Chapter 9
High Dimensional Statistics

This chapter gives some results on high dimensional statistics. Some results for regression were already covered.

9.1 Introduction

Several statistical methods, covered in previous chapters, can be computed using an $n \times n$ matrix or a $p \times p$ matrix, depending on whether n or p is smaller. See Remark 3.14 for ridge regression and Section 9.1 for principle components analysis, which is used for principle components regression.

9.2 Principle Components Analysis

Principle components analysis (PCA) was used for PCR. See Chapter 3.

Suppose $\boldsymbol{W}$ is the standardized $n \times p$ data matrix and $\boldsymbol{T} = \boldsymbol{W}_g/\sqrt{n-g}$. If $n < p$, then the correlation matrix $\boldsymbol{R} = \boldsymbol{T}^T\boldsymbol{T} = \boldsymbol{W}_g^T\boldsymbol{W}_g/(n-g)$ does not have full rank. By singular value decomposition (SVD) theory, the SVD of $\boldsymbol{T}$ is $\boldsymbol{T} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T$ where the positive singular values σ_i are square roots of the positive eigenvalues of both $\boldsymbol{T}^T\boldsymbol{T}$ and of $\boldsymbol{T}\boldsymbol{T}^T$. (The singular values are **not** standard deviations.) Also $\boldsymbol{V} = (\hat{\boldsymbol{e}}_1 \; \hat{\boldsymbol{e}}_2 \; \cdots \; \hat{\boldsymbol{e}}_p)$, and $\boldsymbol{T}^T\boldsymbol{T}\hat{\boldsymbol{e}}_i = \sigma_i^2\hat{\boldsymbol{e}}_i$. Hence classical principal component analysis on the standardized data can be done using $\hat{\boldsymbol{e}}_i$ and $\hat{\lambda}_i = \sigma_i^2$. The SVD of $\boldsymbol{T}^T$ is $\boldsymbol{T}^T = \boldsymbol{V}\boldsymbol{\Lambda}^T\boldsymbol{U}^T$, and

$$\boldsymbol{T}\boldsymbol{T}^T = \frac{1}{n-g}\begin{bmatrix} \boldsymbol{w}_1^T\boldsymbol{w}_1 & \boldsymbol{w}_1^T\boldsymbol{w}_2 & \dots & \boldsymbol{w}_1^T\boldsymbol{w}_n \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{w}_n^T\boldsymbol{w}_1 & \boldsymbol{w}_n^T\boldsymbol{w}_2 & \dots & \boldsymbol{w}_n^T\boldsymbol{w}_n \end{bmatrix}$$

which is the matrix of scalar products divided by n. Similarly, if $\boldsymbol{W}_c$ is the centered data matrix (subtract the means), then $\boldsymbol{T}_c = \boldsymbol{W}_c/\sqrt{n-g}$, and the covariance matrix $\boldsymbol{S} = \boldsymbol{T}_c^T \boldsymbol{T}_c = \boldsymbol{W}_c^T \boldsymbol{W}_c/(n-g)$. For more information about the SVD, see Datta (1995, pp. 552-556) and Fogel et al. (2013).

The following output shows how to do classical PCA with $\boldsymbol{S}$ on a data set using the SVD and $g = 1$. The eigenvectors agree up to sign.

```
x<-cbind(buxx,buxy) # data matrix
mn <- apply(x,2,mean) #sample mean
J <- 0*1:87 + 1  # vector of n ones, n = 87
J <- J%*%t(J)/87 #J%*%x has rows = mn
zc <- x-J%*%x #centered x
yc <- zc/sqrt(87-1) #t(yc) %*% yc = cov(x)
svd(yc)$v             #right eigenvectors of Yc
          [,1]     [,2]     [,3]     [,4]    [,5]
[1,]  0.653883  0.75596 -0.01173  0.00988  0.0268
[2,] -0.001366  0.03980  0.06800 -0.42534 -0.9016
[3,] -0.000489 -0.01276 -0.99161 -0.12775 -0.0151
[4,] -0.000714  0.00251 -0.10890  0.89588 -0.4308
[5,] -0.756594  0.65327 -0.00952  0.00854  0.0252
> svd(t(yc))$u        #left eigenvectors of Yc^T
          [,1]     [,2]     [,3]     [,4]    [,5]
[1,] -0.653883 -0.75596  0.01173 -0.00988 -0.0268
[2,]  0.001366 -0.03980 -0.06800  0.42534  0.9016
[3,]  0.000489  0.01276  0.99161  0.12775  0.0151
[4,]  0.000714 -0.00251  0.10890 -0.89588  0.4308
[5,]  0.756594 -0.65327  0.00952 -0.00854 -0.0252
> prcomp(x)
Standard deviations:
[1] 523.70760  42.50435   6.06073   4.39067   3.80398
Rotation:
               PC1       PC2       PC3       PC4      PC5
len       0.653883  0.75596 -0.01173  0.00988  0.0268
nasal    -0.001366  0.03980  0.06800 -0.42534 -0.9016
bigonal  -0.000489 -0.01276 -0.99161 -0.12775 -0.0151
cephalic -0.000714  0.00251 -0.10890  0.89588 -0.4308
buxy     -0.756594  0.65327 -0.00952  0.00854  0.0252
svd(yc)$d         #singular values = sqrt(eigenvalues)
[1] 523.70760  42.50435   6.06073   4.39067   3.80398
svd(t(yc))$d      #singular values = sqrt(eigenvalues)
[1] 523.70760  42.50435   6.06073   4.39067   3.80398
```

Although PCA can be done if $p > n$, in general need p fixed for the sample eigenvector to be a good estimator of a population eigenvector.

9.3 MANOVA Type Tests

This section reviews Wald type tests and Wald type tests with the wrong dispersion matrix, and uses results from Rajapaksha and Olive (2022). Consider testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ where a $g \times 1$ statistic T_n satisfies $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u} \sim N_g(\boldsymbol{0}, \boldsymbol{\Sigma})$. If $\hat{\boldsymbol{\Sigma}}^{-1} \xrightarrow{P} \boldsymbol{\Sigma}^{-1}$ and H_0 is true, then

$$D_n^2 = D^2_{\boldsymbol{\theta}_0}(T_n, \hat{\boldsymbol{\Sigma}}/n) = n(T_n - \boldsymbol{\theta}_0)^T \hat{\boldsymbol{\Sigma}}^{-1} (T_n - \boldsymbol{\theta}_0) \xrightarrow{D} \boldsymbol{u}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{u} \sim \chi^2_g$$

as $n \to \infty$. Then a Wald type test rejects H_0 at significance level δ 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

