IDS.160 – Mathematical Statistics: A Non-Asymptotic Approach

Lecturer: P Rigollet
Scribe: Chin-Chia Hsu abd Guang-He Lee

Lecture 8
Mar. 3, 2020

Goals: In the last lecture we study the linear regression when the parameter is sparse, considering the sparsity either known or unknown. In this lecture, a variational representation is introduced and can be generalized from HRD estimator to BIC and Lasso estimators. Later we investigate the case when the misspecified linear model in which the regression function is not in the linear form. Finally we move on to Chapter 3– matrix estimation.

1. LINEAR REGRESSION: VARIATIONAL FORM AND GENERALIZATION

One can represent the hard thresholding (HRD) estimator as a minimizer to the following variational formulation:

$$\hat{\theta}^{\mathrm{HRD}} \in \operatorname*{argmin}_{\theta\in\mathbb{R}^d} \left\{ |Y-\theta|_2^2 + 4\tau^2|\theta|_0 \right\} \tag{1.1}$$

or equivalently,

$$\hat{\theta}^{\mathrm{HRD}} \in \operatorname*{argmin}_{\theta\in\mathbb{R}^d} \sum_{i=1}^{d} \left\{ (Y_i-\theta_i)^2 + 4\tau^2 \mathbb{1}(\theta_i \neq 0) \right\} \tag{1.2}$$

To verify (1.2), first given an index i we have

$$(Y_i-\hat{\theta}_i^{\mathrm{HRD}})^2 + 4\tau^2\mathbb{1}(\hat{\theta}_i^{\mathrm{HRD}}\neq 0) = \begin{cases} Y_i^2 & \text{, if } Y_i^2 \le 4\tau^2 (\hat{\theta}_i^{\mathrm{HRD}} = 0) \\ 4\tau^2 & \text{, if } Y_i^2 > 4\tau^2 \end{cases} = \min(Y_i^2, 4\tau^2)$$

Moreover, for any $\theta \in \mathbb{R}^d$,

$$(Y_i-\theta_i)^2 + 4\tau^2\mathbb{1}(\theta_i\neq 0) = \begin{cases} Y_i^2 & \text{, if } \theta_i = 0 \\ (Y_i-\theta_i)^2 + 4\tau^2 & \text{, if } \theta_i \neq 0 \end{cases} \ge \min(Y_i^2, 4\tau^2)$$

which shows that $\hat{\theta}^{\mathrm{HRD}}$ minimizes (1.2).

The formulation (1.1) can be generalized to any design matrix. Here we is the example of BIC estimator.

$$\hat{\theta}^{\mathrm{BIC}} \in \operatorname*{argmin}_{\theta\in\mathbb{R}^d} \left\{ |Y-\mathbb{X}\theta|_2^2 + 4\tau^2|\theta|_0 \right\} \tag{1.3}$$

The rate is the same as hard thresholding linear estimator or ℓ_0-constrained estimator. However, it is NP-hard to compute the BIC estimator in the worst case. In particular, one needs to use the brute force and search among all 2^d sparsity patterns.

By contrast, we can change the problem and replace the ℓ_0-norm by ℓ_1 norm to make it a convex optimization problem.

$$\hat{\theta}^{\mathcal{L}} \in \operatorname*{argmin}_{\theta\in\mathbb{R}^d} \left\{ |Y-\mathbb{X}\theta|_2^2 + 4\tau|\theta|_1 \right\} \tag{1.4}$$

