Some Simple High Dimensional One Sample Tests

Ahlam M. Abid and David J. Olive*
Southern Illinois University

January 18, 2024

Abstract

Consider testing $H_0 : \mu = 0$ versus $H_A : \mu \neq 0$ using a random sample $x_1, \ldots, x_n$ where the $x_i$ are $p \times 1$ random vectors and $p$ may be much larger than $n$. We modify a test and give a new test that has very simple large sample theory.

KEY WORDS: Hotelling’s $T^2$ Test.

*David J. Olive is Professor, School of Mathematical & Statistical Sciences, Southern Illinois University, Carbondale, IL 62901, USA.
1 INTRODUCTION

Consider testing $H_0 : \mu = 0$ versus $H_A : \mu \neq 0$ using independent and identically distributed (iid) $\mathbf{x}_1, \ldots, \mathbf{x}_n$ where the $\mathbf{x}_i$ are $p \times 1$ random vectors and $p$ may be much larger than $n$. Assume the expected value $E(\mathbf{x}_i) = \mu$ and nonsingular covariance matrix $\text{Cov}(\mathbf{x}_i) = \Sigma$. Replace $\mathbf{x}_i$ by $\mathbf{w}_i = \mathbf{x}_i - \mu_0$ to test $H_0 : \mu = \mu_0$ versus $H_A : \mu \neq \mu_0$. This section reviews some tests while the following section gives a new test that has very simple large sample theory.

Suppose $p$ is fixed, and consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ where a $g \times 1$ statistic $T_n$ satisfies $\sqrt{n}(T_n - \theta) \overset{D}{\rightarrow} \mathbf{u} \sim N_g(0, \Sigma)$. If $\hat{\Sigma}^{-1} \overset{P}{\rightarrow} \Sigma^{-1}$ and $H_0$ is true, then

$$D_n^2 = D^2_{\theta_0}(T_n, \Sigma/n) = n(T_n - \theta_0)^T \hat{\Sigma}^{-1} (T_n - \theta_0) \overset{D}{\rightarrow} \mathbf{u}^T \Sigma^{-1} \mathbf{u} \sim \chi^2_g$$

as $n \rightarrow \infty$. Then a Wald type test rejects $H_0$ at significance level $\delta$ if $D_n^2 > \chi^2_{g,1-\delta}$ where $P(X \leq \chi^2_{g,1-\delta}) = 1 - \delta$ if $X \sim \chi^2_g$, a chi-square distribution with $g$ degrees of freedom.

It is common to implement a Wald type test using

$$D_n^2 = D^2_{\theta_0}(T_n, \mathbf{C}_n/n) = n(T_n - \theta_0)^T \mathbf{C}_n^{-1} (T_n - \theta_0) \overset{D}{\rightarrow} \mathbf{u}^T \mathbf{C}^{-1} \mathbf{u}$$

as $n \rightarrow \infty$ if $H_0$ is true, where the $g \times g$ symmetric positive definite matrix $\mathbf{C}_n \overset{P}{\rightarrow} \mathbf{C} \neq \Sigma$. Hence $\mathbf{C}_n$ is the wrong dispersion matrix, and $\mathbf{u}^T \mathbf{C}^{-1} \mathbf{u}$ does not have a $\chi^2_g$ distribution when $H_0$ is true. Often $\mathbf{C}_n$ is a regularized estimator of $\Sigma$, or $\mathbf{C}_n^{-1}$ is a regularized estimator of the precision matrix $\Sigma^{-1}$, such as $\mathbf{C}_n = \text{diag}(\hat{\Sigma})$ or $\mathbf{C}_n = \mathbf{I}_g$, the $g \times g$ identity matrix. Rajapaksha and Olive (2022) showed how to bootstrap Wald tests with the wrong dispersion matrix.

