Roy's largest root test under rank-one alternatives

By I. M. JOHNSTONE

*Department of Statistics, Sequoia Hall, 390 Serra Mall, Stanford University,
California 94305, U.S.A.*

imj@stanford.edu

and B. NADLER

*Department of Computer Science and Applied Mathematics, Weizmann Institute of Science,
P.O. Box 26, Rehovot 76100, Israel*

boaz.nadler@weizmann.ac.il

Summary

Roy's largest root is a common test statistic in multivariate analysis, statistical signal processing and allied fields. Despite its ubiquity, provision of accurate and tractable approximations to its distribution under the alternative has been a longstanding open problem. Assuming Gaussian observations and a rank-one alternative, or concentrated noncentrality, we derive simple yet accurate approximations for the most common low-dimensional settings. These include signal detection in noise, multiple response regression, multivariate analysis of variance and canonical correlation analysis. A small-noise perturbation approach, perhaps underused in statistics, leads to simple combinations of standard univariate distributions, such as central and noncentral χ^2 and F. Our results allow approximate power and sample size calculations for Roy's test for rank-one effects, which is precisely where it is most powerful.

Some key words: Canonical correlation; Concentrated noncentrality; Greatest root statistic; Matrix perturbation; Multivariate analysis of variance; Roy's largest root test.

1. Introduction

Hypothesis testing plays an important role in the analysis of multivariate data. Classical examples include the multiple response linear model, principal component and canonical correlation analysis, and other methods which together form the main focus of standard multivariate texts, e.g., Anderson (2003), Mardia et al. (1979) and Muller & Stewart (2006). They find widespread use in signal processing, social sciences and many other domains.

Under multivariate Gaussian assumptions, in all these cases the associated hypothesis tests can be formulated in terms of either one or two independent Wishart matrices. These are conventionally denoted by H, for hypothesis, and E, for error, depending on whether the covariance matrix Σ is known or unknown; in the latter case E serves to estimate Σ.

James (1964) provided a remarkable five-way classification of the distribution theory associated with these problems. Elements of the classification are indicated in Table 1, along with some representative applications. Departure from the null hypothesis is captured by a matrix Ω, so that the testing problem might be $\mathcal{H}_0 : \Omega = 0$ versus $\mathcal{H}_1 : \Omega \neq 0$. Depending on the particular application, the matrix Ω captures the difference in group means, or the number of

Table 1. *Classification of multivariate problems (James, 1964)*

Case	Multivariate distribution	Distribution for dimension $m = 1$	Dimension $m > 1$	Application
1	$_0F_0$	χ^2	$H \sim W_m(n_H, \Sigma + \Omega)$ Σ known	Signal detection in noise, known covariance matrix
2	$_0F_1$	noncentral χ^2	$H \sim W_m(n_H, \Sigma, \Omega)$ Σ known	Equality of group means, known covariance matrix
3	$_1F_0$	F	$H \sim W_m(n_H, \Sigma + \Omega)$ $E \sim W_m(n_E, \Sigma)$	Signal detection in noise, estimated covariance matrix
4	$_1F_1$	noncentral F	$H \sim W_m(n_H, \Sigma, \Omega)$ $E \sim W_m(n_E, \Sigma)$	Equality of group means, estimated covariance matrix
5	$_2F_1$	Correlation coeff. $r^2/(1 - r^2)$ t-distribution	$H \sim W_p(q, \Sigma, \Omega)$ $E \sim W_p(n - q, \Sigma)$ Ω itself random	Canonical correlation analysis between two groups of sizes $p \leqslant q$.

James's classification of eigenvalue distributions was based on hypergeometric functions $_aF_b$ of matrix argument; their univariate analogues are shown in column 3. Column 4 details the corresponding Wishart assumptions for the sum of squares and cross products matrices; the rightmost column gives a nonexhaustive list of sample applications.

signals or canonical correlations and their strengths. In the absence of detailed knowledge about the structure of Ω under $\mathcal{H}_1$, group invariance arguments show that generic tests depend on the eigenvalues of either $\Sigma^{-1}H$ or $E^{-1}H$, e.g., Muirhead (1982, Ch. 6).

The most common tests fall into two categories. The first consists of linear statistics which depend on all the eigenvalues and are expressible in the form $\sum_i f(\ell_i)$ for some univariate function f. This class includes the three ubiquitous test statistics: Wilks's U, the Bartlett–Nanda–Pillai V and the Lawley–Hotelling W.

