STAT 425: Introduction to Nonparametric Statistics **Winter 2018**

Lecture 14: Learning Theory: Normal Means Model

Instructor: Yen-Chi Chen

Reference: Chapter 7.1 – 7.4 and 7.6 in *All of nonparametric statistics.*

14.1 The Normal Means Model

The normal means model considers the following problem. Assume that we have N independent Gaussian random variables $Z_1, \cdots, Z_N$ such that

$$Z_i \sim N(\mu_i, \sigma^2).$$

Namely, each random variable is a Gaussian that has its own mean and every random variable has a common variance σ^2. We will assume σ^2 is known (we know the noise level).

Example: nonparametric regression. This problem arises naturally in nonparametric regression, where we observe IID pairs $(X_1, Y_1), \cdots, (X_n, Y_n)$ from some distribution. Recall that in basis approach, we model $m(x) = \sum_{\ell=1}^n \theta_\ell \phi_\ell(x)$ for some basis $\phi_1, \phi_2, \cdots$. And our estimator of the coefficients are

$$\widehat{\theta}_\ell = \frac{1}{n}\sum_{i=1}^n Y_i \phi_\ell(X_i).$$

If $X_1, \cdots, X_n$ are IID from $\mathsf{Uni}[0,1]$, each estimator of the coefficient satisfies

$$\widehat{\theta}_\ell \approx N(\theta_\ell, \sigma_\phi^2/n)$$

when n is large. Moreover, the covariance between any distinct pair $\widehat{\theta}_\ell, \widehat{\theta}_j$ is 0, which implies that they are independent asymptotically (this is a property of multivariate Gaussian). Thus, the estimated coefficients $\widehat{\theta}_1, \cdots, \widehat{\theta}_M$ are just like the random variables being studied $Z_1, \cdots, Z_M$ with $\sigma^2 = \sigma_\phi^2/n$.

Example: multiple experiments. Another scenario that the normal means model will be useful is in certain experiments. Assume that we are interested in the average value of M quantities. A simple experiment is that for each quantity, we make n measurements. So totally we have ran $M \times n$ independent experiments. Let T_{ij} denotes the i-th measurement of the j-th quantity. Then the average

$$W_j = \frac{1}{n}\sum_{i=1}^n T_{ij}$$

is a good estimate of the average value of the j-th quantity. Because of the central limit theory and the fact that every measurement is independent,

$$W_j \approx N(\mu_j, \sigma_j^2/n).$$

If we further assume that the variability of measuring every quantity is the same ($\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_M^2$), we obtain the normal means model with $\sigma^2 = \sigma_1^2/n$.

In the normal means model, we want to jointly estimate all parameters $\mu = (\mu_1, \cdots, \mu_M)$ from $Z = (Z_1, \cdots, Z_M)$. A naive estimator is just use the original values of $Z_1, \cdots, Z_M$ as an estimator, which leads to a mean square error

$$R(Z) = \mathbf{MSE}(Z, \mu) = \mathbb{E}\|Z - \mu\|^2 = M\sigma^2.$$

We denote $\widehat{\theta}_{\mathsf{naive}} = Z$ as this estimator. Is this a good estimator? How do we estimate the risk (MSE)?

14.2 Stein's Unbiased Risk Estimator (SURE)

Now we consider any estimator $\widehat{\theta} = \eta(Z)$. And define the function $g : \mathbb{R}^M \mapsto \mathbb{R}^M$ such that $g(Z) = \widehat{\theta} - Z$. When g is differentiable, there is a powerful method for estimating the risk called SURE (Stein's Unbiased Risk Estimator) for the normal means model:

$$\widehat{R}(\widehat{\theta}) = M\sigma^2 + 2\sigma^2 \sum_{i=1}^{M} \frac{\partial g(z_1, \cdots, z_M)}{\partial z_i} + \|g(Z)\|^2. \tag{14.1}$$

In what follows, we will show that

$$\mathbb{E}(\widehat{R}(\widehat{\theta})) = R(\widehat{\theta}). \tag{14.2}$$

Namely, SURE is an unbiased estimator of the risk/MSE.

Before proving that it is an unbiased estimator, we first introduce a useful lemma called **Stein's Lemma**.

