Bootstrapping Analogs of the Two Sample Hotelling's T^2 Test

Hasthika S. Rupasinghe Arachchige Don and Lasanthi C. R. Pelawa Watagoda
Department of Mathematics
Southern Illinois University
Carbondale, Illinois 62901-4408
hasthika@siu.edu, lasanthi@.siu.edu

Keywords Behrens–Fisher problem; bootstrap; prediction region; coordinatewise median; RMVN estimator.

Mathematics Subject Classification Primary 62H15; Secondary 62G09.

Abstract

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, and is a special case of the one way MANOVA model if the two populations are assumed to have the same population covariance matrix. This paper suggests using the Olive (2016, 2017ab) bootstrap technique to develop analogs of Hotelling's T^2 test. The new tests can have considerable outlier resistance, and the tests do not need the population covariance matrices to be equal.

1. Introduction

This paper develops analogs of the two sample Hotelling's T^2 test that use a statistic T_i, such as the coordinatewise median, applied to the ith sample for $i = 1, 2$. 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 the $\boldsymbol{\mu}_i$ are $p \times 1$ vectors. Assume that T_i satisfies a central limit type theorem $\sqrt{n}(T_i - \boldsymbol{\mu}_i) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_i)$ for $i = 1, 2$ where the $\boldsymbol{\Sigma}_i$ are positive definite.

To simplify large sample theory, assume $n_1 = kn_2$ for some positive real number k. Let $\hat{\boldsymbol{\Sigma}}_i$ be a consistent nonsingular estimator of $\boldsymbol{\Sigma}_i$. Then

$$\begin{pmatrix} \sqrt{n_1}\ (T_1 - \boldsymbol{\mu}_1) \\ \sqrt{n_2}\ (T_2 - \boldsymbol{\mu}_2) \end{pmatrix} \xrightarrow{D} N_{2p}\left[\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_2 \end{pmatrix}\right],$$

or

$$\begin{pmatrix} \sqrt{n_2}\ (T_1 - \boldsymbol{\mu}_1) \\ \sqrt{n_2}\ (T_2 - \boldsymbol{\mu}_2) \end{pmatrix} \xrightarrow{D} N_{2p}\left[\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \frac{\boldsymbol{\Sigma}_1}{k} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_2 \end{pmatrix}\right].$$

Hence

$$\sqrt{n_2}\ [(T_1 - T_2) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)] \xrightarrow{D} N_p\left(\mathbf{0}, \frac{\boldsymbol{\Sigma}_1}{k} + \boldsymbol{\Sigma}_2\right).$$

Using $n\boldsymbol{B}^{-1} = \left(\frac{\boldsymbol{B}}{n}\right)^{-1}$ and $n_2 k = n_1$, if $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, then

$$n_2(T_1 - T_2)^T \left(\frac{\boldsymbol{\Sigma}_1}{k} + \boldsymbol{\Sigma}_2\right)^{-1} (T_1 - T_2) =$$

$$(T_1 - T_2)^T \left(\frac{\boldsymbol{\Sigma}_1}{n_1} + \frac{\boldsymbol{\Sigma}_2}{n_2}\right)^{-1} (T_1 - T_2) \xrightarrow{D} \chi^2_p.$$

Hence

$$T_0^2 = (T_1 - T_2)^T \left(\frac{\hat{\boldsymbol{\Sigma}}_1}{n_1} + \frac{\hat{\boldsymbol{\Sigma}}_2}{n_2}\right)^{-1} (T_1 - T_2) \xrightarrow{D} \chi^2_p. \tag{1}$$

Note that k drops out of the above result.

If the sequence of positive integers $d_n \to \infty$ and $Y_n \sim F_{p,d_n}$, then $Y_n \xrightarrow{D} \chi^2_p/p$. Using an F_{p,d_n} distribution instead of a χ^2_p distribution is similar to using a t_{d_n} distribution instead of a standard normal $N(0,1)$ distribution for inference. Instead of rejecting H_0 when $T_0^2 > \chi^2_{p,1-\delta}$, reject H_0 when

$$T_0^2 > pF_{p,d_n,1-\delta} = \frac{pF_{p,d_n,1-\delta}}{\chi^2_{p,1-\delta}} \chi^2_{p,1-\delta}.$$

The term $\frac{pF_{p,d_n,1-\delta}}{\chi^2_{p,1-\delta}}$ can be regarded as a small sample correction factor that improves the test's performance for small samples. For example, use $d_n = \min(n_1 - p, n_2 - p)$. Here $P(Y_n \leq \chi^2_{p,\delta}) = \delta$ if Y_n has a χ^2_p distribution, and $P(Y_n \leq F_{p,d_n,\delta}) = \delta$ if Y_n has an F_{p,d_n} distribution.

The two sample Hotelling's T^2 test is the classical method. If it is not assumed that the population covariance matrices are equal, then this test uses the sample mean and sample covariance matrix $T_i = \overline{\boldsymbol{x}}_i$ and $\hat{\boldsymbol{\Sigma}}_i = \boldsymbol{S}_i$ applied to each sample. This test has considerable robustness to the assumption that both populations have a multivariate normal distribution and to the assumption that the populations have a common population covariance matrix $\boldsymbol{\Sigma}$, but the test can be very poor if outliers are present.