$$D_n^2 = D^2_{\boldsymbol{\theta}_0}(T_n, \boldsymbol{C}_n/n) = n(T_n - \boldsymbol{\theta}_0)^T \boldsymbol{C}_n^{-1} (T_n - \boldsymbol{\theta}_0) \xrightarrow{D} \boldsymbol{u}^T \boldsymbol{C}^{-1} \boldsymbol{u}$$

as $n \to \infty$ if H_0 is true, where the $g \times g$ symmetric positive definite matrix $\boldsymbol{C}_n \xrightarrow{P} \boldsymbol{C} \neq \boldsymbol{\Sigma}$. Hence $\boldsymbol{C}_n$ is the wrong dispersion matrix, and $\boldsymbol{u}^T \boldsymbol{C}^{-1} \boldsymbol{u}$ does not have a χ^2_g distribution when H_0 is true. Often $\boldsymbol{C}_n$ is a regularized estimator of $\boldsymbol{\Sigma}$, or $\boldsymbol{C}_n^{-1}$ is a regularized estimator of the precision matrix $\boldsymbol{\Sigma}^{-1}$, such as $\boldsymbol{C}_n = diag(\hat{\boldsymbol{\Sigma}})$ or $\boldsymbol{C}_n = \boldsymbol{I}_g$, the $g \times g$ identity matrix. Another example is $\boldsymbol{C}_n = \boldsymbol{S}_p$, where $\boldsymbol{S}_p$ is a pooled covariance matrix, and it is assumed that the p groups have the same covariance matrix $\boldsymbol{\Sigma}$. When this assumption is violated, $\boldsymbol{C}_n$ is usually not a consistent estimator of $\boldsymbol{\Sigma}$. When the bootstrap is used, often $\boldsymbol{C}_n = n\boldsymbol{S}^*_T$ where $\boldsymbol{S}^*_T$ is the sample covariance matrix of the bootstrap sample $T^*_1, ..., T^*_B$. The assumption that $n\boldsymbol{S}^*_T$ is a consistent estimator of $\boldsymbol{\Sigma}$ is strong. See, for example, Machado and Parente (2005). Rajapaksha and Olive (2022) showed how to bootstrap Wald tests with the wrong dispersion matrix using the BR and PR bootstrap confidence regions from Definitions 2.19 and 2.20.

Some examples include the pooled t test and one-way ANOVA test. Rupasinghe Arachchige Don and Pelawa Watagoda (2018) and Rupasinghe Arachchige Don and Olive (2019) gave Wald type tests for analogs of the two sample Hotelling's T^2 and one-way MANOVA tests using a consistent estimator $\hat{\boldsymbol{\Sigma}}$ of $\boldsymbol{\Sigma}$. These tests could greatly outperform the classical tests that used the pooled covariance matrix when the sample sizes were large enough to give good estimates of the covariance matrix of each group, but for small sample sizes, the classical tests (with the wrong dispersion matrix) sometimes did better in the simulations.

The bootstrap is useful since if $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$ and $\sqrt{n}(T^*_n - T_n) \xrightarrow{D} \boldsymbol{u}$, then the percentiles of $n(T_n - \boldsymbol{\theta}_0)^T \boldsymbol{C}_n^{-1} (T_n - \boldsymbol{\theta}_0)$ can be estimated with the sample percentiles of $n(T^*_n - T_n)^T \boldsymbol{C}_n^{-1} (T^*_n - T_n)$. See Remark 2.20.

9.3.1 Large Sample Theory

One-way MANOVA type tests give a large class of Wald type tests and Wald type tests with the wrong dispersion matrix. Using double subscripts will be useful for describing these models. Suppose there are independent random samples of size n_i from p different populations (treatments), or n_i cases are randomly assigned to p treatment groups. Then $n = \sum_{i=1}^p n_i$ and the group sample sizes are n_i for $i = 1, ..., p$. Assume that m response variables $\boldsymbol{y}_{ij} = (Y_{ij1}, ..., Y_{ijm})^T$ are measured for the ith treatment group and the jth case in the group. Hence $i = 1, ..., p$ and $j = 1, ..., n_i$. Assume the p treatments have possibly different population location vectors $\boldsymbol{\mu}_i$, such as $E(\boldsymbol{y}_{ij}) = \boldsymbol{\mu}_i$. Coordinatewise population medians and coordinatewise population trimmed means are also useful. Then a one-way MANOVA type test is used to test $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_p$ versus the alternative that not all of the $\boldsymbol{\mu}_i$ are equal.

Large sample theory can be used to derive Wald type tests, although large sample theory is not the only solution. Let $\text{Cov}(\boldsymbol{y}_{ij}) = \boldsymbol{\Sigma}_i$ be the nonsingular population covariance matrix of the ith treatment group or population. To simplify the large sample theory, assume $n_i = \pi_i n$ where $0 < \pi_i < 1$ and $\sum_{i=1}^p \pi_i = 1$. Let T_i be a multivariate location estimator such that $\sqrt{n_i}(T_i - \boldsymbol{\mu}_i) \xrightarrow{D} N_m(\boldsymbol{0}, \boldsymbol{\Sigma}_i)$, and $\sqrt{n}(T_i - \boldsymbol{\mu}_i) \xrightarrow{D} N_m\left(\boldsymbol{0}, \frac{\boldsymbol{\Sigma}_i}{\pi_i}\right)$. Let $\boldsymbol{T} = (T_1^T, T_2^T, ..., T_p^T)^T$, $\boldsymbol{\nu} = (\boldsymbol{\mu}_1^T, \boldsymbol{\mu}_2^T, ..., \boldsymbol{\mu}_p^T)^T$, and $\boldsymbol{A}$ be a full rank $r \times mp$ matrix with rank r, then a large sample test of the form $H_0 : \boldsymbol{A\nu} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{A\nu} \neq \boldsymbol{\theta}_0$ uses

$$\boldsymbol{A}\sqrt{n}(\boldsymbol{T} - \boldsymbol{\nu}) \xrightarrow{D} \boldsymbol{u} \sim N_r\left(\boldsymbol{0}, \boldsymbol{A}\ diag\left(\frac{\boldsymbol{\Sigma}_1}{\pi_1}, \frac{\boldsymbol{\Sigma}_2}{\pi_2}, ..., \frac{\boldsymbol{\Sigma}_p}{\pi_p}\right)\ \boldsymbol{A}^T\right). \tag{9.1}$$

Let the Wald type statistic

$$t_0 = [\boldsymbol{AT} - \boldsymbol{\theta}_0]^T \left[\boldsymbol{A}\ diag\left(\frac{\hat{\boldsymbol{\Sigma}}_1}{n_1}, \frac{\hat{\boldsymbol{\Sigma}}_2}{n_2}, ..., \frac{\hat{\boldsymbol{\Sigma}}_p}{n_p}\right)\ \boldsymbol{A}^T\right]^{-1} [\boldsymbol{AT} - \boldsymbol{\theta}_0]. \tag{9.2}$$

These results prove the following theorem.

