

## APPROXIMATE NULL DISTRIBUTION OF THE LARGEST ROOT IN MULTIVARIATE ANALYSIS<sup>1</sup>

BY IAIN M. JOHNSTONE

*Stanford University*

The greatest root distribution occurs everywhere in classical multivariate analysis, but even under the null hypothesis the exact distribution has required extensive tables or special purpose software. We describe a simple approximation, based on the Tracy–Widom distribution, that in many cases can be used instead of tables or software, at least for initial screening. The quality of approximation is studied, and its use illustrated in a variety of settings.

**1. Introduction.** The greatest root distribution is found everywhere in classical multivariate analysis. It describes the null hypothesis distribution for the union intersection test for any number of classical problems, including multiple response linear regression, MANOVA, canonical correlations, equality of covariance matrices and so on. However, the exact null distribution is difficult to calculate and work with, and so the use of extensive tables or special purpose software has always been necessary.

This paper describes a simple asymptotic approximation, based on the Tracy–Widom distribution. The approximation is not solely asymptotic; we argue that it is reasonably accurate over the entire range of the parameters. “Reasonably accurate” means, for example, less than ten percent relative error in the 95th percentile, even when working with *two* variables and any combination of error and hypothesis degrees of freedom.

This paper focuses on the approximation, its accuracy and its applicability to a range of problems in multivariate analysis. A companion paper [Johnstone (2008)] contains all proofs and additional discussion.

Our main claim is that for many applied purposes, the Tracy–Widom approximation can often, if not quite always, substitute for the elaborate tables and computational procedures that have until now been needed. Our hope is that this paper might facilitate the use of the approximation in applications in conjunction with appropriate software.

1.1. *A textbook example.* To briefly illustrate the Tracy–Widom approximation in action, we revisit the rootstock data, as discussed in Rencher (2002),

---

Received October 2008.

<sup>1</sup>Supported in part by NSF DMS-05-05303, 09-06812 and NIH RO1 EB 001988.

*Key words and phrases.* Canonical correlation, characteristic root, equality of covariance matrices, greatest root statistic, largest eigenvalue, MANOVA, multivariate linear model, Tracy–Widom distribution.

pages 170–173. In a classical experiment carried out from 1918–1934, apple trees of different rootstocks were compared (Andrews and Herzberg [(1985), pages 357–360] has more detail). Rencher (2002) gives data for eight trees from each of six rootstocks. Four variables are measured for each tree:  $Girth4$  = trunk girth at 4 years in mm,  $Growth4$  = extension growth at 4 years in m,  $Girth15$  = trunk girth at 15 years in mm, and  $Wt15$  = weight of tree above ground at 15 years in lb.

	Stock	Girth4	Growth4	Girth15	Wt15
1	I	111	2.569	358	760
2	I	119	2.928	375	821
...					
47	VI	113	3.064	363	707
48	VI	111	2.469	395	952

A one-way multivariate analysis of variance can be used to examine the hypothesis of equality of the four-dimensional vectors of mean values corresponding to each of the six groups (rootstocks). The standard tests are based on the eigenvalues of  $(\mathbf{W} + \mathbf{B})^{-1}\mathbf{B}$ , where  $\mathbf{W}$  and  $\mathbf{B}$  are the sums of squares and products matrices within and between groups respectively. We focus here on the largest eigenvalue, with observed value  $\theta^{obs} = 0.652$ . Critical values of the null distribution depend on parameters, here  $s = 4$ ,  $m = 0$ ,  $n = 18.5$  [using (8) below, along with the conventions of Section 5.1 and Definition  $\theta$ ]. Traditionally these are found by reference to tables or charts. Here, the 0.05 critical value is found—after manual interpolation in those tables—to be  $\theta_{0.05} = 0.377$ . The approximation (6) of this paper yields the approximate 0.05 critical value  $\theta_{0.05}^{TW} = 0.384$ , which clearly serves just as well for rejection of the null hypothesis.

It is more difficult in standard packages to obtain  $p$ -values corresponding to  $\theta^{obs}$ . The default is to use a lower bound based on the  $F$  distribution [see (12)], here  $p_F(\theta^{obs}) = 1.7 \times 10^{-8}$ , which is anti-conservative and several orders of magnitude below the Tracy–Widom approximation given in this paper at (11),  $p_{TW}(\theta^{obs}) = 5.6 \times 10^{-5}$ . The latter is much closer to the formally correct value,<sup>2</sup>  $p(\theta^{obs}) = 3.7 \times 10^{-6}$ . When  $p$ -values are very small, typically only the order of magnitude is of interest. We suggest in Section 2.2 that the Tracy–Widom approximation generally comes close to the correct order of magnitude, whereas the default  $F$  bound is often off by several orders.

1.2. *Organization of paper.* The rest of this introduction provides enough background to state the main Tracy–Widom approximation result. Section 2 focuses on the quality of the approximation by looking both at conventional percentiles and at very small  $p$ -values. The remaining Sections 3–6 describe some

<sup>2</sup>This (actually approximate) value is obtained by interpolation from Koev's function `pmaxeigjacobi` which only handles integer values of  $n$ .

of the classical uses of the largest root test in multivariate analysis, in each case in enough detail to identify the parameters used. Some extra attention is paid in Section 6 to the multivariate linear model, in view of the wide variety of null hypotheses that can be considered.

1.3. *Background.* Our setting is the distribution theory associated with sample draws from the multivariate normal distribution. For definiteness, we use the notation of [Mardia, Kent and Bibby \(1979\)](#), to which we also refer for much standard background material. Thus, if  $\mathbf{x}_1, \dots, \mathbf{x}_n$  denotes a random sample from  $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , a  $p$ -variate Gaussian distribution with mean  $\boldsymbol{\mu}$  and covariance matrix  $\boldsymbol{\Sigma}$ , then we call the  $n \times p$  matrix  $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)'$ , whose  $i$ th row contains the  $i$ th sample  $p$ -vector, a normal data matrix.

A  $p \times p$  matrix  $\mathbf{A}$  that can be written  $\mathbf{A} = \mathbf{X}'\mathbf{X}$  in terms of such a normal data matrix is said to have a Wishart distribution with scale matrix  $\boldsymbol{\Sigma}$  and degrees of freedom parameter  $n$ ,  $\mathbf{A} \sim W_p(\boldsymbol{\Sigma}, n)$ . When  $p = 1$ , this reduces to a scaled chi-squared law  $\sigma^2\chi_{(n)}^2$ .

We consider analogs of the  $F$  and Beta distributions of multivariate analysis, which are based on two independent chi-squared variates. Thus, let  $\mathbf{A} \sim W_p(\boldsymbol{\Sigma}, m)$  be independent of  $\mathbf{B} \sim W_p(\boldsymbol{\Sigma}, n)$ . If  $m \geq p$ , then  $\mathbf{A}^{-1}$  exists and the nonzero eigenvalues of  $\mathbf{A}^{-1}\mathbf{B}$  are quantities of interest that generalize the univariate  $F$  ratio. We remark that the scale matrix  $\boldsymbol{\Sigma}$  has no effect on the distribution of these eigenvalues, and so, without loss of generality, we can suppose that  $\boldsymbol{\Sigma} = \mathbf{I}$ .

