Lecture 3. Inference about multivariate normal distribution

3.1 Point and Interval Estimation

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be i.i.d. $N_p(\boldsymbol{\mu}, \Sigma)$. We are interested in evaluation of the maximum likelihood estimates of $\boldsymbol{\mu}$ and Σ. Recall that the joint density of $\boldsymbol{X}_1$ is

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

for $\boldsymbol{x} \in \mathbb{R}^p$. The negative log likelihood function, given observations $\boldsymbol{x}_1^n := \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, up to an additive constant independent of $\boldsymbol{\mu}$ and Σ, is then

$$\begin{aligned}
\ell^*(\boldsymbol{\mu}, \Sigma \mid \boldsymbol{x}_1^n) &= \frac{n}{2}\log|\Sigma| + \frac{1}{2}\sum_{i=1}^n (\boldsymbol{x}_i - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}) \\
&= \frac{n}{2}\log|\Sigma| + \frac{n}{2}(\bar{\boldsymbol{x}} - \boldsymbol{\mu})'\Sigma^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}) + \frac{1}{2}\sum_{i=1}^n (\boldsymbol{x}_i - \bar{\boldsymbol{x}})'\Sigma^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}).
\end{aligned}$$

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

$$\mathbf{S}_0 = \frac{1}{n}\tilde{\mathbf{X}}\tilde{\mathbf{X}}' = \frac{1}{n}\sum_{i=1}^n \tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i'.$$

Proposition 1. *The MLEs of $\boldsymbol{\mu}$ and Σ, that jointly minimize $\ell^*(\boldsymbol{\mu}, \Sigma \mid \boldsymbol{x}_1^n)$, are*

$$\begin{aligned}
\hat{\boldsymbol{\mu}}^{MLE} &= \bar{\boldsymbol{x}}, \\
\widehat{\Sigma}^{MLE} &= \mathbf{S}_0.
\end{aligned}$$

Note that $\mathbf{S}_0$ is a biased estimator of Σ. The sample variance–covariance matrix $\mathbf{S} = \frac{n}{n-1}\mathbf{S}_0$ is unbiased.

For interval estimation of $\boldsymbol{\mu}$, we largely follow Section 7.1 of Härdle and Simar (2012). First note that since $\boldsymbol{\mu} \in \mathbb{R}^p$, we need to generalize the notion of intervals (primarily defined for $\mathbb{R}^1$) to a higher dimensional space. A simple extension is a direct product of marginal intervals: for example, 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 is a subset of $\mathbb{R}^p$ and it depends on (random) observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. $A = A(\boldsymbol{X}_1^n)$ is a confidence region of size $1 - \alpha \in (0, 1)$ for parameter $\boldsymbol{\mu} \in \mathbb{R}^p$ if

$$P(\boldsymbol{\mu} \in A(\boldsymbol{X}_1^n)) \geq 1 - \alpha$$

where P is with respect to the distribution of $\boldsymbol{X}_1^n$.

(**Elliptical confidence region**) Corollary 7 in lecture 2 provides a pivot which paves a way to construct a confidence region for $\boldsymbol{\mu}$.

Since $\frac{n-p}{p}(\bar{\boldsymbol{X}}-\boldsymbol{\mu})'\mathbf{S}_0^{-1}(\bar{\boldsymbol{X}}-\boldsymbol{\mu}) \sim F_{p,n-p}$ and

$$P\left((\bar{\boldsymbol{X}}-\boldsymbol{\mu})'\mathbf{S}_0^{-1}(\bar{\boldsymbol{X}}-\boldsymbol{\mu}) < \frac{p}{n-p}F_{1-\alpha;p,n-p}\right) = 1-\alpha,$$

we have that

$$A := \left\{\boldsymbol{\nu} \in \mathbb{R}^p : (\bar{\boldsymbol{X}}-\boldsymbol{\nu})'\mathbf{S}_0^{-1}(\bar{\boldsymbol{X}}-\boldsymbol{\nu}) < \frac{p}{n-p}F_{1-\alpha;p,n-p}\right\}$$

is a confidence region of size $1-\alpha$ for parameter $\boldsymbol{\mu}$.

Note: $F_{1-\alpha;p,n-p}$ is defined as the constant such that $P(F < F_{1-\alpha;p,n-p}) = 1-\alpha$ where F is $F_{p,n-p}$ distributed.

The resulting confidence region for a given sample is an ellipse.