Theorem 9.1. Under the above conditions, $t_0 \xrightarrow{D} \chi^2_r$ if H_0 is true.

A useful fact for the F and chi-square distributions is $d_n F_{g,d_n,1-\delta} \to \chi^2_{g,1-\delta}$ as $d_n \to \infty$. Here $P(X \leq F_{g,d_n,1-\delta}) = 1 - \delta$ if $X \sim F_{g,d_n}$. Reject H_0 if $t_0/r > F_{g,d_n,1-\delta}$ where $d_n = min(n_i) = min(n_1, ..., n_p)$.

This one-way MANOVA type test was used by Rupasinghe Arachchige Don and Olive (2019), and a special case was used by Zhang and Liu (2013) and Konietschke et al. (2015) with $T_i = \overline{\boldsymbol{y}}_i$ and $\hat{\boldsymbol{\Sigma}}_i = \boldsymbol{S}_i$, the sample covariance matrix corresponding to the ith treatment group. The $p = 2$ case gives analogs to the two sample Hotelling's T^2 test. See Rupasinghe Arachchige Don and Pelawa Watagoda (2018).

Several tests use the common covariance matrix assumption $\boldsymbol{\Sigma}_i \equiv \boldsymbol{\Sigma}$ for $i = 1, ..., p$. These tests are Wald type tests with the wrong dispersion matrix if the common covariance matrix assumption is wrong. Examples include the pooled t test with $m = p = 1$, the one-way ANOVA test with $m = 1$, the two sample Hotelling's T^2 test (with common covariance matrix) with $p = 2$, and the one-way MANOVA test.

For the Rupasinghe Arachchige Don and Olive (2019) one-way MANOVA type test, let $\boldsymbol{A}$ be the $m(p-1) \times mp$ block matrix

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} & \dots & \text{-}\boldsymbol{I} \\ \boldsymbol{0} & \boldsymbol{I} & \boldsymbol{0} & \dots & \text{-}\boldsymbol{I} \\ \vdots & \vdots & \vdots & & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \dots & \boldsymbol{I} & \text{-}\boldsymbol{I} \end{bmatrix}.$$

Let $\boldsymbol{\mu}_i \equiv \boldsymbol{\mu}$, let $H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_p$ or, equivalently, $H_0 : \boldsymbol{A\nu} = \boldsymbol{0}$, and let

$$\boldsymbol{w} = \boldsymbol{AT} = \begin{bmatrix} T_1 - T_p \\ T_2 - T_p \\ \vdots \\ T_{p-2} - T_p \\ T_{p-1} - T_p \end{bmatrix}. \tag{9.3}$$

Then $\sqrt{n}\boldsymbol{w} \xrightarrow{D} N_{m(p-1)}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{w}})$ if H_0 is true with $\boldsymbol{\Sigma}_{\boldsymbol{w}} = (\boldsymbol{\Sigma}_{ij})$ where $\boldsymbol{\Sigma}_{ij} = \dfrac{\boldsymbol{\Sigma}_p}{\pi_p}$ for $i \neq j$, and $\boldsymbol{\Sigma}_{ii} = \dfrac{\boldsymbol{\Sigma}_i}{\pi_i} + \dfrac{\boldsymbol{\Sigma}_p}{\pi_p}$ for $i = j$. Hence

$$t_0 = n\boldsymbol{w}^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}^{-1} \boldsymbol{w} = \boldsymbol{w}^T \left(\frac{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}}{n} \right)^{-1} \boldsymbol{w} \xrightarrow{D} \chi^2_{m(p-1)}$$

as the $n_i \to \infty$ if H_0 is true. Here $\dfrac{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}}{n}$ is a block matrix where the off diagonal block entries equal $\hat{\boldsymbol{\Sigma}}_p / n_p$ and the ith diagonal block entry is $\dfrac{\hat{\boldsymbol{\Sigma}}_i}{n_i} + \dfrac{\hat{\boldsymbol{\Sigma}}_p}{n_p}$ for $i = 1, ..., (p-1)$. Reject H_0 if

$$t_0 > m(p-1) F_{m(p-1), d_n}(1 - \delta) \tag{9.4}$$

where $d_n = \min(n_1, ..., n_p)$. This Wald type test may start to outperform the one-way MANOVA test if $n \geq (m+p)^2$ and $n_i \geq 40m$ for $i = 1, ..., p$.

If $H_0 : \boldsymbol{A\nu} = \boldsymbol{\theta}_0$ is true, if the $\boldsymbol{\Sigma}_i \equiv \boldsymbol{\Sigma}$ for $i = 1, ..., p$, and if $\hat{\boldsymbol{\Sigma}}$ is a consistent estimator of $\boldsymbol{\Sigma}$, then by Theorem 9.1

$$t_0 = [\boldsymbol{AT} - \boldsymbol{\theta}_0]^T \left[\boldsymbol{A} \, diag\left(\frac{\hat{\boldsymbol{\Sigma}}}{n_1}, \frac{\hat{\boldsymbol{\Sigma}}}{n_2}, ..., \frac{\hat{\boldsymbol{\Sigma}}}{n_p} \right) \boldsymbol{A}^T \right]^{-1} [\boldsymbol{AT} - \boldsymbol{\theta}_0] \xrightarrow{D} \chi^2_r.$$

If H_0 is true but the $\boldsymbol{\Sigma}_i$ are not equal, then we get a bootstrap cutoff by using

$$t^*_{0i} = [\boldsymbol{AT}^*_i - \boldsymbol{AT}]^T \left[\boldsymbol{A} \, diag\left(\frac{\hat{\boldsymbol{\Sigma}}}{n_1}, \frac{\hat{\boldsymbol{\Sigma}}}{n_2}, ..., \frac{\hat{\boldsymbol{\Sigma}}}{n_p} \right) \boldsymbol{A}^T \right]^{-1} [\boldsymbol{AT}^*_i - \boldsymbol{AT}] =$$

$$D^2_{\boldsymbol{AT}^*_i} \left(\boldsymbol{AT}, \boldsymbol{A} \, diag\left(\frac{\hat{\boldsymbol{\Sigma}}}{n_1}, \frac{\hat{\boldsymbol{\Sigma}}}{n_2}, ..., \frac{\hat{\boldsymbol{\Sigma}}}{n_p} \right) \boldsymbol{A}^T \right).$$

Let $F_0 = t_0/r$. Then we can get a bootstrap cutoff using $F^*_{0i} = t^*_{0i}/r$. For $T_i = \overline{\boldsymbol{y}}_i$, let $\hat{\boldsymbol{\Sigma}}$ be the usual pooled covariance matrix estimator.