The matrix analog of a Beta variate is based on the eigenvalues of  $(\mathbf{A} + \mathbf{B})^{-1}\mathbf{B}$ , and leads to the following:

DEFINITION  $\theta$  [[Mardia, Kent and Bibby \(1979\)](#), page 84]. Let  $\mathbf{A} \sim W_p(\mathbf{I}, m)$  be independent of  $\mathbf{B} \sim W_p(\mathbf{I}, n)$ , where  $m \geq p$ . Then the largest eigenvalue  $\theta$  of  $(\mathbf{A} + \mathbf{B})^{-1}\mathbf{B}$  is called the *greatest root statistic* and its distribution is denoted  $\theta(p, m, n)$ .

Since  $\mathbf{A}$  is positive definite, we have  $0 < \theta < 1$ . Clearly  $\theta(p, m, n)$  can also be defined as the largest root of the determinantal equation

$$\det[\mathbf{B} - \theta(\mathbf{A} + \mathbf{B})] = 0.$$

Specific examples will be given below, but in general the parameter  $p$  refers to dimension,  $m$  to the “error” degrees of freedom and  $n$  to the “hypothesis” degrees of freedom. Thus,  $m + n$  represents the “total” degrees of freedom.

There are  $\min(n, p)$  nonzero eigenvalues of  $\mathbf{A}^{-1}\mathbf{B}$  or, equivalently,  $\min(n, p)$  nonzero roots  $\theta = (\theta_i)$  of the determinantal equation above. The joint density function of these roots is given by

$$(1) \quad p(\theta) = C \prod_{i=1}^{\min(n,p)} \theta_i^{(ln-p|-1)/2} (1 - \theta_i)^{(m-p-1)/2} \Delta(\theta),$$

where  $\Delta(\theta) = \prod_{i \neq j} |\theta_i - \theta_j|$  (see, e.g., [Muirhead \[\(1982\), page 112\]](#), or [Anderson \[\(2003\), pages 536–537\]](#)). We shall not need the explicit form of the density in this paper; it is, however, useful sometimes in matching up the various parameter choices used in different references and packages.

The greatest root distribution has the property

$$\theta(p, m, n) \stackrel{\mathcal{D}}{=} \theta(n, m + n - p, p),$$

useful, in particular, in the case when  $n < p$  [e.g. [Mardia, Kent and Bibby \(1979\)](#), page 84].

1.4. *Main result.* Empirical and theoretical investigation has shown that it is useful to develop the approximation in terms of the logit transform of  $\theta$ ; thus, we define

$$(2) \quad W(p, m, n) = \text{logit} \theta(p, m, n) = \log \left( \frac{\theta(p, m, n)}{1 - \theta(p, m, n)} \right).$$

Our main result, stated more formally below, is that with appropriate centering and scaling,  $W$  is approximately Tracy–Widom distributed:

$$(3) \quad \frac{W(p, m, n) - \mu(p, m, n)}{\sigma(p, m, n)} \stackrel{\mathcal{D}}{\Rightarrow} F_1.$$

The centering and scaling parameters are defined by

$$(4) \quad \mu(p, m, n) = 2 \log \tan \left( \frac{\phi + \gamma}{2} \right),$$

$$(5) \quad \sigma^3(p, m, n) = \frac{16}{(m + n - 1)^2} \frac{1}{\sin^2(\phi + \gamma) \sin \phi \sin \gamma},$$

where the angle parameters  $\gamma, \phi$  are defined by

$$\sin^2 \left( \frac{\gamma}{2} \right) = \frac{\min(p, n) - 1/2}{m + n - 1},$$

$$\sin^2 \left( \frac{\phi}{2} \right) = \frac{\max(p, n) - 1/2}{m + n - 1}.$$

1.5. *More on the Tracy–Widom law.* The  $F_1$  distribution, due to [Tracy and Widom \(1996\)](#) and plotted in [Figure 1](#), has its origins in mathematical physics—see [Tracy and Widom \(1996\)](#); [Johnstone \(2001\)](#) for further details. The density is asymmetric, with mean  $\doteq -1.21$  and SD  $\doteq 1.27$ . Both tails have exponential decay, the left tail like  $e^{-|s|^3/24}$  and the right tail like  $e^{-(2/3)s^{3/2}}$ .

For the present paper, what is important is that the  $F_1$  distribution does not depend on any parameters, and the distribution itself, along with its inverse and percentiles, can be tabulated as univariate special functions. These functions play the same role in this paper as the standard normal distribution  $\Phi$ , its inverse  $\Phi^{-1}$  and percentiles  $z_\alpha$  play in traditional statistical application.

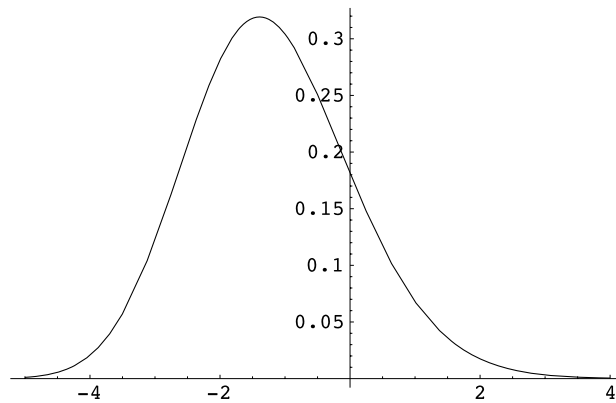


FIG. 1. Density of the Tracy–Widom distribution  $F_1$ .

*Software.* An R package `RMTstat` is available at CRAN ([cran.r-project.org](http://cran.r-project.org)). It facilitates computation of the distributional approximations and largest root tests described in this paper, and the use of percentiles and random draws from the  $F_1$  distribution. Its scope and use is described in more detail in an accompanying report [Johnstone et al. \(2010\)](#). A parallel MATLAB package is in development; it will also contain code to reproduce the figures and table in this paper.

*Percentiles.* Let  $f_\alpha$  denote the  $\alpha$ th percentile of  $F_1$ . For example,

$$f_{0.90} = 0.4501, \quad f_{0.95} = 0.9793, \quad f_{0.99} = 2.0234.$$

Then the  $\alpha$ th percentile of  $\theta(p, m, n)$  is given approximately by

$$(6) \quad \theta_\alpha = e^{\mu + f_\alpha \sigma} / (1 + e^{\mu + f_\alpha \sigma}),$$

where  $\mu = \mu(p, m, n)$ ,  $\sigma = \sigma(p, m, n)$  are given by (4) and (5).

The more formal statement of (3) goes as follows. Assume  $p, m$  and  $n \rightarrow \infty$  together in such a way that

$$(7) \quad \lim \frac{p \wedge n}{m + n} > 0, \quad \lim \frac{m}{p} > 1.$$

For each  $s_0 \in \mathbb{R}$ , there exist  $c, C > 0$  such that for  $s \geq s_0$ ,

$$|P\{W(p, m, n) \leq \mu(p, m, n) + \sigma(p, m, n)s\} - F_1(s)| \leq Cp^{-2/3}e^{-cs}.$$

For the full proof and much more discussion and detail, see the companion paper [[Johnstone \(2008\)](#)].

*REMARKS. Smallest eigenvalue.* If  $\mathbf{A}$  and  $\mathbf{B}$  are as in the definition of  $\theta(p, m, n)$ , then let  $\tilde{\theta}(p, m, n)$  denote the *smallest* eigenvalue of  $(\mathbf{A} + \mathbf{B})^{-1}\mathbf{B}$ . Its distribution is given by

$$\tilde{\theta}(p, m, n) \stackrel{D}{=} 1 - \theta(p, n, m),$$

(note the reversal of  $m$  and  $n$ !). In particular, the Tracy–Widom distribution will give a generally useful approximation to the *lower* tail of  $\tilde{\theta}(p, m, n)$ .

