Mathematical Introduction for Data Analysis **Feb. 20, 2017**

Lecture 3. MLE, Stein's Phenomena and Nonlinear Estimators

Instructor: Yuan Yao, HKUST *Scribe: Xia, Jiacheng; with aid by Dong, Chenyang*

1 Review of risk

Recall that in the previous lectures, we discussed MLE estimator: if $X_i \sim N(\mu, \Sigma)$ (i.i.d.), for $i = 1, 2, ..., n, X_i \in \mathbb{R}^p$, then MLE estimator for mean and variance is:

$$\hat{\mu}_n^{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i, \tag{1}$$

$$\hat{\Sigma}_n^{MLE} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})(X_i - \hat{\mu})^T. \tag{2}$$

With the definition of risk to estimate $\hat{\mu}$:

$$R(\mu) \equiv \mathbb{E}\|\hat{\mu} - \mu\|^2 \tag{3}$$

Although the definition of the risk includes μ and hence made it hard to calculate, this definition can be shown to be in some way equivalent to *prediction error*.
Let the prediction error be $\mathbb{E}_{(y|x_i^n)}\|\hat{\mu} - y\|$, for simplicity $y = \mu + \epsilon$, where $\epsilon \sim N(0, \epsilon)$.

$$\mathbb{E}_{x_i}(\mathbb{E}_{y|x_i}\|(\hat{\mu} - \mu) - \epsilon\|^2) = \mathbb{E}_{x_i}(\mathbb{E}_{\epsilon|x_i}\|(\hat{\mu} - \mu)\|^2 - 2\mathbb{E}_{\epsilon|x_i} < \epsilon, \hat{\mu} - \mu > + \mathbb{E}(\|\epsilon\|)^2) = \mathbb{E}_{(x_i)_1^n}\|\hat{\mu} - \mu\|^2 + Var(\epsilon)$$

The first term of solution is just Risk by definition, and second being constant independent with μ. So minimizing risk R is equivalent to minimizing prediction error.
In the following sections we would discuss the risk of several estimators.

2 MLE estimator

MLE is an unbiased estimator, but not of least risk. In the following part of this note, we would define $P \equiv R(\hat{\mu}^{MLE})$, $Y \equiv \hat{\mu}^{MLE}$. In the previous lecture we have shown that $\hat{\mu}^{MLE}$ is unbiased. i.e. $\mathbb{E}[\hat{\mu}] = \mu$.
If we further assume k-sparse, or energy is concentrating on top k components, $k \equiv \#\{i|\mu_i \neq 0\} \ll p$. Further assume:

$$\sum_{i=1}^{p} \mu_i \equiv \|\mu\| = ck$$

for some constant c. Then

$$R(\hat{\mu}_{MLE}) = \frac{p}{1 + \frac{p}{ck}} = \frac{kp}{k + \frac{p}{c}} \tag{4}$$

The risk would be smaller than p if p is large.

3 Linear estimator

For some constant c,

$$\mu_c \equiv c \cdot Y, C \equiv diag(c_i)$$

for $i \in [1, p]$, and $C \in \mathbb{R}^{p \times p}$. We can use L_1 regularization:

$$\arg\min_{\hat{\mu}} \frac{1}{2}\|\hat{\mu} - Y\| + \lambda\|\hat{\mu}\| \Rightarrow \mu_\lambda = \frac{1}{1+\lambda}Y, c_i = \frac{1}{1+\lambda}$$

Where we can select a special values for C: $C = \frac{1}{1+\lambda} I_p$. In this case:

$$R(\hat{\mu}_c) = \sum_{i=1}^{p} c_i + \sum_{i=1}^{p}(1-c_i)^2\mu_i^2 = \frac{P}{(1+\lambda)^2} + \frac{\lambda^2}{(1+\lambda)^2}\sum_{i=1}^{p} c_i \tag{5}$$

