# Lecture 3. Inference about multivariate normal distribution

## 3.1 Point and Interval Estimation

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ 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 $\mathbf{X}_1$ is

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

for $\mathbf{x} \in \mathbb{R}^p$. The negative log likelihood function, given observations $\mathbf{x}_1^n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, is then

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

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

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

**Proposition 1.** *The m.l.e. of $\mu$ and $\Sigma$, that jointly minimize $\ell(\mu, \Sigma \mid \mathbf{x}_1^n)$, are*

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

Note that $\mathbf{S}_0$ is a biased estimator of $\Sigma$. The sample variance–covariance matrix $\mathbf{S} = \frac{n}{n-1}\mathbf{S}_0$ 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 $\mathbf{X}_1, \ldots, \mathbf{X}_n$. $A = A(\mathbf{X}_1^n)$ 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 $\frac{n-p}{p}(\bar{\mathbf{X}}-\mu)'\mathbf{S}_0^{-1}(\bar{\mathbf{X}}-\mu) \sim F_{p,n-p}$ and $P\left((\bar{\mathbf{X}}-\mu)'\mathbf{S}_0^{-1}(\bar{\mathbf{X}}-\mu) < \frac{p}{n-p}F_{1-\alpha;p,n-p}\right) = 1-\alpha$,

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

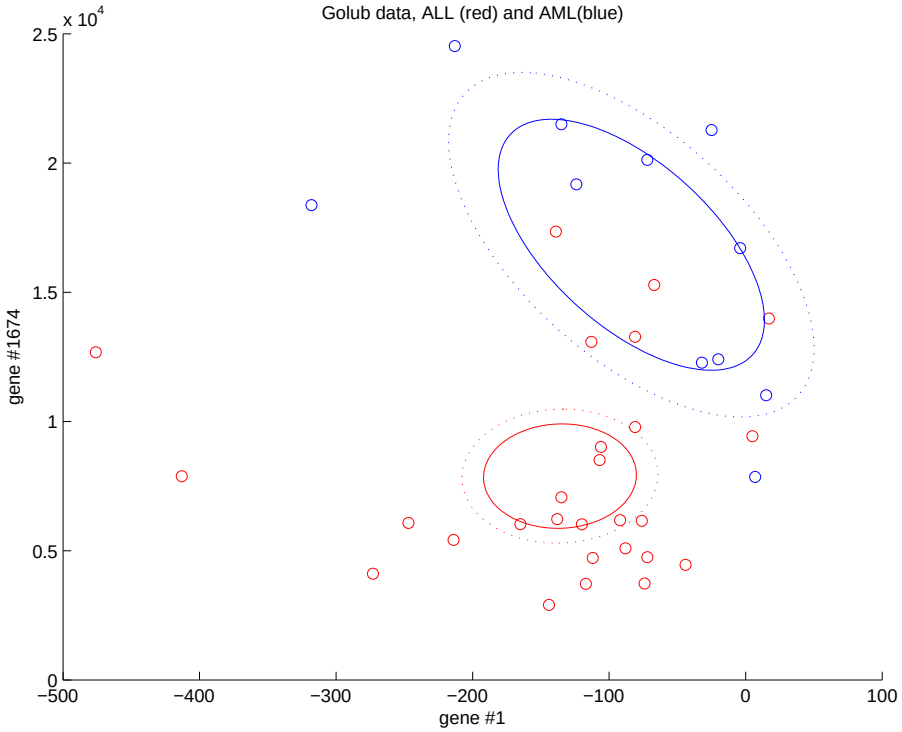


Figure 1: Elliptical confidence regions of size 95% and 99%.

