Biometrics 62, 877–885
September 2006

DOI: 10.1111/j.1541-0420.2006.00533.x

A Multivariate Two-Sample Mean Test for Small Sample Size and Missing Data

Yujun Wu,[1,*] Marc G. Genton,[2] and Leonard A. Stefanski[3]

[1]Department of Biostatistics, School of Public Health, University of Medicine and Dentistry of New Jersey, Piscataway, New Jersey 08854, U.S.A.
[2]Department of Statistics, Texas A&M University, College Station, Texas 77843-3143, U.S.A.
[3]Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695-8203, U.S.A.
[*]*email:* wuy5@umdnj.edu

Summary. We develop a new statistic for testing the equality of two multivariate mean vectors. A scaled chi-squared distribution is proposed as an approximating null distribution. Because the test statistic is based on componentwise statistics, it has the advantage over Hotelling's T^2 test of being applicable to the case where the dimension of an observation exceeds the number of observations. An appealing feature of the new test is its ability to handle missing data by relying on only componentwise sample moments. Monte Carlo studies indicate good power compared to Hotelling's T^2 and a recently proposed test by Srivastava (2004, Technical Report, University of Toronto). The test is applied to drug discovery data.

Key words: Drug discovery; High-dimensional data; Hotelling's T^2; Small n large p.

1. Introduction

Testing the equality of two multivariate mean vectors,

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \quad \text{vs.} \quad H_a : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2, \tag{1}$$

is a fundamental problem. For normally distributed data with common covariance matrix, Hotelling's T^2 test is the method of choice provided sample sizes are sufficiently large (Hotelling, 1931; Muirhead, 1982; Anderson, 1984). Robust variants of Hotelling's T^2 have been studied and the null distribution of T^2 for nonnormal data has been investigated (Kano, 1995; Fujikoshi, 1997; Mudholkar and Srivastava, 2000).

Our interest lies in testing the hypotheses (1) in cases in which the sample sizes n_1 and n_2 of the two groups are small, or the dimension p of the data is large, or the data are compromised by componentwise missing values. When missing values are numerous or when p is larger than $n_1 + n_2 - 2$, the pooled, complete-case covariance matrix is singular and thus calculation of Hotelling's T^2 statistic is not possible. In cases with $p \leq n_1 + n_2 - 2$ with some, but not extensive, missing data, the pooled, complete-case covariance matrix can be nonsingular, and thus T^2 can be calculated and Hotelling's test carried out in the usual fashion using only the complete-case data. However, ignoring incomplete-case data entails a loss of information that can render Hotelling's T^2 test nearly powerless.

Although not considered here, one possible approach to dealing with the large p, small sample size, and missing data problems is to construct a covariance matrix estimate componentwise thus using all of the available data, and force it to be positive definite using methods such as those in Rousseeuw and Molenberghs (1993). The covariance matrix so constructed could then be used to calculate an approximate T^2 statistic. Srivastava (2004) recently proposed and studied a test of (1) for the case of large p and small sample size based on using a generalized inverse of the singular covariance matrix in the construction of a Hotelling T^2-like statistic. Srivastava derived an approximation to the null distribution of the new statistic, thus providing an elegant solution to the testing problem with large p and small n. Srivastava also suggested an imputation method (Srivastava and Carter, 1986) to deal with missing data.

In this article, we propose and study an alternative solution to the testing problem with large p, small n, or missing data. We do not generalize Hotelling's T^2 statistic, but rather we construct a new statistic obtained by summing squared componentwise t-statistics. The pooled component test (PCT) so obtained uses all of the available data and does not require inverting a covariance matrix. We derive the first two moments of the PCT statistic under the null hypothesis and use the moment formulae to approximate the null distribution of the statistic by matching estimated moments to a scaled chi-squared distribution. Simulation results indicate that the null distribution is well approximated in this fashion, and that our test compares favorably to Srivastava's generalized inverse T^2 test, and the complete-case Hotelling T^2 test when it is available.

Our interest in this problem relates to applications in drug discovery in which the relationship between molecular structure and biological activity of chemical compounds is of interest and is used to identify active compounds. Tree-structured approaches are often used to model the

structure-activity relationship (Hawkins, 1982; Hawkins, Young, and Rusinko, 1997). Activities are measured on different proteins-resulting in correlated multivariate continuous responses, and the molecular structure is expressed by a large number of binary covariates indicating presence or absence of specific atom-pairs in the compounds. Thus node splitting rules are based on comparisons of multivariate responses from two populations. One such algorithm, Multivariate Statistical Classification of Activities of Molecules (MultiSCAM) (Keefer, 2001), uses two-sample Hotelling's T^2 tests to judge potential splits, deciding to split on the atom-pair with the most significant test statistic, provided one exists after Bonferroni adjustment. The problem of $p > n$ is inevitable in tree building because node size decreases with increasing tree size, and is compounded by the fact that missing values are common, further limiting the number of complete multivariate responses.

The pooled component test is introduced in Section 2, and its first two moments, used in the chi-squared approximation to its null distribution, are derived in Section 3. Monte Carlo studies of the pooled component test, and comparisons with Hotelling's T^2 and Srivastava's tests, appear in Section 4. Section 5 illustrates the pooled component test in a drug discovery application. Summary comments are given in Section 6, and mathematical details are presented in the Appendix.

2. The Pooled Component Test

Suppose we have independent samples from two multivariate populations

$$\begin{aligned}&\text{Population 1}: Y_{11}, Y_{12}, \ldots, Y_{1n_1}, \quad \text{and}\\ &\text{Population 2}: Y_{21}, Y_{22}, \ldots, Y_{2n_2},\end{aligned}$$

where $Y_{ki} = (Y_{ki1}, Y_{ki2}, \ldots, Y_{kip})^T$, $i = 1, \ldots, n_k, k = 1, 2$, with Y_{kij} being the observation for variable j on subject i from population k, $j = 1, \ldots, p$, and n_k is the sample size of population k. We use the matrix $M_k(n_k \times p)$ to record the pattern of missingness in the sample data from population k. Its (i, j)th element is defined as $M_{kij} = 1$, if Y_{kij} is not missing, and $= 0$ otherwise. Define the componentwise sample mean for variable j in population k as $\bar{Y}_{kj} = (\sum_{l=1}^{n_k} Y_{klj} M_{klj})/n_{kj}$, where $n_{kj} = \sum_{l=1}^{n_k} M_{klj}$ is the number of nonmissing observations for variable j in population k. The pooled componentwise sample variance for variable j and covariance for any two variables i and j are, respectively, defined as

$$S_j^2 = \frac{\sum_{k=1}^{2}\sum_{l=1}^{n_k}(Y_{klj} - \bar{Y}_{kj})^2 M_{klj}}{n_{1j} + n_{2j} - 2} \tag{2}$$

and

$$S_{ij} = \frac{\sum_{k=1}^{2}\sum_{l=1}^{n_k}\left(Y_{kli} - \bar{Y}_{kij}^{(i)}\right)\left(Y_{klj} - \bar{Y}_{kij}^{(j)}\right) M_{kli} M_{klj}}{\sum_{k=1}^{2}\sum_{l=1}^{n_k} M_{kli} M_{klj} - 2}, \tag{3}$$

where $\bar{Y}_{kij}^{(b)} = (\sum_{l=1}^{n_k} Y_{klb} M_{kli} M_{klj})/(\sum_{l=1}^{n_k} M_{1li} M_{1lj}), b = i, j$.