To minimize the risk, $\frac{dR(\hat{\mu}_c)}{d\lambda} = 0$. Again we select a special $\lambda' = \frac{p}{\|\mu\|^2}$ such that

$$R(\mu_{\lambda*}) = \frac{p}{\frac{p}{\|\mu\|^2} + 1}$$

4 Non-linear James-Stein estimator

Previously we also discussed non-linear James-Stein (JS) estimator:

$$\hat{\mu}_{JS} \equiv 1 - \frac{p-2}{\|Y\|^2}Y. \tag{6}$$

We omit the complicated derivation of risk for JS-estimator (can be found at Johnstone [GE]) and draw conclusion here:

$$R(\hat{\mu}_{JS}) = p - \mathbb{E}[\frac{p-2}{\|Y\|^2}] < p. \tag{7}$$

From the equation we can notice that this holds for $p \geq 3$. In fact $R(\hat{\mu}_{JS})$ approximates the risk of MLE as $\|\mu\|$ increases.

4.1 Inadmissibility

We say an estimator of mean $\hat{\mu}$ is inadmissible if there exists another estimator μ^* which can *beat* $\hat{\mu}$ in the following sense.

Definition (Inadmissible)**.** An estimator $\hat{\mu}_n$ of the parameter μ is called **inadmissible** on $\mathbb{R}^p$ with respect to the squared risk if there exists another estimator μ_n^* such that

$$\mathbb{E}\|\mu_n^* - \mu\|^2 \leq \mathbb{E}\|\hat{\mu}_n - \mu\|^2 \qquad \text{for all } \mu \in \mathbb{R}^p,$$

and there exist $\mu_0 \in \mathbb{R}^p$ such that

$$\mathbb{E}\|\mu_n^* - \mu_0\|^2 < \mathbb{E}\|\hat{\mu}_n - \mu_0\|^2.$$

In this case, we also call that μ_n^* **dominates** $\hat{\mu}_n$. Otherwise, the estimator $\hat{\mu}_n$ is called **admissible**.

Intuitively this is a notion that $\hat{\mu}$ can be "beaten" by some other estimator. Previous sections have shown that $\hat{\mu}_{MLE}$ is inadmissible. Also note that $\hat{\mu}_{JS}$ is also inadmissible, we can find a new estimator $\hat{\mu}_{JS^+} \equiv (1 - \frac{p-2}{\|Y-2\|^2}) + Y$ as a counter-example.
However, $\hat{\mu}_{L.E.}$ can be admissible iff:
1. C is symmetric.
2. eigenvalues of C, $0 \leq \rho_i(C) \leq 1$ for all i.
3. $\rho_i(C) = 1$ for at most two i's.
holds for the same time. Again, we refer to Johnstone [GE] for the details of this proof.

5 Lasso / Soft-thresholding

5.1 subgradient

For convex function $f(x)$, the subgradient $\partial f(x) \equiv \{\mu \in \mathbb{R}^p | \forall x \in \mathbb{R}^p, f(x) \geq f(x_0) + < \mu, x - x_0 >\}$. By this definition, if $\nabla f(x_0)$ exists, it is equal to the subgradient, otherwise subgradient would be a set function.

Example:

$$\partial |x| = \begin{cases} \{1\}, & x > 0; \\ [-1, 1], & x = 0; \\ \{-1\}, & x < 0. \end{cases}$$

5.2 risk of Lasso

In this seciton $\|\mu\|_1 \equiv \sum \mu_i$, and

$$\hat{\mu}^{lasso} \equiv \arg\min_{\tilde{\mu}} \frac{1}{2}\|y - \tilde{\mu}\|^2 + \lambda \|\tilde{\mu}\| \tag{8}$$

This satisfies KKT condition of first order:

$$(\tilde{\mu} - y) + \lambda \partial \|\mu\|_1 \ni 0$$

So it has to satisfy $\hat{\mu}_i = y - \lambda \partial \|\mu\|$. If $\mu_i = 0, \lambda \in [-1, 1]$, otherwise $\mu_i = \text{sign}(\mu_i)$.