*Complex-valued data.* There is an entirely analagous result when  $\mathbf{A}$  and  $\mathbf{B}$  follow complex Wishart distributions, with a modified limit distribution  $F_2$ . Details are given in Johnstone (2008).

**2. Quality of approximation.**

2.1. *Comparison with percentiles.* There is a substantial literature computing percentage points of the greatest root distribution for selected parameter values, partially reviewed below. The standard parameterization used in these tables arises from writing the joint density of the roots  $\theta_i$  as

$$p(\theta) = C \prod_{i=1}^s \theta_i^m (1 - \theta_i)^n \Delta(\theta).$$

From this and (1) it is apparent that our “MKB” parameters  $(p, m, n)$  are related to the “Table” parameters  $(s, m, n)$  via

$$(8) \quad \begin{aligned} s &= \min(n, p), & p &= s, \\ m &= (|n - p| - 1)/2, & m &= s + 2n + 1, \\ n &= (m - p - 1)/2, & n &= s + 2m + 1. \end{aligned}$$

In terms of the table parameters and  $N = 2(s + m + n) + 1$ , the centering and scaling constants of the Tracy–Widom approximation are given by

$$\sin^2\left(\frac{\gamma}{2}\right) = \left(s - \frac{1}{2}\right)/N, \quad \sin^2\left(\frac{\phi}{2}\right) = \left(s + 2m + \frac{1}{2}\right)/N$$

and

$$(9) \quad \mu = 2 \log \tan\left(\frac{\phi + \gamma}{2}\right), \quad \sigma^3 = \frac{16}{N^2} \frac{1}{\sin^2(\phi + \gamma) \sin \phi \sin \gamma}.$$

We turn to the comparison of percentage points  $\theta_\alpha^{TW}$  from the Tracy–Widom approximation (6) with the exact values  $\theta_\alpha$  for small values of the table parameters  $(s, m, n)$ . The most extensive tabulations of  $\theta_\alpha(s, m, n)$  have been made by William Chen; he has graciously provided the author with the complete version of the tables excerpted in Chen (2002, 2003, 2004a, 2004b).

Figures 2 and 3 plot  $\theta_\alpha^{TW}$  against  $\theta_\alpha$  at 95th and 90th percentiles for  $s = 2$ . This is the smallest relevant value of  $s$ —otherwise we are in the univariate case covered by  $F$  distributions. The bottom panels, in particular, focus on the *relative error*

$$r = (\theta_\alpha^{TW}/\theta_\alpha) - 1.$$

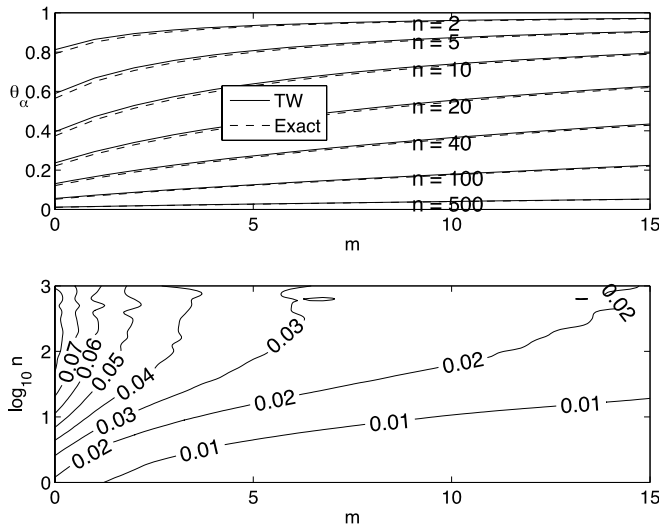


FIG. 2. Comparison of exact and approximate 95th percentiles for  $s = 2$ . Top panel: solid line is the Tracy–Widom approximation  $\theta_{\alpha}^{\text{TW}}(2, m, n)$  plotted as a function of  $m$  for values of  $n$  shown. Dashed lines are the exact percentiles  $\theta_{\alpha}(2, m, n)$  from Chen’s tables. Bottom panel: Contour plots of relative error  $r = (\theta_{\alpha}^{\text{TW}}/\theta_{\alpha}) - 1$ . Horizontal axis is  $m$ , vertical axis is  $\log_{10} n$ , thus covering the range from  $n = 1$  to 1000.

Figure 2 shows that even for  $s = 2$ , the 95th percentile of the TW approximation has a relative error of less than 1 in 20 except in the zone where both  $m \leq 2$  and  $n \geq 10$ , where the relative error is still less than 1 in 10. Note that the relative error is always positive in sign, implying that the approximate critical points yield a conservative test. More extensive contour plots covering  $s = 2(1)6$  and 90th, 95th and 99th percentiles may be found in [Johnstone and Chen \(2007\)](#).

*Work on tables.* There has been a large amount of work to prepare tables or charts for the null distribution of the largest root, much of which is reviewed in [Chen \(2003\)](#). We mention contributions by the following: [Nanda \(1948, 1951\)](#); [Foster and Rees \(1957\)](#); [Foster \(1957, 1958\)](#); [Pillai \(1955, 1956a, 1956b, 1957, 1965, 1967\)](#); [Pillai and Bantegui \(1959\)](#); [Heck \(1960\)](#); [Krishnaiah \(1980\)](#); [Pillai and Flury \(1984\)](#); [Chen \(2002, 2003, 2004a, 2004b\)](#).

Because of the dependence on the three parameters, these tables can run up to 25 pages in typical textbooks, such as those of [Johnson and Wichern \(2002\)](#) and [Morrison \(2005\)](#).

*Code.* [Constantine \(1963\)](#) expresses the c.d.f. of the largest root distribution in terms of a matrix hypergeometric function. [Koev and Edelman \(2006\)](#) have developed efficient algorithms (and a MATLAB package available at <http://www-math.mit.edu/~plamen>) for the evaluation of such matrix hypergeometric functions using recursion formulas from group representation theory.

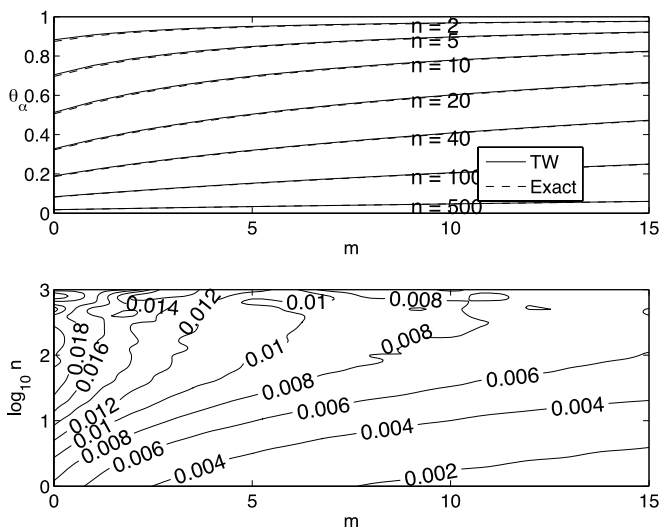


FIG. 3. Comparison of exact and approximate 90th percentiles for  $s = 4$ . Top panel: solid line is the Tracy–Widom approximation  $\theta_\alpha^{\text{TW}}(4, m, n)$  plotted as a function of  $m$  for values of  $n$  shown. Dashed lines are the exact percentiles  $\theta_\alpha(4, m, n)$  from Chen’s tables. Bottom panel: Contour plots of relative error  $r = (\theta_\alpha^{\text{TW}} / \theta_\alpha) - 1$ . Horizontal axis is  $m$ , vertical axis is  $\log_{10} n$ , thus covering the range from  $n = 1$  to 1000.

