Lecture 3. Inference about multivariate normal distribution

3.1 Point and Interval Estimation

Let \( X_1, \ldots, X_n \) be i.i.d. \( N_p(\mu, \Sigma) \). We are interested in evaluation of the maximum likelihood estimates of \( \mu \) and \( \Sigma \). Recall that the joint density of \( X_1 \) is

\[
 f(x) = |2\pi\Sigma|^{-\frac{1}{2}} \exp \left[ -\frac{1}{2}(x - \mu)'\Sigma^{-1}(x - \mu) \right],
\]

for \( x \in \mathbb{R}^p \). The negative log likelihood function, given observations \( x_1^n = \{x_1, \ldots, x_n\} \), is then

\[
 \ell(\mu, \Sigma | x_1^n) = \frac{n}{2} \log |\Sigma| + \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)'\Sigma^{-1}(x_i - \mu) + c
\]

\[
 = \frac{n}{2} \log |\Sigma| + \frac{n}{2}(\bar{x} - \mu)'\Sigma^{-1}(\bar{x} - \mu) + \frac{1}{2} \sum_{i=1}^{n} (x_i - \bar{x})'\Sigma^{-1}(x_i - \bar{x}).
\]

On this end, denote the centered data matrix by \( \tilde{X} = [\tilde{x}_1, \ldots, \tilde{x}_n]_{p \times n} \), where \( \tilde{x}_i = x_i - \bar{x} \).

Let

\[
 S_0 = \frac{1}{n} \tilde{X}\tilde{X}' = \frac{1}{n} \sum_{i=1}^{n} \tilde{x}_i\tilde{x}_i'.
\]

**Proposition 1.** The m.l.e. of \( \mu \) and \( \Sigma \), that jointly minimize \( \ell(\mu, \Sigma | x_1^n) \), are

\[
 \hat{\mu}^{MLE} = \bar{x},
\]

\[
 \hat{\Sigma}^{MLE} = S_0.
\]

Note that \( S_0 \) is a biased estimator of \( \Sigma \). The sample variance–covariance matrix \( S = \frac{n}{n-1}S_0 \) is unbiased.

For interval estimation of \( \mu \), we largely follow Section 7.1 of Härdle and Simar (2012). First note that since \( \mu \in \mathbb{R}^p \), we need to generalize the notion of intervals (primarily defined for \( \mathbb{R}^1 \)) to higher dimension. A simple extension is a direct product of marginal intervals: for intervals \( a < x < b \) and \( c < y < d \), we obtain a rectangular region \( \{(x, y) \in \mathbb{R}^2 : a < x < b, c < y < d \} \).

A confidence region \( A \in \mathbb{R}^p \) is composed of the values of a function of (random) observations \( X_1, \ldots, X_n \). \( A = A(X_1^n) \) is a confidence region of size \( 1 - \alpha \in (0, 1) \) for parameter \( \mu \) if

\[
 P(\mu \in A) \geq 1 - \alpha, \quad \text{for all } \mu \in \mathbb{R}^p.
\]

(Elliptical confidence region) Corollary 7 in lecture 2 provides a pivot which paves a way to construct a confidence region for \( \mu \). Since \( \frac{n-p}{p}(\bar{X} - \mu)'S_0^{-1}(\bar{X} - \mu) \sim F_{p,n-p} \) and

\[
 P \left( (\bar{X} - \mu)'S_0^{-1}(\bar{X} - \mu) < \frac{p}{n-p} F_{1-\alpha,p,n-p} \right) = 1 - \alpha,
\]

\[
 A = \left\{ \mu \in \mathbb{R}^p : (\bar{X} - \mu)'S_0^{-1}(\bar{X} - \mu) < \frac{p}{n-p} F_{1-\alpha,p,n-p} \right\}
\]
is a confidence region of size $1 - \alpha$ for parameter $\mu$.