Alternative statistics to the sample mean can be useful, but large sample tests of the form of (1) need practical consistent estimators $\hat{\boldsymbol{\Sigma}}_i$ of the two asymptotic covariance matrices $\boldsymbol{\Sigma}_i$. Section 2.1 reviews the Olive (2016, 2017ab) method for bootstrapping hypothesis tests. Section 2.2 shows how to apply the bootstrap to test the hypothesis $H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{0}$ versus $H_1 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \neq \boldsymbol{0}$. These tests are useful if the asymptotic covariance matrix is unknown or difficult to estimate. Section 3 gives some simulations and an example.

2. Method

2.1 Bootstrapping hypothesis tests and the prediction region method

Olive (2016, 2017b) shows that there is a useful relationship between prediction regions and confidence regions. Consider predicting a future $p \times 1$ test vector $\boldsymbol{x}_f$, given past training data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$. A *large sample* $100(1-\delta)\%$ *prediction region* is a set $\mathcal{A}_n$ such that $P(\boldsymbol{x}_f \in \mathcal{A}_n) \to 1 - \delta$ while a large sample $100(1-\delta)\%$ confidence region for a parameter $\boldsymbol{\mu}$ is a set $\mathcal{A}_n$ such that $P(\boldsymbol{\mu} \in \mathcal{A}_n) \to 1 - \delta$ as $n \to \infty$. Consider testing $H_0 : \boldsymbol{\mu} = \boldsymbol{c}$ versus $H_1 : \boldsymbol{\mu} \neq \boldsymbol{c}$ where $\boldsymbol{c}$ is a known $p \times 1$ vector.

Some notation is needed to describe the Olive (2013) prediction region for the multivariate location and dispersion model. Let the $p \times 1$ column vector T be a multivariate location estimator, and let the $p \times p$ symmetric positive definite matrix $\boldsymbol{C}$ be a dispersion estimator. Then the ith *squared sample Mahalanobis distance* is the scalar

$$D_i^2 = D_i^2(T, \boldsymbol{C}) = D_{\boldsymbol{x}_i}^2(T, \boldsymbol{C}) = (\boldsymbol{x}_i - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x}_i - T) \qquad (2)$$

for each observation $\boldsymbol{x}_i$. Notice that the Euclidean distance of $\boldsymbol{x}_i$ from the estimate of center T is $D_i(T, \boldsymbol{I}_p)$ where $\boldsymbol{I}_p$ is the $p \times p$ identity matrix. The classical Mahalanobis distance uses

$(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$, the sample mean and sample covariance matrix where

$$\overline{\boldsymbol{x}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \quad \text{and} \quad \boldsymbol{S} = \frac{1}{\mathrm{n}-1} \sum_{\mathrm{i}=1}^{\mathrm{n}} (\boldsymbol{x}_{\mathrm{i}} - \overline{\boldsymbol{x}})(\boldsymbol{x}_{\mathrm{i}} - \overline{\boldsymbol{x}})^{\mathrm{T}}. \tag{3}$$

A large sample $100(1-\delta)\%$ prediction region is the hyperellipsoid

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \le D^2_{(c)}\} = \{\boldsymbol{w} : D_{\boldsymbol{w}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \le D_{(c)}\} \tag{4}$$

for appropriate c. Using $c = \lceil n(1-\delta) \rceil$ covers about $100(1-\delta)\%$ of the training data cases $\boldsymbol{x}_i$, but the prediction region will have coverage lower than the nominal coverage of $1-\delta$ for moderate n. This result is not surprising since empirically statistical methods perform worse on test data. Increasing c will improve the coverage for moderate samples. Let $q_n = \min(1-\delta+0.05, 1-\delta+p/n)$ for $\delta > 0.1$ and

$$q_n = \min(1-\delta/2, 1-\delta+10\delta p/n), \quad \text{otherwise.} \tag{5}$$

If $1-\delta < 0.999$ and $q_n < 1-\delta+0.001$, set $q_n = 1-\delta$.

Let $D_{(U_n)}$ be the $100 q_n$th percentile of the D_i. Then the Olive (2013) large sample $100(1-\delta)\%$ nonparametric prediction region for a future value $\boldsymbol{x}_f$ given iid data $\boldsymbol{x}_1, ..., , \boldsymbol{x}_n$ is

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \le D^2_{(U_n)}\}, \tag{6}$$

while the classical large sample $100(1-\delta)\%$ prediction region is

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \le \chi^2_{p,1-\delta}\}. \tag{7}$$