$$\mu_i = \text{sign}(y_i) \cdot \max(|y_i| - \lambda, 0). \tag{9}$$

For the risk of Lasso, we have:

$$R(\hat{\mu}_{lasso}) = p - \mathbb{E}[2 \sum_{i=1}^{p} Id(|y_i| \leq \lambda) - \sum_{i=1}^{p} y_i^2 \wedge \lambda])$$

Let $\lambda = \sqrt{2 \log p}$, $R(\hat{\mu}_{lasso} \leq 1 + (2 \log p + 1) \sum_{i=1}^{p} 1 \wedge \mu^2$ (here $1 \wedge \mu^2 \equiv \min(1, \mu^2)$). Assume k-sparcity: $k = \#i : \mu_i = 0 \ll p$,

$$R(\hat{\mu}_{lasso}) \sim O(k \log p) \ll p \tag{10}$$

6 Hard-thresholding / L_0-regularization

$$\hat{\mu}_{hard} \equiv \arg\min_{\tilde{\mu}} \|y - \tilde{\mu}\|^2 + \lambda^2 \|\mu\|_0$$

Here $\|\mu\|_0$ is the zero-penalty, numbers such that $\tilde{\mu}_i \neq 0$.
$\hat{\mu}_{hard} = y_i if |y_i| \geq \lambda$, and equals to 0 otherwise. We can make an element-wise analysis:

$$\min Q(\tilde{\mu}_i) \equiv \|y_i - \mu_i\|^2 + \lambda^2 \mathbb{1}(\mu_i \neq 0) = \min(y_i^2, \lambda^2)$$
$$\mathbb{E}(\tilde{\mu}_i^{hard} \neq 0) = \mathbb{E}[y_i] = \mu_i$$

This means that we need a unbiased, non-zero guess. It turn out that the error term of hard-thresholding, if using $\lambda = \sqrt{2 \log p}$, is:

$$R(\hat{\mu}_i^{hard}) \sim O(k \log p + \delta) \ll p \tag{11}$$

This element-wise analysis combined with concentration inequality can give that, if $y_i \sim N(0, I_p)$,

$$Pr(\max_{1 \leq i \leq p} |y_i| \geq \sqrt{2 \log p}) \leq \frac{1}{\sqrt{\pi \log p}}$$

Although hard thresholding provides an "unbiased" estimator, its variance can be greater than lasso, actually it is better when $\hat{\mu}$ is large.

7 Non-convex regularization

$$\hat{\mu}_\lambda^{non-conv} = \arg\min \frac{1}{2}\|y - \mu\|^2 + \lambda \sum_{i=1}^{p} \rho(|\tilde{\mu}_i|) \tag{12}$$

To find minimizer we can use first-order derivative to run iterations:

$$0 = \tilde{\mu}_i - y_i + \lambda \frac{d\rho}{d|\mu_i|}$$

The goal of such regularization is to make $\rho' \to 0$ when $|\hat{\mu}|$ is large, a cusp in the vicinity of 0, and a singularity point at 0 (when $|\hat{\mu}|$ is small.)
However it is NP-hard to find a global optimizer. An empirical approach would be starting from a lasso estimator and run iterations until local convergence, but some further condition on μ is required to achieve global convergence.

