高次元独立性検定とロバストネス
Tests for Independence in High-dimension and Their Robustness

数学専攻 青木 誠
Makoto AOKI

In this paper we deal with the high dimensional data such as micro array data, finance data, and image data. Moreover, on some multivariate analyses we need the condition of independence. We analyze the tests of independence in high-dimension. Let us write population correlation matrix as P, and we consider the hypothesis $H_{01} : P = I_m$.

Let $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N$ be independently and identically distributed as a m-dimensional random vector $\boldsymbol{x}$ which is distributed as multivariate normal with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ, denoted, $\boldsymbol{x} \sim N_p(\boldsymbol{\mu}, \Sigma)$. Let $\bar{\boldsymbol{x}}$ and S denote the sample mean vector and the sample covariance matrix respectively, defined as

$$\bar{\boldsymbol{x}} = \frac{1}{N}\sum_{\alpha=1}^{N} \boldsymbol{x}_\alpha, \qquad N = n + 1,$$

and

$$S \equiv \frac{1}{n}V = \frac{1}{n}\sum_{\alpha=1}^{N}(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}})'$$

respectively. If the (i, j)th element of S is s_{ij}, the sample correlation coefficient between two components of $\boldsymbol{x}$, say, $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, is

$$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}}.$$

The sample correlation matrix, R, is composed of r_{ij}. Let P denote the population matrix. the (i, j)th element of P is ρ_{ij}.

In the tests of independence, there are the inference procedures based on some asymptotic theories. We compare the some procedures by simulation study. Firstly, we look at the likelihood ratio test and Hsu(1949) based on asymptotic theory which has a sample size of up to infinity, while the number of variables is fixed. When we consider the hypothesis $H_{01} : P = I_m$, the statistic of likelihood ratio test,

$$w_{nm} = -(n - \frac{2m + 5}{6}) \log |R|,$$

converges in χ^2 distribution with $m(m-1)/2$ degrees of freedom. Crearly, this procedure is no valid for high-dimensional data since $|R| = 0$ whenever $m > n$. Moreover, when

n is less than eight times of m, we can not keep 5 percent point of the chi-squared distribution.

Let us consider the hypothesis $H_{01} : P = I_m$, then the statistic of test on Hsu(1949) is as follows,

$$v_{nm} = \frac{n\sum_{i<j}^{m} r_{ij}^2 - q}{\sqrt{2q}}, \qquad q = \frac{m(m-1)}{2},$$

which converges to a normal random variable with mean 0 and variance 1. The normal approximation is over significance levels when $m > n$ and m is around n, because this test is based on asymptotic theory which the number of variables is fixed.

Next, we discuss the statistic of Srivastava(2005) for the hypothesis. When the (i, j)th element of Σ is σ_{ij}, the population covariance matrix is equal to diagonal matrix, say, if the (i, j)th element of Σ is σ_{ij}, $H_{02} : \sigma_{ij} = 0$, $i \neq j$. The hypothesis can identify the hypothesis that sample correlation matrix is identity matrix. This asymptotic theory is based on the sample size of up to infinity after the number of variables goes to infinity. Now, we need the following assumptions:
When $a_i = (tr\Sigma^i/m), i = 1, \cdots, 8$,

$$(A) : \text{As } \ m \to \infty, a_i \to a_i^0, 0 < a_i^0 < \infty, i = 1, \cdots, 8$$
$$(B) : n = O(m^\delta), 0 < \delta \leq 1.$$

Under the assumptions, the statistic is

$$u_{nm} = \frac{n(c-1)}{2\sqrt{1 - \left(\frac{1}{m}\right)\left(\frac{a}{b^2}\right)}},$$

where

$$a = \frac{1}{m}\sum_{i=1}^{m} s_{ii}^4, \quad b = \frac{n}{m(n+2)}\sum_{i=1}^{m} s_{ii}^2, \quad d = \frac{n^2}{m(n-1)(n+2)}\left[trS^2 - \frac{1}{n}(trS)^2\right], \quad c = \frac{d}{b}.$$

This statistic converges to a normal random variable with mean 0 and variance 1. The normal approximation is over significance levels when $m > n$ and m is around n. Although this approximation is thought in high dimensional approximation, it is not available, because the asymptotic theory assumes that the sample size and the number of variables separately go to infinity.

On the other hand, the statistic of Schott(2005) for the hypothesis $H_{01} : P = I_m$ is based on asymptotic theory which both sample size and number of variables together go to infinity. The statistic is

$$t_{nm} = \sum_{i=2}^{m}\sum_{j=1}^{i-1} r_{ij}^2 - \frac{m(m-1)}{2n},$$

which converges to a normal random variable with mean 0 and variance $\sigma_{t_{nm}}^2$, where

$$\sigma_{t_{nm}}^2 = \frac{m(m-1)(n-1)}{n^2(n+2)}.$$

Thus $t_{nm}/\sigma_{t_{nm}}$ converges to a normal variable with mean 0 and variance 1. This asymptotic theory is based on $\lim(m/n) = \gamma_1 \in (0, \infty)$. The normal approximation generally yields suitable significance levels.

Sometimes, micro array data and image data are discrete data. We are concerned whether the mentioned procedures can adapt to discrete data. Our meaning of robustness is how far the procedure for normal variables is available to discrete variables. We check for robustness by simulation study. Now, since the statistic of Schott(2005) is available for the test of the independence in high-dimension, we adapt the test of Schott(2005) to discrete variables having uniform distribution. When we run a simulation for binary data, three level categorical data, and four level categorical data, we can not keep significance level in small sample size. But as the number of category of variables increases, the approximation becomes better.