For Theorem 9.2, $(n-p)U = t_0 \xrightarrow{D} \chi^2_{m(p-1)}$ follows trivially from Theorem 9.1, under the equal covariance matrix assumption. Fujikoshi (2002) also showed $(n-p)U \xrightarrow{D} \chi^2_{m(p-1)}$. Kakizawa (2009) also gave large sample theory for some MANOVA tests. Lengthy calculations show $(n-p)U = t_0$. See Rajapaksha (2021) for details.

Theorem 9.2. For the one-way MANOVA test using $\boldsymbol{\theta}_0 = \boldsymbol{0}$, $\boldsymbol{A}$ as defined above Equation (9.3), and $T_i = \overline{\boldsymbol{y}}_i$,

$$(n-p)U = t_0 = [\boldsymbol{AT}]^T \left[\boldsymbol{A} \, diag\left(\frac{\hat{\boldsymbol{\Sigma}}}{n_1}, \frac{\hat{\boldsymbol{\Sigma}}}{n_2}, ..., \frac{\hat{\boldsymbol{\Sigma}}}{n_p} \right) \boldsymbol{A}^T \right]^{-1} [\boldsymbol{AT}]$$

where U is the Hotelling Lawley trace statistic. Hence if the $\boldsymbol{\Sigma}_i \equiv \boldsymbol{\Sigma}$ and $H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_p$ is true, then $(n-p)U = t_0 \xrightarrow{D} \chi^2_{m(p-1)}$.

9.3.2 One Sample Hotelling T^2 Type Tests

Suppose there is a random sample $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ from a population. A common multivariate one sample test of hypotheses is $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ versus $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$ where $\boldsymbol{\mu}$ is a population location measure of the population. When n is much larger than p, the one sample Hotelling (1931) T^2 test is often used. If the $\boldsymbol{x}_i$ are iid with expected value $E(\boldsymbol{x}_i) = \boldsymbol{\mu}$ and nonsingular covariance matrix $Cov(\boldsymbol{x}_i) = \boldsymbol{\Sigma}$, then by the multivariate central limit theorem

$$\sqrt{n}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma}).$$

If H_0 is true, then

$$T_H^2 = n(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0)^T \boldsymbol{S}^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0) \xrightarrow{D} \chi^2_p.$$

The one sample Hotelling's T^2 test rejects H_0 if $T_H^2 > D^2_{1-\delta}$ where $D^2_{1-\delta} = \chi^2_{p,\delta}$ and $P(Y \le \chi^2_{p,\delta}) = \delta$ 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 $P(Y \le F_{p,d,\delta}) = \delta$ if $Y \sim F_{p,d}$. The scaled F cutoff can be used since $T_H^2 \xrightarrow{D} \chi^2_p$ if H_0 holds, and

$$\frac{(n-1)p}{n-p} F_{p,n-p,1-\delta} \to \chi^2_{p,1-\delta}$$

as $n \to \infty$.

Suppose there is a random sample $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$, and that it is desired to test $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ versus $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$ where $\boldsymbol{\mu}$ is a $p \times 1$ vector. We will use $\boldsymbol{\mu} = E(\boldsymbol{x}_i)$. Let the test statistic $T_n = \overline{\boldsymbol{x}}$ and the bootstrapped test statistic $T^* = \overline{\boldsymbol{x}}^*$ where the nonparametric bootstrap is used. Hence n cases are drawn with replacement from the sample to form $\overline{\boldsymbol{x}}^*$. We will also use T_n = the coordinatewise median where $\boldsymbol{\mu}$ is the population coordinatewise median. We will use $\boldsymbol{C}_n = \boldsymbol{C}_n^{-1} = \boldsymbol{I}_p$. Let $\boldsymbol{\theta} = \boldsymbol{\mu}_0 = \boldsymbol{0}$.

The first large sample $100(1-\delta)\%$ confidence region is

$$\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T \boldsymbol{C}_n^{-1} (\boldsymbol{w} - T_n) \le D^2_{(U_{B,T})}\} =$$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{I}) \le D^2_{(U_{B,T})}\} \tag{9.5}$$

where the cutoff $D^2_{(U_{B,T})}$ is the $100(1-\alpha)$th sample quantile of the squared Euclidean distance $D_i^2 = (T_i^* - T_n)^T (T_i^* - T_n)$. Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{0}$ rejects H_0 if $(T_n - \boldsymbol{0})^T (T_n - \boldsymbol{0}) > D^2_{(U_{B,T})}$.

The second large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ is

$$\{\boldsymbol{w} : (\boldsymbol{w} - \overline{T}^*)^T \boldsymbol{C}_n^{-1} (\boldsymbol{w} - \overline{T}^*) \le D^2_{(U_B)}\} =$$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}^*, \boldsymbol{I}) \le D^2_{(U_B)}\} \tag{9.6}$$

where the cutoff $D^2_{(U_B)}$ is the $100(1-\alpha)$th sample quantile of the squared Euclidean distance $D_i^2 = (T_i^* - \overline{T}^*)^T (T_i^* - \overline{T}^*)$ for $i = 1, ..., B$. Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{0}$ rejects H_0 if $(\overline{T}^* - \boldsymbol{0})^T (\overline{T}^* - \boldsymbol{0}) > D^2_{(U_B)}$.

The test uses the result that $\sqrt{n}(\overline{\boldsymbol{x}} - \boldsymbol{u}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$ and $\sqrt{n}(\overline{\boldsymbol{x}}^* - \overline{\boldsymbol{x}}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$. Since $\boldsymbol{I}$ is independent of the bootstrap sample, correction factors for the bootstrap cutoffs were not needed. Since the sample quantile is that of a random variable, B does not need to be large. If $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \boldsymbol{I}$, then

$$(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^T \boldsymbol{I}^{-1} (\overline{\boldsymbol{x}} - \boldsymbol{\mu}) \approx \frac{1}{n} \chi^2_p$$

since

$$n(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^T \boldsymbol{I}^{-1} (\overline{\boldsymbol{x}} - \boldsymbol{\mu}) \xrightarrow{D} \chi^2_p$$

as $n \to \infty$. For high dimensional data with $p \geq n$, we still have $E(\overline{\boldsymbol{x}}) = \boldsymbol{\mu}$, $\text{Cov}(\overline{\boldsymbol{x}}) = \boldsymbol{\Sigma}_{\boldsymbol{x}}/n$, $E(\overline{\boldsymbol{x}}^*) = \overline{\boldsymbol{x}}$, and $\text{Cov}(\overline{\boldsymbol{x}}^*) = (n-1)\boldsymbol{S}/n^2$.