(Simultaneous confidence intervals) Simultaneous confidence intervals for all linear combinations of elements of $\mu$, $a'\mu$ for arbitrary $a \in \mathbb{R}^p$, provides confidence of size $1 - \alpha$ for all intervals covering $a'\mu$ including the marginal means $\mu_1, \ldots, \mu_p$. We are interested in evaluating lower and upper bounds $L(a)$ and $U(a)$ satisfying

$$P(L(a) < a'\mu < U(a), \text{ for all } a \in \mathbb{R}^p) \geq 1 - \alpha, \text{ for all } \mu \in \mathbb{R}^p.$$}

First consider a single confidence interval by fixing a particular vector $a$. To evaluate a confidence interval for $a'\bar{\mu}$, write new random variables $Y_i = a'X_i \sim N_1(a'\mu, a'\Sigma a)$ ($i = 1, \ldots, n$), whose squared t-statistic is $t^2(a) = n\frac{(a'\bar{\mu} - a'\bar{X})^2}{a'Sa} \sim F_{1, n-1}$. Thus, for any fixed $a$,

$$P(t^2(a) \leq F_{1-\alpha, n-1}) = 1 - \alpha. \quad (1)$$

Next, consider many projection vectors $a_1, a_2, \ldots, a_M$ ($M$ is finite only for convenience). The simultaneous confidence intervals of the type similar to (1) are then

$$P \left( \bigcap_{i=1}^{M} \{ t^2(a_i) \leq h(\alpha) \} \right) \geq 1 - \alpha,$$

for some $h(\cdot)$. By collecting some facts,

1. $\max_a t^2(a) \leq h(\alpha)$ implies $t^2(a_i) \leq h(\alpha)$ for all $i$.
2. $\max_a t^2(a) = n(\mu - \bar{X})'S^{-1}(\mu - \bar{X})$,
3. and Corollary 7 in lecture 2,

we have for $h(\alpha) = \frac{n-1}{n} \frac{p}{n-p} F_{1-\alpha, p, n-p}$,

$$P \left( \bigcap_{i=1}^{M} \{ t^2(a_i) \leq h(\alpha) \} \right) \geq P \left( \max_a t^2(a) \leq h(\alpha) \right) = 1 - \alpha.$$

**Proposition 2.** Simultaneously for all $a \in \mathbb{R}^p$, the interval

$$a'\bar{X} \pm \sqrt{h(\alpha)a'Sa}$$

contains $a'\mu$ with probability $1 - \alpha$.

**Example 1.** From the Golub gene expression data, with dimension $d = 7129$, take the first and 1674th variables (genes), to focus on the bivariate case ($p = 2$). There are two populations: 11 observations from AML, 27 from ALL. Figure 1 illustrates the elliptical confidence region of size 95% and 99%. Figure 2 compares the elliptical confidence region with the simultaneous confidence intervals for $a_1 = (1, 0)'$ and $a_2 = (0, 1)'$. 

---

2
3.2 Hypotheses testing

Consider testing a null hypothesis $H_0 : \theta \in \Omega_0$ against an alternative hypothesis $H_1 : \theta \in \Omega_1$. The principle of likelihood ratio test is as follows: Let $L_0$ be the maximized likelihood under
$H_0 : \theta \in \Omega_0$, and $L_1$ be the maximized likelihood under $\theta \in \Omega_0 \cup \Omega_1$. The likelihood ratio statistic, or sometimes called Wilks statistic, is then

$$W = -2 \log \left( \frac{L_0}{L_1} \right) \geq 0$$

The null hypothesis is rejected if the observed value of $W$ is large. In some cases the exact distribution of $W$ under $H_0$ can be evaluated. In other cases, Wilks’ theorem states that for large $n$ (sample size),

$$W \overset{d}{\to} \chi^2_\nu,$$

where $\nu$ is the number of free parameters in $H_1$, not in $H_0$. If the degrees of freedom in $\Omega_0$ is $q$ and the degrees of freedom in $\Omega_0 \cup \Omega_1$ is $r$, then $\nu = r - q$.

Consider testing hypotheses on $\mu$ and $\Sigma$ of multivariate normal distribution, based on $n$-sample $X_1, \ldots, X_n$.

**case I:** $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$, $\Sigma$ is known.

In this case, we know the exact distribution of the likelihood ratio statistic

$$W = n(\bar{x} - \mu_0)'\Sigma^{-1}(\bar{x} - \mu_0) \sim \chi^2_p,$$

under $H_0$.