When $n$ is much larger than $p$, the one sample Hotelling (1931) $T^2$ test is often used to test $H_0 : \mu = \mu_0$ versus $H_A : \mu \neq \mu_0$. The sample mean

$$\overline{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i,$$

and the sample covariance matrix

$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T = (S_{ij}).$$

That is, the $ij$ entry of $\mathbf{S}$ is the sample covariance $S_{ij}$. If the $\mathbf{x}_i$ are iid with expected value $E(\mathbf{x}_i) = \mu$ and nonsingular covariance matrix $\text{Cov}(\mathbf{x}_i) = \Sigma$, then by the multivariate central limit theorem

$$\sqrt{n}(\overline{\mathbf{x}} - \mu) \overset{D}{\rightarrow} N_p(0, \Sigma).$$

If $H_0$ is true, then

$$T^2_H = n(\overline{\mathbf{x}} - \mu_0)^T \mathbf{S}^{-1} (\overline{\mathbf{x}} - \mu_0) = \chi^2_p.$$ 

The one sample Hotelling’s $T^2$ test rejects $H_0$ if $T^2_H > D^2_{1-\delta}$ where $D^2_{1-\delta} = \chi^2_{p,\delta}$ and $P(Y \leq \chi^2_{p,\delta}) = \delta$ if $Y \sim \chi^2_p$. Alternatively, use

$$D^2_{1-\delta} = \frac{(n-1)p}{n-p} F_{p,n-p,1-\delta}.$$
where $P(Y \leq F_{p,d,\delta}) = \delta$ if $Y \sim F_{p,d}$. The scaled $F$ cutoff can be used since $T_H^2 \overset{D}{\to} \chi_p^2$ if $H_0$ holds, and
\[
\frac{(n-1)p}{n-p} F_{p,n-p,1-\delta} \to \chi_{p,1-\delta}^2
\]
as $n \to \infty$.

The next two high dimensional tests are described in Srivastava and Du (2008). Also see Hu and Bai (2015). Let $tr(A)$ be the trace of square matrix $A$. Let $R$ be the sample correlation matrix. Consider testing $H_0 : \mu = 0$ versus $H_A : \mu \neq 0$. Let $D = \text{diag}(S)$.

Let
\[
c_{p,n} = 1 + \frac{tr(R^2)}{p^{3/2}}.
\]
Let $n = O(p^\delta)$ where $0.5 < \delta \leq n$. Then under regularity conditions
\[
Z_1 = \frac{n\overline{x}^TD^{-1}\overline{x} - \frac{(n-1)p}{n-3}D}{2\left(tr(R^2) - \frac{p^2}{n-1}\right)} \overset{D}{\to} N(0,1)
\]
as $n,p \to \infty$. The next test is attributed to Bai and Saranadasa (1996). Suppose $p/n \to c > 0$. Under regularity conditions,
\[
Z_2 = \frac{n\overline{x}^T - tr(S)}{\left[\frac{2(n-1)n}{(n-2)(n+1)}\left(tr(S^2) - \frac{1}{n}[tr(S)]^2\right)\right]^{1/2}} \overset{D}{\to} N(0,1)
\]
as $n,p \to \infty$. Both of these test statistics need $p/n \to c > 0$ or $p/n^2 \to 0$. Hence $p$ cannot be too big.

There are test statistics $T_n$ for testing $H_0 : \mu = 0$ where $p$ can be much larger with
\[
\frac{T_n}{s_n} \overset{D}{\to} N(0,1)
\]
where $T_n$ is relatively simple to compute while $s_n$ is much harder to compute. The following test is due to Chen and Qin (2010). Also see Hu and Bai (2015). Let $a = \sum_{i=1}^n x_i$ and let $X = (x_{ij})$ be the data matrix with $i$th row $= x_i^T$ and $ij$ element $= x_{ij}$. Let $\text{vec}(A)$ stack the columns of matrix $A$ so that $c = \text{vec}(X^T) = [x_1^T, x_2^T, ..., x_n^T]^T$. Then
\[
c^Tc = \sum_{i=1}^n x_i^T x_i = \sum_{i=1}^n ||x_i||^2 = \sum_{i=1}^n \sum_{j=1}^p (x_{ij})^2.
\]
Let
\[
T_n = \frac{1}{n(n-1)}[a^Ta - c^Tc] = \frac{1}{n(n-1)} \sum \sum x_i^T x_j = \frac{1}{n(n-1)} \sum x_i^T x_j.
\]
Then
\[
\frac{T_n}{\sqrt{V(T_n)}} \overset{D}{\to} N(0,1) \quad \text{and} \quad \frac{T_n}{s_n} \overset{D}{\to} N(0,1)
\]
where $s_n$ is rather hard to compute. Here