The second category involves functions of the extreme eigenvalues, the first few largest and smallest eigenvalues. Here we focus on the largest root statistic, based on ℓ_1, which arises systematically in multivariate analysis as the union-intersection test (Roy, 1957). To summarize extensive simulations by Schatzoff (1966) and Olson (1974), Roy's test is most powerful among the common tests when the alternative is of rank-one, i.e., concentrated noncentrality. For fixed dimension, Kritchman & Nadler (2009) showed asymptotic, in sample size, optimality of Roy's test against rank-one alternatives.

We briefly contrast the state of knowledge regarding approximate distributions, both null and alternative, for the two categories of test statistics. Tests based on linear statistics have long had adequate approximations. In particular, for the two-matrix cases, approximations using an F distribution are traditional and widely available in software: central F under the null (SAS, 1999) and noncentral F under the alternative; see Muller & Peterson (1984), Muller et al. (1992) and a 1999 unpublished manuscript by R. G. O'Brien and G. Shieh, discussed in the Supplementary Material. Saddlepoint approximations (Butler & Wood, 2005; Butler & Paige, 2010) are also available.

For Roy's largest root test, the situation is less complete. Under the null hypothesis, evaluation or approximation of the distribution is important, for example, to determine the critical values of Roy's test. Some options are listed in the Supplementary Material.

In contrast, under the alternative, derivation of a simple approximation to the distribution of ℓ_1 has remained a longstanding problem. In principle, the distribution of the largest eigenvalue has an exact representation in terms of a hypergeometric function of matrix argument. Despite recent advances in the numerical evaluation of these special functions (Koev & Edelman, 2006), unless

Table 2. *Power comparison* (%) *of Pillai's and Roy's tests from* 10^6 *simulations with* $p = 6$ *and* $m = 6$; *here* SE $\leqslant 0{\cdot}05$

Power	$\tau = 0{\cdot}09$	$\tau = 0{\cdot}11$	$\tau = 0{\cdot}13$
Pillai trace $\rho = 0$	25·1	46·7	70·3
Largest root $\rho = 0$	34·9	63·8	86·9
Approximation from Proposition 4	30·0	60·1	85·4
Pillai trace $\rho = 0{\cdot}3$	43·8	71·9	91·2
Largest root $\rho = 0{\cdot}3$	60·4	88·1	98·3
Approximation from Proposition 4	56·5	86·8	98·1

the dimension and sample size are less than 15, say, these formulae are challenging to evaluate numerically. In addition, to date, for dimension $m > 2$, no acceptable method has been developed for transforming Roy's largest root test statistic to an F or a χ^2 statistic, and no straightforward method exists for computing powers for Roy's statistic itself, as noted by Anderson (2003, p. 332), Muller et al. (1992), and O'Brien & Shieh.

We develop simple and quite accurate approximations to the distribution of ℓ_1 for the classic problems of multivariate analysis, under a rank-one alternative. Under this concentrated noncentrality alternative, the noncentrality matrix has the form $\Omega = \omega w w^{\mathrm{T}}, \omega > 0$, where $w \in \mathbb{R}^m$ is an arbitrary and unknown unit-norm vector. This setting, in which ℓ_1 is approximately the most powerful test, may be viewed as a specific form of sparsity, indicating that the effect under study can be described by relatively few parameters.

Our approach keeps (m, n_H, n_E) fixed. We study the limit of a large noncentrality parameter or, equivalently, small noise. While small-noise perturbation is a classical method in applied mathematics and mathematical physics, it has apparently seen less use in statistics. Some exceptions, mostly focused on other multivariate problems, include Kadane (1970, 1971), Anderson (1977), Schott (1986), Nadler & Coifman (2005) and Nadler (2008).

Our small-noise analysis uses tools from matrix perturbation theory and yields an approximate stochastic representation for ℓ_1. In concert with standard Wishart results, we deduce its approximate distribution for the five cases of Table 1 in Propositions 1 through 5. The expressions obtained can be readily evaluated numerically, typically via a single univariate integration. Code for the resulting distributions and their power is provided in the Supplementary Material.

The results of this paper can aid power analyses and sample size design in exactly those settings in which Roy's test may be most appropriate, namely when the relevant alternatives are thought to be predominantly of rank-one. Table 2 gives a small illustrative example for the classical problem of comparison of group means, setting 4 in Table 1. Here, we observe 20 multivariate samples from each of $p = 6$ different groups, and test a null hypothesis that their group means are all equal. Under the alternative, the $k = 1, \ldots, p$ group means are assumed to vary as multiples $\mu_k = k\tau\mu_0$ of a fixed vector $\mu_0 = (1, 1, 1, -1, -1, -1) \in \mathbb{R}^6$, with scale factor τ. The noise covariance matrix is $\Sigma = (1 - \rho)I + \rho 11^{\mathrm{T}}$. This setting leads to a rank-one noncentrality matrix for which, assuming Gaussian observations, Proposition 4 of § 3 applies.

