Mathematics for Data Sciences **September 12, 2011**

Lecture 2. Stein's Phenomenon and Shrinkage

Instructor: Yuan Yao, Peking University *Scribe: Hu Sheng*

Books about Learning Theory

1. Vapnik, V. *Statistical Learning Theory*. 1998, Wiley, New York.

   A classic book on Empirical Risk Minimization (ERM), VC-dimension and Support Vector Machine (SVM).

2. L. Györfi, M. Kohler, A. Krzyzak, H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. 2002, Springer-Verlag, New York.

   Nonparametric regression (supervised learning) with random design, emphasis on distribution-free theory (e.g. minimax/individual optimality, adaptation)

3. Tsybakov. *Introduction to Nonparametric Estimation*. 2009, Springer.

   A graduate textbook translated from French version, on nonparametric estimation including density estimation (unsupervised learning).

4. Kearns, M., Vazirani, U. *An introduction to computational learning theory*. 1994, Cambridge, MA: MIT Press.

   A computer science textbook which introduces Probably Approximately Correct (PAC) learning.

Introduction

In this class, we show that when in high-dimensional settings (p large), sample mean is not a good estimation. This is particularly shown by Stein's phenomenon and shrinkage estimators.

Stein's Phenomenon & Shrinkage

In this section, we will discuss some notions related to Stein's phenomenon. First, we will consider multivariate Gaussian model: let $X_1, \dots, X_n \sim \mathcal{N}(\mu, \Sigma)$, $X_i \in \mathbb{R}^p (i = 1 \dots n)$, then the maximum likelihood estimators (MLE) of the parameters (μ and Σ) are as follows:

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \hat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu}_n)(X_i - \hat{\mu}_n)^T$$

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

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

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

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

Otherwise, the estimator μ_n^* is called **admissible**.

Stein (1956) found that if $p \geq 3$, then the MLE estimator $\hat{\mu}_n$ is inadmissible. This property is known as ***Stein's phenomenon***. This phenomenon can be described like:

Problem. Suppose that $X_i \sim \mathcal{N}(\mu, I_p)(i = 1 \dots n)$ are independent p-Gaussian variables, let $y = \hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i \sim \mathcal{N}(\mu, \frac{1}{\sqrt{n}} I_p) = \mathcal{N}(\mu, \varepsilon I_p)$, $\varepsilon = \frac{1}{\sqrt{n}}$, there exist $\tilde{\mu}_n = g(y)y$ where $g : \mathbb{R}^p \to \mathbb{R}$, so that $\mathbb{E}\|\tilde{\mu}_n - \mu\|^2 < \mathbb{E}\|\hat{\mu}_n - \mu\|^2$.

To solve the problem, we start with a preliminary lemma.

Lemma 1 (Stein's lemma)**.** Suppose that a function $f : \mathbb{R}^p \to \mathbb{R}$ satisfies:

(i) $f(u_1, \dots, u_p)$ is absolutely continuous in each coordinate u_i for almost all values (with respect to the Lebesgue measure on $\mathbb{R}^{p-1}$) of other coordinates $(u_j, j \neq i)$

(ii)
$$\mathbb{E}\left|\frac{\partial f(y)}{\partial y_i}\right| < \infty, \qquad i = 1, \dots, p.$$

then

$$\mathbb{E}[(\mu_i - y_i) f(y)] = -\varepsilon^2 \mathbb{E}\left[\frac{\partial f}{\partial y_i}(y)\right], \qquad i = 1, \dots, p.$$

The proof of Stein's lemma can be found on Tsybakov's book (page 157∼158), which essentially exploits the integration by parts. Now, let us look for a function g such that the risk of the estimator $\tilde{\mu}_n = g(y)y$ is smaller than the MLE of y. We have

$$\begin{aligned}
\mathbb{E}\|\tilde{\mu}_n - \mu\|^2 &= \sum_{i=1}^p \mathbb{E}[(g(y)y_i - \mu_i)^2] \\
&= \sum_{i=1}^p \{\mathbb{E}[(y_i - \mu_i)^2] + 2\mathbb{E}[(\mu_i - y_i)(1 - g(y))y_i] \\
&\quad + \mathbb{E}[y_i^2 (1 - g(y))^2]\}.
\end{aligned}$$