$$s_n^2 = \frac{2}{n(n-1)} tr \left[ \sum_{i \neq j} (x_i - \overline{x}_{(i,j)})x_i^T(x_j - \overline{x}_{(i,j)})x_j^T \right]$$

is a consistent estimator of $V(T_n)$ where $\overline{x}_{(i,j)}$ is the sample mean computed without $x_i$ or $x_j$:

$$\overline{x}_{(i,j)} = \frac{1}{n-2} \sum_{k \neq i,j} x_k.$$

The $T_n$ in Equation (1) can be viewed as a modification of $\|\overline{x}\|^2 = \overline{x}^T \overline{x}$ that is a better estimator of $\mu^T \mu$ in high dimensions. Note that $\mu = 0$ iff $\mu^T \mu = 0$ and $E(x_i^T x_j) = \mu^T \mu$ if $x_i$ and $x_j$ are iid with $E(x_i) = \mu$ and $i \neq j$.

As noted by Park and Ayyala (2013), $nT_n = n\overline{x}^T \overline{x} - tr(S)$. This result holds since

$$T_n = \frac{1}{n(n-1)} \left[ \sum_i x_i^T x_j - \sum_i x_i^T x_i \right] = \frac{n^2 \overline{x}^T \overline{x} - \sum_i x_i^T x_i}{n(n-1)}.$$

Now

$$S = \frac{1}{n-1} \left[ \sum_i x_i x_i^T - n \overline{x} \overline{x}^T \right].$$

Thus

$$tr(S) = \frac{1}{n-1} \left[ \sum_i tr(x_i x_i^T) - ntr(\overline{x} \overline{x}^T) \right] = \frac{1}{n-1} \left[ \sum_i x_i^T x_i - n \overline{x} \overline{x}^T \right].$$

Thus

$$n\overline{x}^T \overline{x} - tr(S) = n\overline{x}^T \overline{x} + \frac{n}{n-1} \overline{x} \overline{x}^T - \frac{1}{n-1} \sum_i x_i^T x_i = \frac{n^2 \overline{x}^T \overline{x} - \sum_i x_i^T x_i}{n-1}.$$ 

Section 2 finds estimators $s_n^2$ of $V(T_n)$ that are easier to compute, and gives a new test with very simple large sample theory.

## 2 NEW AND MODIFIED TESTS

Some notation for the simple test is needed. Assume $x_1, ..., x_n$ are iid, $E(x_i) = \mu$ and the variance $V(x_i^T x_j) = \sigma_W^2$ for $i \neq j$. Let $m = \floor{n/2} = \floor{n/2}$ be the integer part of $n/2$. So $\floor{100/2} = \floor{101/2} = 50$. Let the iid random variables $W_i = x_{2i-1}^T x_{2i}$ for $i = 1, ..., m$. Hence $W_1, W_2, ..., W_m = x_1^T x_2, x_3^T x_4, ..., x_{2m-1}^T x_{2m}$. Note that $E(W_i) = \mu^T \mu$ and $V(W_i) = \sigma_W^2$. Let $S_W^2$ be the sample variance of the $W_i$:

$$S_W^2 = \frac{1}{m-1} \sum_{i=1}^m (W_i - \overline{W})^2.$$
If $\sigma^2_W \propto \tau p$, then $n$ may not be large enough for the normal approximation to hold. The following theorem follows from the univariate central limit theorem.

**Theorem 1.** Assume $x_1, \ldots, x_n$ are iid, $E(x_i) = \mu$, and the variance $V(x_i^T x_j) = \sigma^2_W$ for $i \neq j$. Let $W_1, \ldots, W_m$ be defined as above. Then

a) $\sqrt{m}(W - \mu^T \mu) \overset{D}{\to} N(0, \sigma^2_W)$. \\
b) $\frac{\sqrt{m}(W - \mu^T \mu)}{S_W} \overset{D}{\to} N(0, 1)$
as $n \to \infty$.