Koev (2010) collects useful formulas and explains how to use them and `mhg` to compute the exact c.d.f. and percentiles for the largest root distribution over a range of values of the “MKB” parameters corresponding to  $m, n, p \leq 17$ , and  $m, n, p \leq 40$  when  $n - p$  is odd.

SAS/STAT 9.0 made available an option for computing exact  $p$ -values using Davis (1972); Pillai and Flury (1984). There is also some stand-alone software described by Lutz (1992, 2000).

2.2. Accuracy of  $p$ -values. The univariate  $F$  bound. We recall the hypothesis that  $\mathbf{A} \sim W_p(\mathbf{I}, m)$  be distributed independently of  $\mathbf{B} \sim W_p(\mathbf{I}, n)$ , and the characterization of the largest eigenvalue given by

$$(10) \quad \lambda_{\max}(\mathbf{A}^{-1}\mathbf{B}) = \max_{|\mathbf{u}|=1} \frac{\mathbf{u}'\mathbf{B}\mathbf{u}}{\mathbf{u}'\mathbf{A}\mathbf{u}}.$$

For fixed  $\mathbf{u}$  of unit length, the numerator and denominator are distributed as independent  $\chi_{(n)}^2$  and  $\chi_{(m)}^2$  respectively, and so, again for fixed  $\mathbf{u}$ , the ratio has an  $F_{n,m}$  distribution. Consequently, we have the simple bound

$$\frac{m}{n} \lambda_{\max}(\mathbf{A}^{-1}\mathbf{B}) > F \sim F_{n,m}.$$

Using the  $F_{n,m}$  distribution in place of the actual greatest root law yields a lower bound for the significance level, or  $p$ -value. We shall see that this bound can be



anti-conservative by several orders of magnitude, leading to overstatements of the empirical evidence against the null hypothesis. And furthermore, one can expect that the higher the dimension  $p$  of the search space in (10), the worse the bound provided by the  $F$  distribution.

The default  $p$ -value provided in both SAS and R (through package `car`) uses this unsatisfactory distribution bound.

Table 1 attempts to capture a variety of scenarios within the computational range of Koev's software.

Column `Exact` shows a range of significance levels  $\alpha$  covering several orders of magnitude. Column `Largest Root` shows the corresponding quantiles  $\theta_\alpha$  of the largest root distribution, for the given values of  $(s, m, n)$ —these are computed using Koev's MATLAB routine `qmaxeigjacob`. Thus, an observed value of  $\theta(s, m, n) = \theta_\alpha$  would correspond to an exact  $p$ -value  $\alpha$ .

The remaining columns compare the Tracy–Widom approximation and the  $F$  bound. The  $p$ -value obtained from the Tracy–Widom approximation is given by

$$(11) \quad P_{\text{TW}}(\theta_\alpha) = 1 - F_1((\text{logit}(\theta_\alpha) - \mu)/\sigma),$$

where  $\mu$  and  $\sigma$  are computed from (9).

The  $F$  bound on the  $p$ -value is given by

$$(12) \quad P(\theta(s, m, n) > \theta_\alpha) > P_F(\theta_\alpha) = 1 - F_{\nu_1, \nu_2}(\nu_2 \theta_\alpha / (\nu_1 (1 - \theta_\alpha))),$$

where  $\nu_1 = s + 2m + 1$  and  $\nu_2 = s + 2n + 1$  denote the hypothesis and error degrees of freedom respectively.

The two tables consider  $s = 2$  and 6 variables respectively. The values of  $m = -0.5$  and 5 correspond to  $s$  and  $s + 11$  hypothesis degrees of freedom, while the values of  $n = 2$  and 10 translate to  $s + 5$  and  $s + 21$  error degrees of freedom respectively.

At the 10% and 5% levels, the Tracy–Widom approximation is within 20% of the true  $p$ -value at  $s = 6$ , and within 35% of truth at  $s = 2$ . The  $F$ -value is wrong by a factor of four or more at  $s = 2$ , and by three orders of magnitude at  $s = 6$ . At smaller significance levels, the Tracy–Widom approximation generally stays within one order of magnitude of the correct  $p$ -value—except at  $(s, m, n) = (2, -0.5, 10)$ . The  $F$  approximation is off by many orders of magnitude when  $s = 6$ .

In addition, we note that the Tracy–Widom approximation is conservative in nearly all cases, the exception being for  $\theta \geq 0.985$  in the case  $(s, m, n) = (6, -0.5, 2)$ . In contrast, the  $F$  approximation is *always* [cf. (12)] anti-conservative, often badly so.

In applications one is often concerned only with the general order of magnitude of the  $p$ -values associated with tests of the various hypotheses that are entertained—not least because the assumptions of the underlying model are at best approximately true. For this purpose, then, it may be argued that the TW approximate  $p$ -value is often quite adequate over the range of  $(s, m, n)$  values. Of course,

TABLE 1  
 Comparison of the Tracy–Widom approximation and  $F$  bound for cases with  $s = 2$  and  $s = 6$  variables

Largest root	Exact	Tracy–Widom	F	Largest root	Exact	Tracy–Widom	F
$s = 2, m = -0.5, n = 2$				$s = 6, m = -0.5, n = 2$			
0.663	0.1	0.119	0.0223	0.918	0.1	0.115	2.23e-005
0.737	0.05	0.066	0.00933	0.938	0.05	0.0598	4.99e-006
0.850	0.01	0.0169	0.00131	0.966	0.01	0.0116	1.92e-007
0.881	0.005	0.00927	0.000573	0.973	0.005	0.00545	4.96e-008
0.931	0.001	0.00222	8.49e-005	0.985	0.001	0.000839	2.3e-009
0.968	0.0001	0.000251	5.65e-006	0.993	0.0001	4.35e-005	3.1e-011
0.985	1e-005	2.38e-005	3.81e-007	0.997	1e-005	1.64e-006	4.38e-013
0.993	1e-006	1.89e-006	2.58e-008	0.999	1e-006	NaN	6.33e-015
$s = 2, m = -0.5, n = 10$				$s = 6, m = -0.5, n = 10$			
0.268	0.1	0.117	0.0278	0.597	0.1	0.11	0.000206
0.318	0.05	0.0669	0.0123	0.633	0.05	0.0577	6.49e-005
0.418	0.01	0.0214	0.00199	0.698	0.01	0.0134	5.46e-006
0.456	0.005	0.0137	0.000919	0.721	0.005	0.00722	1.99e-006
0.533	0.001	0.00522	0.000157	0.766	0.001	0.00172	2.05e-007
0.624	0.0001	0.00146	1.31e-005	0.816	0.0001	0.000223	8.97e-009
0.696	1e-005	0.000443	1.11e-006	0.854	1e-005	2.86e-005	4.29e-010
0.755	1e-006	0.000141	9.59e-008	0.884	1e-006	3.57e-006	2.17e-011
$s = 2, m = 5, n = 10$				$s = 6, m = 5, n = 10$			
0.592	0.1	0.112	0.0234	0.757	0.1	0.108	0.000117
0.629	0.05	0.0602	0.0103	0.781	0.05	0.0557	3.63e-005
0.697	0.01	0.0149	0.00164	0.823	0.01	0.0119	2.99e-006
0.721	0.005	0.00827	0.000758	0.837	0.005	0.00606	1.08e-006
0.767	0.001	0.00215	0.000129	0.864	0.001	0.00125	1.1e-007
0.817	0.0001	0.000318	1.07e-005	0.894	0.0001	0.000125	4.75e-009
0.855	1e-005	4.71e-005	9.04e-007	0.917	1e-005	1.17e-005	2.25e-010
0.885	1e-006	6.88e-006	7.79e-008	0.934	1e-006	1.03e-006	1.13e-011

if  $(s, m, n)$  is not too large and greater precision is required, then exact  $p$ -values can be sought, using, for example, SAS or Koev’s software.