Suppose now that the function g is such that the assumptions of Lemma 1 hold for the functions $f = f_i$ where $f_i(y) = (1 - g(y))y_i, i = 1, \dots, p$. Then

$$\mathbb{E}[(\mu_i - y_i)(1 - g(y))y_i] = -\varepsilon^2 \mathbb{E}\left[1 - g(y) - y_i \frac{\partial g}{\partial y_i}(y)\right],$$

with

$$\mathbb{E}[(y_i - \mu_i)^2] = \varepsilon^2,$$

we have

$$\mathbb{E}[(\tilde{\mu}_{n,i} - \mu_i)]^2 = \varepsilon^2 - 2\varepsilon^2 \mathbb{E}\left[1 - g(y) - y_i \frac{\partial g}{\partial y_i}(y)\right] + \mathbb{E}[y_i^2 (1 - g(y))^2].$$

Summing over i gives

$$\mathbb{E}\|\tilde{\mu}_n - \mu\|^2 = p\varepsilon^2 + \mathbb{E}[W(y)] = \mathbb{E}\|\hat{\mu}_n - \mu\|^2 + \mathbb{E}[W(y)]$$

with

$$W(y) = -2p\varepsilon^2(1-g(y)) + 2\varepsilon^2 \sum_{i=1}^{p} y_i \frac{\partial g}{\partial y_i}(y) + \|y\|^2(1-g(y))^2.$$

The risk of $\tilde{\mu}_n$ is smaller than that of $\hat{\mu}_n$ if we choose g such that

$$\mathbb{E}[W(y)] < 0.$$

In order to satisfy this inequality, we can search for g among the functions of the form

$$g(y) = 1 - \frac{b}{a + \|y\|^2}$$

with an appropriately chosen constants $a \geq 0$, $b > 0$. Therefore, $W(y)$ can be written as

$$\begin{aligned} W(y) &= -2p\varepsilon^2 \frac{b}{a+\|y\|^2} + 2\varepsilon^2 \sum_{i=1}^{p} \frac{2by_i^2}{(a+\|y\|^2)^2} + \frac{b^2\|y\|^2}{(a+\|y\|^2)^2} \\ &= \frac{1}{a+\|y\|^2}\left(-2pb\varepsilon^2 + \frac{4b\varepsilon^2\|y\|^2}{a+\|y\|^2} + \frac{b^2\|y\|^2}{(a+\|y\|^2)^2}\right) \\ &\leq \frac{1}{a+\|y\|^2}(-2pb\varepsilon^2 + 4b\varepsilon^2 + b^2) \qquad \|y\|^2 \leq a + \|y\|^2 \text{ for } a \geq 0 \\ &= \frac{Q(b)}{a+\|y\|^2}, \qquad Q(b) = b^2 - 2pb\varepsilon^2 + 4b\varepsilon^2. \end{aligned}$$

The minimizer in b of quadratic function $Q(b)$ is equal to

$$b_{opt} = \varepsilon^2(p-2),$$

where the minimum of $W(y)$ satisfies

$$W_{min}(y) \leq -\frac{b_{opt}^2}{a+\|y\|^2} = -\frac{\varepsilon^4(p-2)^2}{a+\|y\|^2} < 0.$$

Note that when $b \in (b_1, b_2)$, *i.e.* between the two roots of $Q(b)$

$$b_1 = 0, \qquad b_2 = 2\varepsilon^2(p-2)$$

we have $W(y) < 0$, which may lead to other estimators having smaller mean square errors than MLE estimator.

When $a = 0$, the function g and the estimator $\tilde{\mu}_n = g(y)y$ associated to this choice of g are given by

$$g(y) = 1 - \frac{\varepsilon^2(p-2)}{\|y\|^2},$$

and

$$\tilde{\mu}_n = \left(1 - \frac{\varepsilon^2(p-2)}{\|y\|^2}\right)y =: \tilde{\mu}_{JS},$$

respectively. $\tilde{\mu}_{JS}$ is called **James-Stein estimator**. If dimension $p \geq 3$ and the norm $\|y\|^2$ is sufficiently large, multiplication of y by $g(y)$ shrinks the value of y to 0. This is called the ***Stein shrinkage***. If $b = b_{opt}$, then