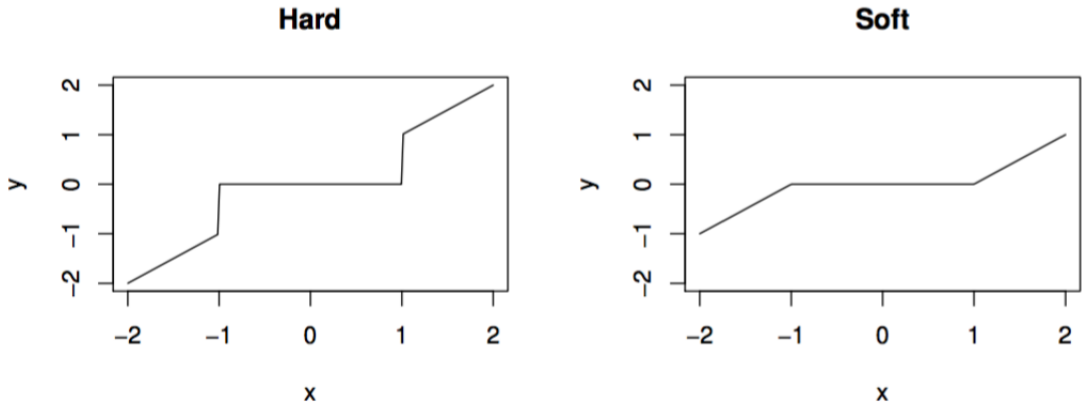

Figure 1: Transformation applied to Y_j with $2\tau = 1$ to obtain the hard (left) and soft (right) thresholding estimators [RH].

It is the Lasso estimator. There exist many efficient algorithms to solve it fast, even on large scale (coordinate decent is one of the popular ways to solve it). We derive "almost" the same rate but pay a cost: we need to assume that $\frac{\mathbb{X}^\top \mathbb{X}}{n} \approx I_d$. Recall that in Pset 1 we define "incoherence" as one measure on the closeness between two matrices. There are many ways for two matrices to be close. For more details, see notes [RH].

What if $\frac{\mathbb{X}^\top \mathbb{X}}{n} = I_d$ in (1.4), does the lead to an intuitive estimator? If $\mathbb{X} = I_d$, in this case $\hat{\theta}^{\mathcal{L}} = \hat{\theta}^{\text{SFT}}$, which is the "soft thresholding estimator," defined as

$$\hat{\theta}_j^{\text{SFT}} = (1 - \frac{2\tau}{|Y_i|})_+ Y_j, \forall j \tag{1.5}$$

Figure 1 illustrates how the soft thresholding function makes the hard thresholding function continuous at $x = |2\tau|$, softening the sharp transition. Basically, this soft thresholding function has the same property. Since we are looking for finite type of results in this course, no constants actually matter, and from this perspective $\hat{\theta}^{\mathcal{L}}$ and $\hat{\theta}^{\text{SFT}}$ are the same estimators. One can check that $\hat{\theta}^{\text{SFT}}$ is indeed the solution to (1.4) by writing the first order condition and using sub-gradient (ℓ_1-norm is not differentiable).

2. MISSPECIFIED LINEAR MODEL

We start from the regression model

$$Y_i = f(X_i) + \varepsilon_i, \quad i = 1, ...n$$

and so far we make the assumption that $f(x) = x^\top \theta^*$ for some $\theta^* \in \mathbb{R}$. What if this assumption is violated but in an approximate way $f(x) \approx x^\top \theta^*$? The technique to solve this scenario is different from what we did in linear regression since $|Y - \mathbb{X}\theta^*|_2$ is no longer equal to ε in the basic inequality $|Y - \mathbb{X}\hat{\theta}|_2 \leq |Y - \mathbb{X}\theta^*|_2$.

First we denote that $Y = \mu + \varepsilon$ and $\mu \approx \mathbb{X}\theta^*$ in the sense that

$$\frac{1}{n}|\mu - \mathbb{X}\theta^*|_2^2 \tag{2.6}$$

is small for some θ^*. How small? We will make this quantity appear in our bound: If it is small, the bound is good. Then denote $K \subset \mathbb{R}^d$ ($K = \mathbb{R}^d, K = B_1$ are two cases to which we paid lots of attention). Formulate the problem

$$\hat{\theta} \in \underset{\theta \in K}{\operatorname{argmin}} |Y - \mathbb{X}\theta|_2^2 \tag{2.7}$$

We are still searching for a solution in the column space of $\mathbb{X}$. What we can do is to compete with the best estimate. The best estimate of μ of this column space is the projection of vector μ on this span. We find the parameters over the set K

$$\theta^* \in \underset{\theta \in K}{\operatorname{argmin}} |\mu - \mathbb{X}\theta|_2^2 \tag{2.8}$$

which are sometimes called the Oracle. It is something that one cannot compute and only ORACLE can. Oracle tells us something about the truth not in a perfect manner: it knows μ but can only answer the closest object to μ in the column space. We are competing with the Oracle, that is, we want $\hat{\theta}$ to achieve as good as θ^* in terms of mean squared error.