For the testing problem (1), we propose a new test statistic by taking the average of squares of the univariate two-sample t-statistics for each individual variable based on componentwise statistics. The test statistic is defined as

$$Q = \frac{1}{p}\sum_{j=1}^{p} a_j Q_j, \tag{4}$$

where the jth component is $Q_j = (\bar{Y}_{1j} - \bar{Y}_{2j})^2/S_j^2$, and $a_j = (n_{1j} n_{2j})/(n_{1j} + n_{2j})$. We call the statistic in (4) a pooled component test statistic and the test based on it a pooled component test (PCT). Because the pooled component test uses only diagonal components of the sample covariance matrix, invertibility of the matrix is not an issue. Dudoit, Fridlyand, and Speed (2002) used a similar idea to construct a classifier based on only diagonal elements of the sample covariance matrix in their classification procedures, but they assumed a common diagonal population covariance matrix while we consider more general conditions with correlations.

In order to decide the rejection region for the hypothesis test (1), we need to determine the null distribution of Q. The exact distribution is complicated, so we approximate it instead. In light of the quadratic-form structure of Q, a scaled chi-squared distribution is a natural candidate to use as an approximation. Figure 1 shows a histogram of 5000 Q-statistics computed under the null hypothesis.

To determine the scale factor and the degrees of freedom of the chi-squared distribution, our strategy is to match moments using approximations to the first two moments of Q under the null hypothesis. Specifically, equating the mean and variance of a scaled chi-squared random variable, $c\chi_d^2$, with the mean and variance of Q results in the equations $E(Q) = cd$ and $\mathrm{Var}(Q) = 2c^2 d$, with solutions

$$c = \frac{\mathrm{Var}(Q)}{2E(Q)} \quad \text{and} \quad d = \frac{2\{E(Q)\}^2}{\mathrm{Var}(Q)}, \tag{5}$$

where c is the scale factor and d is the degrees of freedom of the chi-squared distribution. We can get $\hat{c}$ and $\hat{d}$ by replacing $E(Q)$ and $\mathrm{Var}(Q)$ with the estimators described in the next section. We do not round $\hat{d}$ to an integer, using the Gamma distribution with mean $\hat{d}$ and variance $2\hat{d}$ instead of a chi-squared as the approximating distribution.

3. Estimation of the Mean and Variance of Q

The mean and variance of Q play a critical role in determining the two parameters of the scaled chi-squared distribution. First, we derive expressions for the mean and variance of Q. The exact variance is very complicated due to the possible correlations among the variables and we derive a simpler approximation. Finally, we derive estimates of the mean and approximate variance expressions and use these to calibrate the approximating scaled chi-squared distribution. Lemma 1 and Theorem 1 are key results used in the derivation of $E(Q)$ and $\mathrm{Var}(Q)$.

LEMMA 1: *Suppose that population k has an $N_p(\boldsymbol{\mu}_k, \Sigma)$ distribution, $k = 1, 2$, with the common covariance matrix Σ having diagonal elements σ_i^2 and off-diagonal elements σ_{ij}. Assuming that missingness is completely random,*

$$\begin{aligned}&E\left\{(\bar{Y}_{1i} - \bar{Y}_{2i})^2(\bar{Y}_{1j} - \bar{Y}_{2j})^2\right\}\\ &\quad = \left(\frac{1}{n_{1i}} + \frac{1}{n_{2i}}\right)\left(\frac{1}{n_{1j}} + \frac{1}{n_{2j}}\right)\sigma_i^2\sigma_j^2 + 2\tau_{ij}^2,\end{aligned}$$

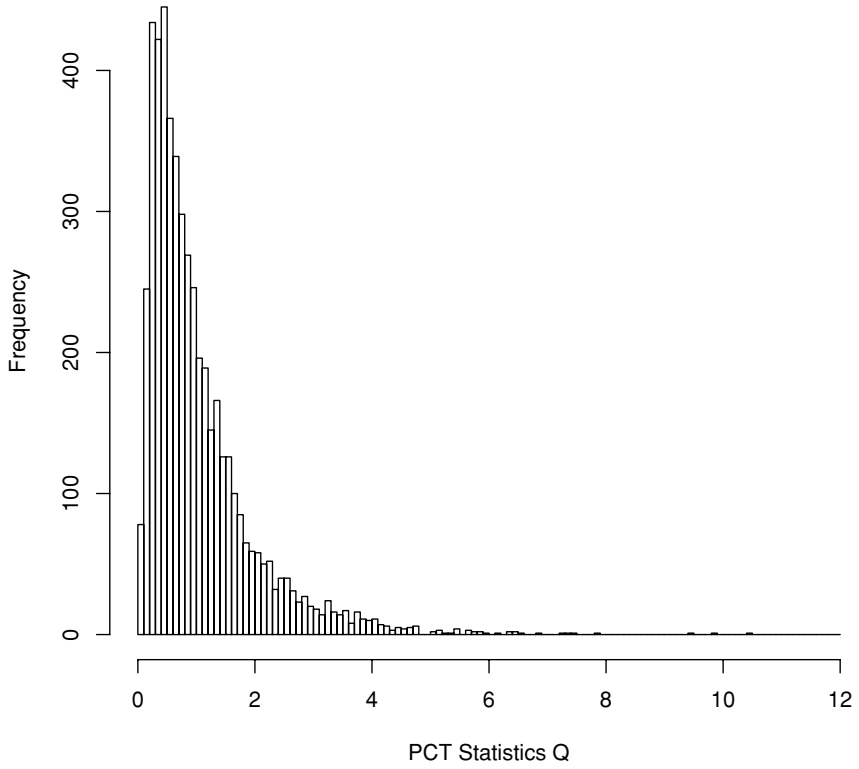

Figure 1. The histogram of 5000 statistics Q under the null hypothesis.

where

$$\tau_{ij} = \frac{\sigma_{ij}}{n_{1i}n_{1j}} \sum_{l=1}^{n_1} M_{1li}M_{1lj} + \frac{\sigma_{ij}}{n_{2i}n_{2j}} \sum_{l=1}^{n_2} M_{2li}M_{2lj}. \quad (6)$$

