Research Colloquium for
Science UG Students 2021-22

Name: Du Liheng
University No.: 3035532870
Major in Statistics
Supervisor: Prof. Jeff Jianfeng Yao

High Dimensional Two Sample Significance Test (Same Wishart Matrix)

Du Liheng

Department of Statistics and Actuarial Science, HKU

ABSTRACT

In this study, we focus on the two samples mean test. With high dimensional dataset, classical Hotelling's T^2 test is undefined. We then examine two more tests proposed by Dempster (1958) and Bai and Saranadasa (1996). A new way to find Dempster's test matrix and non-central parameter is proved and shown. We also conduct a simulation comparison of these methods based on their asymptotic power function to visualize the outcomes.

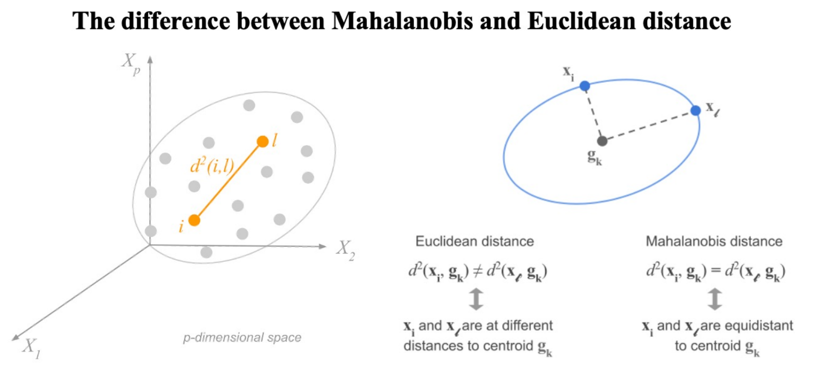
INTRODUCTION

In traditional data analysis, we assume many observations and a few well-selected variables to explain the phenomenon. For multi-linear cases, Hotelling's T^2 test serves as a good tool since it has many robust properties like invariance. $x_1, \cdots, x_{n_1}$ and $y_1, \cdots, y_{n_2}$ are two p dimension samples i.i.d. following $N(\mu_i, \Sigma), i = 1, 2$. To test $H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$, we have

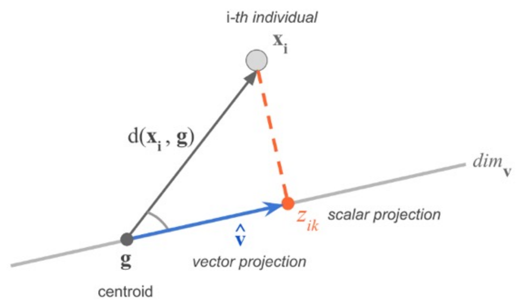
$$T^2 = \frac{n_1 n_2}{n_1 + n_2}(\bar{x} - \bar{y})' S^{-1}(\bar{x} - \bar{y})$$

where the statistic T^2 is the square of the Mahalanobis distance between the two-sample means. Rescheduling it using the Wishart distribution properties, we then can have our result since $\frac{(n_1+n_2-p-1)T^2}{p(n_1+n_2-2)}$ follows a F distribution with d.f. p and $n_1 + n_2 - p - 1$. Under alternative hypothesis, the distribution is non-central with a non-centrality parameter $\lambda = \frac{n_1 n_2}{n_1+n_2}\mu'\Sigma^{-1}\mu$, $\mu = \mu_1 - \mu_2$.

The difference between Mahalanobis and Euclidean distance



However, when it comes to high dimensional cases, growth of dimensionality brings problems like undefined inverse of Wishart matrix. Hotelling's test is undefined since the Wishart matrix is no longer singular. To solve this problem, Dempster has proposed another method. It is mathematically more complex but shares same setting as Hotelling's. It simply replaces the undefined Wishart matrix by creating a new statistic. It is a ratio between the Euclidean distance of two sample means and an average random chosen projection distance.