The Olive (2016, 2017ab) prediction region method obtains a confidence region for $\boldsymbol{\mu}$ by applying the nonparametric prediction region (6) to the bootstrap sample $T_1^*, ..., T_B^*$, and the theory for the method is sketched below. Let $\overline{T}^*$ and $\boldsymbol{S}_T^*$ be the sample mean and sample covariance matrix of the bootstrap sample. Following Bickel and Ren (2001), let the vector of parameters $\boldsymbol{\mu} = T(F)$, the statistic $T_n = T(F_n)$, and $T^* = T(F_n^*)$ where F is the cdf of iid $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$, F_n is the empirical cdf, and F_n^* is the empirical cdf of $\boldsymbol{x}_1^*, ..., \boldsymbol{x}_n^*$, a sample from F_n using the nonparametric bootstrap. If $\sqrt{n}(F_n - F) \xrightarrow{D} \boldsymbol{z}_F$, a Gaussian random process, and if T is sufficiently smooth (Hadamard differentiable with a Hadamard derivative $\dot{T}(F)$), then $\sqrt{n}(T_n - \boldsymbol{\mu}) \xrightarrow{D} \boldsymbol{X}$ and $\sqrt{n}(T_i^* - \overline{T}^*) \xrightarrow{D} \boldsymbol{X}$ with $\boldsymbol{X} = \dot{T}(F)\boldsymbol{z}_F$. Olive (2016, 2017b) uses these results to show that if $\boldsymbol{X} \sim N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_T)$, then $\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{D} \boldsymbol{0}$, $\sqrt{n}(\overline{T}^* - \boldsymbol{\mu}) \xrightarrow{D} \boldsymbol{X}$, and that the prediction region method large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\mu}$ is

$$\{\boldsymbol{w} : (\boldsymbol{w} - \overline{T}^*)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - \overline{T}^*) \le D^2_{(U_B)}\} = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}^*, \boldsymbol{S}_T^*) \le D^2_{(U_B)}\} \tag{8}$$

where $D^2_{(U_B)}$ is computed from $D_i^2 = (T_i^* - \overline{T}^*)^T [\boldsymbol{S}_T^*]^{-1} (T_i^* - \overline{T}^*)$ for $i = 1, ..., B$. Note that the corresponding test for $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ rejects H_0 if $(\overline{T}^* - \boldsymbol{\mu}_0)^T [\boldsymbol{S}_T^*]^{-1} (\overline{T}^* - \boldsymbol{\mu}_0) > D^2_{(U_B)}$. This procedure is basically the one sample Hotelling's T^2 test applied to the T_i^* using $\boldsymbol{S}_T^*$ as the estimated covariance matrix and replacing the $\chi^2_{p,1-\delta}$ cutoff by $D^2_{(U_B)}$.

The prediction region method for testing $H_0 : \boldsymbol{\mu} = \boldsymbol{c}$ versus $H_1 : \boldsymbol{\mu} \neq \boldsymbol{c}$ is simple. Let $\hat{\boldsymbol{\mu}}$ be a consistent estimator of $\boldsymbol{\mu}$ and make a bootstrap sample $\boldsymbol{w}_i = \hat{\boldsymbol{\mu}}_i^* - \boldsymbol{c}$ for $i = 1, ..., B$. Make the nonparametric prediction region (8) for the $\boldsymbol{w}_i$ and fail to reject H_0 if $\boldsymbol{0}$ is in the prediction region, reject H_0 otherwise.

The Bickel and Ren (2001) hypothesis testing method is equivalent to using confidence region (8) with $\overline{T}^*$ replaced by T_n and U_B replaced by $\lceil B(1-\delta) \rceil$. If region (8) or the Bickel and Ren (2001) region is a large sample $100(1-\delta)\%$ confidence region, then so is the other region if $\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{D} \boldsymbol{0}$. Hadamard differentiability and asymptotic normality are sufficient conditions for both regions to be large sample confidence regions if $n\boldsymbol{S}_T^* \xrightarrow{D} \boldsymbol{\Sigma}_T$, but Bickel and Ren (2001) showed that their method can work when Hadamard differentiability fails.

The location model with coordinatewise means, medians, and trimmed means is one example where the Bickel and Ren (2001, p. 96) method works. Since the univariate sample mean, sample median, and sample trimmed mean are Hadamard differentiable and asymptotically normal, each coordinate satisfies $\sqrt{n}(T_{in} - \overline{T}_i^*) \xrightarrow{D} 0$ for $i = 1, ..., p$. Hence $\sqrt{n}(T_n - \overline{T}^*) \xrightarrow{D} \boldsymbol{0}$, and (8) is a large sample $100(1-\delta)\%$ confidence region if T_n is the coordinatewise sample mean, median, or trimmed mean.

Fréchet differentiability implies Hadamard differentiability, and many statistics are shown to be Hadamard differentiable in Bickel and Ren (2001), Clarke (1986, 2000), Fernholtz

(1983), and Gill (1989).