THEOREM 1: *Under the assumptions of Lemma* 1 *the mean and variance of the test statistic* Q *defined in* (4) *are*

$$E(Q) = \frac{1}{p} \sum_{j=1}^{p} \left(\frac{n_{1j} + n_{2j} - 2}{n_{1j} + n_{2j} - 4} \right), \quad (7)$$

and

$$\begin{aligned} &\mathrm{Var}(Q) \\ &= \frac{2}{p^2} \sum_{j=1}^{p} \left\{ \left(\frac{n_{1j} + n_{2j} - 2}{n_{1j} + n_{2j} - 4} \right)^2 \left(\frac{n_{1j} + n_{2j} - 3}{n_{1j} + n_{2j} - 6} \right) \right\} \\ &\quad + \frac{2}{p^2} \sum_{1 \le i < j \le p} \left[\left\{ \sigma_i^2 \sigma_j^2 + \frac{2\tau_{ij}^2}{\left(\frac{1}{n_{1i}} + \frac{1}{n_{2i}} \right) \left(\frac{1}{n_{1j}} + \frac{1}{n_{2j}} \right)} \right\} E\left(\frac{1}{S_i^2 S_j^2} \right) \right. \\ &\quad \left. - \left(\frac{n_{1i} + n_{2i} - 2}{n_{1i} + n_{2i} - 4} \right) \left(\frac{n_{1j} + n_{2j} - 2}{n_{1j} + n_{2j} - 4} \right) \right], \end{aligned} \quad (8)$$

where τ_{ij} *is defined in* (6) *and* $n_{1j} + n_{2j} > 6$ *for any* $j = 1, \ldots, p$.

The proofs of the lemma and theorem are given in the Appendix.

Note that the mean depends only on the sample sizes and p, whereas the variance depends on the unknown population variances and covariances. We need an estimate of $\mathrm{Var}(Q)$ to obtain the desired approximating distribution. A natural estimation strategy is to replace the unknown population variances and covariances with the corresponding sample componentwise moments, and replace $E\{(S_i^2 S_j^2)^{-1}\}$ by the unbiased estimator $(S_i^2 S_j^2)^{-1}$. However, this has the effect of replacing all of the terms $\sigma_i^2 \sigma_j^2 E\{(S_i^2 S_j^2)^{-1}\}$ in the variance expression by 1, when they differ from 1 in general. In fact, under independence, Jensen's Inequality shows that $\sigma_i^2 \sigma_j^2 E\{(S_i^2 S_j^2)^{-1}\} > 1$, thus replacing all such terms by 1 systematically underestimates them. We addressed this problem by replacing $E\{(S_i^2 S_j^2)^{-1}\}$ in the expression for $\mathrm{Var}(Q)$ by the approximation

$$E\left(\frac{1}{S_i^2 S_j^2} \right) \approx \frac{1}{E\left(S_i^2 S_j^2\right)} + \frac{\sigma_i^4 \sigma_j^4}{\left\{ E\left(S_i^2 S_j^2\right) \right\}^3} (w_i w_j - 1), \quad (9)$$

where $w_t = (n_{1t} + n_{2t})/(n_{1t} + n_{2t} - 2)$, $t = 1, \ldots, p$. The approximation in (9) is based on a Taylor series expansion and an independence assumption. Although the independence assumption used here is at odds with our model assumptions, it is important to note that we invoke independence here only to derive an approximation to $E\{(S_i^2 S_j^2)^{-1}\}$, the utility of which is confirmed in our simulation studies. It should also be noted that, when $\min\{n_{1j}, n_{2j}, j = 1, \ldots, p\}$ goes to infinity, both sides in the approximation (9) converge to the same value, $(\sigma_i^2 \sigma_j^2)^{-1}$, with the independence assumption unnecessary. Details of the approximation are given in the Appendix. Substituting the right-hand side of (9) for $E\{(S_i^2 S_j^2)^{-1}\}$ in

expression (8) for $\mathrm{Var}(Q)$ results in the approximate variance formula,

$$\begin{aligned}
&\mathrm{Var}(Q)\\
&\approx \frac{2}{p^2}\sum_{j=1}^{p}\left\{\left(\frac{n_{1j}+n_{2j}-2}{n_{1j}+n_{2j}-4}\right)^2\left(\frac{n_{1j}+n_{2j}-3}{n_{1j}+n_{2j}-6}\right)\right\}\\
&\quad+\frac{2}{p^2}\sum_{1\le i<j\le p}\left[\left\{\sigma_i^2\sigma_j^2+\frac{2\tau_{ij}^2}{\left(\frac{1}{n_{1i}}+\frac{1}{n_{2i}}\right)\left(\frac{1}{n_{1j}}+\frac{1}{n_{2j}}\right)}\right\}\right.\\
&\qquad\times\frac{1}{E\left(S_i^2S_j^2\right)}+\frac{\sigma_i^6\sigma_j^6}{\left\{E\left(S_i^2S_j^2\right)\right\}^3}(w_iw_j-1)\\
&\qquad+\left\{\frac{2\tau_{ij}^2}{\left(\frac{1}{n_{1i}}+\frac{1}{n_{2i}}\right)\left(\frac{1}{n_{1j}}+\frac{1}{n_{2j}}\right)}\right\}\\
&\qquad\times\frac{\sigma_i^4\sigma_j^4}{\left\{E\left(S_i^2S_j^2\right)\right\}^3}(w_iw_j-1)\\
&\qquad\left.-\left(\frac{n_{1i}+n_{2i}-2}{n_{1i}+n_{2i}-4}\right)\left(\frac{n_{1j}+n_{2j}-2}{n_{1j}+n_{2j}-4}\right)\right],
\end{aligned}\tag{10}$$

where τ_{ij} is defined in (6).