To reach this goal, we first arrange all the data into a $p \times n$ matrix $\boldsymbol{Y} = (x_1, \cdots, x_{n_1}, y_1, \cdots, y_{n_2})$. Next, we can define an orthogonal $n \times n$ matrix **H** whose first two columns are $1_n/\sqrt{n}$ and $(n_2 \mathbf{1}_{n_1}', -n_1 \mathbf{1}_{n_2}')/\sqrt{n_1 n_2/n}$, the other columns of **H** are arbitrary orthonormal vectors. Applying this transformation, we have:

$$Z = (z_1, z_2, \cdots, z_n) = \boldsymbol{YH}$$

$$z_1 \sim N\left(\frac{1}{\sqrt{n}}(n_1\mu_1 + n_2\mu_2), \Sigma\right), z_2 \sim N\left(\sqrt{\frac{n_1 n_2}{n}}\mu, \Sigma\right)$$

Finally, the statistic is F and follows a F distribution with d.f. r and $(n_1 + n_2 - 2)r$

$$F = \frac{Q_2}{Q_3 + \cdots + Q_n}, \quad Q = z_i z_i', \quad Q_i \sim m\chi_r^2$$

We have successfully verified that Gram-Schmidt process can be used to find a suitable **H** as the choice is quite arbitrary. Moreover, we prove that Dempster's non-centrality parameter $\Lambda = \sum_{1 \le j \le p} g_j^2 = \frac{n_1 n_2}{n}\mu'\Sigma^{-1}\mu, \mu = \mu_1 - \mu_2$ is the same as the Hotelling's by eigenvalue decomposition.

Dempster's method still requires normality assumption, but Bai-Saranada's test can perform well without it and is mathematically simpler. We first define

$$M_n = \left\|\bar{x}_1 - \bar{x}_2\right\|^2 - \tau tr(S_n), \qquad \tau = \frac{n_1 n_2}{n_1 + n_2}$$

It is then verified that $B_n^2 = \frac{n^2}{(n+2)(n-1)}\left(tr(S_n^2) - \frac{1}{n}(trS_n)^2\right)$ is an ratio-consistent and unbiased estimator for $\sqrt{Var(M_n)}$. By CLT, when n goes to infinity:

$$Z_n = \frac{M_n}{\sqrt{Var(M_n)}} = \frac{\frac{n_1 n_2}{n_1 + n_2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) - tr(S_n)}{\sqrt{\frac{2(n+1)}{n}}B_n} \to N(0,1)$$

is the statistic.

METHODOLOGY

We would like to compare the explanation power difference between the three methods mentioned. To achieve this, we first derive the asymptotical power functions of the three tests. The asymptotic power function of Hotelling's test:

$$\beta_H(\delta) - \Phi\left(-\xi_\alpha + \sqrt{\frac{n(1-y)}{2y}}\kappa(1-\kappa)\left\|\delta\right\|^2\right) \to 0$$

The asymptotic power function of Dempster and Bai-Saranadasa's test is the same:

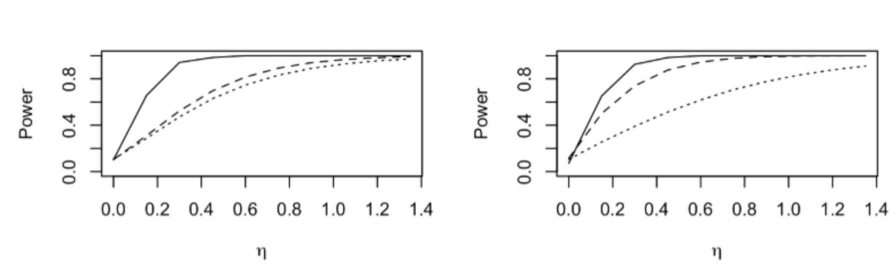
$$\beta_D(\delta) - \Phi\left(-\xi_\alpha + \sqrt{\frac{n\kappa(1-\kappa)\left\|\mu\right\|^2}{\sqrt{2tr(\Sigma^2)}}}\right) \to 0$$

where the parameters satisfy that if $\frac{p}{n_1+n_2} \to y > 0, \frac{n_1}{n_1+n_2} \to \kappa \in (0,1), n = n_1 + n_2 - 2, \left\|\delta\right\|^2 = o(1), \delta = \Sigma^{-\frac{1}{2}}|\mu_1 - \mu_2|$