Lemma 14.1 (Stein's Lemma) *Let $W \sim N(\mu, \sigma^2)$ be a normal random variable and $f : \mathbb{R} \mapsto \mathbb{R}$. Then*

$$\mathbb{E}(f(W)(W - \mu)) = \sigma^2 \mathbb{E}(f'(W)).$$

This lemma can be proved using the integration by parts.

Now we will prove equation (14.2). Let $g_i(z_1, \cdots, z_M)$ be the i-th component of $g(z_1, \cdots, z_M)$. Using the Stein's lemma, we have

$$\sigma^2 \mathbb{E}\left(\frac{\partial g(z_1, \cdots, z_M)}{\partial z_i}\right) = \mathbb{E}\left(g_i(Z)(Z_i - \mu_i)\right).$$

Thus, the expectation of SURE is

$$\begin{aligned}
\mathbb{E}(\widehat{R}(\widehat{\theta})) &= M\sigma^2 + 2\sigma^2 \sum_{i=1}^{M} \mathbb{E}\left(\frac{\partial g(z_1, \cdots, z_n)}{\partial z_i}\right) + \mathbb{E}\|g(Z)\|^2 \\
&= \sum_{i=1}^{M} \mathbb{E}|Z_i - \mu_i|^2 + 2\sum_{i=1}^{M} \mathbb{E}\left(\underbrace{g_i(Z)}_{=\widehat{\theta}_i - Z_i}(Z_i - \mu_i)\right) + \sum_{i=1}^{M} \mathbb{E}|g_i(Z)|^2 \\
&= \sum_{i=1}^{M} \mathbb{E}\left(|Z_i - \mu_i|^2 + 2(\widehat{\theta}_i - Z_i)(Z_i - \mu_i) + |Z_i - \mu_i|^2\right) \\
&= \sum_{i=1}^{M} \mathbb{E}|\widehat{\theta}_i - Z_i + Z_i - \mu_i|^2 \\
&= \sum_{i=1}^{M} \mathbb{E}|\widehat{\theta}_i - \mu_i|^2 \\
&= R(\widehat{\theta}).
\end{aligned}$$

SURE can be applied as long as we know g and σ^2 (noise level). A good news is – we choose the estimator $\widehat{\theta}$ so the function g is often known, making the SURE a powerful tool in practice. If we do not know the noise level σ^2, we may try to estimate it. If the model comes from nonparametric regression and we are using a linear smoother, we can use the theory of linear smoother to construct an estimator of the noise level.

Example: naive estimator. In the naive estimator $\widehat{\theta} = Z$ so $g(z) = 0$. SURE will imply that the risk is $M\sigma^2$, which is what we expect.

Example: soft-thresholding estimator. Now we consider a soft-thresholding estimator. Let $\lambda > 0$ be a constant representing the threshold. Then the soft-thresholding estimator with this threshold is

$$\begin{aligned}\widehat{\theta}_i &= \begin{cases} Z_i + \lambda & \text{if } Z_i < -\lambda \\ 0 & \text{if } -\lambda \leq Z_i \leq \lambda \\ Z_i - \lambda & \text{if } Z_i > \lambda \end{cases} \\ &= \mathsf{sgn}(Z_i)(|Z_i| - \lambda)_+.\end{aligned}$$

Note that $a_+ = \max\{0, a\}$. This estimator is like the one we have encountered in LASSO that shrinks every signal toward 0 by the amount λ and if the signal has an amplitude below λ, we just output 0. In this case,

$$g_i(Z) = \mathsf{sgn}(Z_i)(|Z_i| - \lambda)_+ - Z_i = \begin{cases} \lambda & \text{if } Z_i < -\lambda \\ -Z_i & \text{if } -\lambda \leq Z_i \leq \lambda \\ \lambda & \text{if } Z_i > \lambda \end{cases}$$

SURE will give a risk as

$$R(\widehat{\theta}) = \sum_{i=1}^{M} (\sigma^2 - 2\sigma^2 I(|Z_i| \leq \lambda) + \min\{Z_i^2, \lambda^2\}).$$

Remark

- The SURE can be applied to a more general settings where the random vector $Z \sim N(\mu, \Sigma)$. Again let $\widehat{\theta} = \eta(Z)$ be the estimator and $g(Z) = \widehat{\theta} - Z$ be the function describing the difference between the estimator and Z. Moreover, define an $M \times M$ matrix D such that $D_{ij} = \frac{\partial g_i(z_1, \cdots, z_M)}{\partial z_j}$. Then the SURE is
$$\mathsf{SURE} = \mathsf{tr}(\Sigma) + 2\mathsf{tr}(\Sigma D) + \sum_{i=1}^{M} g_i^2(Z).$$