We consider the simulation study which changes the rate of incidence of random variables when random variables are binary data generated from uniform distribution. The rate of incidence of random variables means that when random variables are 0 or 1, the rate of incidence of 0 is α, $0 < \alpha < 1$ and the rate of incidence of 1 is $1 - \alpha$. When we run a simulation for $\alpha = 0.1, 0.2, 0.3, 0.4, 0.5$, the case of uniform rate of incidence is most available.

Finally, we apply the theory of Schott(2005) to the test of independence for partial correlation matrix. We assume that random variables $\boldsymbol{X}$ and sample covariance matrix S have been partitioned as in

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}^{(1)} \\ \boldsymbol{X}^{(2)} \end{bmatrix}, \quad S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix},$$

where $\boldsymbol{X}^{(1)}$ is p vector, $\boldsymbol{X}^{(2)}$ is q vector, S_{11} is $p \times p$ matrix and S_{22} is $q \times q$ matrix. Now sample covariance matrix of $\boldsymbol{X}^{(1)}$ given $\boldsymbol{X}^{(2)} = \boldsymbol{x}^{(2)}$ is written as

$$S_{11\cdot 2} = S_{11} - S_{12} S_{22}^{-1} S_{21}.$$

Let the (i, j)th element of $S_{11\cdot 2}$ denotes $\sigma_{ij\cdot p+1,\cdots,m}$. Then we can calculate the (i, j)th elements of sample correlation matrix of $\boldsymbol{X}^{(1)}$ given $\boldsymbol{X}^{(2)} = \boldsymbol{x}^{(2)}$, say $R_{11\cdot 2}$, as

$$r_{ij\cdot p+1,\cdots,m} = \frac{\sigma_{ij\cdot p+1,\cdots,m}}{\sqrt{\sigma_{ii\cdot p+1,\cdots,m}}\sqrt{\sigma_{jj\cdot p+1,\cdots,m}}}$$

We assume the (i, j)th element of population correlation matrix $P_{11\cdot 2}$ of $\boldsymbol{X}^{(1)}$ given $\boldsymbol{X}^{(2)} = \boldsymbol{x}^{(2)}$ as $\rho_{ij\cdot p+1,\cdots,m}$, then the hypothesis can be written as $H_{04} : \rho_{ij\cdot p+1,\cdots,m} = 0 (i > j)$. The statistic is made by changing the number of dimension m to p and the sample size n to $n - q$ for the statistic of correlation matrix, because the distribution of partial correlation coefficient is equal to distribution of correlation coefficient changing degrees of freedom and the number of dimension. Thus, the statistic is written as follows

$$t_{nm}^{*} = \sum_{i=2}^{p} \sum_{j=1}^{i-1} r_{ij\cdot p+1,\cdots,m}^{2} - \frac{p(p-1)}{2(n-q)},$$
$$\sigma_{t_{nm}^{*}}^{2} = \frac{p(p-1)(n-q-1)}{(n-q)^2(n-q+2)}.$$

The asymptotic theory is based on condition

$$\lim_{p,n-q\to\infty} \frac{p}{n-q} = \gamma_2 \in (0,\infty), \tag{1}$$

this leads to

$$\lim \sigma^2_{t^*_{nm}} = \lim \frac{p(p-1)(n-q-1)}{(n-q)^2(n-q+2)} = \gamma_2^2.$$

Therefore we have the following theorem.

THEOREM 1. Suppose that the sample correlation matrix $R_{11\cdot 2}$ has been computed from a random sample from a multivariate normal distribution with correlation matrix $P_{11\cdot 2}$. If $P_{11\cdot 2} = I_p$ and condition (1) holds, then t^*_{nm} converges to a normal random variable with mean 0 and variance γ_2^2.

We can check the perfprmances of the null distribution of $t^*_{n-q,p}$ by simulation study. The normal approximation generally yields suitable significance levels. Our future works are applying the theory of Schott(2005) to the test for independence of k sets of variables. For the covariance matrix, the test for independence of sets of variables is introductioned in the tecnical report Schott(2004).

Refferences

[1] ANDERSON, T, W. (1984). An Introduction to Multivariate Statistical Analysis, 3d ed., Jhon Wiley & Sons, Inc., New York.

[2] HALL, P. and HEYDE, C. C. (1980). Martingale Limit Theory and Its Applications. New York: Academic Press.

[3] HSU, P. L. (1949). The limiting distribution of functions of sample means and applications to testing hypothesis, Proceedings of the First Berkeley Symposium of Mathematical Statistics and Probability (ed. J. Neyman), Univ. of California Press, Berkeley and Los Angeles, 359-402.

[4] MORRISON, D. F. (2005). Multivariate Statistical Methods. Belmont, CA:Brooks/Cole.

[5] SCHOTT, J. R. (2005). Miscellanea Testing for complete independence in high dimensions. Biometrika, 92, 951-956.

[6] SCHOTT, J. R. (2004). Testing For Independence in High Dimensions.

[7] SRIVASTAVA, M. S. (2005). Some tests concerning the covariance matrix in high dimensional data. J. Japan Statist. 35, 251-272.