Our estimate of $\mathrm{Var}(Q)$ is obtained by replacing the population variances and covariances in the right-hand side of (10) with sample estimates. We estimate the population variances σ_i^2 by the pooled componentwise sample variances S_i^2 defined in (2). Although S_{ij} defined in (3) is the natural estimator of σ_{ij}, the fact that the number of covariances to be estimated can be large suggests that some form of shrinking would be advantageous in this setting, and we found that to be the case. Threshold shrinkage worked well in preliminary simulations not reported here, and is attractive because of its simplicity. We estimate σ_{ij} with $\hat{\sigma}_{ij}=S_{ij}\Delta_{ij}$, where $\Delta_{ij}=I(p_{ij}<0.05)$, and p_{ij} is the P-value of the usual regression test for zero correlation. Using threshold estimators in the estimator of $\mathrm{Var}(Q)$ reduces both variability and bias and substantially improves its performance as measured in simulations. We estimate $E(S_i^2S_j^2)$ by $S_i^2S_j^2$. The resulting estimator of $\mathrm{Var}(Q)$ is

$$\begin{aligned}
&\widehat{\mathrm{Var}(Q)}\\
&\approx \frac{2}{p^2}\sum_{j=1}^{p}\left\{\left(\frac{n_{1j}+n_{2j}-2}{n_{1j}+n_{2j}-4}\right)^2\left(\frac{n_{1j}+n_{2j}-3}{n_{1j}+n_{2j}-6}\right)\right\}\\
&\quad+\frac{2}{p^2}\sum_{1\le i<j\le p}\left[\left\{1+\frac{2\hat{\tau}_{ij}^2}{\left(\frac{1}{n_{1i}}+\frac{1}{n_{2i}}\right)\left(\frac{1}{n_{1j}}+\frac{1}{n_{2j}}\right)S_i^2S_j^2}\right\}w_iw_j\right.\\
&\qquad\left.-\left(\frac{n_{1i}+n_{2i}-2}{n_{1i}+n_{2i}-4}\right)\left(\frac{n_{1j}+n_{2j}-2}{n_{1j}+n_{2j}-4}\right)\right],
\end{aligned}\tag{11}$$

where

$$\hat{\tau}_{ij}=\frac{\hat{\sigma}_{ij}}{n_{1i}n_{1j}}\sum_{l=1}^{n_1}M_{1li}M_{1lj}+\frac{\hat{\sigma}_{ij}}{n_{2i}n_{2j}}\sum_{l=1}^{n_2}M_{2li}M_{2lj}.$$

4. Simulation Studies

Monte Carlo simulations were carried out in order to assess the quality of the null distribution approximation by the scaled chi-squared distribution for the PCT statistic Q defined by equation (4). For comparison, we also studied the behavior of the exact null distribution of Hotelling's T^2 statistic and the approximate null distribution of Srivastava's test statistic. Furthermore, we studied the power functions of PCT, and compared it with Hotelling's T^2 or Srivastava's tests in different situations. All these studies were conducted for both complete and incomplete data.

4.1 *Simulation Design*

We considered both the cases $p<n$ and $p>n$. When $p<n$, we compared the pooled component test with Hotelling's T^2 test. When $p>n$, we compared it with the test proposed by Srivastava (2004). The data were generated as multivariate normal with common covariance Σ, and mean vectors $\boldsymbol{\mu}_k (k=1,2)$. We considered both a null case ($\boldsymbol{\mu}_1=\boldsymbol{\mu}_2$) and an alternative case ($\boldsymbol{\mu}_1\neq\boldsymbol{\mu}_2$),

- Null case: $\boldsymbol{\mu}_1=\boldsymbol{\mu}_2=(0,0,\ldots,0)^T_{p\times1}$,
- Alternative case: $\boldsymbol{\mu}_1=(0,0,\ldots,0)^T_{p\times1}$; $\boldsymbol{\mu}_2=(0.3,0.3,\ldots,0.3)^T_{p\times1}$.

For both null and alternative cases the covariance matrix Σ had three different structures. In the first, Σ is the $p\times p$ identity matrix. In the second, Σ is the matrix with autoregressive structure, $\Sigma_{ij}=\rho^{|i-j|}$. In the third, $\Sigma_{ii}=1$, and $\Sigma_{ij}=r$ for all $i\neq j$. The three covariance structures are identified as independence (Indep), autoregressive (AR), and equal correlation (EC), respectively. For the EC model $r=0.35$ was chosen to match approximately the average correlation in the drug discovery data analyzed in the next section. For the AR model ρ was chosen so that the average correlation in the AR correlation matrix equaled $r=0.35$.

The total sample sizes were taken to be $n=30$ and $n=80$, and each group's sample has size $n/2$. In the situation $p<n$, we took $p=6$ and have $\rho=0.6$ in the AR structure, whereas, in the situation $p>n$, we let $p=100$ and have $\rho=0.955$. For each combination of n and p, we generated and analyzed 5000 replicated data sets. For each test we computed the proportions of rejections with the significance levels $\alpha=0.1$, 0.05, and 0.01. These proportions estimate the size of the tests under the null and the power of the tests under the alternative.

As for the generation of missing data, based on the complete data generated, we randomly made 20% of them missing. In this situation, Hotelling's T^2 and Srivastava's tests are computed using only complete-case data. To ensure enough information for computation and comparison, we consider only data sets with sample size 80.

4.2 *Simulation Results*

Table 1 exhibits the results in the null case. The Monte Carlo estimated test sizes have standard errors approximately equal to 0.004, 0.003, and 0.001 for the given levels 0.1, 0.05, and

Table 1
The estimated test sizes for complete data under the null case

p	n	Covariance	Methods	0.1*	0.05*	0.01*
6	30	Indep	Hotelling	0.097	0.052	0.011
			PCT	0.101	0.049	0.010
		AR	Hotelling	0.103	0.047	0.007
			PCT	0.096	0.053	0.014
		EC	Hotelling	0.096	0.046	0.012
			PCT	0.095	0.055	0.019
6	80	Indep	Hotelling	0.102	0.052	0.011
			PCT	0.102	0.052	0.010
		AR	Hotelling	0.101	0.053	0.012
			PCT	0.088	0.045	0.012
		EC	Hotelling	0.091	0.045	0.009
			PCT	0.093	0.050	0.013
100	30	Indep	Srivastava	0.081	0.041	0.009
			PCT	0.064	0.026	0.005
		AR	Srivastava	0.000	0.000	0.000
			PCT	0.103	0.06	0.019
		EC	Srivastava	0.000	0.000	0.000
			PCT	0.097	0.069	0.030
100	80	Indep	Srivastava	0.093	0.049	0.011
			PCT	0.078	0.035	0.004
		AR	Srivastava	0.000	0.000	0.000
			PCT	0.097	0.052	0.013
		EC	Srivastava	0.000	0.000	0.000
			PCT	0.086	0.057	0.026

*The nominal significance level.