2.2 Applying the prediction region method to the two sample test

The two sample test of $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ versus $H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ uses $\boldsymbol{\mu} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{c} = \boldsymbol{0}$ with $\boldsymbol{w}_i = T^*_{i1} - T^*_{i2}$ for $i = 1, ..., B$. Make the prediction region (8) where $T^*_i = \boldsymbol{w}_i$. Fail to reject H_0 if $\boldsymbol{0}$ is in the prediction region, reject H_0 otherwise. A sample of size n_i is drawn with replacement from $\boldsymbol{x}_{1,i}, ..., \boldsymbol{x}_{n_i,i}$ for $i = 1, 2$ to obtain the bootstrap sample.

For illustrative purposes, the simulation study will take T_i to be the coordinatewise median, the (Olive (2017b, ch. 4), Olive and Hawkins (2010), and Zhang, Olive, and Ye (2012)) RMVN estimator T_{RMVN}, the sample mean, and the 25% trimmed mean. The asymptotic covariance matrix of the coordinatewise median is difficult to estimate, while that of the RMVN estimator is unknown. The RMVN estimator has been shown to be $\sqrt{n}$ consistent on a large class of elliptically contoured distributions, but has not yet been shown to be asymptotically normal. Hence the bootstrap "test" for the RMVN estimator should be used for exploratory purposes.

The RMVN estimator $(T_{RMVN}, \boldsymbol{C}_{RMVN})$ uses a concentration algorithm. Let $(T_{-1,j}, \boldsymbol{C}_{-1,j})$ be the jth start (initial estimator) and compute all n Mahalanobis distances $D_i(T_{-1,j}, \boldsymbol{C}_{-1,j})$. At the next iteration, the classical estimator $(T_{0,j}, \boldsymbol{C}_{0,j}) = (\overline{\boldsymbol{x}}_{0,j}, \boldsymbol{S}_{0,j})$ is computed from the $c_n \approx n/2$ cases corresponding to the smallest distances. This iteration can be continued for k *concentration steps* resulting in the sequence of estimators $(T_{-1,j}, \boldsymbol{C}_{-1,j}), (T_{0,j}, \boldsymbol{C}_{0,j}), ..., (T_{k,j}, \boldsymbol{C}_{k,j})$. The result of the iteration $(T_{k,j}, \boldsymbol{C}_{k,j})$ is called the jth *attractor*. The algorithm estimator uses one of the attractors. The RMVN estimator uses the same two starts as the Olive (2004) MBA estimator: $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ and $(MED(n), \boldsymbol{I}_p)$ where $MED(n)$ is the coordinatewise median. Then the location estimator T_{RMVN} can be used to test $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$.

3. Simulation and an Example

The simulation used 5000 runs with B bootstrap samples. Olive (2016, 2017b) suggests that the prediction region method can give good results when the number of bootstrap samples $B \geq 50p$ if $n \geq 50p$, and the simulation used various values of B.

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 $\boldsymbol{x}_2 = \sigma\boldsymbol{A}\boldsymbol{w}_2$ where $\boldsymbol{1} = (1,..,1)^T$ is a vector of ones and $\boldsymbol{A} = diag(1,\sqrt{2},...,\sqrt{p})$. 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 $0.6N_p(\boldsymbol{0}, \boldsymbol{I}) + 0.4N_p(\boldsymbol{0}, 25\boldsymbol{I})$, and the multivariate lognormal distribution shifted to have nonzero mean $\boldsymbol{\mu} = 0.649\ \boldsymbol{1}$, but a population coordiatewise median of $\boldsymbol{0}$. Note that Cov($\boldsymbol{x}_2$) $= \sigma^2$ Cov($\boldsymbol{x}_1$), and for the first three distributions, $E(\boldsymbol{x}_i) = E(\boldsymbol{w}_i) = \boldsymbol{0}$ if $\delta = 0$.

Adding the same type and proportion of outliers to groups one and two often resulted in two distributions that were still similar. Hence outliers were added to the first group but not the second, making the covariance structures of the two groups quite different. The outlier proportion was $100\gamma\%$. Let $\boldsymbol{x}_1 = (x_{11}, ..., x_{p1})^T$. The five outlier types for group 1 were type 1: a tight cluster at the major axis $(0, ..., 0, pm)^T$, type 2: a tight cluster at the minor axis $(pm, 0, ..., 0)^T$, type 3: a mean shift $N((pm, ..., pm)^T, diag(1, ..., p))$, type 4: x_{1p} replaced by pm, and type 5: x_{11} replaced by pm. The quantity pm determines how far the outliers are from the clean data.

Let the *coverage* be the proportion of times that H_0 is rejected. We want the *coverage* near 0.05 when H_0 is true and the coverage close to 1.0 for good power when H_0 is false. With 5000 runs, an observed *coverage* inside of (0.04, 0.06) suggests that the true *coverage* is close to the nominal 0.05 coverage when H_0 is true.

3.1 Type I error rates with clean data

Tables 1, 2, and 3 were for clean elliptically contoured distributions (no outliers present), where H_0 is true and the different location estimators estimate $\boldsymbol{\mu} = \boldsymbol{0}$, the point of symmetry for the distribution. The chi–square cutoffs when $p = 5$ and $p = 15$ were 11.071 and 24.996, respectively. The *coverages* were often near the nominal value of 0.05, but the RMVN *coverages* were a bit low for Table 3. The classical Hotelling's T^2 test does not use the bootstrap, and performed poorly when H_0 was true and both the sample sizes and the population covariance matrices were different.