**Theorem 2.** Assume $x_1, \ldots, x_n$ are iid, $E(x_i) = \mu$, and the variance $V(x_i^T x_j) = \sigma^2_W$ for $i \neq j$. Let $W_{ij} = x_i^T x_j$ for $i \neq j$. Then $\theta = \text{Cov}(W_{ij}, W_{id})$ where $j \neq d$, $i < j$, and $i < d$. Then

a) $V(T_n) = \frac{2 \sigma^2_W}{n(n - 1)} + \frac{4(n - 2)\theta}{n(n - 1)}$.

b) If $H_0 : \mu = 0$ is true, then $\theta = 0$ and

$$V(T_n) = \frac{2 \sigma^2_W}{n(n - 1)}.$$

**Proof.** a) To find the variance $V(T_n)$ with $T_n$ from Equation (1), let $W_{ij} = x_i^T x_j = W_{ji}$, and note that

$$T_n = \frac{2}{n(n - 1)} H_n \text{ where } H_n = \sum_{i < j} \sum_k W_{ik} \sum_{d} W_{kd} = \sum_{i < j} \sum_k \sum_{d} W_{ik} W_{kd}.$$ 

Then $V(H_n) = \text{Cov}(H_n, H_n) = \text{Cov}
\left(\sum_{i < j} W_{ij}, \sum_{k < d} W_{kd}\right) = \sum_{i < j} \sum_k \sum_{d} \text{Cov}(W_{ij}, W_{kd}). (3)$

Let $V(W_{ij}) = \sigma^2_W$ for $i \neq j$. The covariances are of 3 types. First, if $(ij) = (kd)$ with $i < j$, then $\text{Cov}(W_{ij}, W_{kd}) = V(W_{ij}) = \sigma^2_W$. Second, if $i, j, k, d$ are distinct with $i < j$ and $k < d$, then $W_{ij}$ and $W_{kd}$ are independent with $\text{Cov}(W_{ij}, W_{kd}) = 0$. Third, there are terms where exactly three of the four subscripts are distinct, which have $\text{Cov}(W_{ij}, W_{id}) = \theta$ where $j \neq d$, $i < j$, and $i < d$ or $\text{Cov}(W_{ij}, W_{kj}) = \theta$ where $i \neq k$, $i < j$, and $k < j$. These covariance terms are all equal to the same number $\theta$ since $W_{ij} = W_{ji}$. The number of ways to get three distinct subscripts is

$$a - b - c = \binom{n}{2}^2 - \binom{n}{2} \binom{n - 2}{2} - \binom{n}{2} = n(n - 1)(n - 2)$$

since $a$ is the number of terms on the right hand side of (3), $b$ is the number of terms where $i, j, k, d$ are distinct with $i < j$ and $k < d$, and $c$ is the number of terms where $(ij) = (kd)$ with $i < j$. [Note that $n(n - 1)$ terms have $i$ and $j$ distinct. Half of these terms have $i < j$ and half have $i > j$. Similarly, $n(n - 1)(n - 2)(n - 3)$ terms have $ijkd$.
distinct, and half of the \( n(n-1) \) terms have \( i < j \), while half of the \( (n-2)(n-3) \) terms have \( k < d \).] Thus

\[
V(H_n) = 0.5n(n-1)\sigma_W^2 + n(n-1)(n-2)\theta.
\]

This calculation was adapted from Lehmann (1975, pp. 336-337). Thus

\[
V(T_n) = \frac{4}{[n(n-1)]^2}V(H_n) = \frac{2\sigma_W^2}{n(n-1)} + \frac{4(n-2)\theta}{n(n-1)}.
\]

b) Now \( \theta = \text{Cov}(x_i^T x_j, x_i^T x_j) \) where \( x_i, x_j, \) and \( x_k \) are iid. Hence \( \theta = \)

\[
\text{Cov}(\sum_d x_{id}x_{jd}, \sum_t x_{it}x_{kt}) = \sum_d \sum_t \text{Cov}(x_{id}x_{jd}, x_{it}x_{kt}) =
\]

\[
\sum_d \sum_t [E(x_{id}x_{jd}x_{it}x_{kt}) - E(x_{id}x_{jd})E(x_{it}x_{kt})] =
\]

\[
\sum_d \sum_t [E(x_{id}x_{it})E(x_{jd})E(x_{kt})] - E(x_{id})E(x_{jd})E(x_{it})E(x_{kt}) =
\]

\[
\sum_d \sum_t [E(x_{jd})E(x_{kt})(E(x_{id}x_{it}) - E(x_{id})E(x_{it}))] =
\]

\[
\sum_d \sum_t [E(x_{jd})E(x_{kt})\text{Cov}(x_{id}, x_{it})].
\]