Beginning from the basic inequality

$$|Y - \mathbb{X}\hat{\theta}|_2 \leq |Y - \mathbb{X}\theta|_2, \quad \forall \theta \in K. \tag{2.9}$$

$$\Rightarrow |\mu + \varepsilon - \mathbb{X}\hat{\theta}|_2 \leq |\mu + \varepsilon - \mathbb{X}\theta^*|_2 \tag{2.10}$$

$$\Rightarrow |\mu - \mathbb{X}\hat{\theta}|_2^2 + 2\langle \mu - \mathbb{X}\hat{\theta}, \varepsilon \rangle + |\varepsilon|_2^2 \leq |\mu - \mathbb{X}\theta^*|_2^2 + 2\langle \mu - \mathbb{X}\theta^*, \varepsilon \rangle + |\varepsilon|_2^2 \tag{2.11}$$

$$\Rightarrow |\mu - \mathbb{X}\hat{\theta}|_2^2 - |\mu - \mathbb{X}\theta^*|_2^2 \leq 2\langle \mathbb{X}\hat{\theta} - \mathbb{X}\theta^*, \varepsilon \rangle \tag{2.12}$$

$$\Rightarrow |\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|_2^2 \leq 2\langle \mathbb{X}\hat{\theta} - \mathbb{X}\theta^*, \varepsilon \rangle \tag{2.13}$$

The last derivation comes from that if the projection of μ can be represented by some $\theta^* \in K$ and Pythagoras theorem. This is the same formula as we have seen before despite the different meanings. We can still apply our tricks in previous lectures. For example, as for least square estimator $\hat{\theta}^{\mathrm{LS}}$,

$$\frac{1}{n} |\mathbb{X}\hat{\theta}^{\mathrm{LS}} - \mathbb{X}\theta^*|_2^2 \lesssim \sigma^2 \frac{\operatorname{rank}(\mathbb{X}^\top \mathbb{X})}{n} \tag{2.14}$$

We add μ and subtract it on the left hand side of (2.14) and expand to obtain

$$\mathbb{E}[\mathsf{MSE}(\mathbb{X}\hat{\theta})] \leq \mathsf{MSE}(\mathbb{X}\theta^*) + C\sigma^2 \frac{\operatorname{rank}(\mathbb{X}^\top \mathbb{X})}{n} \text{ for some constant } C \tag{2.15}$$

$$= \inf_{\theta \in \mathbb{R}^d} \mathsf{MSE}(\mathbb{X}\theta) + C\sigma^2 \frac{\operatorname{rank}(\mathbb{X}^\top \mathbb{X})}{n} \tag{2.16}$$

and we call it Oracle inequality. The term $\inf_{\theta \in \mathbb{R}^d} \mathsf{MSE}(\mathbb{X}\theta)$ is the misspecified error that will goes away if linear model is used. From a statistical perspective, Oracle inequalities are used as devices to guarantee some adaptations to such as smoothness or sparsity. Later when we touch the topics about machine learning, we will see many inequalities that look like the Oracle inequality to bound the risk in a hypothesis class.

Let's consider $K = B_1; |\mathbb{X}_j|_2 \leq \sqrt{n}$. Pythagoras theorem is not valid here. However, with the particular structure of B_1, using Hölder's inequality,

$$|\mathbb{X}\hat{\theta} - \mu|_2^2 \leq |\mathbb{X}\theta^* - \mu|_2^2 + 2\langle \mathbb{X}^\top \varepsilon, \hat{\theta} - \theta^* \rangle \tag{2.17}$$

$$\leq |\mathbb{X}\theta^* - \mu|_2^2 + 2|\mathbb{X}^\top \varepsilon|_\infty |\hat{\theta} - \theta^*|_1 \tag{2.18}$$

where $|\hat{\theta} - \theta^*| \leq 2$ and $|\mathbb{X}^\top \varepsilon|_\infty \lesssim \sigma\sqrt{n \log d}$. Therefore we obtain

$$\mathbb{E}[\mathsf{MSE}(\mathbb{X}\hat{\theta})] \leq \inf_{\theta \in B_1} \mathsf{MSE}(\mathbb{X}\theta) + C\sigma\sqrt{\frac{\log d}{n}} \tag{2.19}$$