Table 2 compares the power of multivariate tests at three signal strengths and two correlation models at level 1%. The Pillai trace V is chosen as representative of the three tests that use all the roots. Especially in the larger two signal settings, Roy's largest root test makes the difference between a plausible experiment and an underpowered study. For a real multivariate example in which Roy's test is argued to be most appropriate, see Hand & Taylor (1987, Study C). In other cases, past experience may suggest that an approximately rank-one alternative is plausible.

Simulations like this can suggest the magnitude of improvement possible in selected cases, but the essence of power and sample size analysis is the comparison of a range of scenarios thought to encompass the likely experimental setting. For this, relatively simple approximate formulas such as those derived in this paper are invaluable.

It is not the purpose of this paper to argue for a general and unexamined use of Roy's test. It is well-established that there is no uniformly best test, and in particular settings issues of robustness with respect to nonnormality already studied by, for example, Olson (1974) may be important. Instead, when there is interest in the performance of the largest root in the rank-one Gaussian cases where it should shine, we provide approximations that have long been lacking.

2. DEFINITIONS AND TWO APPLICATIONS

We present two applications, one from multivariate statistics and the other from signal processing, that illustrate settings 1–4 of Table 1. Following Muirhead (1982, p. 441), we recall that if $z_i \sim N_m(\mu_i, \Sigma)$ for $i = 1, \ldots, n$ are independent, $Z^T = (z_1, \ldots, z_n)$ and $M^T = (\mu_1, \ldots, \mu_n)$, then the $m \times m$ matrix $A = Z^T Z$ is said to have the noncentral Wishart distribution $W_m(n, \Sigma, \Omega)$ with n degrees of freedom, covariance matrix Σ and noncentrality matrix $\Omega = \Sigma^{-1} M^T M$. When $\Omega = 0$, the distribution is a central Wishart, $W_m(n, \Sigma)$. We denote by χ_n^2 a random variate following the chi-squared distribution with n degrees of freedom, and by $\chi_n^2(\delta)$ a variate following the noncentral chi-squared distribution with noncentrality parameter δ.

To introduce settings 1 and 3, signal detection in noise, consider a measurement system consisting of m sensors, such as antennas or microphones. In the signal processing literature, for example Kay (1998), a standard model for the observed samples in the presence of a single signal is

$$x = \rho_s^{1/2} u h + \sigma \xi \tag{1}$$

where h is an unknown m-dimensional vector, assumed fixed during the measurement time window, u is a random variable distributed as $\mathcal{N}(0, 1)$, ρ_s is the signal strength, σ is the noise level and ξ is a random noise vector, independent of u, that is multivariate Gaussian $\mathcal{N}_m(0, \Sigma)$.

In this paper, for the sake of simplicity, we assume real-valued signals and noise. The complex-valued case can be handled in a similar manner; see P. Dharmawansa et al. (arXiv:1411.4226). Thus, let $x_i \in \mathbb{R}^m$ denote n_H independent and identically distributed observations from (1), and let $n_H^{-1} H$ denote their sample covariance matrix, with

$$H = \sum_{i=1}^{n_H} x_i x_i^T \sim W_m(n_H, \sigma^2 \Sigma + \Omega),$$

where $\Omega = \rho_s h h^T$ has rank-one. A fundamental task in signal processing is to test $\mathcal{H}_0 : \rho_s = 0$, no signal present, versus $\mathcal{H}_1 : \rho_s > 0$. If the covariance matrix Σ is known, setting 1 in Table 1, the observed data can be whitened by the transformation $\Sigma^{-1/2} x_i$. Standard detection schemes then depend on the eigenvalues of $\Sigma^{-1} H$ (Wax & Kailath, 1985; Kritchman & Nadler, 2009).

A second important case, setting 3, assumes that the noise covariance matrix Σ is arbitrary and unknown but we have additional independent noise-only observations $z_j \sim \mathcal{N}(0, \Sigma)$ for

$j = 1, \ldots, n_E$. It is then traditional to estimate the noise covariance by $n_E^{-1} E$, where

$$E = \sum_{i=1}^{n_E} z_i z_i^{\mathrm{T}} \sim W_m(n_E, \Sigma),$$

and devise detection schemes using the eigenvalues of $E^{-1} H$. Some representative papers on signal detection in this setting, and the more general scenario with several sources, include Zhao et al. (1986), Zhu et al. (1991), Stoica & Cedervall (1997) and Nadakuditi & Silverstein (2010).