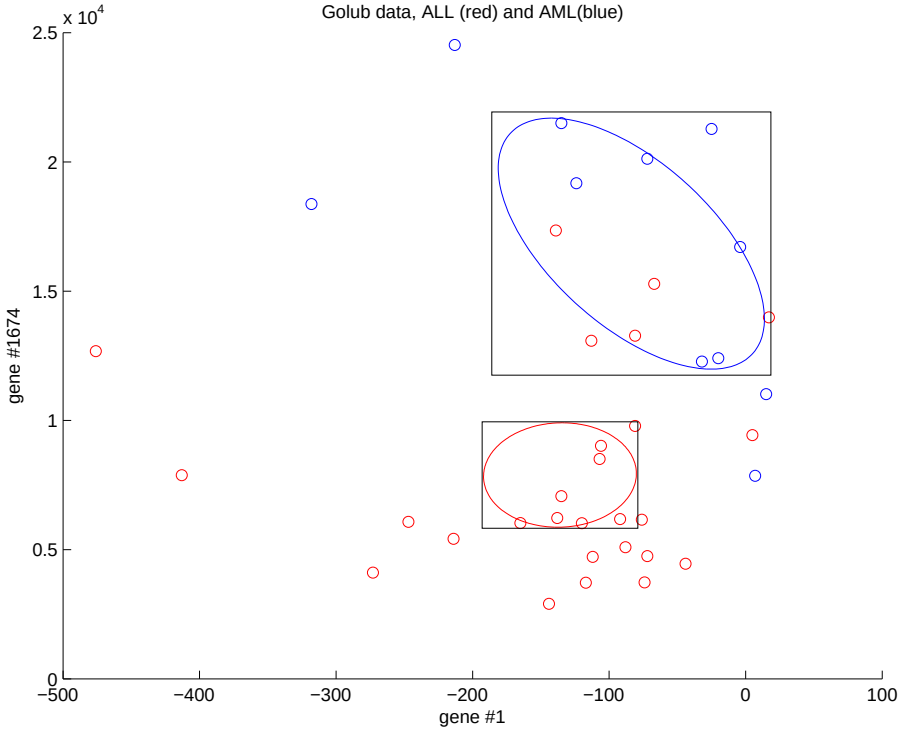


Figure 2: Simultaneous confidence intervals of size 95%

## 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),

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

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 $\mathbf{X}_1, \ldots, \mathbf{X}_n$.

**case I:** $H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0, \quad \Sigma$ is known.
In this case, we know the exact distribution of the likelihood ratio statistic

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

under $H_0$.

**case II:** $H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0, \quad \Sigma$ is unknown.
The m.l.e. under $H_1$ are $\hat{\mu} = \bar{\mathbf{x}}$ and $\widehat{\Sigma} = \mathbf{S}_0$. The restricted m.l.e. of $\Sigma$ under $H_0$ is $\widehat{\Sigma}_{(0)} = \frac{1}{n}\sum_{i=1}^n (\mathbf{x}_i - \mu_0)(\mathbf{x}_i - \mu_0)' = \mathbf{S}_0 + \boldsymbol{\delta}\boldsymbol{\delta}'$, where $\boldsymbol{\delta} = \sqrt{n}(\bar{\mathbf{x}} - \mu_0)$. 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

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

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

**case III:** $H_0 : \Sigma = \Sigma_0, \quad H_1 : \Sigma \neq \Sigma_0, \quad \mu$ 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 $\chi^2_\nu$ with the degrees of freedom $\nu = p(p+1)/2$.

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

$$\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 \sim N_p(\mu_1 - \mu_2, \frac{n_1 + n_2}{n_1 n_2}\Sigma),$$
$$(n_1 + n_2 - 2)\mathbf{S}_P \sim W_p(n_1 + n_2 - 2, \Sigma),$$

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

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

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 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 $\mathbf{S}^{-1}$ in Hotelling's statistic by $(\text{trace}(\mathbf{S})\mathbb{I}_p)^{-1}$. He showed that under $H_0 : \mu = \mathbf{0}$,

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

for $r = \frac{(\text{trace}(\Sigma))^2}{\text{trace}(\Sigma^2)}$, a measure of sphericity of $\Sigma$. 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 $T_B = n\bar{\mathbf{X}}'\bar{\mathbf{X}}$. However $\bar{\mathbf{X}}'\bar{\mathbf{X}}$ is not an unbiased estimator of $\mu'\mu$ since $E(\bar{\mathbf{X}}'\bar{\mathbf{X}}) = \frac{1}{n}\text{trace}(\Sigma) + \mu'\mu$. They showed that the standardized statistic

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

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

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

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

which has asymptotic $N(0,1)$ distribution for $p, n \to \infty$. Here $\mathbf{R} = D_{\mathbf{S}}^{-\frac{1}{2}}\mathbf{S}D_{\mathbf{S}}^{-\frac{1}{2}}$ 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 $H_1 : \mu_1 = \mu_2$, Bai and Saranadasa (1996) proposed to use $T_B = \bar{\mathbf{X}}_1'\bar{\mathbf{X}}_2 - \frac{n_1+n_2}{n_1 n_2}\text{trace}(\mathbf{S}_P)$. The substraction of $\text{trace}(\mathbf{S}_P)$ is to make sure that $E(T_B) = \|\mu_1 - \mu_2\|^2$. Chen and Qin (2010) proposed to not use $\text{trace}(\mathbf{S}_P)$, by considering

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

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 $\mathbf{S}$, denoted by $\mathbf{S}^-$ or $\mathbf{S}^\dagger$, to replace $\mathbf{S}^{-1}$;
3. Estimate $\Sigma$ in a way that is invertible;
4. Reduce the dimension $p$ of the random vector $\mathbf{X}$ by $\mathbf{Z} = h(\mathbf{X}) \in \mathbb{R}^d$, 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 $\Sigma$ 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,*

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

*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,*

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

*and the equality holds when $A$ and $B$ are simulateneously diagonazable.*

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 $S_0 = VDV'$.

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

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