1} \geq \frac{\lambda_{\min}(X_s^\top X_s/n)}{2s\|X_{S^c}^\top X_S/n\|_{\max}}.$$

So far, we have not used the pairwise incoherence. We claim that the pairwise incoherence condition $\delta_{\mathrm{PW}}(X) < \frac{1}{3s}$ makes the right hand side > 1. The key is to observe that $\|X_{S^c}^\top X_S/n\|_{\max} \leq \delta_{\mathrm{PW}}(X)$ and that $\lambda_{\min}(X_s^\top X_s/n) \geq 2/3$ if the pairwise incoherence condition is satisfied. □

1.2.2 The restricted isometry property

Here is another condition that implies the restricted nullspace property.

Proposition 1.2 (Restricted isometry property implies RN(S))**.** *If $\delta_{2s}(X) \leq 1/3$, then X satisfies* RN(S) *for any $|S| \leq s$.*

This is proposition 7.11 in Wainwright's textbook, and we will not provide the proof here.

Remark 1.1. In general, we have the algebraic inequality

$$\delta_{\mathrm{PW}}(X) \leq \delta_S(X) \leq s\delta_{\mathrm{PW}}(X).$$

The pairwise incoherence is computable in polynomial time, while the weaker RIP condition needs time $\sum_{k=1}^{s} \binom{d}{k}$. Here is an exercise which shows that we can satisfy these conditions randomly.

Proposition 1.3. *Let $X \in \mathbb{R}^{n \times d}$ and $X_{i,j} \overset{\text{iid}}{\sim} N(0,1)$. Then*

(a) If $n \gtrsim s^2 \log d$, then $\delta_{\text{PW}}(X) \leq \frac{1}{3s}$ with high probability.

(b) If $n \gtrsim s \log(\frac{ed}{s})$, then $\delta_{2s}(X) \leq \frac{1}{3}$ with high probability.

Here is the idea of the proof.

Proof.

(a) Write

$$\begin{aligned} \delta_{\text{PW}} &= \max_{i,j} |(X^\top X/n - I_d)_{i,j}| \\ &= \max_{i,j} \left| \frac{1}{n} \sum_{k=1}^{n} x_{i,k} x_{j,k} - \delta_{i,j} \right| \end{aligned}$$

Note that $\mathbb{E}[\frac{1}{n} \sum_{i=k}^{n} X_{i,k} X_{j,k}] = \delta_{i,j}$, so $I_{i,j} = \frac{1}{n} \sum_{i=k}^{n} X_{i,k} X_{j,k}] - \delta_{i,j}$ will be $\text{sE}(\frac{1}{\sqrt{n}} \frac{1}{n})$ for fixed i, j. Then Bernstein's inequality gives

$$\mathbb{P}(|I_{i,j}| \geq t) \leq 2 \exp(-cn \min(t, t^2)).$$

Using a union bound, we get

$$\mathbb{P}\left(\max_{i,j} |I_{i,j}| \geq t \right) \leq 2d^2 \exp(-cn \min(t, t^2)).$$

Now, if we let $t = \frac{1}{3s}$, call the right hand side δ, and solve for n, we get the condition $n \gtrsim s^2 \log(d.\delta)$.

(b) The proof is similar, using the matrix version of concentration. □

Remark 1.2. Certain random matrix distributions will satisfy $\text{RN}(S)$ but not the RIP or coherence. For example, we will show later that if $X_i \overset{\text{iid}}{\sim} N(0, \Sigma)$, where $\Sigma = (1 - \mu) I_d + \mu \mathbf{1}\mathbf{1}^\top$, then X still satisfies $\text{RN}(S)$ with high probability. Here is a figure from

Wainwright's textbook:

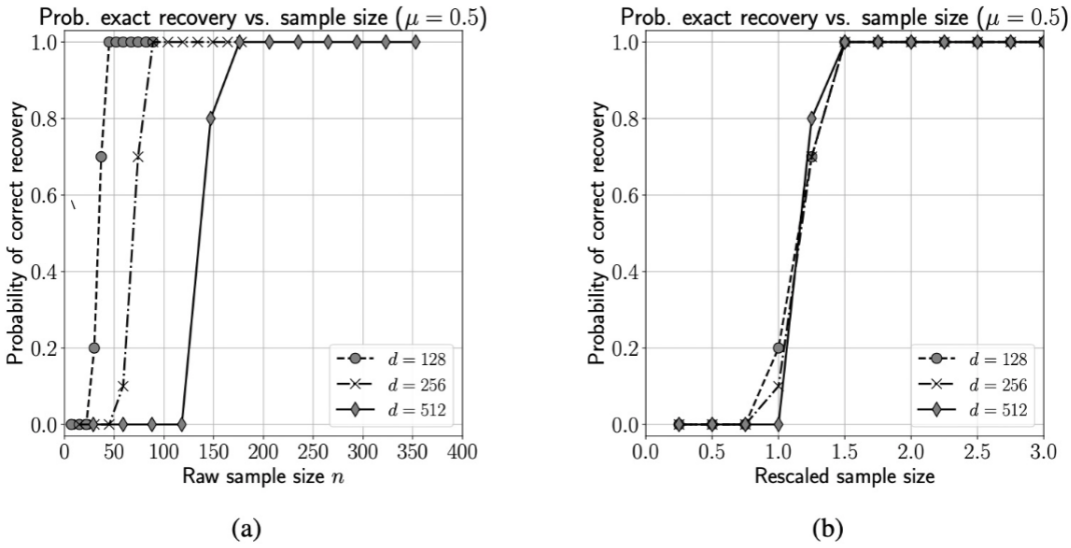

Figure 7.4 (a) Probability of basis pursuit success versus the raw sample size n for random design matrices drawn with i.i.d. rows $X_i \sim \mathcal{N}(0, \mathbf{\Sigma})$, where $\mu = 0.5$ in the model (7.17). Each curve corresponds to a different problem size $d \in \{128, 256, 512\}$ with sparsity $s = \lceil 0.1d \rceil$. (b) The same results replotted versus the rescaled sample size $n/(s \log(ed/s))$. The curves exhibit a phase transition at the same value of this rescaled sample size.

Here, there is a phase transition threshold which needs to be identified with an asymptotic analysis that we will not cover.

1.3 Estimation in the noisy setting

Now we will change our model to $y = X\theta^* + w \in \mathbb{R}^n$, where

$$w = \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix} \in \mathbb{R}^n, \qquad X \in \mathbb{R}^{n \times d}, \qquad X = \begin{bmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{bmatrix}, \qquad \theta^* = \begin{bmatrix} \theta_1^* \\ \vdots \\ \theta_n^* \end{bmatrix}.$$

We assume the sparsity condition $|S(\theta^*)| \leq s$. Given (y, X), we want to estimate θ^*. This time, we want to minimize $\|\theta\|_1$ subject to the constraint that $\|y - X\theta\| \leq b^2$.

Here are three equivalent formulations of the **LASSO problem**, which we use for our estimation:

1. The λ **formulation**:

$$\widehat{\theta} = \arg\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{2n} \|y - X\theta\|_2^2 + \lambda_n \|\theta\|_1 \right\},$$

2. 1**-norm constrained formulation**:

$$\arg\min_{\theta} \left\{ \frac{1}{2n} \|y - X\theta\|_2^2 \right\} \qquad \text{s.t. } \|\theta\|_1 \leq R$$

3. The **error constrained formulation**:

$$\arg\min_{\theta} \{\|\theta\|_1\} \qquad \text{s.t. } \frac{1}{2n} \|y - X\theta\|_2^2 \leq b^2.$$

These are equivalent in the sense that for all $\lambda_n > 0$, there is an $R < \infty$ such that the solution fo the 1-norm constrained formulation with parameter R is a solution of the λ formulation. Similarly, we can go the other way. This equivalence requires a condition on X and is just convex duality.

How can we bound the estimation error? We will discuss this next time.