**case II:** $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$, $\Sigma$ is unknown.

The m.l.e. under $H_1$ are $\hat{\mu} = \bar{x}$ and $\hat{\Sigma} = S_0$. The restricted m.l.e. of $\Sigma$ under $H_0$ is

$$\hat{\Sigma}(0) = \frac{1}{n} \sum_{i=1}^n (x_i - \mu_0)(x_i - \mu_0)' = S_0 + \delta\delta'$$

where $\delta = \sqrt{n}(\bar{x} - \mu_0)$. The likelihood ratio statistic is then

$$W = n \log |S_0 + \delta\delta'| - n \log |S_0|.$$

It turns out that $W$ is a monotone increasing function of

$$\delta' S^{-1} \delta = n(\bar{x} - \mu_0) S^{-1}(\bar{x} - \mu_0),$$

which is the Hotelling’s $T^2(n-1)$ statistic.

**case III:** $H_0 : \Sigma = \Sigma_0$, $H_1 : \Sigma \neq \Sigma_0$, $\mu$ is unknown.

We have the likelihood ratio statistic

$$W = -n \log |\Sigma_0^{-1}S_0| - np + n \text{trace}(\Sigma_0^{-1}S_0).$$

This is the case where the exact distribution of $W$ is difficult to evaluate. For large $n$, use Wilks' theorem to approximate the distribution of $W$ by $\chi^2_p$ with the degrees of freedom $\nu = p(p + 1)/2$.

Next, consider testing the equality of two mean vectors. Let $X_{11}, \ldots, X_{1n_1}$ be i.i.d. $N_p(\mu_1, \Sigma)$ and $X_{21}, \ldots, X_{2n_2}$ be i.i.d. $N_p(\mu_2, \Sigma)$. 

case IV: $H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 \neq \mu_2, \quad \Sigma$ is unknown.
Since

$$\bar{X}_1 - \bar{X}_2 \sim N_p(\mu_1 - \mu_2, \frac{n_1 + n_2}{n_1 n_2} \Sigma),$$

$$(n_1 + n_2 - 2) S_p \sim W_p(n_1 + n_2 - 2, \Sigma),$$

we have Hotelling’s $T^2$ statistic for two-sample problem

$$T^2(n_1 + n_2 - 2) = \frac{n_1 n_2}{n_1 + n_2} (\bar{X}_1 - \bar{X}_2)'S_p^{-1}(\bar{X}_1 - \bar{X}_2),$$

and by Theorem 5 in lecture 2

$$\frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p} T^2(n_1 + n_2 - 2) \sim F_{p,n_1+n_2-p-1}.$$

Similar to case II above, the likelihood ratio statistic is a monotone function of $T^2(n_1 + n_2 - 2)$.

### 3.3 Hypothesis testing when $p > n$

In the high dimensional situation where the dimension $p$ is larger than sample size ($p > n - 1$ or $p > n_1 + n_2 - 2$), the sample covariance $S$ is not invertable, thus the Hotelling’s $T^2$ statistic, which is essential in the testing procedures above, cannot be computed. We survey important proposals for testing hypotheses on means in high dimension, low sample size data.

A basic idea in generalizing a test procedure for the $p > n$ case is to base the test on a computable test statistic which is also an estimator for $\|\mu - \mu_0\|$ or $\|\mu_1 - \mu_2\|$.

In case II (one sample), Dempster (1960) proposed to replace $S^{-1}$ in Hotelling’s statistic by $(\text{trace}(S)I_p)^{-1}$. He showed that under $H_0 : \mu = 0$,

$$T_D = \frac{n \bar{X}'X}{\text{trace}(S)} \sim F_{r,(n-1)r},$$

approximately,

for $r = \frac{(\text{trace}(\Sigma))^2}{\text{trace}(\Sigma^2)}$, a measure of sphericity of $\Sigma$. An estimator $\hat{r}$ of $r$ is used in testing.

Bai and Saranadasa (1996) proposed to simply replace $S^{-1}$ in Hotelling’s statistic by $I_p$, yielding $T_B = n \bar{X}'X$. However $\bar{X}'X$ is not an unbiased estimator of $\mu'\mu$ since $E(\bar{X}'X) = \frac{1}{n} \text{trace}(\Sigma) + \mu'\mu$. They showed that the standardized statistic

$$M_B = \frac{n \bar{X}'X - \text{trace}(S)}{\text{sd}(n \bar{X}'X - \text{trace}(S))} = \frac{n \bar{X}'X - \text{trace}(S)}{\sqrt{(\frac{2(n-1)}{(n-2)(n+1)} (\text{trace}(S^2) - \frac{1}{n} \text{trace}(S))^2)}}$$