Table 2
The estimated powers for complete data under the alternative case

p	n	Covariance	Methods	0.1*	0.05*	0.01*
6	30	Indep	Hotelling	0.332	0.212	0.072
			PCT	0.358	0.243	0.098
		AR	Hotelling	0.179	0.096	0.025
			PCT	0.282	0.194	0.089
		EC	Hotelling	0.182	0.105	0.022
			PCT	0.299	0.220	0.107
6	80	Indep	Hotelling	0.758	0.638	0.384
			PCT	0.778	0.670	0.421
		AR	Hotelling	0.371	0.255	0.095
			PCT	0.558	0.455	0.267
		EC	Hotelling	0.361	0.244	0.087
			PCT	0.589	0.486	0.295
100	30	Indep	Srivastava	0.605	0.468	0.235
			PCT	0.979	0.953	0.832
		AR	Srivastava	—	—	—
			PCT	0.365	0.271	0.149
		EC	Srivastava	—	—	—
			PCT	0.385	0.315	0.210
100	80	Indep	Srivastava	0.966	0.912	0.695
			PCT	1.000	1.000	1.000
		AR	Srivastava	—	—	—
			PCT	0.719	0.628	0.440
		EC	Srivastava	—	—	—
			PCT	0.708	0.626	0.473

*The nominal significance level.

0.01, respectively. If the distribution is derived correctly for the test statistic, then we would anticipate that the estimated test sizes should be close to their nominal significance levels. Clearly, in the case where $p < n$, the table shows that, for the pooled component test, as well as for Hotelling's T^2 test, the estimated test sizes are satisfactory. When $p > n$, because Hotelling's T^2 test is not available, we use Srivastava's test. As seen in the table, when the covariance has Indep structure, Srivastava's test performs well, and the PCT tends to be conservative (has smaller sizes). However, in the other two cases of covariance structure, AR and EC, the chi-squared distribution produces good results in the approximation, while Srivastava's test rejects too often, with all estimated test sizes equal to 0. In fairness, we note that we are using Srivastava's test with covariance matrices (EC) not covered by the supporting asymptotic theory. However, it should also be noted that verifying conditions on the covariance matrix in practice is problematic.

A further study of the chi-squared distribution under the null for the PCT has been conducted in a more extensive simulation with an additional $n = 200$, $p = 30$, and four more significance levels, 0.2, 0.15, 0.07, and 0.03, and the results of estimated test sizes based on 5000 replications are summarized in the Appendix, available at `http://www.tibs.org/biometrics`. The results reveal an overall satisfactory performance of the scaled chi-squared distribution. It turns out that, when the estimated sizes differ significantly from the nominal sizes, they are usually on the conservative side (too small) rather than too large. For example, at the nominal significance levels 0.1 and 0.05, the percentages of cases in which the estimated sizes are significantly too small are 48% and 22%, respectively, while the percentages of cases with significantly too large estimates are 0% and 19%, respectively.

Table 2 shows the results under the alternative. In terms of the estimated powers, in the case where $p < n$, the PCT produces competitive results to Hotelling's T^2 test in the Indep case and performs better in the AR and EC cases. When $p > n$, in the AR and EC cases, because we could not get a reasonable size under the null for Srivastava's test (we got sizes = 0), we did not calculate power. These cases appear as missing values in the table. However, in the Indep case where we got nonzero sizes for Srivastava's test, the PCT has smaller sizes (more conservative) than Srivastava's test as seen in Table 1, yet is seen to be more powerful in Table 2.

Table 3 shows the results when the data have 20% missing values. In this situation, Hotelling's T^2 and Srivastava's tests are based on only complete-case data, and we estimated their test powers conditionally. Specifically, the powers were estimated by using the proportions of rejected samples in the replications for which there are enough complete-case data for computation. Because, when $p = 100$ and $n = 80$, none of the 5000 replications has complete-case data, Srivastava's test statistic is not computed, and we do not list the results in the tables. Examination of the table reveals that the chi-squared distribution approximation still performs satisfactorily. In terms of power, the results show that the pooled component test is consistently superior to Hotelling's T^2 test.

5. Application to Drug Discovery Data

We illustrate the new procedure with quantitative structure-activity data from a drug discovery application. The data

Table 3
The estimated test sizes and powers for incomplete data with 20% *missing values*

Case	p	n	Covariance	Methods	0.1*	0.05*	0.01*
$\mu_1 = \mu_2$	6	80	Indep	Hotelling	0.101	0.050	0.015
				PCT	0.099	0.053	0.011
			AR	Hotelling	0.093	0.046	0.008
				PCT	0.092	0.047	0.013
			EC	Hotelling	0.098	0.049	0.012
				PCT	0.094	0.050	0.011
$\mu_1 = \mu_2$	100	80	Indep	PCT	0.081	0.035	0.004
			AR	PCT	0.097	0.051	0.016
			EC	PCT	0.090	0.062	0.028
$\mu_1 \neq \mu_2$	6	80	Indep	Hotelling	0.237	0.135	0.037
				PCT	0.674	0.548	0.314
			AR	Hotelling	0.181	0.103	0.031
				PCT	0.523	0.421	0.240
			EC	Hotelling	0.175	0.098	0.025
				PCT	0.552	0.455	0.264
$\mu_1 \neq \mu_2$	100	80	Indep	PCT	1.000	1.000	1.000
			AR	PCT	0.712	0.627	0.451
			EC	PCT	0.704	0.627	0.478

*The nominal significance level.

contain 576 chemical compounds, whose chemical structural features are represented by 1024 binary descriptors according to the presence or absence of certain atom-pairs. The biological activity of each compound was measured on 10 target proteins, resulting in a 10-dimensional activity response vector. No missing values exist in this data set. Tree-structured approaches are desired to explore relationships between structure of the compounds and activity and accordingly to identify those aspects of molecular structure that are relevant to a particular biological activity.

Classification trees were built using the MultiSCAM (Keefer, 2001) algorithm, once using Hotelling's T^2 to determine significance of splits, and a second time using the new PCT test. We use 0.05 as a threshold for the Bonferroni-adjusted P-values to judge significance of the splits, that is, if the minimum-adjusted P-value is less than 0.05, the node is split into two subnodes; otherwise, splitting will stop at this node and we call this node a *terminal node*. In the tree built based on PCT, 69 descriptors are selected and there are 75 terminal nodes, whereas the tree based on Hotelling's T^2 identifies 30 descriptors and produced 31 terminal nodes. Clearly, PCT has greater power to detect significant splits compared to Hotelling's T^2 test and hence lower probability of missing important descriptors.