**3. Testing for independence of two sets of variables.** Let  $\mathbf{x}_1, \dots, \mathbf{x}_n$  be a random sample from  $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ . Partition the variables into two sets with dimensions  $p_1$  and  $p_2$  respectively,  $p_1 + p_2 = p$ . Suppose that  $\boldsymbol{\Sigma}$  and the sample covariance matrix  $\mathbf{S}$  are partitioned correspondingly. We consider testing the null hypothesis of independence of the two sets of variables:  $\boldsymbol{\Sigma}_{12} = \mathbf{0}$ . The union-intersection test is based on the largest eigenvalue  $\lambda_1$  of  $\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}$  (Mardia, Kent and Bibby (1979), page 136) and under  $H_0$ ,  $\lambda_1$  has the greatest root distribution  $\theta(p_2, n - 1 - p_1, p_1)$ . Mardia, Kent and Bibby (1979) consider an example test of independence of  $n = 25$  head length and breadth measurements between first sons and second sons, so that  $p_1 = p_2 = 2$ . The observed value  $\lambda_1^{\text{obs}} = 0.622$  exceeds the critical value  $\theta_{0.05} = 0.330$  found by interpolation from the tables. The Tracy–Widom approximation  $\theta_{0.05}^{\text{TW}} = 0.356$  is found from (6) and serves equally well for rejection of  $H_0$  in this case.

**4. Canonical correlation analysis.** Again we have two sets of variables, an  $\mathbf{x}$ -set with  $q$  variables and a  $\mathbf{y}$ -set with  $p$  variables. The goal is to find maxi-

mally correlated linear combinations  $\eta = \mathbf{a}'\mathbf{x}$  and  $\phi = \mathbf{b}'\mathbf{y}$ . We suppose that  $(\mathbf{X}, \mathbf{Y})$  is a data matrix of  $n$  samples (rows) on  $q + p$  variables (columns) such that each row is an independent draw from  $N_{p+q}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ . Again let  $\mathbf{S}$  be the sample covariance matrix, assumed partitioned  $\mathbf{S} = \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{pmatrix}$ . The sample squared canonical correlations  $(r_i^2)$  for  $i = 1, \dots, k = \min(p, q)$  are found as the eigenvalues of  $\mathbf{M}_2 = \mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}$  [Mardia, Kent and Bibby (1979), Sections 10.2.1 and 10.2.2]. The population squared canonical correlations  $\rho_i^2$  are, in turn, the eigenvalues of  $\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$ . In both cases, we assume that the correlations are arranged in decreasing order. The test of the null hypothesis of zero correlation,  $H_0: \rho_1 = \dots = \rho_k = 0$ , is based on the largest eigenvalue  $r_1^2$  of  $\mathbf{M}_2$ . Under  $H_0$ , it is known that  $r_1^2$  has the  $\theta(p, n - q - 1, q)$  distribution, so that the Tracy–Widom approximation can be applied.

*Nonnull cases—a conservative test.* Often it may be apparent that the first  $k$  canonical correlations are nonzero and the main interest focuses on the significance of  $r_{k+1}^2, r_{k+2}^2$ , etc. We let  $H_s$  denote the null hypothesis that  $\rho_{s+1} = \dots = \rho_p = 0$ , and write  $\mathcal{L}(r_k|p, q, n; \boldsymbol{\Sigma})$  for the distribution of the  $r$ th c.c. under population covariance matrix  $\boldsymbol{\Sigma}$ . When the covariance matrix  $\boldsymbol{\Sigma} \in H_s$ , the  $(s + 1)$ st canonical correlation is stochastically smaller than the *largest* canonical correlation in a related null model:

LEMMA 1. *If  $\boldsymbol{\Sigma} \in H_s$ , then*

$$\mathcal{L}(r_{s+1}|p, q, n; \boldsymbol{\Sigma}) \stackrel{st}{<} \mathcal{L}(r_1|p, q - s, n; \mathbf{I}).$$

This nonasymptotic result follows from interlacing properties of the singular value decomposition (Appendix). Since  $\mathcal{L}(r_1^2|p, q - s, n; \mathbf{I})$  is given by the null distribution  $\theta(p, n + s - q - 1, q - s)$ , we may use the latter to provide a conservative  $p$ -value for testing  $H_s$ . In turn, the  $p$ -value for  $\theta(p, n + s - q - 1, q - s)$  can be numerically approximated as in (6) using the Tracy–Widom distribution.

*Example.* Waugh (1942) gave perhaps the first significant illustration of CCA using data on  $n = 138$  samples of Canadian Hard Red Spring wheat and the flour made from each of these samples. The aim was to seek highly correlated indices  $\mathbf{a}'\mathbf{x}$  of wheat quality and  $\mathbf{b}'\mathbf{y}$  of flour quality, since a well correlated grading of raw (wheat) and finished (flour) products was believed to promote fair pricing of each. In all,  $q = 5$  wheat characteristics—kernel texture, test weight, damaged kernels, foreign material, crude protein in wheat—and  $p = 4$  flour characteristics—wheat per bushel of flour, ash in flour, crude protein in flour, gluten quality index—were measured. The resulting squared canonical correlations were  $(r_1^2, r_2^2, r_3^2, r_4^2) = (0.923, 0.554, 0.056, 0.008)$ . The leading correlation

would seem clearly significant and, indeed, from our approximate formula (6),  $\theta_{0.99}^{\text{TW}} = 0.184$ .

To assess the second correlation  $r_2$ , we appeal to the conservative test discussed above based on the null distribution with  $q - 1 = 4$ ,  $p = 4$  and  $n = 138$ . The Tracy–Widom approximation  $\theta_{0.99}^{\text{TW}} \approx \mu + 2.023\sigma \doteq 0.166 \ll 0.554$ , which strongly suggests that this second correlation is significant as well.

Marginal histograms naturally reveal some departures from symmetric Gaussian tails, but they do not seem extreme enough to invalidate the conclusions, which are also confirmed by permutation tests.