has asymptotic $N(0, 1)$ distribution for $p, n \to \infty$. 
Srivastava and Du (2008) proposed to replace $S^{-1}$ in Hotelling’s statistic by $D_S = \text{diag}(S)$. Then $T_S = n\bar{X}'D_S^{-1}\bar{X} - \frac{n-1}{n-3}p$ can be used to estimate $\frac{n(n-1)}{n-3}\|D_S^{\frac{1}{2}}\mu\|^2$, which is zero under $H_0: \mu = 0$. Srivastava and Du’s test statistic is then

$$M_S = \frac{T_S}{\text{sd}(T_S)} = \frac{n\bar{X}'D_S^{-1}\bar{X} - \frac{n-1}{n-3}p}{\sqrt{2\text{trace}(R^2) - \frac{p^2}{n-1}}}$$

which has asymptotic $N(0, 1)$ distribution for $p, n \to \infty$. Here $R = D_S^{-\frac{1}{2}}S^{-\frac{1}{2}}D_S^{-\frac{1}{2}}$ is the sample correlation matrix.

Chen and Qin (2010) improves the two-sample test for mean vectors from that of Bai and Saranadasa (1996). In testing $H_1: \mu_1 = \mu_2$, Bai and Saranadasa (1996) proposed to use $T_B = \bar{X}_1'\bar{X}_2 - \frac{n_1+n_2}{n_1n_2}\text{trace}(S_P)$. The substraction of $\text{trace}(S_P)$ is to make sure that $E(T_B) = \|\mu_1 - \mu_2\|^2$. Chen and Qin (2010) proposed to not use $\text{trace}(S_P)$, by considering

$$T_C = \frac{\sum_{i \neq j}^{n_1} X_{1i}'X_{1j}}{n_1(n_1-1)} + \frac{\sum_{i \neq j}^{n_2} X_{2i}'X_{2j}}{n_2(n_2-1)} - 2\frac{\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} X_{1i}'X_{2j}}{n_1n_2}$$

Since $E(T_C) = \|\mu_1 - \mu_2\|^2$, Chen and Qin proposed to test based on $T_C$.

There are many other ideas, including:

1. The test statistic is essentially the maximum of $p$ normalized marginal mean differences (Cai et al., 2013);
2. Use a generalized inverse of $S$, denoted by $S^-$ or $S^\dagger$, to replace $S^{-1}$;
3. Estimate $\Sigma$ in a way that is invertible;
4. Reduce the dimension $p$ of the random vector $X$ by $Z = h(X) \in \mathbb{R}^d$, for $d < n$, then apply the traditional theory of hypothesis testing.

__________

Next lecture is on linear dimension reduction—principal component analysis.

References


In evaluating the MLE of $\Sigma$ for MVN, one can use the following famous result.

**Lemma 3** (von Neumann). For any $m \times m$ symmetric matrices $A$ and $B$ with eigenvalues $\sigma_A = (\sigma_{A1}, \ldots, \sigma_{Am})'$ and $\sigma_B = (\sigma_{B1}, \ldots, \sigma_{Bm})'$ in decreasing order,

$$|\text{trace}(A'B)| \leq \sigma_A'\sigma_B,$$

and the equality holds when $A$ and $B$ have the same eigenvectors.

A general version of von Neumann inequality is:

**Lemma 4** (von Neumann). For any $m \times n$ matrices $A$ and $B$ with vectors of singular values $\sigma_A$ and $\sigma_B$ in decreasing order,

$$|\text{trace}(A'B)| \leq \sigma_A'\sigma_B,$$

and the equality holds when $A$ and $B$ are simultaneously diagonizable.

The problem was to minimize the negative log-likelihood $\log |\Sigma| + \text{trace}(\Sigma^{-1}S_0)$. The parameter, assumed to be nonnegative definite, is eigen-decomposed into $U\Lambda U'$. Likewise, we can eigen-decompose the real symmetric matrix $S_0 = VD V'$.

$$\log |\Sigma| + \text{trace}(\Sigma^{-1}S_0) = \log |\Lambda| + \text{trace}(U\Lambda^{-1}U'VDV') \geq \log |\Lambda| + \text{trace}(\Lambda^{-1}D) = \sum_{i=1}^{p} \log(\lambda_i) + \text{trace}(d_i/\lambda_i) = \sum_{i=1}^{p} (a_i - \log(a_i)) + \log(d_i),$$

where $a_i = d_i/\lambda_i$, and minimized when $a_i = 1$. 

7