Statistical Inference and Learning- Final Exam

Yiqing Wang and Boaz Nadler

Date: Thu, Feb. 4, 2016, 9:30am-1pm (3.5 hours)

Please answer **ONLY 3 out of the 4** following questions. The total number of points is more than 100 !

Q1 Let Y be a random variable taking values in a finite interval $[-b, b]$. Its mean is zero, namely, $\mathbb{E}[Y] = 0$. We observe n i.i.d. samples $(X_1, \cdots, X_n)$ from the random variable $X = Y + \mu$, where $\mu \in \mathbb{R}$ is unknown. Our task is to estimate the true value of μ.

1. Show that the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ is an unbiased estimator of μ

$$\mathbb{E}[\bar{X}_n] = \mu.$$

Is this the only unbiased estimator of μ ? Explain your answer.

2. Under the square loss, the estimator $\bar{X}_n$'s risk at μ is its MSE

$$R(\bar{X}_n, \mu) := \mathbb{E}[(\bar{X}_n - \mu)^2].$$

Show that the maximum risk $\sup_{\mu\in\mathbb{R}} R(\bar{X}_n, \mu)$ decreases at rate $O(1/n)$.

3. Show that for all $c > 0$

$$\mathbb{P}\left(|\bar{X}_n - \mu| \geq cn^{-1/4}\right) \leq 2\exp\left(-\frac{\sqrt{n}c^2}{2b^2}\right). \tag{1}$$

Hint: use Hoeffding's inequality.

4. In some situations, there is a-priori knowledge or belief that μ may be equal to zero with non-negligible probability. One option to take such knowledge into account is to consider the following estimator of μ, which outputs precisely a value of zero if $\bar{X}_n$ is small

$$\hat{\mu} = T_n = \begin{cases} \bar{X}_n, & \text{if } |\bar{X}_n| > n^{-1/4} \\ 0, & \text{if } |\bar{X}_n| \leq n^{-1/4}. \end{cases}$$

Show that for all $k \geq 1$, $n^k R(T_n, 0)$ converges to zero. (Hint: use Cauchy-Schwartz's inequality and Inequality (1)). This implies that T_n is *super-efficient* at $\mu = 0$, since its MSE at $\mu = 0$ decreases at a much faster rate than $O(1/n)$.

Remark: One can show that for any fixed $\mu \neq 0$, its MSE decay rate remains at $O(1/n)$.

5*. Section 4 above seems to imply that T_n is a better estimator than the sample mean $\bar{X}_n$. This, however, is not true and its *super-efficiency* at $\mu = 0$ comes at a price: the improvement of the estimator at $\mu = 0$ results in an increase in risk at other values of μ. In this section your task is to show that its maximum risk $\sup_{\mu\in\mathbb{R}} R(T_n, \mu)$ actually decreases at rate $O(1/\sqrt{n})$, hence much slower than $O(1/n)$.

Specifically, let $\mu_n = \frac{1}{2}n^{-1/4}$. Show that (i) as $n \to \infty$, $\mathbb{P}\left(|\bar{X}_n| > n^{-1/4}\right) \to 0$; (ii) use the result in (i) to show that $\lim_{n\to\infty} n^{1/2} R(T_n, \mu_n) = 1/4$.

Q2 Consider a system which can only be in one of two possible states $S \in \{0, 1\}$. If $S = 0$, it emits a R^d-valued fixed signal $\boldsymbol{a} = (a_1, \cdots, a_d)$ whereas if $S = 1$, the fixed output signal is $\boldsymbol{b} = (b_1, \cdots, b_d)$. The signal vector is only observed after it has been corrupted by additive Gaussian noise N which satisfies

$$\mathbb{E}[N] = \mathbf{0}, \quad \mathbb{E}[NN^T] = \Sigma.$$

The covariance matrix Σ is assumed to be of full rank. Therefore, the noisy signal vector can be expressed as

$$Y = \begin{cases} \boldsymbol{a} + N, & \text{if } S = 0 \\ \boldsymbol{b} + N, & \text{if } S = 1. \end{cases}$$

We want to detect the state S from the noisy signal Y by testing two hypotheses

$$H_0 : S = 0 \quad \text{versus} \quad H_1 : S = 1.$$

1. Show that the likelihood ratio test statistic is

$$(\boldsymbol{b} - \boldsymbol{a})^T \Sigma^{-1} Y.$$

2. Construct the optimal rejection region so that its resulting type I error is equal to α.

3. Find the rejection region whose type I error equals its type II error. Such a rejection region is said to be minimax optimal.

4*. Define $\gamma = \sqrt{(\boldsymbol{b} - \boldsymbol{a})^T \Sigma^{-1} (\boldsymbol{b} - \boldsymbol{a})}$ as the signal-to-noise ratio (SNR) associated to this detection problem. How do the errors of the minimax optimal rejection region depend on the SNR γ ?