For clean multivariate lognormal data, H_0 is true when $\sigma = 1$ (identical distributions for both groups), but H_0 is not true for the population mean when $\sigma = 2$. For $\sigma = 2$, the coordinatewise median had *coverages* near the nominal, while the sample mean had good power with *coverages* near 1. The RMVN coverage was a bit low when $\sigma = 1$ with power that was often less than that of the sample mean when $\sigma = 2$. See Table 4. The simulated cutoffs were quite similar to the chi-square cutoffs for Tables 1 through 4.

Table 1: *coverages* for clean multivariate normal data

p	n_1	n_2	σ	B	Median	Mean	Tr.Me	RMVN	Class
5	250	250	1	250	0.0470	0.0554	0.0568	0.0402	0.0560
				1000	0.0440	0.0606	0.0540	0.0414	
			2	250	0.0472	0.0550	0.0574	0.0422	0.0498
				1000	0.0420	0.0568	0.0538	0.0392	
5	250	500	1	250	0.0490	0.0524	0.0496	0.0394	0.0552
				1000	0.0462	0.0588	0.0584	0.0448	
			2	250	0.0460	0.0540	0.0524	0.0436	0.0070
				1000	0.0470	0.0500	0.0534	0.0386	
15	750	750	1	750	0.0462	0.0626	0.0622	0.0466	0.0450
				1000	0.0390	0.0514	0.0470	0.0378	
			2	750	0.0492	0.0598	0.0608	0.0464	0.0516
				1000	0.0474	0.0556	0.0568	0.0446	
15	750	1500	1	750	0.0466	0.0538	0.0550	0.0466	0.0480
				1000	0.0492	0.0556	0.0548	0.0444	
			2	750	0.0424	0.0538	0.0520	0.0454	0.0014
				1000	0.0514	0.0532	0.0542	0.0426	

3.2 Type I error rates with contaminated data

Table 5 illustrates the simulated results where group 1 had outliers. The coordinatewise median worked with a little higher type I error rate (around 0.08) than the nominal level of 0.05 for the mixture, multivariate t, and multivariate log normal distributions, but failed for the multivariate normal data when $\gamma = 0.4$. The sample mean (classical and bootstrap) and

Table 2: *coverages* for clean $0.6N_p(\mathbf{0}, \boldsymbol{I}) + 0.4N_p(\mathbf{0}, 25\boldsymbol{I})$ data

p	n_1	n_2	σ	B	Median	Mean	Tr.Me	RMVN	Class
5	250	250	1	250	0.0420	0.0560	0.0480	0.0394	0.0462
				1000	0.0386	0.0532	0.0464	0.0336	
			2	250	0.0454	0.0550	0.0476	0.0416	0.0476
				1000	0.037	0.0484	0.0400	0.0368	
	250	500	1	250	0.0460	0.0542	0.0538	0.0416	0.0470
				1000	0.0368	0.0502	0.0416	0.0404	
			2	250	0.0480	0.0600	0.0474	0.0390	0.0060
				1000	0.0416	0.0598	0.0498	0.0416	
15	750	750	1	750	0.0434	0.0536	0.0540	0.0448	0.0496
				1000	0.0406	0.0598	0.0474	0.0396	
			2	750	0.0468	0.0626	0.0518	0.0456	0.0464
				1000	0.0456	0.0566	0.0490	0.0454	
15	750	1500	1	750	0.0456	0.0584	0.0568	0.0488	0.0502
				1000	0.0426	0.0550	0.0478	0.0438	
			2	750	0.0456	0.0576	0.0508	0.0442	0.0004
				1000	0.0416	0.0572	0.0488	0.0510	

Table 3: *coverages* for clean multivariate t_4 data

p	n_1	n_2	σ	B	Median	Mean	Tr.Me	RMVN	Class
5	250	250	1	250	0.0442	0.0574	0.0570	0.0266	0.0456
				1000	0.0426	0.0570	0.0530	0.0282	
			2	250	0.0496	0.0618	0.0614	0.0328	0.0542
				1000	0.0480	0.0558	0.0578	0.0292	
5	250	500	1	250	0.0484	0.0512	0.0540	0.0346	0.0504
				1000	0.0420	0.0488	0.0494	0.0310	
			2	250	0.0408	0.0580	0.0526	0.0348	0.0058
				1000	0.0410	0.0492	0.0510	0.0348	
15	750	750	1	750	0.0470	0.0550	0.0562	0.0232	0.0414
				1000	0.0382	0.0526	0.0476	0.0228	
			2	750	0.0472	0.0572	0.0542	0.0248	0.0442
				1000	0.0502	0.0496	0.0556	0.0258	
15	750	1500	1	750	0.0482	0.0556	0.0528	0.0224	0.0446
				1000	0.0464	0.0496	0.0528	0.0254	
			2	750	0.0442	0.0534	0.0502	0.0314	0.0016
				1000	0.0452	0.0508	0.0554	0.0262	