A simulation is conducted to verify the asymptotic of the power functions. Three settings are generated which corresponds to A: $n = 45 \gg p = 4, B$: $n = 45 > p = 40, C$: $p = [20,200] > n = 45$. For each of them, both normal and non-normal datasets are generated. For normal sets, the covariance matrix $\Sigma = (1-\rho)I_p + \rho J_p, \rho = 0, 0.5$. For non-normal sets, $X_{ijk} = U_{ijk} + \rho U_{i,j+1,k} + \mu_{j,k}, (j = 1, \cdots, p; i = 1, \cdots, N_k; k = 1,2)$, $U_{ijk} \sim \Gamma(4,1)$ is generated by a moving factor model. All tests are then conducted under size α = 0.05 with 1000 repetitions.

RESULTS

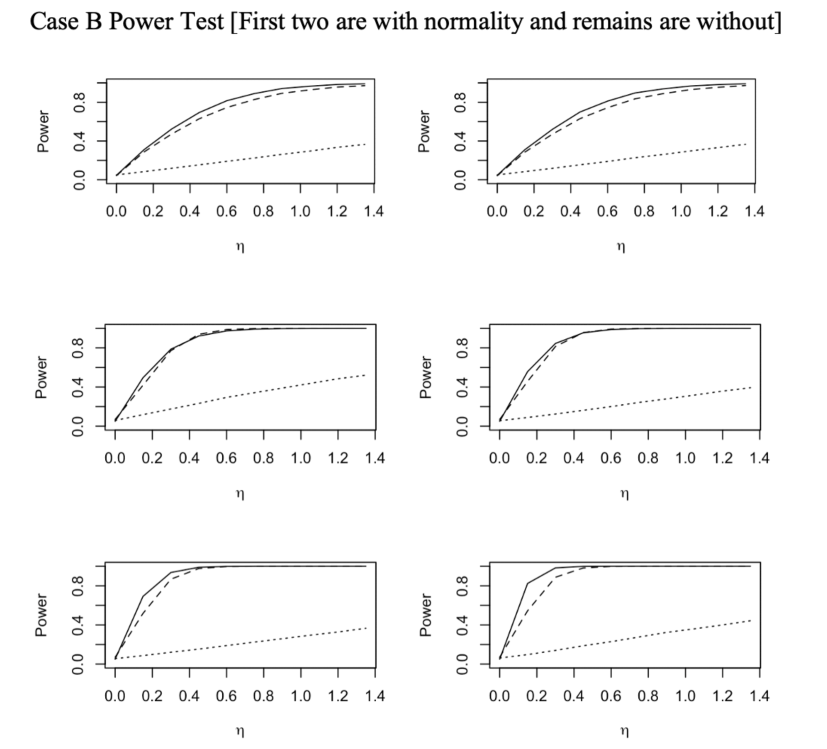
Low Dimensional Case A Power Test



H: Hotelling's T2 test (Dotted line), D: Dempster's non exact test (Dashed line), BS: Bai and Saranadasa's normal test (Solid line)

The power of Hotelling's T^2 test remains increasing in a rather fast speed, though still over-performed by the other two tests. Meanwhile, there still exists a certain amount of gap between Dempster's non exact test and Bai and Saranadasa's test when both n and p are not large enough to show the asymptotic convergence of their power function.