$\boldsymbol{C}_n^{-1} = \boldsymbol{I}$ can be replaced by $\boldsymbol{C}_n^{-1} = diag(1/S_1^2, ..., 1/S_p^2)$ where $S_i^2 = S_{ii}$ when the sample covariance matrix $\boldsymbol{S} = (S_{ij})$. Other choices of $\boldsymbol{C}_n$ can be used as long as the computational complexity of $\boldsymbol{C}_n^{-1}$ is not too high.

The `mpack` function *hdhot1wsim* was used for the simulation.

The argument `xtype` gives the multivariate distribution of $\boldsymbol{x}$ where $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$. Hence `xtype` $= 1$ for $\boldsymbol{x} \sim N_p(\boldsymbol{0}, \boldsymbol{I})$, `xtype` $= 2$ for a mixture distribution $\boldsymbol{x} \sim 0.6N_p(\boldsymbol{0}, \boldsymbol{I}) + 0.4N_p(\boldsymbol{0}, 25\boldsymbol{I})$ for the default argument `eps` $= 0.4$, `xtype` $= 3$ for a multivariate t_4 distribution for the default argument `dd` $= 4$, and `xtype` $= 4$ for a multivariate lognormal distribution where $\boldsymbol{x} = (x_1, ..., x_p)$ with $w_i = \exp(Z)$ where $Z \sim N(0, 1)$ and $x_i = w_i - E(w_i)$ where $E(w_i) = \exp(0.5)$. The argument `covtyp` $= 1$ if $\boldsymbol{A} = \boldsymbol{I}$ so, and `covtyp` $= 2$ if $\boldsymbol{A} = diag(\sqrt{1}, ..., \sqrt{p})$. When `covtyp` $= 3$, $\text{cor}(Y_i, Y_j) = \rho$ where $\rho = 0$ if $\psi = 0$, $\rho \to 1/(c+1)$ as $p \to \infty$ if $\psi = 1/\sqrt{cp}$ where $c > 0$, and $\rho \to 1$ as $p \to \infty$ if $\psi \in (0, 1)$ is a constant. $E(\boldsymbol{x}) = \delta\boldsymbol{1}$ where $\boldsymbol{1}$ is the $p \times 1$ vector of ones. Then the argument `delta` $= \delta$.

The first three distributions have mean $\boldsymbol{\mu} = E(\boldsymbol{x})$ equal to the population coordinatewise median since the distributions are elliptically contoured distributions with center $\boldsymbol{\mu}$. The fourth distribution does not have $E(\boldsymbol{x}) =$ the population coordinatewise median. Hence if $H_0 : \boldsymbol{\mu} = \boldsymbol{0}$ is true for $\boldsymbol{\mu} = E(\boldsymbol{x})$, then H_0 is false if $\boldsymbol{\mu}$ is the population coordinatewise median.

The simulation used 5000 runs, the 4 xtypes, and the 3 covtyps. We used $n = 100$ and $p = 10, 100, 200, 400$. For covty=3, we used $\psi = 1/\sqrt{p}$. We used delta $= 0$ and delta $= 1$. For $\delta = 0$, expect coverage to be less than 0.1 as p increases.

Consider testing $H_0 : \boldsymbol{\mu} = \boldsymbol{0}$ versus $H_A : \boldsymbol{\mu} \neq \boldsymbol{0}$ using independent and identically distributed (iid) $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ where the $\boldsymbol{x}_i$ are $p \times 1$ random vectors and p may be much larger than n. Replace $\boldsymbol{x}_i$ by $\boldsymbol{w}_i = \boldsymbol{x}_i - \boldsymbol{\mu}_0$ to test $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ versus $H_A : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$.

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

$$c_{p,n} = 1 + \frac{tr(\boldsymbol{R}^2)}{p^{3/2}}.$$

Let $n = O(p^\delta)$ where $0.5 < \delta \leq n$. Then under regularity conditions

$$Z_1 = \frac{n\overline{\boldsymbol{x}}^T \boldsymbol{D}^{-1} \overline{\boldsymbol{x}} - \frac{(n-1)p}{n-3}}{2\left(tr(\boldsymbol{R}^2) - \frac{p^2}{n-1}\right)} \xrightarrow{D} N(0,1)$$

as $n, p \to \infty$. The next test is attributed to Bai and Saranadasa (1996). Suppose $p/n \to c > 0$. Under regularity conditions,

$$Z_2 = \frac{n\overline{\boldsymbol{x}}^T \overline{\boldsymbol{x}} - tr(\boldsymbol{S})}{\left[\frac{2(n-1)n}{(n-2)(n+1)}\left(tr(\boldsymbol{S}^2) - \frac{1}{n}[tr(\boldsymbol{S})]^2\right)\right]^{1/2}} \xrightarrow{D} N(0,1)$$

as $n, p \to \infty$. Both of these test statistics needed $p/n \to c > 0$ or $p/n^2 \to 0$. Hence p can not be too big.

There are test statistics T_n for testing $H_0 : \boldsymbol{\mu} = \boldsymbol{0}$ where p can be much larger with

$$\frac{T_n}{s_n} \xrightarrow{D} N(0,1)$$

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 $\boldsymbol{a} = \sum_{i=1}^n \boldsymbol{x}_i$ and let $\boldsymbol{X} = (x_{ij})$ be the data matrix with ith row $= \boldsymbol{x}_i^T$ and ij element $= x_{ij}$. Let $vec(\boldsymbol{A})$ stack the columns of matrix $\boldsymbol{A}$ so that $\boldsymbol{c} = vec(\boldsymbol{X}^T) = [\boldsymbol{x}_1^T, \boldsymbol{x}_2^T, ..., \boldsymbol{x}_n^T]^T$. Then

$$\boldsymbol{c}^T \boldsymbol{c} = \sum_{i=1}^n \boldsymbol{x}_i^T \boldsymbol{x}_i = \sum_{i=1}^n \|\boldsymbol{x}_i\|^2 = \sum_{i=1}^n \sum_{j=1}^p (x_{ij})^2.$$

Let $T_n =$

$$\frac{1}{n(n-1)}[\boldsymbol{a}^T \boldsymbol{a} - \boldsymbol{c}^T \boldsymbol{c}] = \frac{1}{n(n-1)} \sum \sum_{i \neq j} \boldsymbol{x}_i^T \boldsymbol{x}_j = \frac{1}{n(n-1)} \sum_{i \neq j} \boldsymbol{x}_i^T \boldsymbol{x}_j. \quad (9.7)$$

