Lecture 3

Basic Knowledge. Multivariate normal (distribution, properties, estimation, bound, CLT, large deviation...maybe 2 lectures).

1 Normal Distribution

Normal distribution \( N(\mu, \sigma^2) \): random variable \( X \) follows normal distribution with parameters \( \mu \) and \( \sigma^2 \) if the density function of \( X \) is given by

\[
f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}.\]

It is known that \( E(X) = \mu \) and \( Var(X) = \sigma^2 \).

If \( X \sim N(\mu, \sigma^2) \) then for any \( a, b \in \mathbb{R} \), \( aX + b \sim N(a\mu + b, a^2\sigma^2) \).

Suppose \( X = (X_1, \cdots, X_k) \) and \( X_i \) are i.i.d. standard normal random variables. Then it is obviously that

\[
E(X) = (0, 0, \cdots, 0), \quad COV(X) = I_k.
\]

Then for a \( n \) dimensional vector \( \mu \) and \( n \times k \) matrix \( A \)

\[
E(\mu + AX) = \mu, COV(\mu + AX) = AA^T.
\]

Denote \( AA^T \) by \( \Sigma \), we have the following definition.

**Definition 1** The distribution of random vector \( AX \) is called a multivariate normal distribution with covariance matrix \( \Sigma \) and is denoted by \( N(0, \Sigma) \). And the distribution of \( \mu + AX \) is called a multivariate normal distribution with mean \( \mu \) and covariance matrix \( \Sigma \), \( N(\mu, \Sigma) \).

To make the definition valid, we need to verify that the distribution of \( AX \) depend on \( A \) only through \( AA^T \). We can use the moment generating function to do this.
Suppose the moment generating function of $X$ is $M(t)$, we know that $M(t) = e^{t^T t/2}$.

Therefore the m.g.f. of $AX$ is

$$M_{AX}(t) = E(e^{t^T AX}) = M(t^T A) = e^{t^T A A^T t}.$$ 

This means the m.g.f. of $AX$ depend on $A$ only through $A A^T$, so the distribution of $AX$ only depends on $A A^T$.

Based on the definition, we can also calculate the joint pdf of $N(\mu, \Sigma)$ (when $\Sigma$ is full rank),

$$f(x) = \left(\frac{1}{\sqrt{2\pi}}\right)^n (\det|\Sigma|)^{-1/2} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

where $x = (x_1, \cdots, x_n)^T$ is a $n$ dimensional vector. We can also see that if $Y$ follows $N(\mu, \Sigma)$ distribution then for any matrix $B$

$$BY \sim N(B\mu, B\Sigma B^T).$$

Suppose $X_1, X_2, \cdots, X_n$ are i.i.d. standard normal random variables, then $Y = \sum_{i=1}^{n} X_i^2$ follows $\chi^2$ distribution with $n$ degrees of freedom (also $\Gamma$ distribution with parameter $n/2$ and $1/2$). It is easy to verify that $E(Y) = n$, $Var(Y) = 2n$.

2 Central Limit Theorem

**Theorem 1** Suppose $X_1, X_2, \cdots, X_n$ are i.i.d. random variables with finite mean $\mu$ and variance $\sigma^2 > 0$. Let $Z_n = \frac{\sum_{i=1}^{n} X_i - n \mu}{\sigma \sqrt{n}}$, then the distribution of $Z_n$ converge to $N(0, 1)$ in distribution, i.e.

$$\lim_{n \to \infty} P(Z_n \leq z) = \Phi(z).$$

We also have the following Theorem to control the tail probability.
Theorem 2 Suppose $\epsilon_i, i = 1, 2, \cdots$, are independent random variables with $E\epsilon_i = 0$, $Var(\epsilon_i) = v_i \leq v_0$ for all $i$. Moreover, suppose the moment generating function $M_i(x) \triangleq E(\exp(x\epsilon_i))$ exists when $|x| < \rho$ for some $\rho > 0$ and all $i$. Let

$$Z_m = \frac{1}{\sqrt{v_0}} \sum_{i=1}^{m} a_{mi} \epsilon_i$$

with $\sum_{i=1}^{m} a_{mi}^2 = 1$ and $|a_{mi}| \leq c_0/\sqrt{m}$ for some constant $c_0$, then for $\lambda = o(m^{-1/4})$ and sufficiently large $m$

$$\frac{P(|Z_m| > \sigma_m \lambda)}{2(1 - \Phi(\lambda))} \leq \exp(C \frac{\lambda^3}{m^{1/4}}(1 + O(m^{-1/4})))$$

where $\sigma_m^2 = \sum a_{mi}^2 v_i/v_0$ and $C > 0$ is a constant.

3 Estimation of Parameters

In this part we introduce the parameter estimate of normal distribution. Suppose we observe $x = (x_1, x_2, \cdots, x_p)^T \sim N(\mu, \Sigma)$ where $\mu = (\mu_1, \cdots, \mu_p)^T$. Our primary goal is estimation of the mean vector. Let $\delta : R^n \to R^n$ denote a estimator of $\theta$.

Quadratic loss is the most commonly used loss function

$$L(\mu, \delta) = \|\mu - \delta\|^2_2.$$ 

Corresponding to any loss function such as the quadratic loss there is also a risk function defined by

$$R(\mu, \delta) = E_{\mu, \Sigma}(L(\mu, \delta)).$$

This risk function depends on both the true parameter and the estimator.
A procedure $\delta_1$ is inadmissible if

$$R(\mu, \delta_2) \leq R(\mu, \delta_1), \text{ for any } \mu, \text{ and } < \text{ for some } \mu.$$ 

A procedure is admissible if it is not inadmissible.

The supremum of the risk of a procedure is denoted by $\bar{R}(\delta) = \sup_{\mu, \Sigma} R(\mu, \delta)$. A procedure is called minimax if this value is as small as possible. That is, $\delta_m$ is minimax if

$$\bar{R}(\delta_m) = \inf_\delta \bar{R}(\delta_m).$$

Now we look at this problem from a Bayes point of view. Suppose there is a prior distribution $P_\mu$ of $\mu$. Then the Bayes risk of a procedure is defined as

$$R_B(\delta, P_\mu) = \int R(\mu, \delta) dP_\mu.$$ 

A procedure is called Bayes procedure if it minimizes the Bayes risk (with respect to some prior $P_\mu$).

**Theorem 3** *Under the quadratic loss, the ordinary estimator $\delta_0(x) = x$ is minimax.*