High Dimensional Statistics

Yanbo Tang

Imperial College London

March. 2025

Content of this week

- LASSO MSE
- SLOPE
- LASSO sparse set recovery

Fixed versus random design

Random design

$$\mathbb{E}_{X \sim P}\left[\|\hat{f}(X) - f(X)\|_2^2\right]$$

Fixed design

$$\frac{1}{n}\sum_{i=1}^{n} \|\hat{f}(x_i) - f(x_i)\|_2^2 = \mathbb{E}_{X \sim \bar{P}_n}\left[\|\hat{f}(X) - f(X)\|_2^2\right]$$

L^0 performance

θ is k sparse

Denote by $B_0(k)$ the ℓ_0 "ball" of $\mathbb{R}^d$, i.e., the set of k-sparse vectors, defined by

$$B_0(k) = \{\theta \in \mathbb{R}^d : |\theta|_0 \leq k\}.$$

Our goal is to control the MSE of $\hat{\theta}^{LS}_K$ when $K = B_0(k)$. Note that computing $\hat{\theta}^{LS}_{B_0(k)}$ is to find:

$$\hat{\theta}^{\mathrm{LS}}_{B_0(k)} \in \operatorname*{argmin}_{\theta \in B_0(k)} |Y - \mathbb{X}\theta|_2^2,$$

but this would require computing $\binom{d}{k}$ least squares estimators.

Best subset selection

Theorem

Fix a positive integer $k \leq d/2$. Let $K = B_0(k)$ be set of k-sparse vectors of $\mathbb{R}^d$, assume that $\theta^* \in B_0(k)$ and assume that the linear model with $\varepsilon \sim subG_n(\sigma^2)$. Then, for any $\delta > 0$, with probability $1 - \delta$, it holds

$$MSE(\mathbb{X}\hat{\theta}^{IS}_{B_0(k)}) \lesssim \frac{k\sigma^2}{n} \log\left(\frac{ed}{2k}\right) + \log(6)\frac{\sigma^2 k}{n} + \frac{\sigma^2}{n}\log(1/\delta).$$

Note that if we knew the exact sparsity level k_0, then this bound would be essentially optimal.

Matching rates for LASSO?

Assumption INC(k) We say that the design matrix $\mathbb{X}$ has incoherence k for some integer $k > 0$ if

$$\left| \frac{\mathbb{X}^T \mathbb{X}}{n} - I_d \right|_\infty \leq \frac{1}{32k}$$

where the $|A|_\infty$ denotes the largest element of A in absolute value. Equivalently,

1. For all $j = 1, \ldots, d$,
$$\left| \frac{\|\mathbb{X}_j\|_2^2}{n} - 1 \right| \leq \frac{1}{32k}.$$
2. For all $1 \leq i, j \leq d, i \neq j$, we have
$$\frac{|\mathbb{X}_i^T \mathbb{X}_j|}{n} \leq \frac{1}{32k}.$$

Technical Lemma

Lemma

Fix a positive integer $k \leq d$ and assume that $\mathbb{X}$ satisfies assumption INC(k). Then, for any $S \in \{1, \ldots, d\}$ such that $|S| \leq k$ and any $\theta \in \mathbb{R}^d$ that satisfies the cone condition

$$|\theta_{S^c}|_1 \leq 3|\theta_S|_1,$$

it holds

$$|\theta|_2^2 \leq 2\frac{|\mathbb{X}\theta|_2^2}{n}.$$

We will interpret the cone condition when we consider sparse recovery.

LASSO

Theorem

Fix $n \geq 2$. Assume that the linear model with $\varepsilon \sim subG_n(\sigma^2)$. Moreover, assume that $\|\theta^*\|_0 \leq k$ and that X satisfies assumption INC(k). Then the Lasso estimator $\hat{\theta}^{\mathcal{L}}$ with regularization parameter defined by

$$2\tau = 8\sigma\sqrt{\frac{\log(2d)}{n}} + 8\sigma\sqrt{\frac{\log(1/\delta)}{n}}$$

satisfies

$$MSE(X\hat{\theta}^{\mathcal{L}}) = \frac{1}{n}\|X\hat{\theta}^{\mathcal{L}} - X\theta^*\|_2^2 \lesssim k\sigma^2\frac{\log(2d/\delta)}{n}.$$

and

$$\|\hat{\theta}^{\mathcal{L}} - \theta^*\|_2^2 \lesssim k\sigma^2\frac{\log(2d/\delta)}{n}.$$

with probability at least $1-\delta$.

