High-dimensional Statistical Analysis

Pradeep Ravikumar
UT Austin

Outline

1. Lasso; High-dimensional Statistical Analysis
2. Structure Recovery: Sparsistency
3. Parameter Error Bounds

Recall: High-dimensional Statistical Analysis

Typical Statistical Consistency Analysis: Holding model size (p) fixed, as number of samples goes to infinity, estimated parameter $\widehat{\theta}$ approaches the true parameter θ^*.

Meaningless in finite sample cases where $p \gg n$!

Need a new breed of modern statistical analysis: both model size p **and** sample size n go to infinity!

Typical Statistical Guarantees of Interest for an estimate $\widehat{\theta}$:

- Structure Recovery e.g. is sparsity pattern of $\widehat{\theta}$ same as of θ^*?
- Parameter Bounds: $\|\widehat{\theta} - \theta^*\|$ (e.g. ℓ_2 error bounds)
- Risk (Loss) Bounds: difference in expected loss

Recall: Lasso

Estimator: Lasso program

$$\widehat{\theta} \in \arg\min_{\theta} \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\theta)^2 + \lambda_n \sum_{j=1}^{p} |\theta_j|$$

Some past work: Tibshirani, 1996; Chen et al., 1998; Donoho/Xuo, 2001; Tropp, 2004; Fuchs, 2004; Meinshausen/Buhlmann, 2005; Candes/Tao, 2005; Donoho, 2005; Haupt & Nowak, 2006; Zhao/Yu, 2006; Wainwright, 2006; Zou, 2006; Koltchinskii, 2007; Meinshausen/Yu, 2007; Tsybakov et al., 2008

Statistical Assumption: (x_i, y_i) from Linear Model:

$y_i = x_i^T\theta^* + w_i$, with $w_i \sim N(0, \sigma^2)$.

Sparsistency

Theorem. Suppose the design matrix X satisfies some conditions (to be specified later), and suppose we solve the Lasso problem with regularization penalty

$$\lambda_n > \frac{2}{\gamma}\sqrt{\frac{2\sigma^2 \log p}{n}}.$$

Then for some $c_1 > 0$, the following properties hold with probability at least $1 - 4\exp(-c_1 n\lambda_n^2) \to 1$:

- The Lasso problem has unique solution $\widehat{\theta}$ with support contained with the true support: $S(\widehat{\theta}) \subseteq S(\theta^*)$.
- If $\theta^*_{\min} = \min_{j \in S(\theta^*)} |\theta^*_j| > c_2 \lambda_n$ for some $c_2 > 0$, then $S(\widehat{\theta}) = S(\theta^*)$.

(Wainwright 2008; Zhao and Yu, 2006;...)

Sufficient Conditions: Dependency Bound

$$\lambda_{\mathsf{min}}\left(\frac{1}{n}X_S^T X_S\right) \geq C_{\mathsf{min}} > 0.$$

$$\lambda_{\mathsf{max}}\left(\frac{1}{n}X_S^T X_S\right) \leq D_{\mathsf{max}} < \infty.$$

Ensures that the relevant covariates are not "too dependent".

Sufficient Conditions: Incoherence

$\|X_{S^c}^T X_S (X_S^T X_S)^{-1}\|_\infty \leq 1 - \gamma,$

for some $\gamma > 0$.

Equivalent:

$\max_{j \in S^c} \|X_j^T X_S (X_S^T X_S)^{-1}\|_1 \leq 1 - \gamma.$

Weaker form of orthogonality:

LHS equal to zero if all columns are orthogonal (which is not possible when $p > n$).

Sufficient Conditions: Gaussian Design

Suppose X has $i.i.d$ rows, with $X_i \sim N(0, \Sigma)$. Then the sufficient conditions stated earlier are satisfied if:

- $\lambda_{\min}(\Sigma_{SS}) \geq C_{\min} > 0$.
  $\lambda_{\max}(\Sigma_{SS}) \leq D_{\max} < \infty$.

- $\|\|\Sigma_{S^c S}(\Sigma_{SS})^{-1}\|\|_\infty \leq 1 - \gamma$,
  for some $\gamma > 0$.

- Sample Scaling: $n > K s \log p$, for some $K > 0$.

Proof: One can show that under sample scaling, population conditons imply the sample conditions.

Proof of Sparsistency

Stationary Condition:

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

where $\widehat{z} \in \partial\|\widehat{\theta}\|_1$ is the sub-gradient of $\|\widehat{\theta}\|_1$.

Sub-gradient : equal to derivative when the function is differentiable; otherwise a set.