## 5. Tests of common means or variances.

5.1. *Equality of means for common covariance.* Suppose that we have  $k$  populations with independent data matrices  $\mathbf{X}_i$  consisting of  $n_i$  observations drawn from an  $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$  and put  $n = \sum n_i$ . This is the one-way multivariate analysis of variance illustrated in Example 1.1. For testing the null hypothesis of equality of means  $H_0: \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_k$ , we form, for each population, the sample mean  $\bar{\mathbf{x}}_i$  and covariance matrix  $\mathbf{S}_i$ , normalized so that  $n_i \mathbf{S}_i \sim W_p(\boldsymbol{\Sigma}, n_i - 1)$ . The basic quantities are the within groups sum of squares  $\mathbf{W} = \sum n_i \mathbf{S}_i \sim W_p(\boldsymbol{\Sigma}, n - k)$  and the between group sum of squares  $\mathbf{B} = \sum n_i (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})' \sim W_p(\boldsymbol{\Sigma}, k - 1)$  under  $H_0$ , independently of  $\mathbf{W}$ . The union-intersection test of  $H_0$  uses the largest root of  $\mathbf{W}^{-1}\mathbf{B}$  or, equivalently, that of  $(\mathbf{W} + \mathbf{B})^{-1}\mathbf{B}$ , and the latter has, under  $H_0$ , the  $\theta(p, n - k, k - 1)$  distribution.

5.2. *Equality of covariance matrices.* Suppose that independent samples from two normal distributions  $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$  and  $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$  lead to covariance estimates  $\mathbf{S}_i$  which are independent and Wishart distributed on  $n_i$  degrees of freedom:  $n_i \mathbf{S}_i \sim W_p(n_i, \boldsymbol{\Sigma}_i)$  for  $i = 1, 2$ . Then the largest root test of the null hypothesis  $H_0: \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$  is based on the largest eigenvalue  $\theta$  of  $(n_1 \mathbf{S}_1 + n_2 \mathbf{S}_2)^{-1} n_2 \mathbf{S}_2$ , which under  $H_0$  has the  $\theta(p, n_1, n_2)$  distribution Muirhead (1982), page 332.

6. **Multivariate linear model.** The multivariate linear model blends ideas well known from the univariate setting with new elements introduced by correlated multiple responses. In view of the breadth of models covered, and the variety of notation in the literature and in the software, we review the setting in a little more detail, beginning with the familiar model for a single response

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}.$$

Here  $\mathbf{y}$  is an  $n \times 1$  column vector of observations on a response variable,  $\mathbf{X}$  is an  $n \times q$  model matrix, and  $\mathbf{u}$  is an  $n \times 1$  column vector of errors, assumed here to be independent and identically distributed as  $N(0, \sigma^2)$ . The  $q \times 1$  vector  $\boldsymbol{\beta}$  of

unknown parameters has the least squares estimate—when  $\mathbf{X}$  has full rank—given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

The error sum of squares  $SS_E = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{y}'\mathbf{P}\mathbf{y}$ , where  $\mathbf{P}$  denotes orthogonal projection onto the subspace orthogonal to the columns of  $\mathbf{X}$ , it has rank  $n - q$ , and so  $SS_E \sim \chi_{(n-q)}^2$ .

Consider the linear hypothesis  $H_0 : \mathbf{C}_1\boldsymbol{\beta} = 0$ , where  $\mathbf{C}_1$  is a  $g \times q$  matrix of rank  $g$ . In the simplest example,  $\mathbf{C}_1 = [\mathbf{I}_g \ \mathbf{0}]$  extracts the first  $g$  elements of  $\boldsymbol{\beta}$ ; more generally, the rows of  $\mathbf{C}_1$  are often contrasts among the components of  $\boldsymbol{\beta}$ . To describe the standard  $F$ -test of  $H_0$ , let  $\mathbf{C}_2$  be any  $(q - g) \times q$  matrix such that  $\mathbf{C} = \begin{pmatrix} \mathbf{C}_1 \\ \mathbf{C}_2 \end{pmatrix}$  becomes an invertible  $q \times q$  matrix. We may then write

$$\mathbf{X}\boldsymbol{\beta} = [\mathbf{X}\mathbf{C}^{(1)} \quad \mathbf{X}\mathbf{C}^{(2)}] \begin{pmatrix} \mathbf{C}_1\boldsymbol{\beta} \\ \mathbf{C}_2\boldsymbol{\beta} \end{pmatrix},$$

where we have partitioned  $\mathbf{C}^{-1} = [\mathbf{C}^{(1)} \quad \mathbf{C}^{(2)}]$  into blocks with  $g$  and  $q - g$  columns respectively.

Let  $\mathbf{P}_1$  denote the orthogonal projection onto the subspace orthogonal to the columns of  $\mathbf{X}\mathbf{C}^{(2)}$ . We have the sum of squares decomposition

$$\mathbf{y}'\mathbf{P}_1\mathbf{y} = \mathbf{y}'\mathbf{P}\mathbf{y} + \mathbf{y}'(\mathbf{P}_1 - \mathbf{P})\mathbf{y}$$

and the hypothesis sum of squares for testing  $H_0 : \mathbf{C}_1\boldsymbol{\beta} = 0$  is given by  $SS_H = \mathbf{y}'\mathbf{P}_2\mathbf{y}$ , with  $\mathbf{P}_2 = \mathbf{P}_1 - \mathbf{P}$ . The projection  $\mathbf{P}_2$  has rank  $g$  and so under  $H_0$ ,  $SS_H \sim \chi_{(g)}^2$ . The projections  $\mathbf{P}$  and  $\mathbf{P}_2$  are orthogonal and so the sums of squares have independent chi-squared distributions, and under  $H_0$  the traditional  $F$ -statistic

$$F = \frac{SS_H/g}{SS_E/(n - q)} \sim F_{g,n-q}.$$

Explicit expressions for the sums of squares are given by

$$SS_E = \mathbf{y}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y},$$

$$SS_H = (\mathbf{C}_1\hat{\boldsymbol{\beta}})'[\mathbf{C}_1(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}_1']^{-1}\mathbf{C}_1\hat{\boldsymbol{\beta}}.$$

In the multivariate linear model,

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{U},$$

the single response  $\mathbf{y}$  is replaced by  $p$  response vectors, organized as columns of the  $n \times p$  matrix  $\mathbf{Y}$ . The model (or design) matrix  $\mathbf{X}$  remains the same for each response; however, there are separate vectors of unknown coefficients and errors for each response; these are organized into a  $q \times p$  matrix  $\mathbf{B}$  of regression coefficients and an  $n \times p$  matrix  $\mathbf{E}$  of errors. The multivariate aspect of the model is the assumption that the rows of  $\mathbf{U}$  are independent, with multivariate normal distribution

having mean  $\mathbf{0}$  and common covariance matrix  $\Sigma$ . Thus,  $\mathbf{U}$  is a normal data matrix of  $n$  samples from  $N_p(\mathbf{0}, \Sigma)$ . Assuming for now that the model matrix  $\mathbf{X}$  has full rank, the least squares estimator

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

The linear hypothesis becomes

$$H_0: \mathbf{C}_1\mathbf{B} = \mathbf{0}.$$

The sums of squares of the univariate case are replaced by hypothesis and error sums of squares and products matrices:

$$(13) \quad \begin{aligned} E &= \mathbf{Y}'\mathbf{P}\mathbf{Y} = \mathbf{Y}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y}, \\ H &= \mathbf{Y}'\mathbf{P}_2\mathbf{Y} = (\mathbf{C}_1\hat{\mathbf{B}})'[\mathbf{C}_1(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}_1']^{-1}\mathbf{C}_1\hat{\mathbf{B}}, \end{aligned}$$

in which the univariate vectors  $\mathbf{y}$  and  $\hat{\boldsymbol{\beta}}$  are simply replaced by their multivariate analogs  $\mathbf{Y}$  and  $\hat{\mathbf{B}}$ . It follows that  $\mathbf{E} \sim W_p(\Sigma, n - q)$  and that under  $H_0$ ,  $\mathbf{H} \sim W_p(\Sigma, g)$ ; furthermore,  $\mathbf{E}$  and  $\mathbf{H}$  are independent. Generalizations of the  $F$ -test are obtained from the eigenvalues  $(\lambda_i)$  of the matrix  $\mathbf{E}^{-1}\mathbf{H}$  or, equivalently, the eigenvalues  $\theta_i$  of  $(\mathbf{H} + \mathbf{E})^{-1}\mathbf{H}$ .