Proof

From definition of $\hat{\theta}_2$ (trying to show the cone condition)

$$\frac{1}{n}\|Y - X\hat{\theta}_2\| \le \frac{1}{n}\|Y - X\theta^*\|_2^2 + 2\tau\|\theta^*\|_1 - 2\tau\|\hat{\theta}_2\|_1$$

Add $\tau\|\hat{\theta}_2 - \theta_*\|_1$ and multiply by n on both side

$$\|X\hat{\theta}_2 - X\theta^*\|_2^2 + n\tau\|\hat{\theta}_2 - \theta^*\|_1 \le 2\varepsilon^T X(\hat{\theta}_2 - \theta^*) + n\tau\|\hat{\theta}_2 - \theta^*\|_1 + 2n\tau\|\theta^*\|_1 - 2n\tau\|\hat{\theta}_2\|_1 \quad (1)$$

$$\varepsilon^T X(\hat{\theta}_2 - \theta^*) \le \|\varepsilon^T X\|_\infty \|\hat{\theta}_2 - \theta^*\|_1 \le \frac{n\tau}{2}\|\hat{\theta}_2 - \theta^*\|_1 \quad \text{w.p } 1-\delta$$

(as $P(|X^T\varepsilon|_\infty \ge \epsilon) \le P(\max_{j=1,\dots,d} |X_j^T\varepsilon| > t) \le 2d\exp(-\frac{\epsilon^2}{4n\delta^2})$

where we used $|X_j|_2^2 \le nf\frac{1}{32k} < 2n$

take $\epsilon = n\tau$ in

Proof Cont.

$$\|X\hat{\theta}^L - X\theta^*\|_2^2 + n\tau\|\hat{\theta}^L - \theta^*\|_1 \qquad \|\theta\|_1 = \|\theta_S\|_1 + \|\theta_{S^c}\|_1$$

$$\le 2n\tau\|\hat{\theta}^L - \theta^*\|_1 + 2n\tau\|\theta^*\|_1 - 2n\tau\|\hat{\theta}^L\|_1$$

$$= 2n\tau\|\hat{\theta}^L_S - \theta^*\|_1 + 2n\tau\|\theta^*\|_1 - 2n\tau\|\hat{\theta}^L_S\|_1$$

$$\le 4n\tau\|\hat{\theta}^L_S - \theta^*\|_1 \quad (2)$$

(2) implies the cone condition

$$\|X\hat{\theta}^L - X\theta^*\|_2^2 + n\tau\|\hat{\theta}^L_{S^c} - \theta^*_{S^c}\|_1 + n\tau\|\hat{\theta}^L_S - \theta^*_S\|_1$$

$$\le 4n\tau\|\hat{\theta}^L_S - \theta^*\|_1$$

$$\frac{\|X\hat{\theta}^L - X\theta^*\|_2^2}{n\tau} + \|\hat{\theta}^L_{S^c} - \theta^*_{S^c}\|_1 \le 3\|\hat{\theta}^L_S - \theta^*_S\|_1$$

this shows the cone condition.

$$\|\hat{\theta}^L_S - \theta^*\|_1 \le \sqrt{|S|}\,\|\hat{\theta}^L_S - \theta^*\|_2 \le \sqrt{|S|}\,\|\hat{\theta}^L - \theta^*\|_2$$

$$\le \sqrt{\frac{2k}{n}}\,\|X\hat{\theta}^L - X\theta^*\|_2$$

Optimal rate (MSE) - SLOPE

Definition

(*Slope estimator*). Let $\lambda = (\lambda_1, \ldots, \lambda_d)$ be a non-increasing sequence of positive real numbers, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d > 0$. For $\theta = (\theta_1, \ldots, \theta_d) \in \mathbb{R}^d$, let $(\theta_1^*, \ldots, \theta_d^*)$ be a non-increasing rearrangement of the modulus of the entries, $|\theta_1|, \ldots, |\theta_d|$. We define the *sorted* ℓ_1 *norm* of θ as

$$|\theta|_* = \sum_{j=1}^{d} \lambda_j |\theta_j^*|,$$

or equivalently as

$$|\theta|_* = \max_{\phi \in S_d} \sum_{j=1}^{d} \lambda_j |\theta_{\phi(j)}|.$$

Cont.

Definition

The *Slope estimator* is then given by

$$\hat{\theta}^S \in \arg\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{n} \| Y - X\theta \|_2^2 + 2\tau |\theta|_* \right\}$$

for a choice of tuning parameters λ and $\tau > 0$.

