Two Sample Problem for High-dimensional Data with Unequal Covariance Matrices

Takahiro Nishiyama, Masashi Hyodo

Abstract — We consider a two-sample test for the mean vectors of high-dimensional data when the dimension is large compared to the sample size. In this talk, we discuss the multivariate Behrens-Fisher problem, that is, we assume that the variance-covariance matrices are not homogeneous across groups. For these situations, we propose a Dempster type test statistic. Also, we derive asymptotic null distribution and asymptotic expansion for the upper percentiles of this statistic when both the sample size and the dimension tend to infinity. Finally, we evaluate the accuracy of approximation by Monte Carlo simulation.

Keyword: Asymptotic expansion; Behrens-Fisher problem; high-dimensional data; Monte Carlo simulation.

1 Introduction

Let \( x^{(1)}_1, x^{(1)}_2, \ldots, x^{(2)}_{N_2} \) be independent and identically distributed \( p \)-dimensional random vectors that have the multivariate normal distribution with mean vector \( \mu_k \) and covariance matrix \( \Sigma_k \). We consider testing the following hypothesis:

\[
H_0 : \mu_1 = \mu_2 \quad \text{vs.} \quad H_1 : \mu_1 \neq \mu_2 . \tag{1.1}
\]

When \( \Sigma_1 \neq \Sigma_2 \), (1.1) is well-known multivariate Behrens-Fisher problem (see, e.g., Bennett (1951), Yanagihara and Yuan (2005)).

Let \( N_1 \leq N_2 \), then we define

\[
y_i = x^{(1)}_i - \frac{N_1}{N_2} x^{(2)}_i - \frac{1}{N_1 N_2} \sum_{j=1}^{N_1} x^{(2)}_j - \frac{1}{N_2} \sum_{i=1}^{N_2} x^{(2)}_i \quad (i = 1, \ldots, N_1).
\]

The expected value of \( y_i \) and covariance matrix of \( y_i \) and \( y_j \) are

\[
E(y_i) = \mu_1 - \mu_2 \quad \text{and} \quad \text{Cov}(y_i, y_j) = \delta_{ij} \left( \Sigma_1 + \frac{N_1}{N_2} \Sigma_2 \right) \quad (\equiv \Sigma),
\]

respectively (where \( \delta_{ij} \) is Kronecker's delta). Therefore \( y_i \sim N_p(\mu, \Sigma) \) and then (1.1) is equivalent to testing the following hypothesis:

\[
H_0 : \mu = 0 \quad \text{vs.} \quad H_1 : \mu \neq 0 . \tag{1.2}
\]

For testing (1.2), when \( p < n_1 = N_1 - 1 \), following \( T^2 \) statistic is used:

\[
T^2 = N_1 \bar{y} S^{-1} \bar{y} \tag{1.3}
\]

where

\[
\bar{y} = \frac{1}{N_1} \sum_{i=1}^{N_1} y_i, \quad S = \frac{1}{n_1} \sum_{i=1}^{n_1} (y_i - \bar{y})(y_i - \bar{y})'.
\]
2 Asymptotic null distribution for high-dimensional case

In this section, we assume that $p \geq n_1$. In this case, $S$ becomes singular, and it will be impossible to use $T^2$ statistic (1.3). For such cases, a non-exact test was first proposed by Dempster (1958, 1960) for one and two sample cases. Although some statistics based on $T^2$ type statistics are proposed for this problem, we consider the statistics based on Dempster trace criterion (D-criterion) which is no restrictions about relation between the dimension $p$ and sample size $N$. However, we cannot apply D-criterion directly in these situations. So, we apply D-criterion to these situations by using the technique described in Section1 (see, Bennett (1951)). For testing (1.2), we propose the following corresponding test statistic:

$$T_D = N_1 \frac{\bar{y} \bar{y}^T}{\text{tr}S}. \quad (2.1)$$

In this talk, we discuss null distribution of the statistic (2.1) under a following high dimensional framework:

$$A1 : p, n_1 \rightarrow \infty, \quad p/n_1 \rightarrow c \in (0, \infty).$$

Further, in addition to (A1), we assume that

$$A2 : \lim_{p \rightarrow \infty} \frac{\text{tr} \Sigma^i}{p} \rightarrow c \in (0, \infty), \quad i = 1, \ldots, 4.$$

Then under the null hypothesis, we obtain the following theorem.

**Theorem 1.** Under the asymptotic framework (A1) and the assumption (A2), it holds that

$$\frac{T_D}{\sigma_D} \overset{d}{\rightarrow} N(0, 1),$$

where $\overset{d}{\rightarrow}$ denotes convergence in distribution, and

$$\overline{T}_D = \sqrt{p} \left\{ \frac{N_1 \bar{y} \bar{y}^T}{\text{tr}S} - 1 \right\}, \quad \sigma_D^2 = \frac{2a_2}{a_1^2} = \frac{\sqrt{2p \Sigma^2 / p}}{\text{tr}S^2 / p}.$$

When $\Sigma$ is unknown, we note that use following consistent estimators instead of $a_1$ and $a_2$:

$$\overline{a}_1 = \frac{\text{tr}S}{p}, \quad \overline{a}_2 = \frac{n_1^2 \left[ \text{tr}S^2 - (\text{tr}S)^2 / n_1 \right]}{p(n_1 - 1)(n_1 + 2)}.$$

Also, we derive asymptotic expansion for the upper percentiles of $\overline{T}_D$ statistic for the cases that $\Sigma$ is known and unknown, and we evaluate the accuracy of approximation by Monte Carlo simulation.

**References**