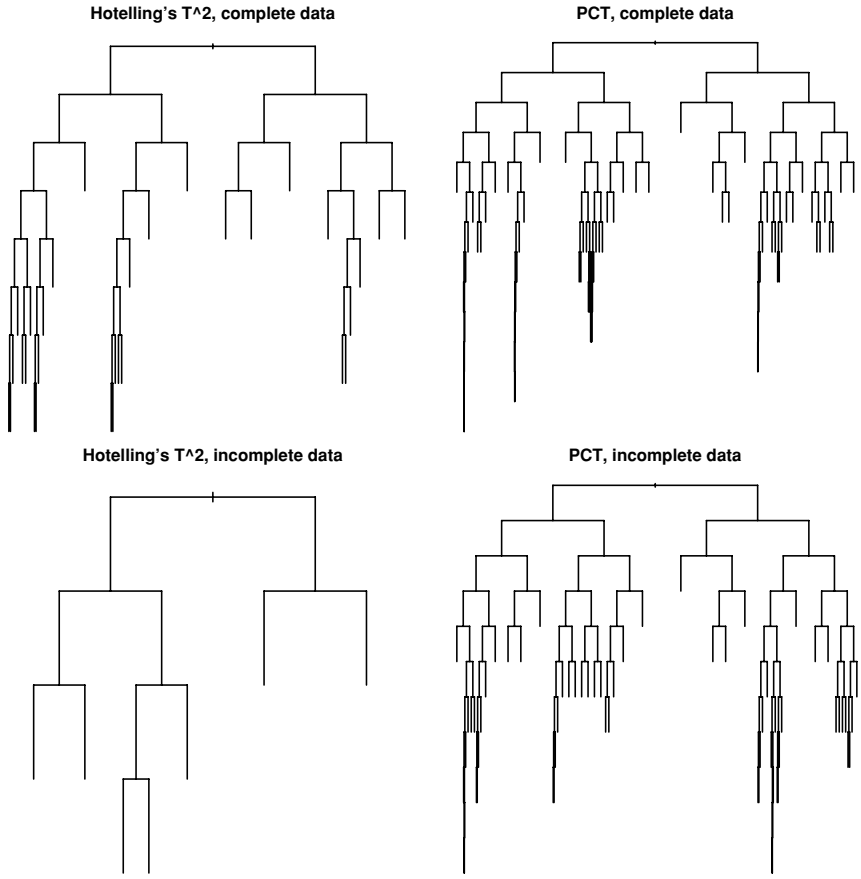

Figure 2. The tree dendrograms built by Hotelling's T^2 and PCT for both the complete and incomplete (15% missing) drug discovery data.

To assess the performance of PCT for missing data, we randomly deleted 15% of the activity data and rebuilt the trees with the two tests again (Hotelling's T^2 test is using the complete-case data). The tree by PCT selects 58 descriptors and has 64 terminal nodes, whereas the tree by Hotelling's T^2 test identifies only 6 descriptors and has 7 terminal nodes. Thus with missing values, Hotelling's T^2 test becomes almost powerless. The resulting dendrograms based on Hotelling's T^2 test and PCT for both complete and incomplete data are presented in Figure 2. A striking feature of the figure is the similarity of the trees built by PCT for the complete and incomplete data, compared to the relative dissimilarity of the trees built using Hotelling's T^2. The figure also illustrates the greater power of PCT for splitting nodes with smaller sample sizes.

6. Conclusions

The pooled component test statistic is built on component-wise statistics, and hence avoids the problems of $p > n$ and missing data, which is an advantage over the well-known Hotelling's T^2 test. Furthermore, our simulations indicate that the null distribution of the pooled component test statistic is well approximated by the scaled chi-squared distribution, making it easy to apply.

The simulation results show that for missing data, the pooled component test is significantly better than Hotelling's T^2 test in terms of power. Even for complete data, the pooled component test performs comparably to Hotelling's T^2. Srivastava's test is designed to address the problem of $p > n$, and is justified asymptotically under certain conditions on the covariance matrix. However, in our simulation studies its finite-sample performance was lacking in some cases; also, its application with missing data is problematic as is calculating the generalized inverse when p is very large.

Acknowledgements

We thank the editor and referees for several helpful comments on an earlier draft of the article that substantially improved both content and composition. We also thank GlaxoSmithKline, and Chris Keefer and Chris Bizon in particular, for encouragement, discussions, and financial support. Finally, the first author would like to thank the Biometrics Society David P. Byar Young Investigator Award Committee for recognizing his research contribution with the 2005 Byar Award.

References

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*. New York: Wiley.

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association* **97,** 77–87.

Fujikoshi, Y. (1997). An asymptotic expansion for the distribution of Hotelling's T^2-statistic under nonnormality. *Journal of Multivariate Analysis* **61,** 187–193.

Hawkins, D. M. E., ed. (1982). *Topics in Applied Multivariate Analysis*. Cambridge, U.K.: Cambridge University Press.

Hawkins, D. M., Young, S. S., and Rusinko, A., III (1997). Analysis of a large structure-activity data set using recursive partitioning. *Quantitative Structure-Activity Relationships* **16,** 296–302.

Hotelling, H. (1931). The generalization of Student's ratio. *Annals of Mathematical Statistics* **2,** 360–378.

Kano, Y. (1995). An asymptotic null and nonnull distribution of Hotelling's T^2-statistic under general distributions. *American Journal of Mathematical and Management Sciences* **15,** 317–341.

Keefer, C. E. (2001). *Use of multivariate data mining techniques in pharmaceutical systems based research*. Abstract of papers, 222nd ACS National Meeting, Chicago.

Mudholkar, G. S. and Srivastava, D. K. (2000). Robust analogs of Hotelling's two-sample T^2. *Communications in Statistics—Theory and Methods* **29,** 2717–2749.

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*. New York: Wiley.

Rousseeuw, P. J. and Molenberghs, G. (1993). Transformation of non positive semidefinite correlation matrices. *Communications in Statistics—Theory and Methods* **22,** 965–984.

Srivastava, M. S. (2004). *Multivariate theory for analyzing high-dimensional data*. Technical Report, University of Toronto, Toronto, Canada.

Srivastava, M. S. and Carter, E. M. (1986). The maximum likelihood method for non-response in sample survey. *Survey Methodology* **12,** 61–72.

Received July 2004. *Revised November* 2005.
Accepted November 2005.

Appendix

A.1 *Proof of Lemma* 1

Based on the multivariate normality assumption, under the null hypothesis $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, we have

$$\begin{pmatrix} \bar{Y}_{1i} - \bar{Y}_{2i} \\ \bar{Y}_{1j} - \bar{Y}_{2j} \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_i^2 & \tau_{ij} \\ \tau_{ij} & \tau_j^2 \end{pmatrix} \right),$$

where

$$\tau_i^2 = \mathrm{Var}(\bar{Y}_{1i} - \bar{Y}_{2i}) = \left(\frac{1}{n_{1i}} + \frac{1}{n_{2i}} \right) \sigma_i^2,$$

$$\tau_j^2 = \mathrm{Var}(\bar{Y}_{1j} - \bar{Y}_{2j}) = \left(\frac{1}{n_{1j}} + \frac{1}{n_{2j}} \right) \sigma_j^2,$$

and

$$\begin{aligned} \tau_{ij} &= \mathrm{Cov}\{(\bar{Y}_{1i} - \bar{Y}_{2i}), (\bar{Y}_{1j} - \bar{Y}_{2j})\} \\ &= \mathrm{Cov}(\bar{Y}_{1i}, \bar{Y}_{1j}) + \mathrm{Cov}(\bar{Y}_{2i}, \bar{Y}_{2j}) \\ &= \frac{\sigma_{ij}}{n_{1i} n_{1j}} \sum_{l=1}^{n_1} M_{1li} M_{1lj} + \frac{\sigma_{ij}}{n_{2i} n_{2j}} \sum_{l=1}^{n_2} M_{2li} M_{2lj}. \end{aligned}$$