3. MATRIX ESTIMATION: BASICS

Let's first remind ourselves of the (sub)Gaussian sequence model:

$$Y = \theta^* + \varepsilon \in \mathbb{R}^d. \tag{3.20}$$

We can always reorganize these vectors to matrices. For example, we can divide the vector into 3 chunks, put the 3 chunks as 3 columns in a matrix, and then we have a matrix estimation problem.

$$Y = \theta^* + \varepsilon \in \mathbb{R}^{(d/3)\times 3}. \tag{3.21}$$

Honestly, if we don't impose any structure on the data other than being a matrix, then the estimation problem is exactly the same as the original (sub)Gaussian sequence model. For example, if θ^* is sparse, we can use $\hat{\theta}^{\mathrm{HRD}}$ as the estimator: we look at the matrix, keep the entries with high magnitudes and kill the entries with low magnitudes. Then we immediately obtain a matrix estimator, assuming that it is sparse. Essentially, even though we have a matrix here, the estimator only concerns the sparsity without considering other properties of the matrix. We can also impose some nice structures on the matrix. For example, we will talk about covariance matrix estimation. We could assume that these are covariances of things that observed in different points in time. Therefore, as things are spread in time, it is natural to assume that the covariance matrix is concentrated on the diagonal. The quintessential low dimensional structure that we can impose on matrix is governed by the rank of the matrix. We could ask, for example, what is the rate of estimation for a low-rank matrix with additive noises.

The matrix estimation problem is motivated by Netflix prize (2006-2011)[1], a 1 million grand prize for estimating a matrix for Netflix. Concretely, the researchers are given a sparse matrix, where the rows correspond to users and the columns correspond to movies. The matrix M contains sparse rating observations $M_{i,j} \in \{1, 2, 3, 4, 5\}$ for the movie j from the user i. There are at least two characteristics of such matrix. First, the observations are "noisy", as an integer value clearly does not well-calibrate the rating in the user's mind. Second, lots of the entries are missing, since each user only watch and rate a small portion of entries. The second problem is termed as deletion noise by the lecturer.

The simplest thing that we can do is to assume each rating $M_{i,j}$ can be

$$M_{i,j} = u_i v_j, \tag{3.22}$$

where $u_i, v_j \in \mathbb{R}$. This entails that the resulting estimation Θ^* is a rank 1 matrix:

$$\Theta^* = uv^\top. \tag{3.23}$$

We can generalize the rank 1 matrix to a rank r matrix as:

$$\Theta^* = \sum_{j=1}^{r} a_j u_j v_j^\top, \tag{3.24}$$

[1] https://www.netflixprize.com

where a_j is a scalar, and u_j, v_j are vectors. To ensure that everyone is on the same page, below we review some basic facts about matrices.

- Eigenvalue and eigenvectors for a square matrix A satisfy the folllowing equation:

$$Au = \lambda u, \tag{3.25}$$

  where u is an eigenvector and λ is the corresponding eigenvalue. If $A = A^\top$, then we have n real eigenvalues $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$. In this course, we always assume that the eigenvectors have unit ℓ_2 norm $|u|_2 = 1$. Moreover the eigenvectors of A for an orthonormal basis of the linear span of the columns of A.

- Singular value decomposition (SVD): given $A \in \mathbb{R}^{m\times n}$, it can be factorized by SVD as:

$$A = UDV^\top, U \in \mathbb{R}^{m\times r}, D \in \mathbb{R}^{r\times r}, V \in \mathbb{R}^{n\times r} \tag{3.26}$$

  where r is the rank of A, $U^\top U = I_r, V^\top V = I_r$, and D is a diagonal matrix with singular values. The SVD can also be written in vector from:

$$A = \sum_{j=1}^{r} \lambda_j u_j v_j^\top, \lambda_j \in \mathbb{R}, u_j \in \mathbb{R}^m, v_j \in \mathbb{R}^n. \tag{3.27}$$