$$W_{min}(y) = -\frac{\varepsilon^4(p-2)^2}{\|y\|^2}.$$

Lemma 2. Let $p \geq 3$. Then, for all $\mu \in \mathbb{R}^p$,

$$0 < \mathbb{E}\left(\frac{1}{\|y\|^2}\right) < \infty.$$

The proof of Lemma 2 can be found on Tsybakov's book (page 158∼159). For the function W, Lemma 2 implies $-\infty < \mathbb{E}[W(y)] < 0$, provided that $p \geq 3$. Therefore, if $p \geq 3$, the risk of the estimator $\tilde{\mu}_n$ satisfies

$$\mathbb{E}\|\tilde{\mu}_n - \mu\|^2 = p\varepsilon^2 - \mathbb{E}\left(\frac{\varepsilon^4(p-2)^2}{\|y\|^2}\right) < \mathbb{E}\|\hat{\mu}_n - \mu\|^2$$

for all $\mu \in \mathbb{R}^p$.

Besides James-Stein estimator, there are other estimators having smaller mean square errors than MLE $\hat{mu}_n$.

- *Stein estimator*: $a = 0, b = \varepsilon^2 p$,
$$\tilde{\mu}_S := \left(1 - \frac{\varepsilon^2 p}{\|y\|^2}\right) y$$
- *Positive part James-Stein estimator*:
$$\tilde{\mu}_{JS+} := \left(1 - \frac{\varepsilon^2 (p-2)}{\|y\|^2}\right)_+ y$$
- *Positive part Stein estimator*:
$$\tilde{\mu}_{S+} := \left(1 - \frac{\varepsilon^2 p}{\|y\|^2}\right)_+ y$$

where $(x)_+ = \min(0, x)$. Denote the mean square error by $MSE(\tilde{\mu}) = \mathbb{E}\|\tilde{\mu} - \mu\|^2$, then we have

$$MSE(\tilde{\mu}_{JS+}) < MSE(\tilde{\mu}_{JS}) < MSE(\hat{\mu}_n), \qquad MSE(\tilde{\mu}_{S+}) < MSE(\tilde{\mu}_S) < MSE(\hat{\mu}_n).$$

Summary

Stein's phenomenon firstly shows that in high dimensional estimation, shrinkage may lead to better performance than MLE, the sample mean. This opens a new era for modern high dimensional statistics. In fact discussions above study independent random variables in p-dimensional space, concentration of measure tells us some priori knowledge about the estimator distribution – samples are concentrating around certain point. Shrinkage toward such point may naturally lead to better performance.

However, after Stein's phenomenon firstly proposed in 1956, for many years researchers have not found the expected revolution in practice. Mostly because Stein's type estimators are too complicated in real applications and very small gain can be achieved in many cases. Researchers struggle to show real application examples where one can benefit greatly from Stein's estimators. For example, Efron-Morris (1974) showed three examples that JS-estimator significantly improves the multivariate estimation. On other other hand, deeper understanding on Shrinkage-type estimators has been pursued from various aspects in statistics.

The situation changes dramatically when LASSO-type estimators by Tibshirani, also called Basis Pursuit by Donoho et al. are studied around 1996. This brings sparsity and L1-regularization into the central theme of high dimensional statistics and leads to a new type of shrinkage estimator, thresholding. For example,

$$\min_{\tilde{\mu}} I = \min_{\tilde{\mu}} \frac{1}{2}\|\tilde{\mu} - \mu\|^2 + \lambda\|\tilde{\mu}\|_1$$

Subgradients of I over $\tilde{\mu}$ leads to

$$0 \in \partial_{\tilde{\mu}_j} I = (\tilde{\mu}_j - \mu_j) + \lambda \text{signset}(\tilde{\mu}_j) \Rightarrow \tilde{\mu}_j = \text{sign}(\mu_j)(|\mu_j| - \lambda)_+$$

where the set-valued map $\text{signset}(x) = 1$ if $x > 0$, $\text{signset}(x) = -1$ if $x < 0$, and $\text{signset}(x) = [-1, 1]$ if $x = 0$, is the subgradient of absolute function $|x|$. Under this new framework shrinkage estimators achieves a new peak with an ubiquitous spread in data analysis with high dimensionality.