Definition: For any convex function g, its sub-gradient at a point x, denoted by $\partial g(x)$ is the set of all points z such that, for all $y \neq x$:

$g(y) - g(x) \geq z^T(y - x)$.

For ℓ_1 norm: $z \in \partial\| \ \theta\|_1$ if:

$z_j = \text{sign}(\theta_j)$, if $\theta_j \neq 0$,

$|z_j| \leq 1$, if $\theta_j = 0$.

Proof of Sparsistency

Stationary Condition:

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

where $\widehat{z} \in \partial\|\widehat{\theta}\|_1$ is the sub-gradient of $\|\widehat{\theta}\|_1$.

Have to show: $\widehat{\theta}_{S^c} = 0$!

Easier to show inequalities (can bound terms), than equalities! Way out: "Witness" proof technique.

We will explicitly *construct* a $(\tilde{\theta}, \tilde{z})$ which satisfy the stationary condition, and for which $\tilde{\theta}_{S^c} = 0$!

Catch: Have to show $\tilde{z} \in \partial\|\tilde{\theta}\|_1$ (which we will show holds with high-probability).

Proof of Sparsistency

Set $\tilde{\theta}$ as the solution of an "oracle" problem:

$\tilde{\theta} = \arg\min_{\{\theta:\, \theta_{S^c}=0\}} \left\{ \frac{1}{n}\|y - X\theta\|_2^2 + \lambda_n \|\theta\|_1 \right\}.$

Set $\tilde{z}_S = \partial\|\tilde{\theta}_S\|_1$.

Set $\tilde{z}_{S^c} = -\frac{1}{\lambda_n}\left\{\frac{1}{n}X_{S^c}^T(X_S\tilde{\theta}_S - y)\right\}$.

$(\tilde{\theta}, \tilde{z})$ satisfies stationary condition of original problem:

Stationary Condition of Oracle Problem:
$\frac{1}{n}X_S^T(X_S\tilde{\theta}_S - y) + \lambda_n \tilde{z}_S = 0$.

Construction: $\frac{1}{n}X_{S^c}^T(X_S\tilde{\theta}_S - y) + \lambda_n \tilde{z}_{S^c} = 0$.

Proof of Sparsistency

Remains to show that $\tilde{z} \in \partial\|\tilde{\theta}\|_1$!

Construction: $\tilde{z}_S \in \partial\|\tilde{\theta}_S\|_1$.

Have to show: $\tilde{z}_{S^c} \in \partial\|\tilde{\theta}_{S^c}\|_1$.

By construction: $\tilde{\theta}_{S^c} = 0$. So have to show: $|z_j| \leq 1$, for all $j \in S^c$.

Equivalently: $\|z_{S^c}\|_\infty \leq 1$.

Proof of Sparsistency

Notation: $\Delta = \tilde{\theta} - \theta^*$; $W = y - X\theta^*$.

Stationary Condition: $\frac{1}{n} X_S^T (X_S \tilde{\theta}_S - y) + \lambda_n \tilde{z}_S = 0$.

Rewritten: $\left(\frac{1}{n} X_S^T X_S\right) \Delta_S + \frac{1}{n} X_S^T W + \lambda \tilde{z}_S = 0$.

Hence: $\Delta_S = \left(\frac{1}{n} X_S^T X_S\right)^{-1} [-\lambda \tilde{z}_S - \frac{1}{n} X_S^T W]$.

Construction:

$$\begin{aligned}
\lambda_n \tilde{z}_{S^c} &= -\frac{1}{n} X_{S^c}^T X_S \Delta_S - \frac{1}{n} X_{S^c}^T W \\
&= \left(\frac{1}{n} X_{S^c}^T X_S\right) \left(\frac{1}{n} X_S^T X_S\right)^{-1} [-\lambda \tilde{z}_S - \frac{1}{n} X_S^T W] - \frac{1}{n} X_{S^c}^T W.
\end{aligned}$$

Let $c_n = \|X^T W\|_\infty$. Recall: $|||X_{S^c}^T X_S (X_S^T X_S)^{-1}|||_\infty \leq 1 - \gamma$.

Then $\lambda_n \|\tilde{z}_{S^c}\|_\infty \leq (1-\gamma)(\lambda_n + c_n) + c_n \leq (2-\gamma)c_n + (1-\gamma)\lambda_n < \lambda_n$,

provided $c_n < \gamma/(2-\gamma)\lambda_n$: we show this holds *with high probability*.

Whence: $\|\tilde{z}_{S^c}\|_\infty < 1$, as required, *with high probability*.