  The vector form can be extended to the largest possible rank $\min(m, n)$ by letting $\lambda_j = 0, \forall j > r$ as:

$$A = \sum_{j=1}^{\min(m,n)} \lambda_j u_j v_j^\top, \lambda_j \in \mathbb{R}, u_j \in \mathbb{R}^m, v_j \in \mathbb{R}^n. \tag{3.28}$$

  Note that we have

$$AA^\top u_j = \lambda_j^2 u_j, A^\top A v_j = \lambda_j^2 v_j. \tag{3.29}$$

  If A is positive semi-definite (PSD; $A = A^\top$ and $u^\top Au \geq 0, \forall u \in \mathbb{R}^n$), the eigenvalues are equal to the singular values.

  We use the largest singular value to define the matrix operator norm:

$$\|A\|_{\text{op}} = \lambda_{\max}(A) = \max_{x\in\mathbb{R}^n : x\neq 0} \frac{|Ax|_2}{|x|_2} = \max_{y\in B_2(\mathbb{R}^m), x\in B_2(\mathbb{R}^n)} y^\top Ax. \tag{3.30}$$

  This is called the operator norm as A is a linear operator from $(\mathbb{R}^n, |\cdot|_2)$ to $(\mathbb{R}^m, |\cdot|_2)$.

  If A is PSD,

$$\|A\|_{\text{op}} = \lambda_{\max}(A) = \max_{x\in B_2(\mathbb{R}^m)} x^\top Ax. \tag{3.31}$$

- vector norms and inner product. Let $A = a_{ij}, B = b_{ij}$.
  - $|A|_q = (\sum_{i,j} |a_{ij}|^q)^{1/q}, q > 0$.
  - $|A|_\infty = \max_{i,j} |a_{ij}|$.

- $|A|_0 = \sum_{i,j} \mathbf{1}(a_{ij} \neq 0)$.
- (Frobenius norm) $\|A\|_F = |A|_2 = \sqrt{\sum_{i,j} a_{ij}^2} = \sqrt{\mathrm{Tr}(A^\top A)}$.
- (inner product) $\langle A, B\rangle = \mathrm{Tr}(A^\top B) = \mathrm{Tr}(AB^T)$.

- Spectral norms. Let $\lambda = (\lambda_1, \ldots, \lambda_r)$ be singular values of A.
  - (Schatten q-norm) $\|A\|_q = |\lambda|_q$.
  - If $q = 2$, $\|A\|_2^2 = \|A\|_F^2 = \mathrm{Tr}(A^\top A) = \mathrm{Tr}(VDU^\top UDV^\top) = \mathrm{Tr}(D^2) = \sum_{j=1}^r \lambda_j^2$.
  - If $q = 1$, $\|A\|_1 = \|A\|_*$ (called nuclear norm or trace norm).
  - If $q = \infty$, $\|A\|_\infty = \lambda_{\max}(A) = \|A\|_{\mathrm{op}}$

- Useful matrix inequalities. Let $A, B \in \mathbb{R}^{n\times m}, n \leq m$. Let $\lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_n(A) \geq 0$ denote the singular values of A, and $\lambda_1(B) \geq \lambda_2(B) \geq \cdots \geq \lambda_n(B) \geq 0$ the singular values of B.
  - (Weyl' 12): $\max_j |\lambda_j(A) - \lambda_j(B)| \leq \|A - B\|_{\mathrm{op}}$.
  - (Hoffman-Wielandt '53): $\sum_k |\lambda_k(A) - \lambda_k(B)|^2 \leq \|A - B\|_F^2$.
  - (Hölder): $\langle A, B\rangle \leq \|A\|_p \|A\|_q$, where $p, q \geq 1$ and $1/p + 1/q = 1$. Note that we also have the vector version $\langle A, B\rangle \leq |A|_p |A|_q$.
  - Lemma (Eckart–Young). Given $A = \sum_{j=1}^r \lambda_j u_j v_j^\top, \forall k < r$, we let $A_k = \sum_{j=1}^k \lambda_j u_j v_j^\top$ be the truncated SVD. Then, $\|A - A_k\|_F^2 = \inf_{B:\mathrm{rank}(B)\leq k} \|A - B\|_F^2 = \sum_{j=k+1}^r \lambda_j^2$.