Thus, under the null hypothesis  $\mathbf{C}_1\mathbf{B} = \mathbf{0}$ , Roy's maximum root statistic  $\theta_1$  has null distribution

$$(14) \quad \begin{aligned} \theta_1 &\sim \theta(p, n - q, g), & \text{where} \\ p &= \text{dimension}, & g = \text{rank}(\mathbf{C}_1), \\ q &= \text{rank}(\mathbf{X}), & n = \text{sample size.} \end{aligned}$$

*Two extensions.* (a) *X not of full rank.* This situation routinely occurs when redundant parameterizations are used, for example, when dealing with factors in analysis of variance models. One approach (e.g., MKB, Section 6.4) is to rearrange the columns of  $\mathbf{X}$  and partition  $\mathbf{X} = [\mathbf{X}_1 \mathbf{X}_2]$  so that  $\mathbf{X}_1$  has full rank. We must also assume that the matrix  $\mathbf{C}_1$  is *testable* in the sense that, as a function of  $\mathbf{B}$ ,  $\mathbf{X}\mathbf{B} = \mathbf{0}$  implies  $\mathbf{C}_1\mathbf{B} = \mathbf{0}$ . In such cases, if we partition  $\mathbf{C}_1 = [\mathbf{C}_{11} \mathbf{C}_{12}]$  conformally with  $\mathbf{X}$ , then  $\mathbf{C}_{12} = \mathbf{C}_{11}(\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1\mathbf{X}_2$  is determined from  $\mathbf{C}_{11}$ .

With these assumptions, we use  $\mathbf{X}_1$  and  $\mathbf{C}_{11}$  in (13) and (14) in place of  $\mathbf{X}$  and  $\mathbf{C}_1$ .

(b) *Intra-subject hypotheses.* A straightforward extension is possible in order to test null hypotheses of the form

$$\mathbf{C}_1\mathbf{B}\mathbf{M}_1 = \mathbf{0},$$

where  $\mathbf{M}_1$  is  $p \times r$  of rank  $r$ . The columns of  $\mathbf{M}_1$  capture particular linear combinations of the dependent variables—for an example, see, e.g., Morrison (2005), Chapter 3.6.

We simply consider a modified linear model

$$\mathbf{Y}\mathbf{M}_1 = \mathbf{X}\mathbf{B}\mathbf{M}_1 + \mathbf{U}\mathbf{M}_1.$$

An important point is that the rows of  $\mathbf{U}\mathbf{M}_1$  are still independent, now distributed as  $N_p(\mathbf{0}, \mathbf{M}_1' \boldsymbol{\Sigma} \mathbf{M}_1)$ . So we may simply apply the above analysis, replacing  $\mathbf{Y}$ ,  $\mathbf{U}$  and  $\mathbf{B}$  by  $\mathbf{Y}\mathbf{M}_1$ ,  $\mathbf{U}\mathbf{M}_1$  and  $\mathbf{B}\mathbf{M}_1$  respectively. In particular, the greatest root statistic now has null distribution given by

$$\theta_1 \sim \theta(r, n - q, g).$$

*Linear hypotheses in SAS.* Analyses involving the four multivariate tests are provided in a number of SAS routines, such as GLM and CANCORR. The parameterization used here can be translated into that used in SAS by means of the documentation given in the SAS/STAT Users Guide—we refer to the section on Multivariate Tests in version 9.1, page 48 ff. The linear hypotheses correspond to MKB notation via

MKB	SAS
$\mathbf{C}_1$	$\mathbf{L}$
$\mathbf{B}$	$\boldsymbol{\beta}$
$\mathbf{M}_1$	$\mathbf{M}$

while the parameters of the greatest root distribution are given by

	MKB		SAS	
dimension	$r$	rank( $\mathbf{M}_1$ )	rank( $\mathbf{M}$ )	$p$
hypothesis	$g$	rank( $\mathbf{C}_1$ )	rank( $\mathbf{L}$ )	$q$
error	$n - q$			$v$

(Note: we use sans serif font for the SAS parameters!) Finally, the SAS printouts use the following parameters:

$$\begin{aligned} s &= p \wedge q, \\ m &= (|p - q| - 1)/2, \\ n &= (v - p - 1)/2. \end{aligned}$$

**7. Concluding discussion.** We have described the Tracy–Widom approximation to the null distribution for the largest root test for a variety of classical multivariate procedures. These procedures exhibit varying degrees of sensitivity to the assumption of normality, independence etc. Documenting the sensitivity/robustness of the T–W approximation is clearly an important issue for further work. Two brief remarks can be made. In the corresponding single Wishart setting [e.g., [Johnstone \(2001\)](#)], the largest eigenvalue can be shown, under the null distribution, to still have the T–W limit if the original data have “light tails” (i.e., sub-Gaussian) [see [Soshnikov \(2002\)](#); [Péché \(2009\)](#)]. In the double Wishart settings,

simulations for canonical correlation analysis with  $n = 100$  samples on  $q = 20$  and  $p = 10$  variables, each following i.i.d.  $t_{(5)}$  or i.i.d. random sign distributions, showed that the T–W distribution for the leading correlation  $r_1^2$  still holds in the central 99% of the distribution.

#### APPENDIX: PROOF OF LEMMA

If  $\Sigma \in H_s$ , there are at most  $s$  nonzero canonical correlations, and we may suppose without loss of generality that the  $q$   $\mathbf{x}$ -variables have been transformed so that only the last  $s$  of them have any correlation with  $\mathbf{Y}$ . We employ the singular value decomposition (SVD) description of CCA, cf. Golub and Van Loan (1996), Section 12.4.3. Using QR decompositions, write

$$\mathbf{X} = Q_X R_X, \quad \mathbf{Y} = Q_Y R_Y.$$

Let  $C = Q_X^T Q_Y$  and form the SVD  $C = U R V^T$ . Then the diagonal elements  $r_1 \geq r_2 \geq \dots \geq r_{\min(p,q)}$  of  $R$  contain the sample canonical correlations.

Now consider the reduced  $n \times (q - s)$  matrix  $\mathbf{X}^-$  obtained by dropping the last  $s$  columns from  $\mathbf{X}$ . Form the QR decomposition  $\mathbf{X}^- = Q_{X^-} R_{X^-}$ . From the nature of the decomposition, we have  $Q_X = [Q_{X^-} \ Q^+]$ , that is,  $Q_{X^-}$  represents the first  $q - s$  columns of  $Q_X$ . Consequently,  $C_- = Q_{X^-}^T Q_Y$  forms the first  $q - s$  rows of  $C$ . Our lemma now follows from the interlacing property of singular values [e.g., Golub and Van Loan (1996), Corollary 8.6.3].

$$\sigma_{s+1}(C) \leq \sigma_1(C_-).$$

Indeed, our earlier discussion implies that  $\mathbf{X}^-$  and  $\mathbf{Y}$  are independent, and so  $\sigma_1(C_-)$  has the null distribution  $\mathcal{L}(r_1 | p, q - s, n; \mathbf{I})$ .