(**Simultaneous confidence intervals**) Simultaneous confidence intervals for all linear combinations of elements of $\boldsymbol{\mu}$, i.e., $\boldsymbol{a}'\boldsymbol{\mu}$ for arbitrary $\boldsymbol{a} \in \mathbb{R}^p$, provide confidence of size $1-\alpha$ for any interval in this framework, which covers $\boldsymbol{a}'\boldsymbol{\mu}$. This includes the marginal means $\mu_1, \dots, \mu_p$ (when choosing $\boldsymbol{a}$ wisely). We are interested in evaluating lower and upper bounds $L(\boldsymbol{a})$ and $U(\boldsymbol{a})$ satisfying

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

Note: the lower and upper limits of the interval depend on the data (hence is random) and is also specified by $\boldsymbol{a}$.

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

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

Next, consider many projection vectors $\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_M$ (M is finite only for convenience). The simultaneous confidence intervals of the type similar to (1) are then

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

for some $h(\cdot)$. We collect some facts below.

1. $\max_{\boldsymbol{a}} t^2(\boldsymbol{a}) \leq h(\alpha)$ implies $t^2(\boldsymbol{a}_i) \leq h(\alpha)$ for all i.

2. $\max_{\boldsymbol{a}} t^2(\boldsymbol{a}) = n(\boldsymbol{\mu}-\bar{\boldsymbol{X}})'\mathbf{S}^{-1}(\boldsymbol{\mu}-\bar{\boldsymbol{X}})$, HS Theorem 2.5 (note the rank)

3. Corollary 7 in lecture 2.

These suggest that we have $h(\alpha) = \frac{(n-1)p}{n-p} F_{1-\alpha;p,n-p}$ and

$$P\left(\bigcap_{i=1}^{M}\{t^2(\boldsymbol{a}_i) \le h(\alpha)\}\right) \ge P\left(\max_{\boldsymbol{a}} t^2(\boldsymbol{a}) \le h(\alpha)\right) = 1 - \alpha.$$

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

$$\boldsymbol{a}'\bar{\boldsymbol{X}} \pm \sqrt{h(\alpha)\boldsymbol{a}'\boldsymbol{S}\boldsymbol{a}/n}$$

contains $\boldsymbol{a}'\boldsymbol{\mu}$ *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 $\boldsymbol{a}_1 = (1, 0)'$ and $\boldsymbol{a}_2 = (0, 1)'$.

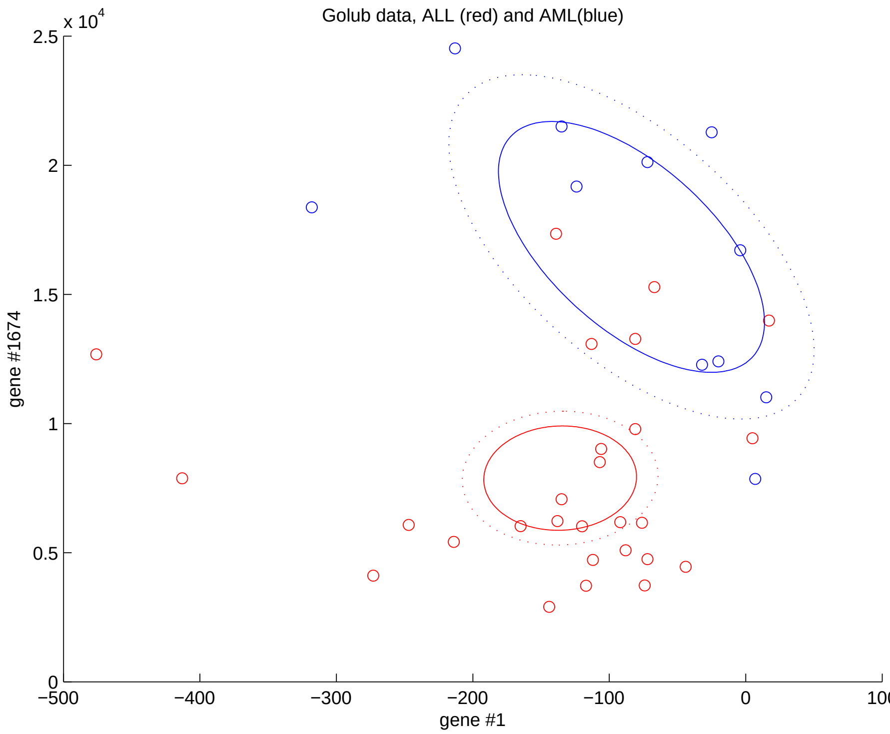

Figure 1: Elliptical confidence regions of size 95% and 99%.