Table 4: *coverages* for clean lognormal data

p	n_1	n_2	σ	B	Median	Mean	Tr.Me	RMVN	Class
5	250	250	1	250	0.0408	0.0460	0.0514	0.0274	0.0470
				1000	0.0388	0.0494	0.0474	0.0254	
			2	250	0.0436	0.9816	0.0858	0.1108	0.9968
				1000	0.0398	0.9846	0.0788	0.1168	
5	250	500	1	250	0.0398	0.0540	0.0496	0.0316	0.0472
				1000	0.0368	0.0588	0.0446	0.0292	
			2	250	0.0418	0.9998	0.1192	0.2492	0.9964
				1000	0.0424	0.9994	0.1158	0.2520	
15	750	750	1	750	0.0402	0.0506	0.0480	0.0216	0.0502
				1000	0.0410	0.0444	0.0490	0.0238	
			2	750	0.0506	1.0000	0.3670	1.0000	1.0000
				1000	0.0510	1.0000	0.3748	1.0000	
15	750	1500	1	750	0.0420	0.0580	0.0514	0.0258	0.0514
				1000	0.0478	0.0558	0.0608	0.0284	
			2	750	0.0446	1.0000	0.6110	1.0000	1.0000
				1000	0.0464	1.0000	0.6256	1.0000	

25% trimmed mean failed to achieve the nominal level with any of the distributions used when H_0 was true for the clean data. The RMVN estimator worked with all four distributions with a better type I error rate compared to the other estimators. The chi–square cutoff was 9.488 since $p = 4$.

The coordinatewise median can achieve better coverages for smaller proportions of outliers with higher values of pm (not shown in the tables), i.e. the outliers had to be far from the clean data compared to the RMVN estimator. The RMVN estimator can handle higher proportions of outliers as shown in the Table 5.

3.3 Power simulation

In the power simulation, $\delta > 0$ was used. Hence for the first three distributions $\boldsymbol{\mu}_2 = \mathbf{0}$ and $\boldsymbol{\mu}_1 = \delta(1, ..., 1)^T$. Then the Euclidean distance between the two means was$\sqrt{p}\delta$, where p is the number of parameters. Therefore the distance increases as p increase. The value of δ had to be fairly small so that the simulated power was not always 1. Also see Table 4 with $\sigma = 2$.

For Table 6, the sample mean (bootstrap and classical) had the best power while the sample median had the worst power. For Table 5, the RMVN estimator had the best power while the sample mean has the worst power. The trimmed mean had the best power for Table 7. For Table 8, the RMVN estimator had poor power when $p = 5$, $n = 250$, and $\sigma = 2$. No method was always best or worst.

3.4. Real data example

The Johnson (1996) STATLIB bodyfat data consists of 252 observations on 15 variables including the density determined from underwater weighing and the percent body fat measurement. Consider these two variables with two age groups: age ≤ 50 and age > 50. The test with the RMVN estimator had $D_0 = 1.78$ while the test with the coordinatewise median had $D_0 = 1.35$. Both tests had cutoffs near 2.37 and fail to reject H_0. The classical two sample Hotelling's T^2 test rejects H_0 with a test statistic of 4.74 and a p-value of 0.001.

The DD plots, shown in Figures 1 and 2, reveal five outliers. After deleting the outliers, the three tests all fail to reject H_0. The RMVN test had $D_0 = 1.63$ with cutoff 2.25, the

Table 5: *Coverages* and cutoffs with outliers: $p = 4, n_1 = n_2 = 200, B = 200$

Dist.	Otype	γ	pm		Med	Mean	Tr.Me	RMVN	Class
MVN	1	0.4	10	Cov	0.6946	1.0000	1.0000	0.0330	1.0000
				cut	10.158	9.769	9.798	10.701	
	2	0.4	20	Cov	0.5232	1.0000	1.0000	0.0382	1.0000
				cut	9.836	9.776	9.809	9.268	
	3	0.4	20	Cov	0.8578	1.0000	1.0000	0.0402	1.0000
				cut	10.214	9.761	9.760	9.288	
	4	0.1	10	Cov	0.0980	0.8654	0.1450	0.0382	0.8684
				cut	9.898	9.771	9.777	9.851	
Mix	2	0.4	20	Cov	0.0828	1.0000	1.0000	0.0144	1.0000
				cut	10.542	9.788	9.878	11.300	
	5	0.1	10	Cov	0.0820	0.5306	0.1228	0.0184	0.5276
				cut	9.933	9.779	9.881	11.056	
MVT	1	0.4	10	Cov	0.0854	0.6700	0.1548	0.0204	1.0000
				cut	10.232	9.799	9.787	10.200	
	5	0.1	20	Cov	0.0864	1.0000	0.1418	0.0304	1.0000
				cut	9.924	9.795	9.795	9.830	
Log	3	0.4	20	Cov	0.0778	1.0000	1.0000	0.0162	1.0000
				cut	13.689	9.822	9.827	12.607	
	4	0.1	10	Cov	0.0842	0.3158	0.1482	0.0234	0.3044
				cut	10.013	9.875	9.872	10.416	