8 Linearized Bregman Iteration

$$0 = \tilde{\mu}_i - y_i + \lambda p(\tilde{\mu}) \tag{13}$$

Given $t = \lambda$, we have $\frac{p(\tilde{\mu})}{t} = -(\tilde{\mu} - y) = -\nabla(\frac{1}{2}\|y - \tilde{\mu}\|^2)$, where $p(\tilde{\mu}) \in \partial(\tilde{\mu})$.
If $\hat{\mu}_i \neq 0$, $p(\hat{\mu}_i) = \pm 1$, here p is constant.

$$\frac{dp_i}{dt} = -\hat{\mu}_i + y_i$$

or $\hat{\mu}_i = y_i$, $\hat{\mu}$ is unbiased estimator. Otherwise we get piecewise constant. The equation above is called Bragman ISS, it selects variable consistently under nearly same conditions as Lasso. To solve it, one can either use Simple Discretization (LBI):

$$\frac{dp(\tilde{\mu}(t))}{dt} + \frac{1}{k}\frac{d\tilde{\mu}(t)}{dt} = -(\tilde{\mu}(t) - y) = -\nabla l(\tilde{\mu}(t))$$

Or use Euler Discretization:

$$\begin{cases} \frac{W_{t+1} - W_t}{\alpha} & = -\nabla l(\tilde{\mu}(t)) \\ W_t & = p_t(\tilde{\mu}(t)) + \frac{\tilde{\mu}(t)}{k}, \quad p_t(\tilde{\mu}(t)) \in \partial \|\tilde{\mu}(t)\|_1 \end{cases}$$

Which gives $\tilde{\mu}(t) = k \cdot shrink(W_t, 1) = k \cdot \text{sign}(W_t) \max(|W_t| - 1, 0)$

9 Consistency of variable selection

For high dimensional statistics, i.e.

$$\mu = X^{n\times p}\beta; n < p$$

We further assume sparsity($\#\beta_i \neq 0 = k \ll n < p$).
There are two scenarios for this problem:

- Noise free: $Y = X\beta \rightarrow$ compressed sensing
- Noise: $Y = X\beta + \epsilon$, where ϵ is the subgaussian noise.

There are two purposes for consistency: L_2 consistency and model selection consistency.

9.1 L_2 consistency

We now discuss restricted eigenvalue condition for Lasso. The Hessian of loss is defined as:

$$\hat{\Sigma}_n \equiv \frac{1}{n}X^T X \geq 0$$

where $n < p, rank(\hat{\Sigma}_n) < p$, it is convex but not strongly convex.

$$\Delta = \hat{\beta}_{lasso} - \beta; \{\Delta^T \Sigma_n \Delta \geq \gamma\|\Delta\|^2 : \gamma > 0, \|\Delta_{S^\perp}\|_1 \leq c\|\Delta_S\|_1\}$$

. Here $\|\hat{\beta}^{lasso}_{S^\perp} - \beta_{S^\perp}\|$ has to be small, and $S^\perp$ is the non-support set.
It is minimax optimal up to $\log p$, $R(\hat{\mu}_{lasso}) \leq O(\sqrt{\frac{k\log p}{n}})$

9.2 Model selection consistency

Model selection consistency concerns with the question that under what conditions an estimator $\hat{\beta}$ recovers the sparsity pattern of the ground truth β, i.e. $\text{supp}(\hat{(\beta)}) = \text{supp}(\beta)$ or a slightly stronger one, $\text{sign}(\hat{\beta}) = \text{sign}(\beta)$. The latter is often called *sign-consistency*. The following two conditions are necessary and sufficient for Lasso's model selection consistency, as well as many other algorithms including Linearized Bregman Iterations.

9.2.1 Irrepresentable(Incoherence) Condition (IRR)

$$\forall j \in S_1^\perp, \|(\frac{1}{n}X_S^T X_S)^{-1} X_S^T X_j\|_1 \leq 1 - \eta(\eta > 0)$$

IRR is stronger than RE.

9.2.2 Strong signal

$\min|\beta_i| > O(\sqrt{\frac{k\log p}{n}}) \approxeq \hat{\lambda}_n$. $supp(\beta^{lasso}_{\hat{\lambda}_n}) = S; \text{sign}(\beta^{lasso}_{\hat{\lambda}_n}) = \text{sign}(\beta)$ holds with very high probability.