To introduce settings 2 and 4 of Table 1, consider the common problem of comparison of means from p groups, the one-way multivariate analysis of variance. Let I_k index observations in the kth group and assume a model

$$y_i = \mu_k + \xi_i, \quad i \in I_k \quad (k = 1, \ldots, p).$$

Here $\xi_i \sim \mathcal{N}_m(0, \Sigma)$ are independent with the error covariance Σ assumed to be the same for all groups, and the indices $\{1, \ldots, n\} = I_1 \cup \cdots \cup I_p$, with $n_k = |I_k|$ and $n = n_1 + \cdots + n_p$. We test the equality of group means, $\mathcal{H}_0 : \mu_1 = \cdots = \mu_p$, versus the alternative $\mathcal{H}_1$ that the μ_k are not all equal. Known Σ leads to setting 2. When Σ is unknown we obtain setting 4.

The standard approach is then to form between-groups and within-groups covariance matrices,

$$H = \sum_{k=1}^{p} n_k (\bar{y}_k - \bar{y})(\bar{y}_k - \bar{y})^{\mathrm{T}}, \quad E = \sum_k \sum_{i \in I_k} (y_i - \bar{y}_k)(y_i - \bar{y}_k)^{\mathrm{T}},$$

where $\bar{y}_k$ and $\bar{y}$ are the group and overall sample means respectively. The independent Wishart distributions of H and E appear in rows 2 and 4 of Table 1. The degrees of freedom are then $n_H = p - 1$ and $n_E = n - p$. The noncentrality matrix is $\Omega = \Sigma^{-1} \sum_1^p n_k (\mu_k - \bar{\mu})(\mu_k - \bar{\mu})^{\mathrm{T}}$, where $\bar{\mu} = n^{-1} \sum n_k \bar{\mu}_k$ is the overall population mean.

A rank-one noncentrality matrix is obtained if we assume that under the alternative, the means of the different groups are all proportional to the same unknown vector μ_0, with each multiplied by a group-dependent strength parameter, that is, $\mu_k = s_k \mu_0$. This yields a rank-one noncentrality matrix $\Omega = \omega \Sigma^{-1} \mu_0 \mu_0^{\mathrm{T}}$, where $\bar{s} = n^{-1} \sum_k n_k s_k$ and $\omega = \sum_{k=1}^p n_k (s_k - \bar{s})^2$. As discussed in the Supplementary Material, comparison of group means is but a particular example of the more general and ubiquitous multiple linear regression setting, to which, under rank-one alternatives and Gaussian observations, our results apply.

3. Main results for rank-one alternatives

3·1. *Setting*

Let ℓ_1 be the largest eigenvalue of either $\Sigma^{-1} H$ or $E^{-1} H$, depending on the specific setting. Roy's test rejects the null if $\ell_1 > t_\alpha$, where t_α is the threshold corresponding to a false alarm or Type I error rate α. The probability of detection, or the power of Roy's test, is defined as

$$P_D = P_{D,\Omega} = \mathrm{pr}\,(\ell_1 > t_\alpha \,; \mathcal{H}_1). \tag{2}$$

Under the null hypothesis in all four settings, $\Omega = 0$ and thus H has a central $W_m(n_H, \Sigma)$ distribution. Settings (1, 2) and (3, 4) differ by the presence or absence of $E \sim W_m(n_E, \Sigma)$.

An exact or approximate threshold t_α may be found by the methods referenced in the Supplementary Material.

As previously discussed, the focus of this paper is on the power P_D under rank-one alternatives. To this end, we present, in five propositions, simple approximate expressions for the distribution of Roy's largest root statistic ℓ_1 for all five settings described in Table 1, under a rank-one alternative. Detailed proofs appear in the Supplementary Material.

3·2. *Single Wishart matrix*

We begin with settings 1 and 2, where the matrix Σ is assumed to be known. Then, without loss of generality we study the largest eigenvalue of $\Sigma^{-1/2} H \Sigma^{-1/2}$. In setting 1, this matrix is distributed as $W_m(n_H, \sigma^2 I + \lambda_H w w^{\mathrm{T}})$, for some suitable vector $w \in \mathbb{R}^m$.