14.3 Shrinkage Estimator

Let $b \in [0, 1]$ be a constant. We now consider an estimator of the form $\widehat{\theta}_b = bZ$. Such an estimator is called **linear** estimator. Naively, $b = 1$ seems to be a good estimator. As we have seen, $b = 1$ gives an unbiased estimator of the parameter of interest μ. But is this the best estimator among all linear estimators?

To talk about optimality, we just consider the risk (MSE). What is the risk (MSE) of this estimator? Simple calculation shows that

$$\begin{aligned}R(\widehat{\theta}_b) &= \mathbb{E}\|bZ - \mu\|^2 \\ &= \mathbf{bias}^2(bZ) + \mathsf{Var(bZ)} \\ &= (1 - b)^2 \sum_{i=1}^{M} \mu_i^2 + Mb^2\sigma^2 \\ &= (1 - b)^2 \|\mu\|^2 + Mb^2\sigma^2.\end{aligned}$$

It is in a quadratic form of b, so we take derivative and optimize it. The optimal choice is

$$b^* = \frac{\|\mu\|^2}{\|\mu\|^2 + M\sigma^2}$$

with an optimal risk

$$R(\widehat{\theta}_b^*) = \frac{M\sigma^2\|\mu\|^2}{\|\mu\|^2 + M\sigma^2} < M\sigma^2 = \text{ the risk of the naive estimator.}$$

Namely, we should shrink the value of Z a little bit to obtain the optimal estimator.

From a mathematical point of view, this seems to be a reasonable choice – we are finding an estimator that has the minimal risk for estimating μ.

However, if we try to interpret the result, this seems to be something that is very counter intuitive. Now we think about a common scenario where we will face the normal means model – the situations that we have multiple independent experiments (the second example in Section 14.1). When we are analyzing the M quantities of interest, they can be completely unrelated quantities. For instance, the first quantity may be the average temperature of Chicago and the second quantity is the average GPA of all UW students and the third quantity is the average height of residents in New York City. All these quantities are unrelated to each other but somehow when we estimate them jointly, the optimal estimator would use all information to estimate one quantity. Namely, the average height of the residents in NYC will affect how we are estimating the average GPA of UW students!

This phenomenon is called Stein's phenomenon and was discovered by Prof. Charles Stein, a famous statistician at UC Berkeley, in 1956 and it has shocked the entire statistical community because people at that time believed that the unbiased estimator should be a very good estimator. Moreover, the naive estimator Z is not only an unbiased estimator but also the maximum likelihood estimator so most people believe that it should be the best estimator. In addition, what astonished people more was the counter intuitive example we just talked about – unrelated experiments somehow interact each other in the optimal estimator when we want to jointly estimate them.

Although the choice b^* minimizes the MSE, it cannot be used in practice because it involves unknown quantity μ. Now we will introduce a famous estimator called the **James-Stein (JS) estimator**:

$$\widehat{\mu}_{JS} = \left(1 - \frac{(M-2)\sigma^2}{\|Z\|^2}\right) Z.$$

Namely, the JS estimator uses $b = 1 - \frac{(M-2)\sigma^2}{\|Z\|}$. You can prove that such an estimator has a smaller risk than the naive estimator Z.

14.4 Many Normal Random Variables

What happens if there is no signal at all in our data but we still try to fit a model? As we have seen, the coefficients will behave like a normal distribution. When there is no signal at all, then the expectations $\mu_1, \cdots, \mu_N$ will all be 0. Thus, we are observing

$$X_1, \cdots, X_N \sim N(0, \sigma^2).$$

Even all these random variables have mean 0, just by chance we will see some large number as long as N is large. So we want to quantify the magnitude of such a large number caused by randomness. A simple approach is to consider the maximum deviation

$$Y_N = \max\{X_1, \cdots, X_n\} = \max_{i=1,\cdots,N} X_i.$$

And we want to study $\mathbb{E}(Y_N)$, the expected value of such a maximum deviation caused by purely random noises.