In what follows, we use

$$\lambda_j = \sqrt{\log(2d/j)}, \quad j = 1, \ldots, d.$$

MSE - SLOPE

Theorem

Fix $n \geq 2$. Assume that the linear model holds where $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I_n)$. Moreover, assume that $|\theta^*|_0 \leq k$ and that $\mathbb{X}$ satisfies assumption INC(k') with $k' \geq 4k \log(2de/k)$. Then the Slope estimator $\hat{\theta}^S$ with regularization parameter defined by

$$\tau = 8\sqrt{2}\sigma\sqrt{\frac{\log(1/\delta)}{n}}$$

satisfies

$$MSE(\mathbb{X}\hat{\theta}^S) = \frac{1}{n}\|\mathbb{X}\hat{\theta}^S - \mathbb{X}\theta^*\|_2^2 \lesssim \sigma^2 \frac{k \log(2d/k\delta)}{n}$$

and

$$\|\hat{\theta}^S - \theta^*\|_2^2 \lesssim \sigma^2 \frac{k \log(2d/k) \log(1/\delta)}{n}.$$

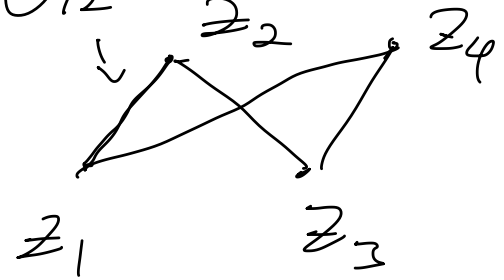
with probability at least $1 - \delta$.

Why sparsity - selection of Gaussian graphical models

$(Z_1, \dots, Z_d)$ $\quad \frac{1}{\sqrt{(2\pi)^d \det(\Theta^{-1})}} \exp\left(-\frac{1}{2} z^T \Theta z\right)$

$\Theta = \Sigma^{-1}$

$\Theta_{12} \neq 0$



$$\begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix}$$

$S \subseteq V := \{1, 2, \dots, d\}$ $\quad N(s) = \{t \in V \mid \Theta_{st} \neq 0\}$

$Z_s = \langle Z_{\setminus s}, \theta^* \rangle + \varepsilon_s \quad Z_{\setminus s} = \begin{pmatrix} Z_2 \\ Z_3 \\ Z_4 \\ \vdots \\ Z_d \end{pmatrix}$

$\varepsilon \sim N(0, \sigma^2)$

the sparsity pattern for θ^* matches that of Θ_S.

Basis pursuit

Before we begin with the sparse recovery of the noisy version of the problem, let us think about the deterministic version of the problem. Suppose that you are asked to solve the following system of equations for θ:

$$X\theta = Y. \qquad X \in \mathbb{R}^{n \times d}, \quad Y \in \mathbb{R}^{n}$$

if $d > n$ impossible

if additionally $\|\theta\|_0 \le k$ is known to be a sol^{n}

$$\min_{\theta \in \mathbb{R}^d} \|\theta\|_0 \ : \ X\theta = Y$$

$$\min_{\theta \in \mathbb{R}^d} \|\theta\|_1 \ : \ X\theta = Y$$

Restricted nullspace

$$\text{null}(X) = \{\Delta \in \mathbb{R}^d : X\Delta = 0\}$$

$$\forall \Delta \in \text{null}(X) \quad X\theta^* = y \Rightarrow X(\theta^* + \Delta) = y$$

$$\underline{\theta} \quad T(\theta^*) = \{\Delta \in \mathbb{R}^d \mid \|\theta^* + \Delta t\|_1 \le \|\theta^*\|_1 \text{ for some } t > 0\}$$

$$\|\theta^*\|_1 \ge \|\theta^* + t\Delta\|_1 = \|\theta^* + t\Delta_S\|_1 + \|t\Delta_{S^c}\|_1$$