Table 6: *Coverages* when H_0 is false for MVN data.

p	$n_1 = n_2$	σ	B	δ	Med	Mean	Tr.Me	RMVN	Class
5	250	1	250	0.35	0.9598	0.9990	0.9928	0.9942	0.9988
			1000	0.35	0.9684	0.9994	0.9970	0.9978	
		2	250	0.35	0.5958	0.8442	0.7672	0.7604	0.8402
			1000	0.35	0.5832	0.8346	0.7438	0.7470	
15	750	1	750	0.15	0.7394	0.9552	0.9012	0.9268	0.9556
			1000	0.15	0.7474	0.9522	0.8984	0.9178	
		2	750	0.15	0.3078	0.5318	0.4550	0.4468	0.5156
			1000	0.15	0.3118	0.5218	0.4430	0.4464	

Table 7: *Coverages* when H_0 is false for mixture data.

p	$n_1 = n_2$	σ	B	δ	Med	Mean	Tr.Me	RMVN	Class
5	250	1	250	0.45	0.8826	0.4062	0.9304	0.9938	0.4032
			1000	0.45	0.8858	0.4058	0.9338	0.9948	
		2	250	0.45	0.4458	0.1910	0.5222	0.7454	0.1642
			1000	0.45	0.4656	0.1890	0.5386	0.7626	
15	750	1	750	0.20	0.6204	0.2274	0.7148	0.9492	0.2114
			1000	0.20	0.6316	0.2228	0.7190	0.9494	
		2	750	0.20	0.2318	0.1154	0.2894	0.5034	0.1042
			1000	0.20	0.2438	0.1092	0.2916	0.4980	

Table 8: *Coverages* when H_0 is false for multivariate t_4 data.

p	$n_1 = n_2$	σ	B	δ	Med	Mean	Tr.Me	RMVN	Class
5	250	1	250	0.38	0.9642	0.9562	0.9916	0.9878	0.9548
			1000	0.38	0.9728	0.9572	0.9944	0.9880	
		2	250	0.38	0.5958	0.5960	0.7198	0.6488	0.6074
			1000	0.38	0.6188	0.6152	0.7490	0.6636	
15	750	1	750	0.20	0.9418	0.9270	0.9868	0.9714	0.9232
			1000	0.20	0.9422	0.9304	0.9860	0.9724	
		2	750	0.20	0.4934	0.4932	0.6422	0.5384	0.4754
			1000	0.20	0.4842	0.4916	0.6362	0.5252	

Table 9: *Coverages* when H_0 is false for lognormal data.

p	$n_1 = n_2$	σ	B	δ	Median	Mean	Tr.Me	RMVN	Class
5	250	1	250	0.45	0.9982	0.8256	0.9994	0.879	0.8208
			1000	0.45	0.9980	0.8324	0.9996	0.883	
		2	250	0.45	0.8210	0.4704	0.6488	0.0914	0.4630
			1000	0.45	0.8378	0.4646	0.6624	0.1038	
15	750	1	750	0.30	1.0000	0.9186	1.0000	0.8514	0.9120
			1000	0.30	1.0000	0.9178	1.0000	0.8544	
		2	750	0.30	0.9436	1.0000	0.5042	0.9438	1.0000
			1000	0.30	0.9484	1.0000	0.5022	0.9424	

coordinatewise median test had $D_0 = 1.22$ with cutoff 2.38, and the classical test had test statistic 2.39 with a p-value of 0.09.

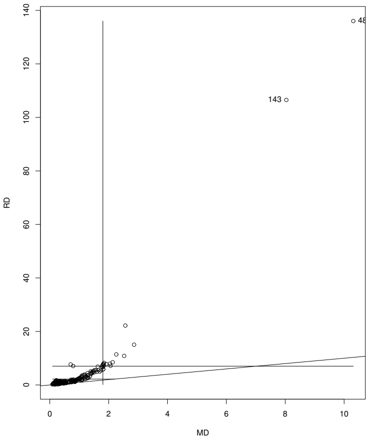

Figure 1: DD plot for the age ≤ 50 group.

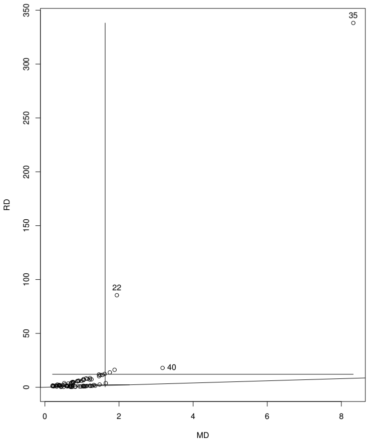

Figure 2: DD plot for the age > 50 group.