Before we deriving a bound on Y_N, we need to introduce a very useful inequality. Recall that a function $g : \mathbb{R} \mapsto \mathbb{R}$ is called a **convex function** if for any two points x, y and any number $\alpha \in [0, 1]$,

$$g(\alpha x + (1 - \alpha)y) \leq \alpha g(x) + (1 - \alpha)g(y).$$

Examples of convex functions are $g(x) = x^2$, $g(x) = e^x$. Note that a function h is called a **concave function** if $-h$ is a convex function.

A simple way to picture a convex function is: for any point x, we can define a tangent line passing through $(x, g(x))$ and is tangent to the function $g(x)$. Then $g(x)$ is always above or equal to such a line.

A more mathematical way to think about a convex function is: if $g''(x) \geq 0$, then $g(x)$ is convex.

Lemma 14.2 (Jensen's inequality) *If g is convex, then*

$$\mathbb{E}(g(X)) \geq g(\mathbb{E}(X)).$$

If g is concave, then

$$\mathbb{E}(g(X)) \leq g(\mathbb{E}(X)).$$

Proof: Consider a line $L(x) = ax + b$ such that it it tangent to $g(x)$ at $(\mathbb{E}(X), g(\mathbb{E}(X)))$. By the property of a convex function.

$$\mathbb{E}(g(X)) \geq \mathbb{E}(L(X)) = \mathbb{E}(a + bX) = a + b\mathbb{E}(X) = L(\mathbb{E}(X)) = g(\mathbb{E}(X)).$$

■

Moreover, we define the moment generating function (MGF) for a random variable X as

$$\phi_X(t) = \mathbb{E}(e^{tX})$$

for $t > 0$. The MGF is also called the Laplace transform in Physical sciences. If X follows a normal distribution $N(\mu, \sigma^2)$, then its MGF is

$$\phi_{N(\mu,\sigma^2)}(t) = e^{\mu t + \sigma^2 t^2/2}.$$

Using Jensen's inequality and the MGF of normal, we can bound the expectation of the maximum fluctuation $\mathbb{E}(Y_N)$.

Theorem 14.3 *Assume that*

$$X_1, \cdots, X_N \sim N(0, \sigma^2)$$

and define $Y_N = \max\{X_1, \cdots, X_N\} = \max_{i=1,\cdots,N} Y_i$. Then

$$\mathbb{E}(Y_N) \leq \sigma\sqrt{2 \log N}.$$

Proof: Because e^x is a convex function, for any positive number t, Jensen's inequality implies

$$\begin{aligned}
\exp\left(t\ \mathbb{E}\left(\max_{i=1,\cdots,N} X_i\right)\right) &\leq \mathbb{E}\left(\exp\left(t\max_{i=1,\cdots,N} X_i\right)\right) \quad \text{(Jensen's inequality)}\\
&= \mathbb{E}\left(\max_{i=1,\cdots,N} \exp\left(tX_i\right)\right)\\
&\leq \mathbb{E}\left(\sum_{i=1}^N \exp\left(tX_i\right)\right)\\
&= \sum_{i=1}^N \mathbb{E}\left(\exp\left(tX_i\right)\right)\\
&= Ne^{t^2\sigma^2/2} \quad \text{(MGF of normal).}
\end{aligned}$$

Thus, taking logarithm in both sides, we obtain

$$t\mathbb{E}(Y_N) = t\ \mathbb{E}\left(\max_{i=1,\cdots,N} X_i\right) \leq \log N + \frac{t^2\sigma^2}{2}.$$

Diving both sides by t, we have

$$\mathbb{E}(Y_N) \leq \frac{\log N}{t} + \frac{t\sigma^2}{2}$$

for every $t > 0$. Now by optimizing t, which occurs at $t = \sqrt{\frac{2\log N}{\sigma^2}}$, we obtain

$$\mathbb{E}(Y_N) \leq \sigma\sqrt{2\log N}.$$

■

Namely, the fluctuation caused by random noises is at the order of $O(\sqrt{\log N})$. Although it diverges when we have many coefficients ($N \to \infty$), the rate of divergence is quiet slow.

Remark.

- Although we only consider normal distributions in Theorem 14.3, the same result applies to any random variables with a moment generating function

$$\phi_X(t) \leq e^{\frac{1}{2}\sigma^2t^2}$$

  for some σ^2 and for every positive t.