The terms in $\boldsymbol{c}^T \boldsymbol{c} = \sum_{i=1}^n \boldsymbol{x}_i^T \boldsymbol{x}_i$ are the terms that cause the restriction on p for asymptotic normality for the previous two tests. Under $H_0 : \boldsymbol{\mu} = \boldsymbol{0}$ and additional regularity conditions,

$$\frac{T_n}{s_n} \xrightarrow{D} N(0,1)$$

where s_n is rather hard to compute. Here

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

where $\overline{\boldsymbol{x}}_{(i,j)}$ is the sample mean computed without $\boldsymbol{x}_i$ or $\boldsymbol{x}_j$:

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

The T_n in Equation (9.7) can be viewed as a modification of $\|\overline{\boldsymbol{x}}\|^2 = \overline{\boldsymbol{x}}^T \overline{\boldsymbol{x}}$ that is a better estimator of $\boldsymbol{\mu}^T \boldsymbol{\mu}$ in high dimensions. Note that $\boldsymbol{\mu} = \boldsymbol{0}$ iff $\boldsymbol{\mu}^T \boldsymbol{\mu} = 0$ and $E(T_n) = E(\boldsymbol{x}_i^T \boldsymbol{x}_j) = \boldsymbol{\mu}^T \boldsymbol{\mu}$ if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are iid with $E(\boldsymbol{x}_i) = \boldsymbol{\mu}$ and $i \neq j$.

The bootstrap often works well on such statistics, but the nonparametric bootstrap fails because terms like $\boldsymbol{x}_j^T \boldsymbol{x}_j$ 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 $T_m^* \approx N(0, s_m^2)$ 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 σ_m^2 which may be less than s_m^2. Here $\overline{T}^{**}$ is the sample mean of all $\binom{n}{m}$ 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 $m/n \to 0$ with $m \to \infty$. Sampling without replacement is like sampling with replacement when $n >> m$, and sampling with replacement leads to iid T_m^* with respect to the bootstrap distribution. Heuristically, the T_m^* may be approximately iid $N(\overline{T}^{**}, s_m^2)$ if $m/n \to 0$ and $m \to \infty$. The *slpack* program `hdhot1sim` uses $m = floor(2n/3)$ and worked well in simulations. This choice of m gives an ad hoc test unless theory can be given for the test.

Let W_i be an indicator random variable with $W_i = 1$ if $\boldsymbol{x}_i^*$ 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 $W_{ij} = W_i W_j$ with $i \neq j$. Again, the W_{ij} are binary and identically distributed. $P(W_{ij} = 1) =$ P(ordered pair $(\boldsymbol{x}_i, \boldsymbol{x}_j)$) 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

$$T_m^* = \frac{1}{m(m-1)} \sum \sum_{k \neq d} \boldsymbol{x}_{i_k}^T \boldsymbol{x}_{i_d} = \frac{1}{m(m-1)} \sum \sum_{i \neq j} W_i W_j \boldsymbol{x}_i^T \boldsymbol{x}_j$$

where the $\boldsymbol{x}_{i_1}, ..., \boldsymbol{x}_{i_m}$ are the m vectors $\boldsymbol{x}_i$ selected in the sample. The first double sum has $m(m-1)$ terms while the second double sum has $n(n-1)$ terms. Hence

$$E(T_m^*) = \frac{1}{m(m-1)} \sum \sum_{i \neq j} E[W_i W_j] \boldsymbol{x}_i^T \boldsymbol{x}_j = T_n.$$

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

To find the variance $V(T_n)$ from Equation (9.7), let $W_{ij} = \boldsymbol{x}_i^T \boldsymbol{x}_j = W_{ji}$, and note that

$$T_n = \frac{2}{n(n-1)} H_n \quad \text{where} \quad \mathrm{H_n} = \sum_{\mathrm{i} <} \sum_{\mathrm{j}} \boldsymbol{x}_\mathrm{i}^\mathrm{T} \boldsymbol{x}_\mathrm{j} = \sum_{\mathrm{i<j}} \boldsymbol{x}_\mathrm{i}^\mathrm{T} \boldsymbol{x}_\mathrm{j}.$$

Then $V(H_n) = \text{Cov}(H_n, H_n) =$

$$\text{Cov}(\sum_{i <} \sum_{j} W_{ij}, \sum_{k <} \sum_{d} W_{kd}) = \sum_{i <} \sum_{j} \sum_{k <} \sum_{d} \text{Cov}(W_{ij}, W_{kd}). \tag{9.8}$$

Let $V(W_{ij}) = \sigma^2_W$ for $i \neq j$. The covariances are of 3 types. First, if $(ij) = (kd)$, then $\text{Cov}(W_{ij}, W_{kd}) = V(W_{ij}) = \sigma^2_W$. There are $n(n-1)/2$ such terms. Second, if i, j, k, d are distinct, 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 θ 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 (9.8), b is the number of terms where i, j, k, d are distinct, and c is the number of terms where $(ij) = (kd)$.

$$V(H_n) = 0.5n(n-1)\sigma^2_W + n(n-1)(n-2)\theta.$$

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

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

It can be shown that $\theta = 0$ if $\boldsymbol{\mu} = \mathbf{0}$. Hence the test based on (9.7) can be good if $\sqrt{2\sigma^2_W/n^2}$ is small where σ^2_W does depend on p.

The following test has simple large sample theory, and can be good if $\sqrt{\sigma^2_W/n}$ is small. Hence we expect the test based on (9.7) to be better. Some notation for the simple test is needed. Assume $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid, $E(\boldsymbol{x}_i) = \boldsymbol{\mu}$ and the variance $V(\boldsymbol{x}_i^T \boldsymbol{x}_j) = \sigma^2_W$ for $i \neq j$. Let $m = \text{floor(n/2)} = \lfloor \text{n/2} \rfloor$ be the integer part of $n/2$. So floor(100/2) = floor(101/2) = 50. Let the iid random variables $W_i = \boldsymbol{x}_{2i-1}^T \boldsymbol{x}_{2i}$ for $i = 1, ..., m$. Hence $W_1, W_2, ..., W_m = \boldsymbol{x}_1^T \boldsymbol{x}_2, \boldsymbol{x}_3^T \boldsymbol{x}_4, ..., \boldsymbol{x}_{2m-1}^T \boldsymbol{x}_{2m}$. Note that $E(W_i) = \boldsymbol{\mu}^T \boldsymbol{\mu}$ and $V(W_i) = \sigma^2_W$. Let S^2_W 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 $\sigma_W^2 \propto \tau p$, then n may not be large enough for the normal approximation to hold. The following theorem follows from the univariate central limit theorem.

Theorem 9.1. Assume $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid, $E(\boldsymbol{x}_i) = \boldsymbol{\mu}$, and the variance $V(\boldsymbol{x}_i^T \boldsymbol{x}_j) = \sigma_W^2$ for $i \neq j$. Let $W_1, ..., W_m$ be defined as above. Then
a) $\sqrt{m}(\overline{W} - \boldsymbol{\mu}^T \boldsymbol{\mu}) \xrightarrow{D} N(0, \sigma_W^2)$.

$$b) \ \frac{\sqrt{m}(\overline{W} - \boldsymbol{\mu}^T \boldsymbol{\mu})}{S_W} \xrightarrow{D} N(0,1)$$

as $n \to \infty$.