Under \( H_0, \mu = 0 \) and thus \( E(x_{jd}) = E(x_{it}) = 0 \). Hence \( \theta = 0 \). \( \square \)

The bootstrap often works well on statistics such as \( T_n \), but the nonparametric bootstrap fails because terms like \( x_i^T x_j \) need to be avoided, and the nonparametric bootstrap has replicates: the proportion of cases in the bootstrap sample that are not replicates is about \( 1 - e^{1} \approx 2/3 \approx 7/11 \). The \( m \) out of \( n \) bootstrap without replacement draws a sample of size \( m \) without replacement from the \( n \) cases. For \( B = 1 \), this is a data splitting estimator, and \( T_{m}^* \approx N(0, \sigma_m^2) \) for large enough \( m \) and \( p \). If \( B \) is larger, the data cloud has correlated \( T_{m,1}^*, ..., T_{m,B}^* \) centered at \( \overline{T}^* \) with variance \( \sigma_m^2 \) which may be less than \( s_m^2 \). Here \( \overline{T}^* \) is the sample mean of all \( (n\choose 2) \) statistics \( T_m^* \) obtained by drawing a sample of size \( m \) with replacement from \( n \). Theory for the \( m \) out of \( n \) bootstrap often has \( m/n \rightarrow 0 \) with \( m \rightarrow \infty \). Sampling without replacement is like sampling with replacement when \( n \gg m \), and sampling with replacement leads to iid \( T_m^* \) with respect to the bootstrap distribution. Heuristically, the \( T_m^* \) may be approximately iid \( N(\overline{T}^*, \sigma_m^2) \) if \( m/n \rightarrow 0 \) and \( m \rightarrow \infty \). Using \( m = \text{floor}(2n/3) \) and worked well in simulations.

Let \( W_i \) be an indicator random variable with \( W_i = 1 \) if \( x_i^T \) is in the sample and \( W_i = 0 \), otherwise, for \( i = 1, ..., n \). The \( W_i \) are binary and identically distributed, but not independent. Hence \( P(W_i = 1) = m/n \). Let \( W_{ij} = W_iW_j \) with \( i \neq j \). Again, the \( W_{ij} \) are binary and identically distributed. \( P(W_{ij} = 1) = P(\text{ordered pair} (x_i, x_j)) \) was selected in the sample. Hence \( P(W_{ij} = 1) = m(m-1)/[n(n-1)] \) since \( m(m-1) \) ordered pairs were selected out of \( n(n-1) \) possible ordered pairs. Then

\[
T_m^* = \frac{1}{m(m-1)} \sum_{k \neq d} x_{ik}^T x_{id} = \frac{1}{m(m-1)} \sum_{i \neq j} W_iW_jx_i^T x_j
\]
where the $x_{i1}, \ldots, x_{im}$ are the m vectors $x_i$ selected in the sample. The first double sum has $m(m-1)$ terms while the second double sum has $n(n-1)$ terms. Hence

$$E(T_m^*) = \frac{1}{m(m-1)} \sum_{i \neq j} E[W_i W_j] x_i^T x_j = T_n.$$ 

See similar calculations in Buja and Stuetzle (2006). Note that $V(T_m^*) = E([T_m^*]^2) - [T_n]^2 = Cov(T_m^*, T_m^*)$.

3 SIMULATIONS

In the simulations, we examined four tests. The first two tests used Theorem 2 b) and Equation 2): $T_n / s_n \sim N(0, 1)$ if $s_n^2$ is a consistent estimator of $V(T_n)$. The first test used $\hat{\sigma}_W^2 = S^2$ based on Theorem 1. The second test used

$$\hat{\sigma}_W^2 = \frac{1}{n(n-1)} \sum_{i \neq j} (x_i^T x_j - T_n)^2 = \frac{1}{n(n-1)} \sum_{i \neq j} (W_{ij} - T_n)^2.$$ 

If the denominator $n(n-1)$ was replaced by $n(n-1) - 1$, this statistic would be the usual sample variance of the $W_{ij}$, which are not independent.