For simplicity of notation, let us use V_i to denote $(\bar{Y}_{1i} - \bar{Y}_{2i})$ and V_j to denote $(\bar{Y}_{1j} - \bar{Y}_{2j})$. Then the expectation of the product can be written as

$$
\begin{aligned}
&E\{(\bar{Y}_{1i}-\bar{Y}_{2i})^2(\bar{Y}_{1j}-\bar{Y}_{2j})^2\}\\
&\quad=E(V_i^2V_j^2)\\
&\quad=E\{V_j^2E(V_i^2\,|\,V_j)\}\\
&\quad=E(V_j^2[\{E(V_i\,|\,V_j)\}^2+\mathrm{Var}(V_i\,|\,V_j)])\\
&\quad=E[V_j^2\{E(V_i\,|\,V_j)\}^2]+E\{V_j^2\mathrm{Var}(V_i\,|\,V_j)\}.
\end{aligned}
$$

Simple computations yield

$$
E[V_j^2\{E(V_i\,|\,V_j)\}^2]=E\left\{\left(\frac{\tau_{ij}}{\tau_j^2}\right)^2V_j^4\right\}=3\tau_{ij}^2,
$$

and $E\{V_j^2\mathrm{Var}(V_i\,|\,V_j)\}=\tau_i^2\tau_j^2-\tau_{ij}^2$. Therefore, we have

$$
\begin{aligned}
&E\{(\bar{Y}_{1i}-\bar{Y}_{2i})^2(\bar{Y}_{1j}-\bar{Y}_{2j})^2\}\\
&\quad=3\tau_{ij}^2+\tau_i^2\tau_j^2-\tau_{ij}^2\\
&\quad=\tau_i^2\tau_j^2+2\tau_{ij}^2\\
&\quad=\left(\frac{1}{n_{1i}}+\frac{1}{n_{2i}}\right)\left(\frac{1}{n_{1j}}+\frac{1}{n_{2j}}\right)\sigma_i^2\sigma_j^2+2\tau_{ij}^2.
\end{aligned}
$$

A.2 *Proof of Theorem* 1

(1) It is straightforward to see that $a_jQ_j\sim F_{1,n_{1j}+n_{2j}-2}$. Hence, we have

$$
E(Q)=\frac{1}{p}\sum_{j=1}^{p}E(a_jQ_j)=\frac{1}{p}\sum_{j=1}^{p}\left(\frac{n_{1j}+n_{2j}-2}{n_{1j}+n_{2j}-4}\right).
$$

(2) According to the formula of variance, we have

$$
\mathrm{Var}(Q)=\frac{1}{p^2}\sum_{j=1}^{p}\mathrm{Var}(a_jQ_j)+\frac{2}{p^2}\sum_{1\le i<j\le p}\mathrm{Cov}(a_iQ_i,a_jQ_j).
$$

Here

$$
\mathrm{Var}(a_jQ_j)=2\left(\frac{n_{1j}+n_{2j}-2}{n_{1j}+n_{2j}-4}\right)^2\left(\frac{n_{1j}+n_{2j}-3}{n_{1j}+n_{2j}-6}\right),
$$

and

$$
\begin{aligned}
&\mathrm{Cov}(a_iQ_i,a_jQ_j)\\
&\quad=a_ia_jE(Q_iQ_j)-E(a_iQ_i)E(a_jQ_j)\\
&\quad=a_ia_jE\{(\bar{Y}_{1i}-\bar{Y}_{2i})^2(\bar{Y}_{1j}-\bar{Y}_{2j})^2\}E\left(\frac{1}{S_i^2S_j^2}\right)\\
&\qquad-\left(\frac{n_{1i}+n_{2i}-2}{n_{1i}+n_{2i}-4}\right)\left(\frac{n_{1j}+n_{2j}-2}{n_{1j}+n_{2j}-4}\right)\\
&\quad=\left\{\sigma_i^2\sigma_j^2+\frac{2\tau_{ij}^2}{\left(\frac{1}{n_{1i}}+\frac{1}{n_{2i}}\right)\left(\frac{1}{n_{1j}}+\frac{1}{n_{2j}}\right)}\right\}E\left(\frac{1}{S_i^2S_j^2}\right)\\
&\qquad-\left(\frac{n_{1i}+n_{2i}-2}{n_{1i}+n_{2i}-4}\right)\left(\frac{n_{1j}+n_{2j}-2}{n_{1j}+n_{2j}-4}\right).
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\mathrm{Var}(Q)\\
&=\frac{2}{p^2}\sum_{j=1}^{p}\left\{\left(\frac{n_{1j}+n_{2j}-2}{n_{1j}+n_{2j}-4}\right)^2\left(\frac{n_{1j}+n_{2j}-3}{n_{1j}+n_{2j}-6}\right)\right\}\\
&\quad+\frac{2}{p^2}\sum_{1\le i<j\le p}\left[\left\{\sigma_i^2\sigma_j^2+\frac{2\tau_{ij}^2}{\left(\frac{1}{n_{1i}}+\frac{1}{n_{2i}}\right)\left(\frac{1}{n_{1j}}+\frac{1}{n_{2j}}\right)}\right\}\right.\\
&\qquad\left.\times E\left(\frac{1}{S_i^2S_j^2}\right)-\left(\frac{n_{1i}+n_{2i}-2}{n_{1i}+n_{2i}-4}\right)\left(\frac{n_{1j}+n_{2j}-2}{n_{1j}+n_{2j}-4}\right)\right].
\end{aligned}
$$

A.3 *Approximating* $E\{(S_i^2S_j^2)^{-1}\}$

According to a Taylor expansion, we have

$$
\begin{aligned}
\frac{1}{S_i^2S_j^2}\approx\;&\frac{1}{E(S_i^2S_j^2)}-\frac{1}{\{E(S_i^2S_j^2)\}^2}\{S_i^2S_j^2-E(S_i^2S_j^2)\}\\
&+\frac{1}{\{E(S_i^2S_j^2)\}^3}\{S_i^2S_j^2-E(S_i^2S_j^2)\}^2,
\end{aligned}
$$

which yields that

$$
E\left(\frac{1}{S_i^2S_j^2}\right)\approx\frac{1}{E(S_i^2S_j^2)}+\frac{1}{\{E(S_i^2S_j^2)\}^3}\mathrm{Var}(S_i^2S_j^2). \quad (A.1)
$$

The computation of $\mathrm{Var}(S_i^2S_j^2)$ is very challenging due to the possible correlation between S_i^2 and S_j^2. However, under the simple assumption of independence between S_i^2 and S_j^2, we have the following theorem.