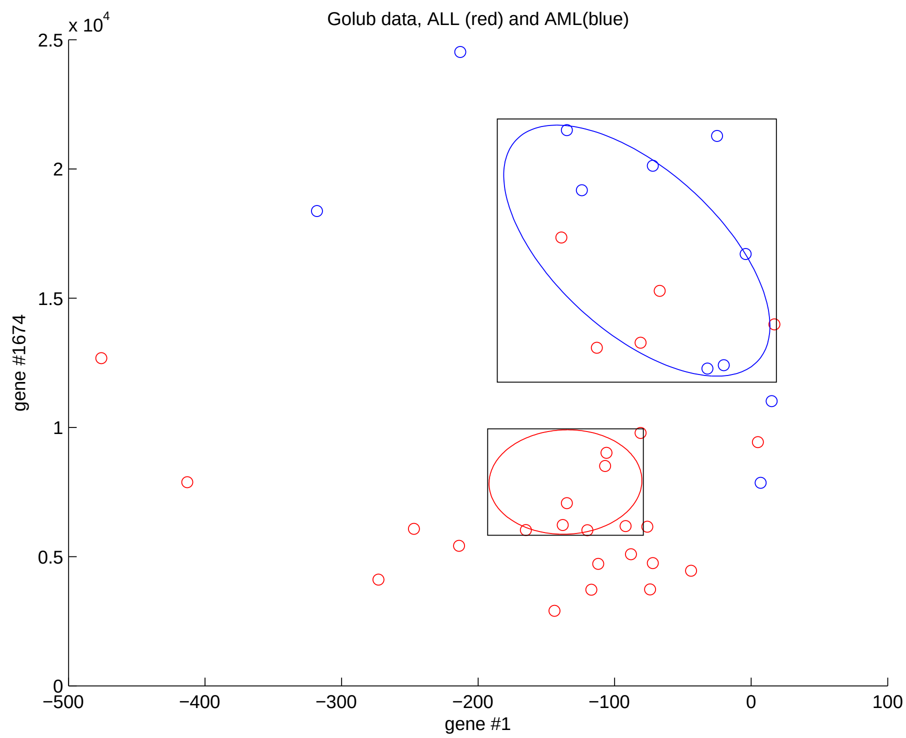

Figure 2: Simultaneous confidence intervals of size 95%

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

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

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),

$$W \overset{\mathcal{L}}{\Longrightarrow} \chi^2_\nu,$$

where ν is the number of free parameters in H_1 but not in H_0. If the degrees of freedom in Ω_0 is q and the degrees of freedom in $\Omega_0 \cup \Omega_1$ is r, then $\nu = r - q$.

Consider testing hypotheses on $\boldsymbol{\mu}$ and Σ of multivariate normal distribution, based on n-sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$.

case I: $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0, \quad H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0, \quad \Sigma$ is known.

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

$$W = n(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)'\Sigma^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0) \sim \chi^2_p,$$

under H_0.

[Show on blackboard.] It is worthwhile to first write $-2\log(L(\boldsymbol{x};\boldsymbol{\mu},\Sigma)) =$

$$-2\ell(\boldsymbol{x};\boldsymbol{\mu},\Sigma) = n\log|2\pi\Sigma| + n\cdot tr\{\Sigma^{-1}\mathbf{S}\} + n(\bar{\boldsymbol{x}}-\boldsymbol{\mu})^T\Sigma^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})$$

case II: $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0, \quad H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0, \quad \Sigma$ is unknown.

The MLEs under H_1 are $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}}$ and $\widehat{\Sigma} = \mathbf{S}_0$. The restricted MLE of Σ under H_0 is $\widehat{\Sigma}_{(0)} = \frac{1}{n}\sum_{i=1}^n(\boldsymbol{x}_i-\boldsymbol{\mu}_0)(\boldsymbol{x}_i-\boldsymbol{\mu}_0)' = \mathbf{S}_0 + \boldsymbol{\delta}\boldsymbol{\delta}'$, where $\boldsymbol{\delta} = \sqrt{n}(\bar{\boldsymbol{x}}-\boldsymbol{\mu}_0)$. The likelihood ratio statistic is then

$$W = n\log|\mathbf{S}_0 + \boldsymbol{\delta}\boldsymbol{\delta}'| - n\log|\mathbf{S}_0|.$$

It turns out that W is a monotone increasing function of

$$\boldsymbol{\delta}'\mathbf{S}^{-1}\boldsymbol{\delta} = n(\bar{\boldsymbol{x}}-\boldsymbol{\mu}_0)\mathbf{S}^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu}_0),$$

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

case III: $H_0 : \Sigma = \Sigma_0, \quad H_1 : \Sigma \neq \Sigma_0, \quad \boldsymbol{\mu}$ is unknown.

We have the likelihood ratio statistic

$$W = -n\log|\Sigma_0^{-1}\mathbf{S}_0| - np + n\text{trace}(\Sigma_0^{-1}\mathbf{S}_0).$$

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 χ^2_ν with the degrees of freedom $\nu = p(p+1)/2$.