PROPOSITION 1. *Let $H \sim W_m(n_H, \sigma^2 I + \lambda_H w w^{\mathrm{T}})$ with $\|w\| = 1$ and $\lambda_H > 0$ and let ℓ_1 be its largest eigenvalue. Then, with (m, n_H, λ_H) fixed, as $\sigma \to 0$,*

$$\ell_1 = (\lambda_H + \sigma^2)\chi^2_{n_H} + \chi^2_{m-1}\sigma^2 + \frac{\chi^2_{m-1}\chi^2_{n_H-1}}{(\lambda_H + \sigma^2)\chi^2_{n_H}}\sigma^4 + o_{\mathrm{p}}(\sigma^4), \tag{3}$$

where $\chi^2_{n_H}, \chi^2_{m-1}$ and $\chi^2_{n_H-1}$ are independent χ^2 variates.

PROPOSITION 2. *Let $H \sim W_m(n_H, \sigma^2 I, (\omega/\sigma^2) w w^{\mathrm{T}})$ with $\|w\| = 1, \omega > 0$ and let ℓ_1 be its largest eigenvalue. Then, with (m, n_H, ω) fixed, as $\sigma \to 0$*

$$\ell_1 = \sigma^2 \chi^2_{n_H}(\omega/\sigma^2) + \chi^2_{m-1}\sigma^2 + \frac{\chi^2_{m-1}\chi^2_{n_H-1}}{\sigma^2 \chi^2_{n_H}(\omega/\sigma^2)}\sigma^4 + o_{\mathrm{p}}(\sigma^4), \tag{4}$$

where the three variates $\chi^2_{n_H}(\omega/\sigma^2)$, χ^2_{m-1} and $\chi^2_{n_H-1}$ are independent.

Remark 1. If in the above propositions σ^2 is held fixed along with m and n_H, and instead we suppose that $\lambda_H \to \infty$ or $\omega \to \infty$, then the same expansions hold, but now with error terms $o_{\mathrm{p}}(1/\lambda_H)$ or $o_{\mathrm{p}}(1/\omega)$. Lest it be thought unrealistic to base approximations on large λ_H, small σ or, in setting 5, ρ near 1, in standard simulation situations they lead to levels of power conventionally regarded as desirable, see § 4; indeed in these cases, weaker signals would not be acceptably detectable.

Remark 2. Propositions 1 and 2 directly provide approximate power for Roy's test, with the following error bounds. Let $\ell_1 = \ell_1(H)$ be the exact largest eigenvalue, and ℓ its stochastic approximation (3) or (4). Then, uniformly for all x bounded away from zero, if $n_H \geqslant 3$,

$$\mathrm{pr}(\ell_1 > x) = \mathrm{pr}(\ell > x) + O(\sigma^5 \log^5 \sigma^{-1}). \tag{5}$$

Approximations to the moments of ℓ_1 follow directly. From (3), independence of the chi-squared variates and standard moment formulas, we have that as $\sigma \to 0$,

$$E(\ell_1) = n_H \lambda_H + (m - 1 + n_H)\sigma^2 + \frac{(m-1)(n_H-1)}{(\lambda_H + \sigma^2)(n_H - 2)}\sigma^4 + o(\sigma^4). \tag{6}$$

In setting 2, we simply replace $n_H \lambda_H$ by ω in the first term, and the denominator of the third term by $\omega + \sigma^2(n_H - 2)$.

To compare the variances $\mathrm{var}(\ell_1)$ in settings 1 and 2, it is natural to set $\omega = \lambda_H n_H$, so that the means match to the leading two orders. If $\sigma = 1$ and $\lambda_H = \omega/n_H$ is large, then

$$\mathrm{var}(\ell_1) = \begin{cases} 2n_H\lambda_H^2 + 4n_H\lambda_H + 2(m - 1 + n_H) + o(1), & \text{setting 1,} \\ 4n_H\lambda_H + 2(m - 1 + n_H) + o(1), & \text{setting 2.} \end{cases} \tag{7}$$

Thus, for $\lambda_H \gg 1$, the fluctuations of ℓ_1 in setting 2 are significantly smaller. While beyond the scope of this paper, this result has implications for the detection power of Gaussian signals versus those of constant modulus.

3·3. *Two Wishart matrices*

Next, we consider the two-matrix case, where Σ is estimated from data.