Q3 Let $(x_1, \ldots, x_n)$ be n i.i.d. observations from a probability distribution with density $p(x)$. Recall that in class we considered kernel density estimator of the form

$$\hat{p}(x) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{x - x_j}{h}\right) \tag{2}$$

with a suitably chosen kernel function K.

1. Suppose that $p(x)$ is a smooth density such that its second derivative is smooth and bounded, and in particular satisfies

$$|p''(x) - p''(y)| \leq L|x - y| \qquad \forall x, y \in \mathbb{R}$$

What is then an upper bound on the mean squared error $\mathbb{E}[(\hat{p}(x_0) - p(x_0))^2]$ at some fixed point x_0 and how does it depend on n? What are the conditions that the kernel K must satisfy for this upper bound to hold ?

2. In practice we need to estimate the bandwidth h. A common method is leave-one-out cross-validation. Explain this method and the resulting formula for estimating the bandwidth.

3. In some cases, we know a-priori that the density $p(x)$ has a compact support in an interval I. For example, if x is a physical quantity that cannot be negative then $x \geq 0$, and $I = [0, \infty)$. Let us study what happens to the kernel density estimate (2) for points near the boundary, when the kernel K is symmetric and supported on [-1,1].

To this end, write $x = hz$, where $z \in [0, 1]$. Show that

$$\mathbb{E}[\hat{p}(hz)] = a_0(z)p(0) - h(a_1(z) - za_0(z))p'(0) + O(h^2).$$

where $a_j(z) = \int_{-1}^{z} u^j K(u) du$.

4. In particular what is $\mathbb{E}[\hat{p}(0)]$? Is it a consistent estimator of $p(0)$ as $n \to \infty$ and $h \to 0$? Suggest a simple correction method to get a consistent estimate of $p(0)$.

Q4 Let $\mathbf{x}$ be a d-dimensional Gaussian random vector with distribution $\mathcal{N}(\mu, \mathbf{\Sigma})$. We wish to find a deterministic projection vector $\mathbf{w}$, such that it is of unit length $\|\mathbf{w}\|_2 = 1$ and the variance of $\mathbf{w}^T\mathbf{x}$, denoted by $Var[\mathbf{w}^T\mathbf{x}]$, is maximal. Such a projection vector is called the first principal component.

1. Show that up to a minus sign, $\mathbf{w}$ is the eigenvector with largest eigenvalue of $\mathbf{\Sigma}$. (Hint: Prove that $Var[\mathbf{w}^T\mathbf{x}] = \mathbb{E}[\mathbf{w}^T(\mathbf{x} - \mu)(\mathbf{x} - \mu)^T\mathbf{w}]$)

2. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be n i.i.d. samples from the Gaussian distribution $\mathcal{N}(0, \mathbf{\Sigma})$, and let S be the sample covariance matrix (assuming it is known that the mean is zero),

$$S = \frac{1}{n} \sum_{j=1}^{n} \mathbf{x}_j \mathbf{x}_j^T$$

Suppose that $d = 5$ and $n > d$. It turns out that the matrix S has 3 eigenvalues which are *exactly* zero. What does this mean ? Is it possible that $\mathbf{\Sigma}$ is invertible, and why ? (no need for rigorous proof here).

3. Suppose that $\mathbf{\Sigma}$ is a rank one perturbation of the identity matrix,

$$\mathbf{\Sigma} = \lambda \mathbf{v}\mathbf{v}^T + I_d \tag{3}$$

with $\|\mathbf{v}\|_2 = 1$. What is the first principal component $\mathbf{w}$ of $\mathbf{\Sigma}$ and what is the corresponding variance $Var[\mathbf{w}^T\mathbf{x}]$?

4. Assume the covariance structure of Eq. (3). Show that the sample variance in the direction of $\mathbf{v}$, namely $\mathbf{v}^T S \mathbf{v}$ is a random variable with distribution $\frac{(\lambda+1)}{n}\chi_n^2$.
(Recall that if $Z_1, \ldots, Z_n$ are i.i.d. $\mathcal{N}(0, 1)$ random variables then $\sum_{i=1}^{n} Z_i^2 \sim \chi_n^2$).

5. Eq. (3) is a common model of the covariance structure in multi-antenna communication channels, where λ is then the signal-to-noise ratio, which is a measure of the quality of the channel. Suppose that we observe two datasets, $\{\mathbf{x}_i\}_{i=1}^n$ and $\{\mathbf{y}_i\}_{i=1}^n$, both of which follow the covariance model Eq. (3), with $n = 200$ and $d = 5$. For example, the $\mathbf{x}_i$ are samples in one time slot and $\mathbf{y}_i$ are samples in the next time slot. We compute the sample covariance of both datasets, S_x and S_y and their eigenvalues. We observe that eigenvalues 2,3,4,5 in both cases are close to 1 (as expected) but the largest eigenvalue of S_x, $\lambda_1^y = 46.3$, whereas $\lambda_1^x = 51.2$. This might indicate that the communication channel SNR has deteriorated, but may also be due to random fluctuations around a value of say $\lambda = 50$. How likely is the second scenario? What if $n = 20,000$ instead ? Explain your answer.