**Acknowledgment.** William Chen graciously provided an electronic copy of his tables for the distribution of the largest root.

#### REFERENCES

- ANDERSON, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Wiley, Hoboken, NJ. MR1990662
- ANDREWS, D. F. and HERZBERG, A. M. (1985). *Data*. Springer, New York.
- CHEN, W. W. (2002). Some new tables of the largest root of a matrix in multivariate analysis: A computer approach from 2 to 6. Presented at the 2002 American Statistical Association.
- CHEN, W. W. (2003). Table for upper percentage points of the largest root of a determinantal equation with five roots. *InterStat* (5). Available at [interstat.statjournals.net](http://interstat.statjournals.net).
- CHEN, W. W. (2004a). The new table for upper percentage points of the largest root of a determinantal equation with seven roots. *InterStat* (1). Available at [interstat.statjournals.net](http://interstat.statjournals.net).
- CHEN, W. W. (2004b). Some new tables for the upper probability points of the largest root of a determinantal equation with seven and eight roots. In *Special Studies in Federal Tax Statistics. Statistics of Income Division, Internal Revenue Service* (J. Dalton and B. Kilss, eds.) 113–116.
- CONSTANTINE, A. G. (1963). Some non-central distribution problems in multivariate analysis. *Ann. Math. Statist.* **34** 1270–1285. MR0181056



- DAVIS, A. W. (1972). On the marginal distributions of the latent roots of the multivariate beta matrix. *Ann. Math. Statist.* **43** 1664–1670. [MR0343465](#)
- FOSTER, F. G. (1957). Upper percentage points of the generalized Beta distribution. II. *Biometrika* **44** 441–453. [MR0090199](#)
- FOSTER, F. G. (1958). Upper percentage points of the generalized Beta distribution. III. *Biometrika* **45** 492–503. [MR0100366](#)
- FOSTER, F. G. and REES, D. H. (1957). Upper percentage points of the generalized Beta distribution. I. *Biometrika* **44** 237–247. [MR0086462](#)
- GOLUB, G. H. and VAN LOAN, C. F. (1996). *Matrix Computations*, 3rd ed. Johns Hopkins Univ. Press, Baltimore. [MR1417720](#)
- HECK, D. L. (1960). Charts of some upper percentage points of the distribution of the largest characteristic root. *Ann. Math. Statist.* **31** 625–642. [MR0119301](#)
- JOHNSON, R. A. and WICHERN, D. W. (2002). *Applied Multivariate Statistical Analysis*, 6th ed. Pearson Prentice Hall, Upper Saddle River, NJ.
- JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295–327. [MR1863961](#)
- JOHNSTONE, I. M. (2008). Multivariate analysis and Jacobi ensembles: Largest eigenvalue, Tracy–Widom limits and rates of convergence. *Ann. Statist.* **36** 2638–2716. [MR2485010](#)
- JOHNSTONE, I. M. and CHEN, W. W. (2007). Finite sample accuracy of Tracy–Widom approximation for multivariate analysis. In *2007 JSM Proceedings* 1161–1166. Amer. Statist. Assoc., Alexandria, VA.
- JOHNSTONE, I. M., MA, Z., PERRY, P. O. and SHAHRAM, M. (2010). RMTstat: Distributions, statistics and tests derived from random matrix theory. Manuscript in preparation.
- KOEV, P. (2010). Computing multivariate statistics. Manuscript in preparation.
- KOEV, P. and EDELMAN, A. (2006). The efficient evaluation of the hypergeometric function of a matrix argument. *Math. Comp.* **75** 833–846 (electronic). [MR2196994](#)
- KRISHNAIAH, P. R. (1980). Computations of some multivariate distributions. In *Handbook of Statistics, Volume 1—Analysis of Variance* (P. R. Krishnaiah, ed.) 745–971. North-Holland, Amsterdam.
- LUTZ, J. G. (1992). A Turbo Pascal unit for approximating the cumulative distribution function of Roy’s largest root criterion. *Educational and Psychological Measurement* **52** 899–904.
- LUTZ, J. G. (2000). Roy table: A program for generating tables of critical values for Roy’s largest root criterion. *Educational and Psychological Measurement* **60** 644–647.
- MARDIA, K. V., KENT, J. T. and BIBBY, J. M. (1979). *Multivariate Analysis*. Academic Press, London. [MR0560319](#)
- MORRISON, D. F. (2005). *Multivariate Statistical Methods*, 4th ed. Thomson, Belmont, CA.
- MUIRHEAD, R. J. (1982). *Aspects of Multivariate Statistical Theory*. Wiley, New York. [MR0652932](#)
- NANDA, D. N. (1948). Distribution of a root of a determinantal equation. *Ann. Math. Statist.* **19** 47–57. [MR0024106](#)
- NANDA, D. N. (1951). Probability distribution tables of the largest root of a determinantal equation with two roots. *J. Indian Soc. Agricultural Statist.* **3** 175–177. [MR0044788](#)
- PÉCHÉ, S. (2009). Universality results for largest eigenvalues of some sample covariance matrices ensembles. *Probab. Theory Related Fields* **143** 481–516.
- PILLAI, K. C. S. (1955). Some new test criteria in multivariate analysis. *Ann. Math. Statist.* **26** 117–121. [MR0067429](#)
- PILLAI, K. C. S. (1956a). On the distribution of the largest or smallest root of a matrix in multivariate analysis. *Biometrika* **43** 122–127. [MR0077052](#)
- PILLAI, K. C. S. (1956b). Some results useful in multivariate analysis. *Ann. Math. Statist.* **27** 1106–1114. [MR0081852](#)
- PILLAI, K. C. S. (1957). *Concise Tables for Statisticians*. The Statistical Center, Univ. of the Philippines, Manila.

- PILLAI, K. C. S. (1965). On the distribution of the largest characteristic root of a matrix in multivariate analysis. *Biometrika* **52** 405–414. [MR0205391](#)
- PILLAI, K. C. S. (1967). Upper percentage points of the largest root of a matrix in multivariate analysis. *Biometrika* **54** 189–194. [MR0215433](#)
- PILLAI, K. C. S. and BANTEGUI, C. G. (1959). On the distribution of the largest of six roots of a matrix in multivariate analysis. *Biometrika* **46** 237–240. [MR0102152](#)
- PILLAI, K. C. S. and FLURY, B. N. (1984). Percentage points of the largest characteristic root of the multivariate beta matrix. *Commun. Statist. Part A* **13** 2199–2237. [MR0754832](#)
- RENCHER, A. C. (2002). *Methods of Multivariate Analysis*, 2nd ed. Wiley, New York. [MR1885894](#)
- SOSHNIKOV, A. (2002). A note on universality of the distribution of the largest eigenvalues in certain classes of sample covariance matrices. *J. Statist. Phys.* **108** 1033–1056. [MR1933444](#)
- TRACY, C. A. and WIDOM, H. (1996). On orthogonal and symplectic matrix ensembles. *Commun. Math. Phys.* **177** 727–754. [MR1385083](#)
- WAUGH, F. V. (1942). Regressions between sets of variables. *Econometrica* **10** 290–310.

DEPARTMENT OF STATISTICS  
STANFORD UNIVERSITY  
STANFORD, CALIFORNIA 94305  
USA  
E-MAIL: [imj@stanford.edu](mailto:imj@stanford.edu)