THEOREM 2: *If* $Y_{kl1},\ldots,Y_{klp}$ *are independent,* $k=1,2$, *and* $l=1,\ldots,n_k$, *then*

$$
\mathrm{Var}(S_i^2S_j^2)=(w_iw_j-1)\sigma_i^4\sigma_j^4,
$$

where $w_t=(n_{1t}+n_{2t})/(n_{1t}+n_{2t}-2)$, $t=1,\ldots,p$.

See next section for its proof. Although the derivation of $\mathrm{Var}(S_i^2S_j^2)$ in Theorem 2 requires independence among variables $Y_{ki1},\ldots,Y_{kip}$, we may still use this result in the usual nonindependent cases. Thus, approximately,

$$
E\left(\frac{1}{S_i^2S_j^2}\right)\approx\frac{1}{E(S_i^2S_j^2)}+\frac{\sigma_i^4\sigma_j^4}{\{E(S_i^2S_j^2)\}^3}(w_iw_j-1). \quad (A.2)
$$

A.4 *Proof of Theorem* 2

By definition and according to the independence assumption, $\mathrm{Var}(S_i^2S_j^2)=E(S_i^2S_j^2)^2-\{E(S_i^2S_j^2)\}^2=E(S_i^4)E(S_j^4)-\{E(S_i^2)E(S_j^2)\}^2$. On the other hand, we have

$$
\frac{(n_{1i}+n_{2i}-2)}{\sigma_i^2}S_i^2\sim\chi^2_{n_{1i}+n_{2i}-2}.
$$

Thus,

$$
E\left\{\frac{(n_{1i}+n_{2i}-2)}{\sigma_i^2}S_i^2\right\}=n_{1i}+n_{2i}-2,
$$

and

$$
\mathrm{Var}\left\{\frac{(n_{1i}+n_{2i}-2)}{\sigma_i^2}S_i^2\right\}=2(n_{1i}+n_{2i}-2).
$$

Then, it is easy to obtain $E(S_i^2) = \sigma_i^2$, and

$$\begin{aligned} E(S_i^4) &= \left\{E\left(S_i^2\right)\right\}^2 + \mathrm{Var}\left(S_i^2\right) \\ &= \sigma_i^4 + \left\{\frac{\sigma_i^2}{(n_{1i}+n_{2i}-2)}\right\}^2 \mathrm{Var}\left\{\frac{(n_{1i}+n_{2i}-2)}{\sigma_i^2} S_i^2\right\} \\ &= \sigma_i^4 + \frac{2\sigma_i^4}{(n_{1i}+n_{2i}-2)} \\ &= \frac{(n_{1i}+n_{2i})\sigma_i^4}{(n_{1i}+n_{2i}-2)}. \end{aligned}$$

By the same argument, we have

$$E\left(S_j^2\right) = \sigma_j^2 \quad \text{and} \quad E\left(S_j^4\right) = \frac{(n_{1j}+n_{2j})\sigma_j^4}{(n_{1j}+n_{2j}-2)}.$$

Therefore,

$$\begin{aligned} \mathrm{Var}\left(S_i^2 S_j^2\right) &= \left\{\frac{(n_{1i}+n_{2i})\sigma_i^4}{(n_{1i}+n_{2i}-2)}\right\}\left\{\frac{(n_{1j}+n_{2j})\sigma_j^4}{(n_{1j}+n_{2j}-2)}\right\} - \sigma_i^4\sigma_j^4 \\ &= \left\{\left(\frac{n_{1i}+n_{2i}}{n_{1i}+n_{2i}-2}\right)\left(\frac{n_{1j}+n_{2j}}{n_{1j}+n_{2j}-2}\right) - 1\right\}\sigma_i^4\sigma_j^4 \\ &= (w_i w_j - 1)\sigma_i^4\sigma_j^4. \end{aligned}$$

The estimated test sizes for PCT under the null case in a more extensive investigation

p	n	Covariance	0.2⋆	0.15⋆	0.1⋆	0.07⋆	0.05⋆	0.03⋆	0.01⋆
6	30	Indep	0.195	0.146	0.101	0.071	0.049	0.031	0.010
		AR	0.185	0.136	0.096	0.069	0.053	0.036	0.014
		EC	0.182	0.135	0.095	0.072	0.055	0.038	0.019
	80	Indep	0.203	0.154	0.102	0.071	0.052	0.031	0.010
		AR	0.179	0.133	0.088	0.063	0.044	0.030	0.012
		EC	0.180	0.135	0.093	0.069	0.050	0.035	0.013
	200	Indep	0.188	0.141	0.091	0.063	0.047	0.030	0.010
		AR	0.182	0.131	0.088	0.062	0.045	0.028	0.013
		EC	0.185	0.138	0.090	0.061	0.046	0.032	0.013
30	30	Indep	0.192	0.141	0.093	0.065	0.043	0.022	0.007
		AR	0.184	0.138	0.101	0.078	0.062	0.042	0.018
		EC	0.162	0.129	0.099	0.075	0.063	0.046	0.025
	80	Indep	0.194	0.142	0.089	0.062	0.043	0.025	0.008
		AR	0.174	0.136	0.091	0.067	0.051	0.035	0.014
		EC	0.157	0.128	0.093	0.070	0.055	0.043	0.020
	200	Indep	0.189	0.141	0.092	0.064	0.043	0.024	0.007
		AR	0.170	0.126	0.082	0.061	0.048	0.032	0.014
		EC	0.148	0.113	0.089	0.068	0.050	0.034	0.018
100	30	Indep	0.170	0.116	0.064	0.043	0.026	0.014	0.005
		AR	0.191	0.142	0.103	0.076	0.060	0.041	0.019
		EC	0.157	0.127	0.097	0.081	0.070	0.053	0.030
	80	Indep	0.167	0.116	0.078	0.049	0.035	0.018	0.004
		AR	0.183	0.142	0.097	0.070	0.052	0.033	0.013
		EC	0.149	0.118	0.086	0.070	0.057	0.043	0.026
	200	Indep	0.185	0.135	0.090	0.061	0.041	0.023	0.007
		AR	0.181	0.142	0.099	0.070	0.051	0.034	0.016
		EC	0.145	0.120	0.089	0.072	0.056	0.043	0.025

⋆The nominal significance level.