$$\ge \|\theta^*\|_1 - t\|\Delta_S\|_1 + t\|\Delta_{S^c}\|_1$$

$$\Rightarrow \|\Delta_{S^c}\|_1 \le \|\Delta_S\|_1$$

Restricted nullspace illustration

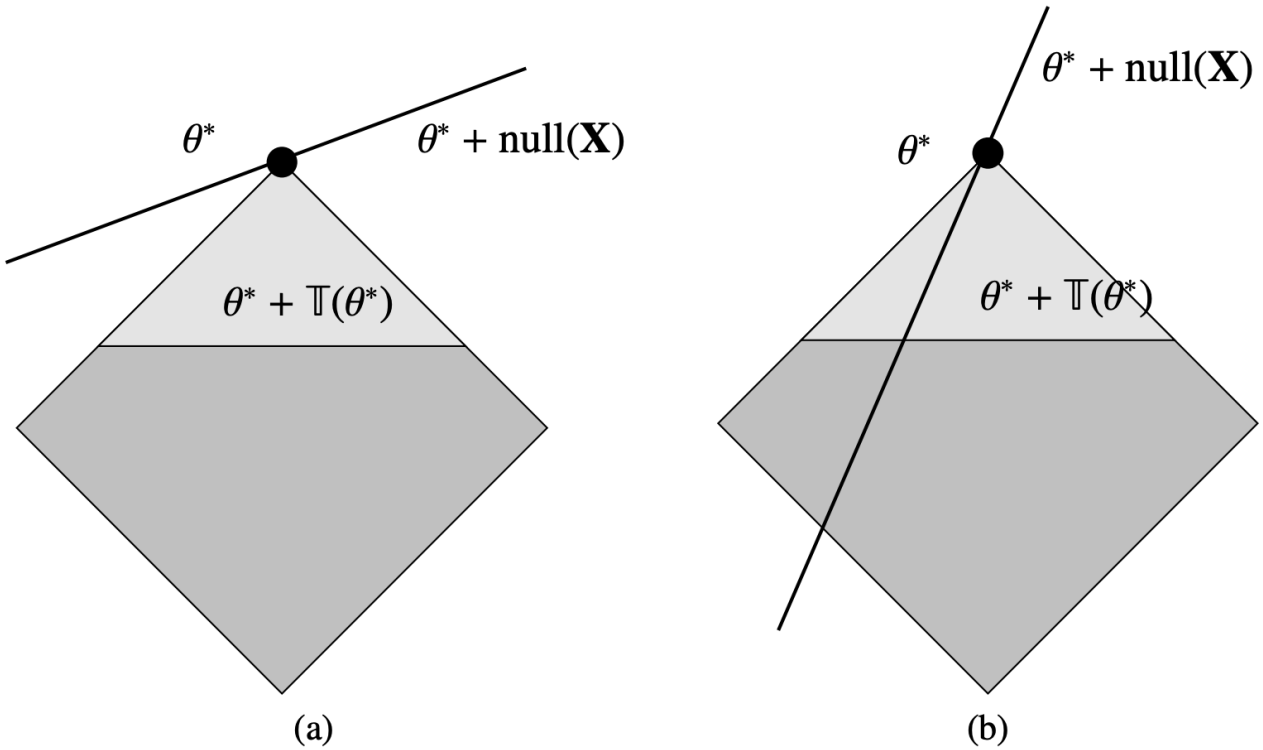


Figure 7.2 Geometry of the tangent cone and restricted nullspace property in $d = 2$ dimensions. (a) The favorable case in which the set $\theta^* + \text{null}(\mathbf{X})$ intersects the tangent cone only at θ^*. (b) The unfavorable setting in which the set $\theta^* + \text{null}(\mathbf{X})$ passes directly through the tangent cone.

Image taken from High-Dimensional Statistics by Martin Wainwright.

Restricted null space

Let:

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

Definition

Restricted nullspace The matrix X satisfies the restricted nullspace property with respect to S if $C(S) \cap \text{null}(X) = \{0\}$.

Unique solution - basis pursuit

Theorem

The following two properties are equivalent:

(a) *For any vector $\theta^* \in \mathbb{R}^d$ with support S, the basis pursuit program (7.9) applied with $y = X\theta^*$ has unique solution $\hat{\theta} = \theta^*$.*

(b) *The matrix X satisfies the restricted nullspace property with respect to S.*

b ⇒ a $\hat{\theta}$ and θ^* are solutions

(1) $\|\hat{\theta}\|_1 \le \|\theta^*\|_1$ define $\tilde{\Delta} := \hat{\theta} - \theta^*$

$$\|\theta^*_S\|_1 = \|\theta^*\|_1 \ge \|\theta^* + \tilde{\Delta}\|_1 = \|\theta^*_S + \tilde{\Delta}_S\|_1 + \|\tilde{\Delta}_{S^c}\|_1$$

$$\ge \|\theta^*\|_1 - \|\tilde{\Delta}_S\|_1 + \|\tilde{\Delta}_{S^c}\|_1$$

$$\|\tilde{\Delta}_S\| \ge \|\tilde{\Delta}_{S^c}\|_1$$

Proof $a \Rightarrow b$

$\forall S$ sparse pattern.

$\text{null}(X) \setminus \{0\} \cap C(S)$ is empty.