PROPOSITION 3. *Let $H \sim W_m(n_H, \Sigma + \lambda_H ww^{\mathrm{T}})$ and $E \sim W_m(n_E, \Sigma)$ be independent Wishart matrices, with $w^{\mathrm{T}}\Sigma^{-1}w = 1$. If m, n_H and n_E are fixed and $\lambda_H \to \infty$, then*

$$\ell_1(E^{-1}H) \approx c_1(\lambda_H + 1)F_{a_1,b_1} + c_2F_{a_2,b_2} + c_3, \tag{8}$$

where the F-variates are independent, and with $v = n_E - m > 1$, the parameters a_i, b_i, c_i are

$$a_1 = n_H, \qquad b_1 = v + 1, \quad a_2 = m - 1, \quad b_2 = v + 2, \tag{9}$$

$$c_1 = a_1/b_1, \quad c_2 = a_2/b_2, \quad c_3 = a_2/\{v(v - 1)\}. \tag{10}$$

PROPOSITION 4. *Let $H \sim W_m(n_H, \Sigma, \Omega)$ and $E \sim W_m(n_E, \Sigma)$ be independent Wishart matrices with $v = n_E - m > 1$. Assume that the noncentrality matrix has rank-one, $\Omega = \omega\mu\mu^{\mathrm{T}}$, where $\|\mu\| = 1$. If m, n_H and n_E remain fixed and $\omega \to \infty$, then*

$$\ell_1(E^{-1}H) \approx c_1F_{a_1,b_1}(\omega) + c_2F_{a_2,b_2} + c_3, \tag{11}$$

where the F-variates are independent and the parameters a_i, b_i, c_i are given by (9) *and* (10).

In the $n_E \to \infty$ limit, the F-variates in (8) and (11) converge to χ^2 variates and we recover the first two terms in the approximations of Propositions 1 and 2, with $\sigma^2 = 1$ held fixed.

In (8) and (11), $\approx$ denotes a stochastic approximation whose nature we discuss briefly here. When $m = 1$, we have $c_2 = c_3 = 0$ and the first term gives the exact distribution of H/E for both Propositions 3 and 4. For $m > 1$, in setting 4, we note that to leading order $\ell_1 = O_{\mathrm{p}}\{(\omega + n_H)/n_E\}$, whereas the errors arise from ignoring terms $O_{\mathrm{p}}(\omega^{-1/2})$ and higher in an eigenvalue expansion and replacing stochastic terms of order $O_{\mathrm{p}}\{(\omega + n_H)^{1/2}m^{1/2}n_E^{-3/2}\}$ by their expectations. The corresponding statements apply to setting 3 if we replace $\omega + n_H$ by $\lambda_H n_H$ and $\omega^{-1/2}$ by $\lambda_H^{-1/2}$. Detailed discussion appears in the Supplementary Material.

We now turn to expressions for $E(\ell_1)$ and $\mathrm{var}(\ell_1)$ in settings 3 and 4, analogous to (6) and (7).

COROLLARY 1. *Consider setting 4, with $\ell_1 = \ell_1(E^{-1}H)$ and $p_3(\nu) = \nu^2(\nu - 2)$. Then*

$$E(\ell_1) = \frac{\omega + n_H}{n_E - m - 1} + \frac{m - 1}{n_E - m} + O(\omega^{-1/2}), \tag{12}$$

$$\text{var}(\ell_1) = \frac{2\{\omega^2 + \nu n_H(n_H + 2\omega)\}}{p_3(\nu - 1)} + \frac{2(m-1)(n_E - 1)}{p_3(\nu)} + O(\omega^{-1}). \tag{13}$$

In setting 3, ω is replaced by $\lambda_H n_H$ and in (13) the term $n_H + 2\omega$ is increased to $n_H(\lambda_H + 1)^2$.

Let $\hat{\Sigma} = n_E^{-1}E$ be an unbiased estimator of Σ. Comparison with Propositions 1 and 2 shows that in expectation $\ell_1(\hat{\Sigma}^{-1}H)$ exceeds $\ell_1(\Sigma^{-1}H)$ by a multiplicative factor close to $n_E/(n_E - m - 1)$. Hence, the largest eigenvalue of $n_E E^{-1}H$ is typically larger than that of $\Sigma^{-1}H$. Again, the fluctuations of ℓ_1 in setting 4 are smaller than for signal detection, setting 3.

Nadakuditi & Silverstein (2010) studied the large-parameter limiting value, but not the distribution, of $\ell_1(E^{-1}H)$ as $m/n_H \to c_E$ and as $m/n_H \to c_H$, also in non-Gaussian cases. In this limit, our formula (12) agrees, to leading order terms, with the large-λ_H limit of their expression (equation (23)). Hence, our analysis shows that the limits for the mean of $\ell_1(E^{-1}H)$ are quite accurate even at smallish values of m, n_E and n_H. This is also reflected in our simulations in the Supplementary Material.