9.3.3 Two Sample Hotelling T^2 Type Tests

Suppose there are two independent random samples from two populations or groups. A common multivariate two sample test of hypotheses is $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ versus $H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ where $\boldsymbol{\mu}_i$ is a population location measure of the ith population for $i = 1, 2$. The two sample Hotelling's T^2 test is the classical method for the test.

Suppose there are two independent random samples $\boldsymbol{x}_{1,1}, ..., \boldsymbol{x}_{n_1,1}$ and $\boldsymbol{x}_{1,2}, ..., \boldsymbol{x}_{n_2,2}$ from two populations or groups, and that it is desired to test $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ versus $H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ where $\boldsymbol{\mu}_i$ are $m \times 1$ vectors. Let $n = n_1 + n_2$.

The classical test uses

$$T_C^2 = (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^T \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\hat{\boldsymbol{\Sigma}}_{pool}\right]^{-1} (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)$$

where

$$\hat{\boldsymbol{\Sigma}}_{pool} = \frac{(n_1 - 1)\boldsymbol{S}_1 + (n_2 - 1)\boldsymbol{S}_2}{n-2}.$$

Then reject H_0 if $T_C^2 > mF_{m,n-2,1-\alpha}$.

The large sample test uses

$$T_L^2 = (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^T \left(\frac{\boldsymbol{S}_1}{n_1} + \frac{\boldsymbol{S}_2}{n_2}\right)^{-1} (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2).$$

Let $d_n = \min(n_1 - p, n_2 - p)$. Then reject H_0 if $T_L^2 > mF_{m,d_n,1-\alpha}$.

Note that $T_C^2 \approx T_L^2$ if $n_1 \approx n_2 \geq 20m$ and the two tests are asymptotically equivalent if $n_i/n \to 0.5$ as $n_1, n_2 \to \infty$. The BR bootstrap cutoff for the classical test uses

$$D_i^2 = (T_i^* - T_n)^T \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) \hat{\boldsymbol{\Sigma}}_{pool} \right]^{-1} (T_i^* - T_n)$$

where $T_n = (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)$ and $T_i^* = (\overline{\boldsymbol{x}}_{1i}^* - \overline{\boldsymbol{x}}_{2i}^*)$. We also use the PR and BR bootstrap tests for the test statistic

$$(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^T (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)$$

that uses $C_n = \boldsymbol{I}$. These two tests are also used in Section 9.

The data distributions in the simulation are the same as those described in Section 9.3.2, but $n_i \geq 10m$. For the classical test, there are distributions where T_C^2 is too large compared to the cutoff, resulting in large type I error, and there are distributions where T_C^2 is too small compared to the cutoff, resulting in small type I error. For highly skewed data, large n_i were often needed before the large sample test had type I error close to the nominal, but the type I error tended to be less than 0.12 when the nominal type I error was 0.05. The tests using $\boldsymbol{C}_n$ tended to have type I error close to the nominal, at the cost of producing a confidence region that has a large volume.

Suppose there are two independent random samples from two populations or groups. A common multivariate two sample test of hypotheses is $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ versus $H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ where $\boldsymbol{\mu}_i$ is a population location measure of the ith population for $i = 1, 2$. The two sample Hotelling's T^2 test is the classical method for the test.

Suppose there are two independent random samples $\boldsymbol{x}_{1,1}, ..., \boldsymbol{x}_{n_1,1}$ and $\boldsymbol{x}_{1,2}, ..., \boldsymbol{x}_{n_2,2}$ from two populations or groups, and that it is desired to test $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ versus $H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ where $\boldsymbol{\mu}_i$ are $m \times 1$ vectors. We will use $\boldsymbol{\mu}_i = E(\boldsymbol{x}_i)$, and $p > n_i$ is possible. Let the test statistic $T_n = \overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2$ and the bootstrapped test statistic $T^* = \overline{\boldsymbol{x}}_1^* - \overline{\boldsymbol{x}}_2^*$ where the nonparametric bootstrap is used. Hence n_i cases are drawn with replacement from sample i to form $\overline{\boldsymbol{x}}_i^*$. We will use $\boldsymbol{C}_n = \boldsymbol{C}_n^{-1} = \boldsymbol{I}_m$. Let $\boldsymbol{\theta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$.

The first large sample $100(1-\delta)\%$ confidence region is

$$\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T \boldsymbol{C}_n^{-1} (\boldsymbol{w} - T_n) \leq D^2_{(U_B,T)}\} = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{I}) \leq D^2_{(U_B,T)}\} \tag{9.9}$$

where the cutoff $D^2_{(U_B,T)}$ is the $100(1-\alpha)$th sample quantile of the squared Euclidean distance $D_i^2 = (T_i^* - T_n)^T (T_i^* - T_n)$. Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{0}$ rejects H_0 if $(T_n - \boldsymbol{0})^T (T_n - \boldsymbol{0}) > D^2_{(U_B,T)}$.

The second large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ is

$$\{\boldsymbol{w} : (\boldsymbol{w} - \overline{T}^*)^T \boldsymbol{C}_n^{-1} (\boldsymbol{w} - \overline{T}^*) \leq D^2_{(U_B)}\} = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}^*, \boldsymbol{I}) \leq D^2_{(U_B)}\} \tag{9.10}$$

where the cutoff $D^2_{(U_B)}$ is the $100(1-\alpha)$th sample quantile of the squared Euclidean distance $D_i^2 = (T_i^* - \overline{T}^*)^T (T_i^* - \overline{T}^*)$ for $i = 1, ..., B$. Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{0}$ rejects H_0 if $(\overline{T}^* - \boldsymbol{0})^T (\overline{T}^* - \boldsymbol{0}) > D^2_{(U_B)}$.

The test uses the result that $\sqrt{n}(\overline{\boldsymbol{x}} - \boldsymbol{u}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$ and $\sqrt{n}(\overline{\boldsymbol{x}}^* - \overline{\boldsymbol{x}}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$. Since $\boldsymbol{I}$ is independent of the bootstrap sample, correction factors for the bootstrap cutoffs were not needed. Since the sample quantile is that of a random variable, B does not need to be large. If $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}_{\boldsymbol{x}_i} = \boldsymbol{I}$, and $n_1 = n_2 = k$, then

$$(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^T \boldsymbol{I}^{-1} (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2) \approx \frac{2}{k} \chi^2_m$$

since

$$(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^T (2\boldsymbol{I}/k)^{-1} (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2) \xrightarrow{D} \chi^2_m$$

as $k \to \infty$.