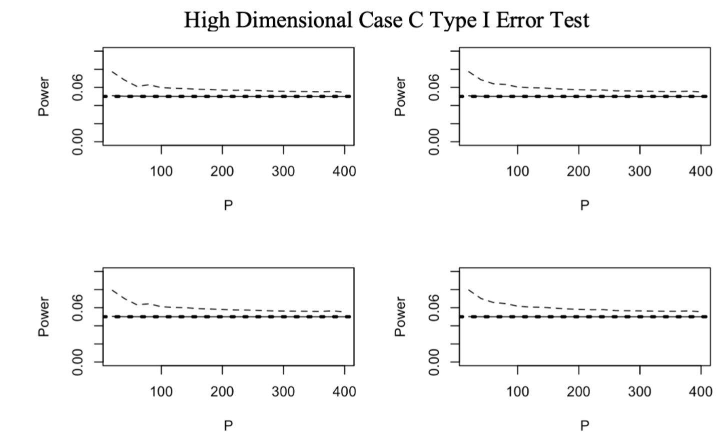
Case B Power Test [First two are with normality and remains are without]



H: Hotelling's T2 test (Dotted line), D: Dempster's non exact test (Dashed line), BS: Bai and Saranadasa's normal test (Solid line)

Hotelling's curve increases much slower in case B. Meanwhile, Bai and Saranadasa's test has almost the same significance level with Dempster's, which proves the theoretical asymptotic property.

High Dimensional Case C Type I Error Test



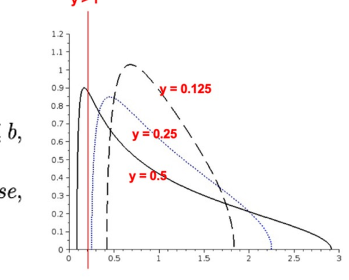
Dempster's non exact test (Dashed line), Bai and Saranadasa's normal test (Solid line)

Both tests stay around the set size α. It's worth noticing that when p is not high enough, Dempster's test has higher chance of having type I error. This difference can be explained since Dempster's estimation relies on higher dimension to provide accuracy.

DISCUSSION

From the results above, Dempster and Bai-Saranada's test both outperform Hotelling's. Numerically speaking, the main difference of increasing speed is due to the $\sqrt{1-y}$ in the Hotelling's asymptotic power functions shown above. This extra parameter limits the increase of Hotelling's power function. A more fundamental reason is the skewing of the Wishart matrix from Σ when p is high. In this case, the ratio between the maximum and the minimum eigenvalue of Wishart distribution will converge to $\frac{1+\sqrt{y}}{1-\sqrt{y}}$ for y ∈ (0,1) so that if y is close to 1 then the gap between the maximum and minimum eigenvalue could be extremely high which makes the error serious. To clearly see this, the distribution of the eigenvalues of the Wishart distribution has been proved by Yao (2015) to be:

$$F_{y,\sigma^2}(x) = \begin{cases} \frac{1}{2\pi xy\sigma^2}\sqrt{(b-x)(x-a)}, & a \leqslant x \leqslant b, \\ 0, & otherwise, \end{cases}$$



where $a = \sigma^2(1-\sqrt{y})^2$ and $b = \sigma^2(1+\sqrt{y})^2$ with an additional point mass of the value $1 - \frac{1}{y}$ at the origin if y > 1.

REFERENCES

- Bai, Z. and Saranadasa, H. (1996). Effect of high dimension: by an example of a two sample problem. Statistica Sinica, pages 311–329.
- Dempster, A. P. (1958). A high dimensional two sample significance test. The Annals of Mathe- matical Statistics, pages 995–1010.
- Dempster, A. P. (1960). A significance test for the separation of two highly multivariate small samples. Biometrics, 16(1):41–50.
- Yao, J., Zheng, S., and Bai, Z. (2015). Sample covariance matrices and high-dimensional data analysis. Cambridge University Press Cambridge.