    Proof.

$$\|A - A_k\|_F^2 = \|\sum_{j=k+1}^r \lambda_j u_j v_j^\top\|_F^2 = \sum_{i,j=k+1}^r \lambda_i \lambda_j \mathrm{Tr}(u_i v_i^\top v_j u_j^\top) \tag{3.32}$$

$$= \sum_{i,j=k+1}^r \lambda_i \lambda_j \mathrm{Tr}(v_i^\top v_j u_j^\top u_i) = \sum_{i=1}^r \lambda_i^2, \tag{3.33}$$

    where the last equality is due to $u_j^\top u_i = v_i^\top v_j = \delta_{ij}$ Now we take B with singular values $\sigma_1 \geq \cdots \geq \sigma_k \geq 0 \geq \cdots \geq 0$.

$$\|A - B\|_F^2 \geq \sum_{i=1}^r (\lambda_j - \sigma_j)^2 = \sum_{i=1}^k (\lambda_j - \sigma_j)^2 + \sum_{i=1}^k \lambda_j^2, \tag{3.34}$$

    where the inequality is due to Hoffman-Wielandt. □

Summary:

- Linear regression
  1. Hard thresholding and variational form

$$\hat{\theta}^{\mathrm{HRD}} \in \operatorname*{argmin}_{\theta\in\mathbb{R}^d} \left\{|Y - \theta|_2^2 + 4\tau^2 |\theta|_0\right\}$$

2. BIC estimator

$$\hat{\theta}^{\mathrm{BIC}} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left\{ |Y - \mathbb{X}\theta|_2^2 + 4\tau^2 |\theta|_0 \right\}$$

3. Lasso and soft thresholding estimator

$$\hat{\theta}^{\mathcal{L}} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left\{ |Y - \mathbb{X}\theta|_2^2 + 4\tau |\theta|_1 \right\}$$

If $\mathbb{X} = I_d$, $\hat{\theta}^{\mathcal{L}} = \hat{\theta}^{\mathrm{SFT}}$ where

$$\hat{\theta}_j^{\mathrm{SFT}} = (1 - \frac{2\tau}{|Y_i|})_+ Y_j, \forall j$$

- Misspecified linear model and oracle inequality: When $\mu \approx \mathbb{X}\theta^*$ for some $\theta* \in K$, we derived the oracle inequality for expected MSE of two estimators

1. $K = \mathbb{R}^d$:

$$\mathbb{E}[\mathsf{MSE}(\mathbb{X}\hat{\theta})] \leq \inf_{\theta \in \mathbb{R}^d} \mathsf{MSE}(\mathbb{X}\theta) + C\sigma^2 \frac{\mathrm{rank}(\mathbb{X}^\top \mathbb{X})}{n}$$

2. $K = B_1; |\mathbb{X}_j|_2 \leq \sqrt{n}$:

$$\mathbb{E}[\mathsf{MSE}(\mathbb{X}\hat{\theta})] \leq \inf_{\theta \in B_1} \mathsf{MSE}(\mathbb{X}\theta) + C\sigma \sqrt{\frac{\log d}{n}}$$

- Matrix basics:

1. Eigenvalues and eigenvectors: $Au = \lambda u$.
2. SVD: $A = UDV^\top$, where $D = \mathrm{diag}(\lambda)$.
3. Operator norm: $\|A\|_{\mathrm{op}} = \lambda_{\max(A)}$.
4. Vector norms $|A|_q = (\sum_{i,j} |a_{ij}^q|)^{1/q}, q > 0$.
5. Spectral norms: $\|A\|_q = |\lambda|_q$.
6. (Weyl' 12): $\max_j |\lambda_j(A) - \lambda_j(B)| \leq \|A - B\|_{\mathrm{op}}$.
7. (Hoffman-Wielandt '53): $\sum_k |\lambda_k(A) - \lambda_k(B)|^2 \leq \|A - B\|_F^2$.
8. (Hölder): $\langle A, B \rangle \leq \|A\|_p \|A\|_q$, where $p, q \geq 1$ and $1/p + 1/q = 1$.
9. Truncated SVD is the best rank k approximation in Frobenius norm.