CENTRAL LIMIT THEOREM FOR HOTELLING'S T^2 STATISTIC UNDER LARGE DIMENSION

GUANGMING PAN, WANG ZHOU

The sample covariance matrix is defined by

$$\mathcal{S} = \frac{1}{n}\sum_{j=1}^{n}(\mathbf{s}_j - \bar{\mathbf{s}})(\mathbf{s}_j - \bar{\mathbf{s}})^T,$$

where $\bar{\mathbf{s}} = n^{-1}\sum_{j=1}^{n}\mathbf{s}_j$ and $\mathbf{s}_j = (X_{1j}, \cdots, X_{pj})^T$. Here $\{X_{ij}\}$, $i, j = \cdots$, is a double array of independent and identically distributed (i.i.d.) real r.v.'s with $EX_{11} = 0$ and $EX_{11}^2 = 1$.

Sample covariance matrices are also of essential importance in multivariate statistical analysis because many test statistics involve their eigenvalues and/or eigenvectors. The typical example is T^2 statistic, which was proposed by Hotelling [2]. We refer to [1] and [3] for various uses of the T^2 statistic.

The T^2 statistic, which is the origin of multivariate linear hypothesis tests and the associated confidence sets, is defined by

$$T^2 = n(\bar{\mathbf{s}} - \boldsymbol{\mu}_\mathbf{0})^T \boldsymbol{\mathcal{S}}^{\mathbf{-1}}(\bar{\mathbf{s}} - \boldsymbol{\mu}_\mathbf{0}), \tag{1}$$

whose distribution is invariant if each $\mathbf{s}_j$ is replaced by $\mathbf{\Sigma}^{1/2}\mathbf{s}_j$ with $\mathbf{\Sigma}$ any non-singular p by p matrix when $\boldsymbol{\mu}_0 = 0$. If $\{\mathbf{s}_1, \cdots, \mathbf{s}_n\}$ is a sample from the p-dimensional population $N(\boldsymbol{\mu}, \mathbf{\Sigma})$, then $\left[T^2\ /(n-1)\right]\left[(n-p)/p\right]$ follows a noncentral F distribution and moreover, the F distribution is central if $\boldsymbol{\mu} = \boldsymbol{\mu}_\mathbf{0}$. When p is fixed, the limiting distribution of T^2 for $\boldsymbol{\mu} = \boldsymbol{\mu}_\mathbf{0}$ is the χ^2-distribution even if the parent distribution is not normal.

In recent three or four decades, in many research areas, including signal processing, network security, image processing, genetics, stock marketing and other economic problems, people are interested in the case where p is quite large or proportional to the sample size. Thus it will be desirable if one can obtain the asymptotic distribution of the famous Hotelling T^2 statistic when the dimension of the random vectors is proportional to the sample size. It is the aim of this work.

Published in *Annals of Applied Probability.* **21,** 1860-1910.

The main results are then presented in the following theorems.

Theorem 1. *Suppose that:*
(1) For each n $X_{ij} = X_{ij}^n, i, j = 1, 2, \cdots$, are i.i.d. real r.v.'s with $EX_{11} = \mu, EX_{11}^2 = 1$ and $EX_{11}^4 < \infty$.
(2) $p \leq n, c_n = p/n \to c \in (0, 1)$ as $n \to \infty$.
Then, when $\boldsymbol{\mu_0} = (\mu, \cdots, \mu)^T$,

$$\frac{\sqrt{n}}{\sqrt{2c_n(1-c_n)^{-3}}}\big(\frac{T^2}{n} - c_n(1-c_n)^{-1}\big) \xrightarrow{D} N(0, 1),$$

where $F_{c_n}(x)$ denotes $F_c(x)$ by substituting c_n for c.

One typical application of Theorem 1 lies in making inference on the large dimensional mean vector of the multivariate model

$$\boldsymbol{Z}_j = \Gamma \mathbf{s}_j + \boldsymbol{\mu}, \quad E\mathbf{s}_j = 0, \quad j = 1, \cdots, n,$$

where Γ is an m by p matrix, $m \leq p$. This model means that each $\boldsymbol{Z}_j$ is a linear transformation of some p-variate random vector $\mathbf{s}_j$. It can generate a rich collection of $\boldsymbol{Z}_j$ from $\mathbf{s}_j$ with the given covariance matrix $\boldsymbol{\Sigma} = \Gamma\Gamma^T$. Most important, it includes the multivariate normal model.

We will prove Theorem 1 by establishing Theorem 2 which presents asymptotic distributions of random quadratic forms involving sample means and sample covariance matrices.

For any analytic function $f(\cdot)$, define

$$f(\boldsymbol{\mathcal{S}}) = \mathbf{U}^T diag(f(\lambda_1), \cdots, f(\lambda_p))\mathbf{U},$$

where $\mathbf{U}^T diag(\lambda_1, \cdots, \lambda_p)\mathbf{U}$ denotes the spectral decomposition of the matrix $\boldsymbol{\mathcal{S}}$.

Theorem 2. *In addition to the assumption (1) of Theorem 1, suppose that $c_n = p/n \to c > 0$, $EX_{11} = 0$, $g(x)$ is a function with a continuous first derivative in a neighborhood of c, and $f(x)$ is analytic on an open region containing the interval*

$$[I_{(0,1)}(c)(1-\sqrt{c})^2, (1+\sqrt{c})^2]. \tag{2}$$

Then,

$$\left(\sqrt{n}[\frac{\bar{\mathbf{s}}^T f(\boldsymbol{\mathcal{S}})\bar{\mathbf{s}}}{||\bar{\mathbf{s}}||^2} - \int f(x)dF_{c_n}(x)], \sqrt{n}(g(\bar{\mathbf{s}}^T\bar{\mathbf{s}}) - g(c_n))\right) \xrightarrow{D} (X, Y),$$

where $Y \sim N(0, 2c(g'(c))^2)$, which is independent of X, a Gaussian r.v. with $EX = 0$ and

$$Var(X) = \frac{2}{c}\left(\int f^2(x)dF_c(x) - (\int f(x)dF_c(x))^2\right). \tag{3}$$

References

[1] Anderson, T. W. (1984). *An introduction to multivariate statistical analysis.* 2nd edition, John Wiley & Sons.

[2] Hotelling, H. (1931). The generalization of Student's ratio. *Ann. Math. Statist.* **2** 360-378.

[3] Lehmann, E. L. and Romano, J. P. (2005). *Testing statistical hypotheses.* 3rd edition, Springer.