3·4. *Canonical correlation analysis*

Let $\{x_i\}_{i=1}^{n+1}$ denote $n + 1$ multivariate Gaussian observations on $m = p + q$ variables with unknown mean μ and covariance matrix Σ, and let S denote the mean-centred sample covariance. Assume without loss of generality that $p \leqslant q$ and decompose Σ and S as

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \quad S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix},$$

where Σ_{11} and Σ_{22} are square matrices of sizes $p \times p$ and $q \times q$, respectively. We might alternatively assume that $\mu = 0$ is known and that we have n independent observations. In either case, the parameter n denotes the degrees of freedom of the Wishart matrix nS.

The population and sample canonical correlation coefficients, denoted by $\rho_1, \ldots, \rho_p$ and $r_1, \ldots, r_p$, are the positive square roots of the eigenvalues of $\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ and $S_{11}^{-1}S_{12}S_{22}^{-1}S_{21}$. We study the distribution of the largest sample canonical correlation, in the presence of a single large population correlation coefficient, $\rho_1 > 0$ and $\rho_2 = \cdots = \rho_p = 0$.

To state our final proposition, we need a modification of the noncentral F distribution that is related to the squared multiple correlation coefficient.

DEFINITION 1. *A random variable U follows a χ_n^2-weighted noncentral F distribution, with parameters a, b, c and n, written as $F_{a,b}^{\chi}(c, n)$, if it has the form*

$$U = \frac{\chi_a^2(Z)/a}{\chi_b^2/b},$$

where the noncentrality parameter $Z \sim c\chi_n^2$ is itself a random variable and all three chi-squared variates are independent.

If $c = 0$, the F^χ distribution of U reduces to a central F. For $c > 0$, the distribution is easily evaluated numerically via either of the representations

$$\operatorname{pr}(U \leqslant u) = \int_0^\infty p_n(t) F_{a,b;ct}(u)\, \mathrm{d}t = \sum_{k=0}^\infty p_K(k) F_{a+2k,b}\{au/(a+2k)\}.$$

In the first, $F_{a,b;\omega}$ is the noncentral F distribution with noncentrality parameter ω and p_n is the density of χ_n^2: this is just the definition. In the second, p_K is the discrete probability density function of a negative binomial variate with parameters $(n/2, c)$: this is an analogue of the more familiar representation of noncentral $F_{a,b;\omega}$ as a mixture of $F_{a+2k,b}$ with Poisson weights with parameter $\omega/2$. The equality above may be verified directly or from Muirhead (1982, p. 175ff), which also gives an expression for the F^χ distribution in terms of the Gauss hypergeometric function ${}_2F_1$.

PROPOSITION 5. *Let $\ell_1 = r_1^2/(1 - r_1^2)$, where r_1 is the largest sample canonical correlation between two groups of sizes $p \leqslant q$ computed from $n + 1$ independent multivariate Gaussian observations, with $v = n - p - q > 1$. Then in the presence of a single population correlation coefficient ρ between the two groups, asymptotically as $\rho \to 1$,*

$$\ell_1 \approx c_1 F_{q,v+1}^\chi(c,n) + c_2 F_{p-1,v+2} + c_3, \tag{14}$$

with $c = \rho^2/(1 - \rho^2)$ and

$$c_1 = \frac{q}{v+1}, \quad c_2 = \frac{p-1}{v+2}, \quad c_3 = \frac{p-1}{v(v-1)}.$$

When $p = 1$, the quantity r_1^2 reduces to the squared multiple correlation coefficient, or coefficient of determination, between a single response variable and q predictor variables. Equation (14) then reduces to a single term $\{q/(n-q)\}F_{q,n-q}^\chi(c,n)$, which is in fact the exact distribution of r_1^2 in this setting (Muirhead, 1982, p. 173).

By (14), the largest empirical canonical correlation coefficient is biased upwards,

$$E(\ell_1) = \frac{n}{n-p-q-1}\frac{\rho^2}{1-\rho^2} + \frac{p+q-1}{n-p-q-1} + O\{(1-\rho^2)^{1/2}\},$$

by both a multiplicative factor $n/(n-p-q-1)$ and an additive factor. This bias may be significant for small sample sizes.

4. SIMULATIONS

4·1. *Empirical densities*

We present a series of simulations that support our theoretical analysis and illustrate the accuracy of our approximations. For different signal strengths we make $2\,500\,000$ independent random realizations of the two matrices E and H, and record the largest eigenvalue ℓ_1. We also compute its approximate density and power of Roy's test via our Propositions 1–5.