Next, consider testing the equality of two mean vectors. Let $\boldsymbol{X}_{11},\ldots,\boldsymbol{X}_{1n_1}$ be i.i.d. $N_p(\boldsymbol{\mu}_1,\Sigma)$ and $\boldsymbol{X}_{21},\ldots,\boldsymbol{X}_{2n_2}$ be i.i.d. $N_p(\boldsymbol{\mu}_2,\Sigma)$.

case IV: $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2, \quad H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2, \quad \Sigma$ is unknown.

Since

$$\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2 \sim N_p(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2, \frac{n_1+n_2}{n_1 n_2}\Sigma),$$
$$(n_1+n_2-2)\mathbf{S}_P \sim W_p(n_1+n_2-2,\Sigma),$$

where $(n_1+n_2-2)\mathbf{S}_P := (n_1-1)\mathbf{S}_1 + (n_2-1)\mathbf{S}_2$, we have Hotelling's T^2 statistic for two-sample problem

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

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 $\mathbf{S}$ 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 (HDLSS) 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 $\|\boldsymbol{\mu} - \boldsymbol{\mu}_0\|$ or $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|$.

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

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

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

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

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

has asymptotic $N(0,1)$ distribution for $p, n \to \infty$.

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

$$M_S = \frac{T_S}{\widehat{\text{sd}}(T_S)} = \frac{n\bar{\boldsymbol{X}}'D_{\mathbf{S}}^{-1}\bar{\boldsymbol{X}} - \frac{n-1}{n-3}p}{\sqrt{2\text{trace}(\boldsymbol{R}^2) - \frac{p^2}{n-1}}},$$

which has asymptotic $N(0,1)$ distribution for $p, n \to \infty$. Here $\boldsymbol{R} = D_{\mathbf{S}}^{-\frac{1}{2}}\mathbf{S}D_{\mathbf{S}}^{-\frac{1}{2}}$ 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 $H_1 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, Bai and Saranadasa (1996) proposed to use $T_B = \bar{\boldsymbol{X}}_1'\bar{\boldsymbol{X}}_2 - \frac{n_1+n_2}{n_1n_2}\text{trace}(\mathbf{S}_P)$. The substraction of trace($\mathbf{S}_P$) is to make sure that $E(T_B) = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2$. Chen and Qin (2010) proposed to not use trace($\mathbf{S}_P$), by considering

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

Since $E(T_C) = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2$, 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 $\mathbf{S}$, denoted by $\mathbf{S}^-$ or $\mathbf{S}^\dagger$, to replace $\mathbf{S}^{-1}$;

3. Estimate Σ in a way that is invertible;

4. Reduce the dimension p of the random vector $\boldsymbol{X}$ by $\boldsymbol{Z} = h(\boldsymbol{X}) \in \mathbb{R}^d$, for $d < n$, then apply the traditional theory of hypothesis testing.

____________Next lecture is on linear dimension reduction–principal component analysis.

References

Bai, Z. and Saranadasa, H. (1996), "Effect of high dimension: by an example of a two sample problem," *Statist. Sinica*, 6, 311–329.

Cai, T. T., Liu, W., and Xia, Y. (2013), "Two-Sample Test of High Dimensional Means under Dependency," *To appear in Journal of the Royal Statistical Society: Series B.*

Chen, S. X. and Qin, Y.-L. (2010), "A two-sample test for high-dimensional data with applications to gene-set testing," *The Annals of Statistics*, 38, 808–835.

Dempster, A. P. (1960), "A significance test for the separation of two highly multivariate small samples," *Biometrics*, 16, 41–50.

Srivastava, M. S. and Du, M. (2008), "A test for the mean vector with fewer observations than the dimension," *Journal of Multivariate Analysis*, 99, 386–402.

In evaluating the MLE of Σ for MVN, one can use the following famous result.

Note:

- $\frac{\partial |\mathbf{Y}|}{\partial \mathbf{Y}} = |\mathbf{Y}|(\mathbf{Y}^{-1})^T$
- $\frac{\partial}{\partial \mathbf{Y}} \text{trace}(\mathbf{A}\mathbf{Y}^{-1}\mathbf{B}) = -(\mathbf{Y}^{-1}\mathbf{B}\mathbf{A}\mathbf{Y}^{-1})^T$

$$\frac{\partial}{\partial \Sigma}\{\log|\Sigma| + \text{trace}(\Sigma^{-1}\mathbf{S}_0)\} = |\Sigma|^{-1}|\Sigma|(\Sigma^{-1}) - (\Sigma^{-1}\mathbf{S}_0\Sigma^{-1}) = (\Sigma^{-1}) - (\Sigma^{-1}\mathbf{S}_0\Sigma^{-1})$$

The first order condition leads to $\hat{\Sigma}^{MLE} = \mathbf{S}_0$