$\forall \theta \in \text{null}(X) \setminus \{0\}$ Consider $\min_{\beta \in \mathbb{R}^d} \|\beta\|_1 : X_\beta = X \begin{bmatrix} \theta_S^* \\ 0 \end{bmatrix}$

the unique solution is $\hat{\beta} = [\theta_S^* \; 0]$.

as $X\theta^* = 0 \Rightarrow [0 \; -\theta_{S^c}^*]^T$ is also a solution

$$X \begin{pmatrix} \theta_S^* \\ \theta_{S^c} \end{pmatrix} = 0 \quad \text{i.e.} \quad X \begin{pmatrix} \theta_S^* \\ 0 \end{pmatrix} = -X \begin{pmatrix} 0 \\ \theta_{S^c} \end{pmatrix}$$

but $[\theta_S^* \; 0]^T$ is the unique minimizer

$$\text{i.e.} \quad \|\theta_S^*\|_1 < \|\theta_{S^c}^*\|_1$$

$$\text{i.e.} \quad \theta^* \notin C(S)$$

Restricted isometry

$\hat{\Sigma}_n = \frac{X^T X}{n}$

Proposition

If

$$\left| \frac{\mathbb{X}^T \mathbb{X}}{n} - I_d \right|_\infty \leq \frac{1}{3k},$$

then the restrict null space condition holds for all subsets of cardinality at most k.

RIP

Definition

For an integer k, $X \in \mathbb{R}^{n \times d}$ satisfies a restricted isometry property of order k with constant $\delta_k(X) > 0$ if

$$\left\| \frac{X_S^\top X_S}{n} - I_k \right\|_{op} \leq \delta_k(X)$$

for all subsets of size at most k.

Proposition

If the RIP constant of order $2k$ is bounded as $\delta_{2k} < 1/3$, then the restricted null space condition holds for any subset S of cardinality $|S| \leq k$.

Assumptions

Assumption

The smallest eigenvalue of the sample covariance submatrix indexed by S is bounded below:

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

Assumption

There exists some $\alpha \in [0, 1)$ such that mutual incoherence

$$\max_{j \in S_c} ||(X_S^T X_S)^{-1} X_S^T X_j||_1 \leq \alpha.$$

want to predict X_j using X_S

$\forall j \in S^c$ $\quad \hat{w} = \operatorname{argmin} ||X_j - X_S w||_2^2 = (X_S^T X_S)^{-1} X_S^T X_j.$

LASSO

Theorem

Consider an S-sparse linear regression model for which the design matrix satisfies Assumptions 2 and 3. Then for any choice of regularization parameter such that

$$\lambda_n \geq \frac{2}{1-\alpha} \left\| X_S^T \Pi_{S^\perp}(\mathbf{X}) \frac{\epsilon}{n} \right\|_\infty, \tag{1}$$

the Lasso program 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.*

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

$$\|\hat{\theta}_S - \theta^*_S\|_\infty \leq \underbrace{\left\| \left(\frac{X_S^T X_S}{n} \right)^{-1} X_S^T \frac{\epsilon}{n} \right\|_\infty + \left\| \left(\frac{X_S^T X_S}{n} \right)^{-1} \right\|_\infty \lambda_n}_{B(\lambda_n, X)}, \tag{2}$$

where $\|A\|_\infty = \max_{i=1,\ldots,s} \sum_j |A_{i,j}|$ is the matrix ℓ_∞-norm.

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

Corollary

For a S-sparse linear model based on a noise vector ϵ with zero-mean i.i.d. σ-sub-Gaussian entries, and a deterministic design matrix X that satisfies Assumptions 2 and 3, as well as the C-column normalization condition $\max_{j=1,\ldots,d} \|\mathsf{X}_j\|_2/\sqrt{n} \leq C$. Suppose that we solve the Lasso program with regularization parameter

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

for some $\delta > 0$. Then the optimal solution $\widehat{\theta}$ is unique with its support contained within S, and satisfies the ℓ_∞-error bound

$$\|\widehat{\theta}_S - \theta^*_S\|_\infty \leq \frac{\sigma}{\sqrt{c_{\min}}}\left(\sqrt{\frac{2\log s}{n}} + \delta\right) + \left\|\left(\frac{\mathsf{X}_S^T\mathsf{X}_S}{n}\right)^{-1}\right\|_\infty \lambda_n, \qquad (3)$$

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

Proof

Sub-gradients

Primal Dual Witness

Definition

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*

$$\widehat{\theta}_S \in \arg\min_{\theta_S \in \mathbb{R}^s} \left\{ \underbrace{\frac{1}{2n}\|y - X_S\theta_S\|_2^2 + \lambda_n\|\theta_S\|_1}_{=:f(\theta_S)} \right\}, \tag{4}$$

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

Witness for LASSO

Lemma

If the lower eigenvalue condition 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

Proof of main theorem