Spring 2018: STA 6448
Advanced Probability and Inference II
Lecture 23

Yun Yang

- High-dimensional linear regression
- Non-parametric least squares

Variable selection consistency for the Lasso

Assume the design matrix X to be deterministic.

Conditions:

(A3) Lower eigenvalue:

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

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

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

Variable selection consistency for the Lasso

Let $\Pi_{S^\perp} = I_n - X_S(X_S^T X_S)^{-1} X_S^T$ denote an orthogonal projection matrix.

Theorem

Under conditions (A3) and (A4), if $\lambda_n \geq \frac{2}{1-\alpha} \| X_{S^c}^T \Pi_{S^\perp} \frac{w}{n} \|_\infty$, then

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

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

(c) *ℓ_∞-bounds:*

$$\|\widehat{\theta}_S - \theta^*_S\|_\infty \leq \underbrace{\left\| \left(\frac{X_S^T X_S}{n} \right)^{-1} X_S^T \frac{w}{n} \right\|_\infty + \left|\!\left|\!\left| \left(\frac{X_S^T X_S}{n} \right)^{-1} \right|\!\right|\!\right|_\infty \lambda_n}_{B(\lambda_n; X)} .$$

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

Variable selection consistency for the Lasso

Corollary

Suppose the noise vector w has zero-mean i.i.d. σ-sub-Gaussian entries, and X satisfies (A3) and (A4), and is C-column normalized. If for some $\delta > 0$,

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

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

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

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

Non-parametric least squares: Problem setup

- Problem: use observations of predictors or covariates $x \in \mathcal{X}$ in to predict a response variable $y \in \mathcal{Y}$
- Goal: estimate a function $f : \mathcal{X} \mapsto \mathcal{Y}$ such that the error $y - f(x)$ is as small as possible over some range of pairs (x, y).
- In the random design scenario, both the response and covariate are random quantities. We measure the quality of f in terms of its mean-squared error (MSE)

$$\bar{\mathcal{L}}_f := \mathbb{E}_{X,Y}\left[(Y - f(X))^2\right].$$

- The function f^* minimizing this criterion is known as the *Bayes' least-squares estimate* or the *regression function*, and it is given by the conditional expectation

$$f^*(x) = \mathbb{E}[Y \mid X = x].$$

Non-parametric least squares: Problem setup

- In practice, the expectation defining the MSE cannot be computed, since the distribution over (X,Y) is not known.
- Instead, we are given a collection of samples $\{(x_i,\, y_i)\}_{i=1}^n$, which can be used to compute an empirical analogue of the mean-squared error

$$\widehat{\mathcal{L}}_f := \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2.$$

- The method of non-parametric least squares is based on minimizing this least-squares criterion over some suitably controlled function class.

Different measures of prediction quality

- Given an estimate f of the regression function, it is natural to measure its quality in terms of the excess risk

$$\bar{\mathcal{L}}_f - \bar{\mathcal{L}}_{f^*} = \mathbb{E}_X\left[(f(X) - f^*(X))^2\right] = \|f - f^*\|^2_{L^2(\mathbb{P})}.$$

  where $\mathbb{P}$ denotes the distribution over the covariates. We will adopt the shorthand notation $\|f - f^*\|_2^2$.

- We will measure the error using a closely related but slightly different measure by replacing $\mathbb{P}$ with the empirical distribution $\mathbb{P}_n = n^{-1}\sum_{i=1}^n \delta_{x_i}$,

$$\|f - f^*\|_{L^2(\mathcal{P}_n)} := \left[\frac{1}{n}\sum_{i=1}^n \left(f(x_i) - f^*(x_i)\right)^2\right].$$

  We will adopt the shorthand notation $\|f - f^*\|_n$.