4. Discussion

This paper suggests a practical method to perform a multivariate two sample test when the asymptotic covariance matrix of the statistic T_i is difficult to estimate. Such tests may be useful when the data distribution is unknown or outliers are present. The method was illustrated with the coordinatewise median, sample mean, 25% trimmed mean, and RMVN estimators. All four estimators work well when the prediction region method was applied to the clean data, although care needs to be taken with the multivariate lognormal distribution where the four estimators T_i are estimating different parameters $\boldsymbol{\mu}_{T_i}$.

Both the sample mean and the 25% trimmed mean failed to achieve the nominal coverage when H_0 is true with the contaminated data. The coordinatewise median could handle up to 10% outliers, while the RMVN estimator could handle up to 40% outliers. Both estimators were robust to the equal covariance assumption.

Konietschke, Bathke, Harrar, and Pauly (2015) suggest a method for bootstrapping the MANOVA model, and Willems, Pison, Rousseeuw, and Van Aelst (2002) suggest a robust one sample Hotelling's T^2 type test. References for robust one way MANOVA tests are in Finch and French (2013), Todorov and Filzmoser (2010), Van Aelst and Willems (2011), and Wilcox (1995).

The *R* software was used in the simulation. See R Core Team (2016). Programs are in the Olive (2017b) collection of *R* functions *mpack.txt* available from (http://lagrange.math.siu.edu/Olive/mpack.txt). The function `hot2sim` was used to simulate the tests of hypotheses, and `predreg` computes the confidence region given the bootstrap values from `rhot2boot`. The Curran (2013) *R* package `Hotelling` was used to perform the classical 2 sample Hotelling's T^2 test.

Acknowledgments

The authors thank Dr. Olive and a referee for comments that improved this article.

References

Bickel, P. J., Ren, J. –J. (2001). The bootstrap in hypothesis testing. In: de Gunst, M., Klaassen, C., van der Vaart, A. Eds. *State of the Art in Probability and Statistics: Festschrift for William R. van Zwet.* The Institute of Mathematical Statistics, CA: Hayward, pp. 91-112.

Clarke, B. R. (1986). Nonsmooth analysis and Fréchet differentiability of M functionals.

Probability Theory and Related Fields 73:137-209.

Clarke, B. R. (2000). A review of differentiability in relation to robustness with an application to seismic data analysis. *Proceedings of the Indian National Science Academy, A* 66:467-482.

Curran, J. M. (2013). *Hotelling: Hotelling's T-squared test and variants. R* Package version 1.0-2, (https://cran.r-project.org/package=Hotelling).

Fernholtz, L. T. (1983). *von Mises Calculus for Statistical Functionals.* New York: Springer.

Finch, H., French, B. (2013). A "Monte Carlo" comparison of robust MANOVA test statistics. *Journal of Modern Applied Statistical Methods* 12:35-81.

Gill, R. D. (1989). Non- and semi-parametric maximum likelihood estimators and the von Mises method, part 1. *Scandinavian Journal of Statistics* 16:97-128.

Johnson, R. W. (1996). Fitting percentage of body fat to simple body measurements. *Journal of Statistics Education* 4, online at (www.amstat.org/publications/jse/).

Konietschke, F., Bathke, A. C., Harrar, S. W., Pauly, M. (2015). Parametric and nonparametric bootstrap methods for general MANOVA. *Journal of Multivariate Analysis* 140:291-301.

Olive, D. J. (2004). A resistant estimator of multivariate location and dispersion. *Computational Statistics & Data Analysis* 46:99-102.

Olive, D. J. (2013). Asymptotically optimal regression prediction intervals and prediction regions for multivariate data. *International Journal of Statistics and Probability* 2:90-100.

Olive, D. J. (2016). Bootstrapping hypothesis tests and confidence regions. Unpublished Manuscript with the bootstrap material from Olive (2017b) at (http://lagrange.math.siu.edu /Olive/ppvselboot.pdf).

Olive, D. J. (2017a). Applications of hyperellipsoidal prediction regions. *Statistical Papers.* To appear.

Olive, D. J. (2017b). *Robust Multivariate Analysis.* New York: Springer. To appear.

Olive, D. J., Hawkins, D. M. (2010). Robust multivariate location and dispersion. Unpublished Manuscript available from (http://lagrange.math.siu.edu/Olive/pphbmld.pdf).

R Core Team (2016). R: *A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing. (www.R-project.org).

Todorov, V., Filzmoser, P. (2010). Robust statistics for the one-way MANOVA. *Computational Statistics & Data Analysis* 54:37-48.

Van Aelst, S., Willems, G. (2011). Robust and efficient one-way MANOVA tests. *Journal of the American Statistical Association* 106:706-718.

Wilcox, R. R. (1995). Simulation results on solutions to the multivariate "Behrens-Fisher" problem via trimmed means. *The Statistician* 44:213-225.

Willems, G., Pison, G., Rousseeuw, P. J., Van Aelst, S. (2002). A robust "Hotelling" test. *Metrika* 55:125-138.

Zhang, J., Olive, D. J., Ye, P. (2012). Robust covariance matrix estimation with canonical correlation analysis. *International Journal of Statistics and Probability* 1:119-136.