Four types of data distributions $\boldsymbol{w}_i$ were considered that were identical for $i = 1, 2$. Then $\boldsymbol{x}_1 = \boldsymbol{A}\boldsymbol{w}_1 + \delta\boldsymbol{1}$ and $\boldsymbol{x}_2 = \sigma \boldsymbol{B}\boldsymbol{w}_2$ where $\boldsymbol{1} = (1, .., 1)^T$ is a vector of ones. We used $\boldsymbol{A} = \boldsymbol{B} = diag(1, \sqrt{2}, ..., \sqrt{m})$, $\boldsymbol{A} = \boldsymbol{B} = \boldsymbol{I}$, and $\boldsymbol{A} = \boldsymbol{I}$ with $\boldsymbol{B} = diag(1, \sqrt{2}, ..., \sqrt{m})$. The $\boldsymbol{w}_i$ distributions were the multivariate normal distribution $N_p(\boldsymbol{0}, \boldsymbol{I})$, the multivariate t distribution with 4 degrees of freedom, the mixture distribution $0.6N_m(\boldsymbol{0}, \boldsymbol{I}) + 0.4N_m(\boldsymbol{0}, 25\boldsymbol{I})$, and the multivariate lognormal distribution shifted to have zero mean. Note that $\text{Cov}(\boldsymbol{x}_2) = \sigma^2 \text{ Cov}(\boldsymbol{x}_1)$ when $\boldsymbol{A} = \boldsymbol{B}$, and $E(\boldsymbol{x}_i) = E(\boldsymbol{w}_i) = \boldsymbol{0}$ if $\delta = 0$.

The `mpack` function *hdhot2wsim* was used for the simulation.

There are test statistics T_n for testing $H_0 : \mu_1 = \mu_2$ where p can be much larger with

$$\frac{T_n}{s_n} \xrightarrow{D} N(0, 1)$$

where T_n is relatively simple to compute while s_n is much harder to compute. Let $\boldsymbol{a} = \sum_{i=1}^{n_1} \boldsymbol{x}_{1i}$ and let $\boldsymbol{X}_1 = (x_{1ij})$ be the data matrix with ith row $= \boldsymbol{x}_{1i}^T$ and ij element $= x_{1ij}$. Let $vec(\boldsymbol{A})$ stack the columns of matrix $\boldsymbol{A}$ so that $\boldsymbol{c} = vec(\boldsymbol{X}_1^T) = [\boldsymbol{x}_{11}^T, \boldsymbol{x}_{12}^T, ..., \boldsymbol{x}_{1n_1}^T]^T$. Then

$$\boldsymbol{c}^T \boldsymbol{c} = \sum_{i=1}^{n1} \boldsymbol{x}_{1i}^T \boldsymbol{x}_{1i} = \sum_{i=1}^{n_1} \|\boldsymbol{x}_{1i}\|^2 = \sum_{i=1}^{n_1} \sum_{j=1}^{p} (x_{1ij})^2.$$

Let $\boldsymbol{b} = \sum_{i=1}^{n_2} \boldsymbol{x}_{2i}$ and let $\boldsymbol{X}_2 = (x_{2ij})$ be the data matrix with ith row $= \boldsymbol{x}_{2i}^T$ and ij element $= x_{2ij}$. Let $\boldsymbol{d} = vec(\boldsymbol{X}_2^T) = [\boldsymbol{x}_{21}^T, \boldsymbol{x}_{22}^T, ..., \boldsymbol{x}_{2n_2}^T]^T$. Then

$$\boldsymbol{d}^T \boldsymbol{d} = \sum_{i=1}^{n_2} \boldsymbol{x}_{2i}^T \boldsymbol{x}_{2i} = \sum_{i=1}^{n_2} \|\boldsymbol{x}_{2i}\|^2 = \sum_{i=1}^{n_2} \sum_{j=1}^{p} (x_{2ij})^2.$$

Note that $\|\boldsymbol{a} - \boldsymbol{b}\|^2 = \boldsymbol{a}^T \boldsymbol{a} + \boldsymbol{b}^T \boldsymbol{b} - 2\boldsymbol{a}^T \boldsymbol{a}$, and let

$$T_n = \frac{1}{n_1(n_1 - 1)} [\boldsymbol{a}^T \boldsymbol{a} - \boldsymbol{c}^T \boldsymbol{c}] \; + \; \frac{1}{n_2(n_2 - 1)} [\boldsymbol{b}^T \boldsymbol{b} - \boldsymbol{d}^T \boldsymbol{d}] \; - \; \frac{2\boldsymbol{a}^T \boldsymbol{b}}{n_1 n_2}.$$

The terms in $\boldsymbol{c}^T\boldsymbol{c}$ and $\boldsymbol{d}^T\boldsymbol{d}$ are the terms that cause the restriction on p for asymptotic normality. Under $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ and additional regularity conditions,

$$\frac{T_n}{s_n} \xrightarrow{D} N(0,1)$$

where s_n is rather hard to compute. See Hu and Bai (2015) and Chen and Qin (2010).

The m out of n bootstrap without replacement draws a sample of size m_i without replacement from the n_i cases, $i = 1, 2$. For $B = 1$, this is a data splitting estimator, and $T_m^* \approx N(0, s_m^2)$ 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 σ_m^2 which may be less than s_m^2. Here $\overline{T}^{**}$ is the sample mean of all $\binom{n_1}{m_1} + \binom{n_2}{m_2}$ statistics T_m^* obtained by drawing a sample of size m_i with replacement from n_i. Heuristically, the T_m^* may be approximately iid $N(\overline{T}^{**}, s_m^2)$ if $m_i/n \to 0$ and $m_i \to \infty$.

The *slpack* program `hdhot2sim` uses $m_i = floor(2n_i/3)$ and worked well in simulations. This choice of m_i gives an ad hoc test unless theory can be given for the test.

9.4 One Way MANOVA Type Tests

9.5 Summary

9.6 Complements

Jolliffe (2010) is an authoritative text on PCA. Mϕller et al. (2005) discussed PCA, principal component regression, and drawbacks of M estimators. Olive (2017b) discussed outlier resistant PCA methods. Koch (2014) has some interesting results on high dimensional PCA.

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

Some high dimensional two sample tests include Feng et al. (2015), Feng and Sun (2015), and Gregory et al. (2015). Tests that assume $\boldsymbol{\Sigma}_{\boldsymbol{x}_1} = \boldsymbol{\Sigma}_{\boldsymbol{x}_2}$ can have nice large sample theory, but the equal covariance matrix assumption is too strong.

9.7 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

9.1. Consider the data set 6, 3, 8, 5, and 2. Show work.