These tests computed intervals

$$[T_n - t_{1-\alpha/2,m-1} \sqrt{2\hat{\sigma}_W^2 / [n(n-1)]}, T_n + t_{1-\alpha/2,m-1} \sqrt{2\hat{\sigma}_W^2 / [n(n-1)]}].$$

The third test computed the usual $t$ confidence interval

$$[\bar{W} - t_{1-\alpha/2,m-1} S_W / \sqrt{m}, \bar{W} + t_{1-\alpha/2,m-1} S_W / \sqrt{m}]$$

for $\mu^T \mu$ based on the $W_i$ from Theorem 1. The fourth “test” used the $m$ out of $n$ bootstrap to compute $T^*_1, \ldots, T^*_B$ with $B = 100$. We used the shorth bootstrap “confidence interval” described in Olive (2023, chapter 2) and Pelawa Wata goda and Olive (2021). All four tests rejected $H_0$ if 0 was not in the interval. The fourth “test” is ad hoc since it has not yet been proven to have level $\alpha$. Since $nT_n = n\bar{x}^T \bar{x} - tr(S)$, the bootstrap test is also a competitor for the test based on $Z_2$.

Adapting an argument from Lehmann (1999, pp. 367-368), let $Z(a) = E(a^T x) = a^T \mu$. Then it can be shown that $\theta = V(Z(x_i)) = V(x_i^T \mu) \geq 0$. Thus we expect the first two tests to have good power because the estimated variance $\hat{V}(T_n)$, that is consistent under $H_0$, is too small when $H_0$ is not true.

The simulation used four distribution types where $x = Ay + \delta 1$ with $E(x) = \delta 1$ where $1$ is the $p \times 1$ vector of ones. Type 1 used $y \sim N_p(0, I)$, type 2 used a mixture distribution $y \sim 0.6 N_p(0, I) + 0.4 N_p(0, 25 I)$, type 3 for a multivariate $t_4$ distribution, and type 4 for a multivariate lognormal distribution where $y = (y_1, \ldots, y_p)$ with $w_i = \exp(Z)$ where $Z \sim N(0, 1)$ and $y_i = w_i - E(w_i)$ where $E(w_i) = \exp(0.5)$. The covariance matrix type depended on the matrix $A$. Type 1 used $A = I_p$, type 2 used $A = \text{diag}(\sqrt{1}, \ldots, \sqrt{p})$, and type 3 used $A = \psi^T I_p + (1 - \psi) I_p$ giving $\text{cor}(x_{ij}, x_{ik}) = \rho$ for $j \neq k$ where $\rho = 0$ if $\psi = 0$, $\rho \to 1/(c + 1)$ as $p \to \infty$ if $\psi = 1/\sqrt{p}$ where $c > 0$, and $\rho \to 1$ as $p \to \infty$ if $\psi \in (0, 1)$ is a constant. We used $\delta = 0$ and $\delta = 1$. The simulation used 5000 runs, the 4 $x$ distributions, and the 3 matrices $A$. For the third $A$, we used $\psi = 1/\sqrt{p}$. 

7
4 CONCLUSIONS

Some high dimensional one sample tests include Chen et al. (2011), Hyodo and Nishiyama (2017), Park and Ayyala (2013), Srivastava, and Du (2008), Wang, Peng, and Li (2015), and Zhao (2017). Hu and Bai (2015) also describes some tests.

Simulations were done in R. See R Core Team (2020). The collection of Olive (2023) R functions slpack, available from (http://parker.ad.siu.edu/Olive/slpack.txt), has some useful functions for the inference. The function hdhot1sim was used to simulate the four tests, while the function hdhot1sim2 simulates the first test, which is rather fast. This function avoided \( p \times p \) matrices to lessen memory problems.

5 References


Olive, D.J. (2023), Prediction and Statistical Learning, online course notes, see (http://parker.ad.siu.edu/Olive/slearnbk.htm).