Proof of Sparsistency

Gaussian Tail Bounds: If $W_i \sim N(0, \sigma^2)$, then:

$$\mathbb{P}[|X_j^T W| > \alpha] \leq \exp(-cn\alpha^2).$$

Then, by an application of the union bound:

$$\mathbb{P}[\sup_{j=1}^{p} |X_j^T W| > \alpha] \leq p \exp(-cn\alpha^2) = \exp(-cn\alpha^2 + \log p).$$

Thus, for $\lambda_n = c_1 \sqrt{\frac{\log p}{n}}$,

$\|X^T W\|_\infty = \sup_{j=1}^{p} |X_j^T W| \leq c_2 \lambda_n$ with probability at least $1 - \exp(-c' n \lambda_n^2)$.

Parameter Error Bounds

Restricted Eigenvalue:

Let $\mathcal{C} = \{\Delta : \|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1\}$:

Then for all $\Delta \in \mathcal{C}$:

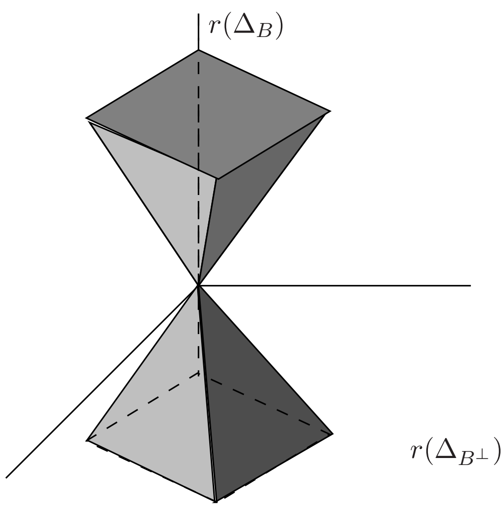
$\|X\Delta\|_2^2 \geq \kappa\|\Delta\|_2^2$, for some $\kappa > 0$.

Theorem: Suppose the design matrix X satisfies the restricted eigenvalue condition. Then the Lasso solution $\widehat{\theta}$ satisfies:

$\|\widehat{\theta} - \theta^*\|_2 \leq c\sqrt{\frac{s \log p}{n}}$.

Parameter Error Bounds

Lemma: The solution to the Lasso problem $\widehat{\theta} = \theta^* + \Delta$ satisfies the following cone condition: $\|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1$.



Here: $S = 3; S^c = 1, 2$.

Parameter Error Bounds

Let $L(\theta) = \frac{1}{n}\|X\theta - y\|_2^2$.

Then, by optimality of Lasso solution $\widehat{\theta}$:

$L(\widehat{\theta}) + \lambda\|\widehat{\theta}\|_1 \leq L(\theta^*) + \lambda\|\theta^*\|_1$.

Convexity:

$L(\widehat{\theta}) \geq L(\theta^*) + \nabla L(\theta^*) \cdot \Delta \geq L(\theta^*) - \|\nabla L(\theta^*)\|_\infty \|\Delta\|_1$.

If we set $\lambda \geq 2\|\nabla L(\theta^*)\|_\infty = 2\|X^T W\|_\infty$, then:

$-\frac{\lambda}{2}\|\Delta\|_1 + \lambda\|\widehat{\theta}\|_1 \leq \lambda\|\theta^*\|_1$.

Noting that

$$\begin{aligned}
\|\widehat{\theta}\|_1 = \|\theta^* + \Delta\|_1 &= \|\theta^*_S + \Delta_S + \Delta_{S^c}\|_1 \\
&= \|\Delta_{S^c}\|_1 + \|\theta^*_S + \Delta_S\|_1 \\
&\geq \|\Delta_{S^c}\|_1 + \|\theta^*_S\|_1 - \|\Delta_S\|_1,
\end{aligned}$$

and rearranging terms, we get:

$\|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1$.

Parameter Error Bounds

Again, by optimality of Lasso solution $\widehat{\theta}$:

$L(\widehat{\theta}) + \lambda\|\widehat{\theta}\|_1 \leq L(\theta^*) + \lambda\|\theta^*\|_1$.

Suppose, over the restricted set $\{\Delta : \|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1\}$:

$L(\theta) \geq L(\theta^*) + \nabla L(\theta^*) \cdot \Delta + \kappa\|\Delta\|_2^2$.

Then, by re-arranging terms as earlier, we get:

$\kappa\|\Delta\|_2^2 \leq 3\lambda\|\Delta_S\|_1 \leq 3\sqrt{s}\|\Delta_S\|_2 \leq 3\lambda\sqrt{s}\|\Delta\|_2$.

Hence:

$\|\Delta\|_2 \leq \frac{3}{\kappa}\lambda\sqrt{s}$.

Thus, $\|\Delta\|_2 \leq c\sqrt{\frac{s \log p}{n}}$.