The top row of Fig. 1 compares the empirical density of $\{\ell_1(H) - E(\ell_1)\}/\sigma(\ell_1)$ in the signal detection and multivariate analysis of variance cases to the theoretical formulas, (3)
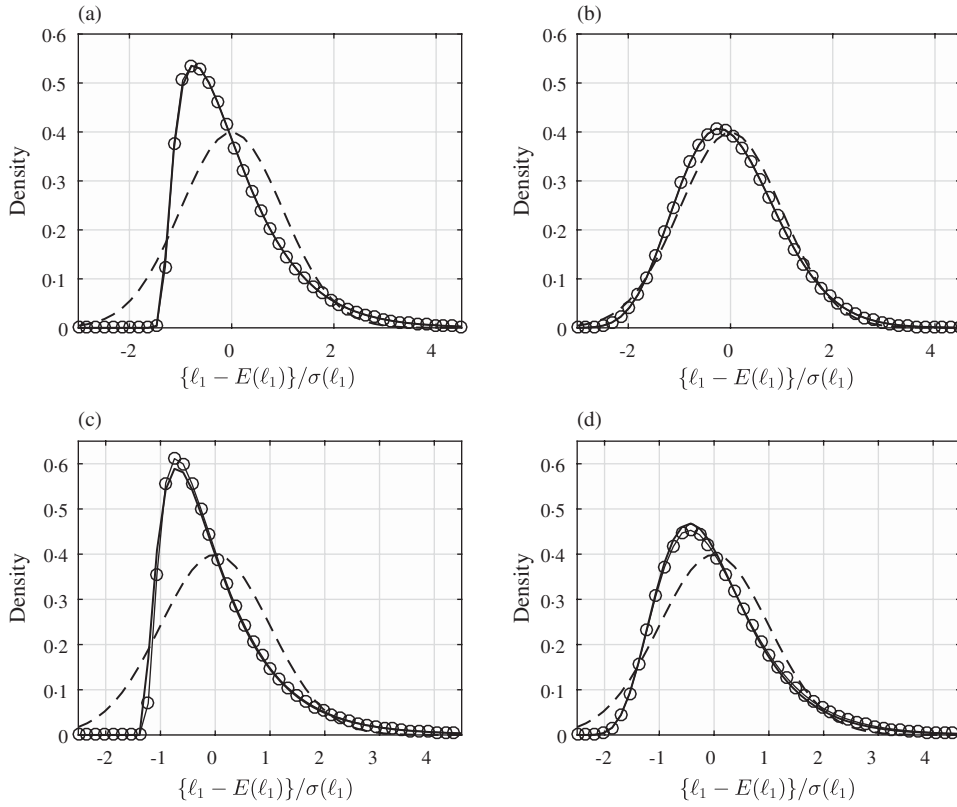
Fig. 1. Density of largest eigenvalue $\ell_1(H)$ (top row) and of $\ell_1(E^{-1}H)$ (bottom row). (a) Setting 1 (signal detection) with $m = 5, n_H = 4, \lambda_H = 10, \sigma = 1$. (b) Setting 2 (multivariate analysis of variance) with $m = 5$, $p = 5$ groups and $n_k = 8$ observations per group, and $\omega = 40$. (c) Setting 3 with parameters as above and $n_E = 35$. (d) Setting 4, also with $n_E = 35$. Comparison of empirical density (circles) to the theoretical approximations from Propositions 1–4, equations (3), (4), (8) and (11), respectively (solid line). For reference, the dashed curve is the density of a standard Gaussian.

and (4), respectively. In this simulation, where all parameter values are small, the theoretical approximation is remarkably accurate, far more so than the classical asymptotic Gaussian approximation. The latter would be valid in the large-parameter limit with $m, n_H \rightarrow \infty$ and $m/n_H \rightarrow c > 0$, so long as $\lambda_H > c^{1/2}$ (Baik et al., 2005; Paul, 2007).

Further figures showing the relative accuracy of these approximations appear in the Supplementary Material. The relative error in the top panels of Fig. 1 is less than 0·02 except in the left tail, which is excluded in approximations to power, $G_1(x) = \mathrm{pr}(\ell_1 > x)$. This is consistent with the theoretical analysis in the proof of the power approximation (5) and with plots of the relative error in the approximation to $G_1(x)$, both given in the Supplementary Material.

The bottom row of Fig. 1 corresponds to the two-matrix case and the approximate density of $\ell_1(E^{-1}H)$, after normalization, in settings 3 and 4. As expected from the analysis, the density is skewed. The approximated density matches quite closely the empirical one, and both are far from the nominal limiting Gaussian density. The relative error in the bottom panels is less than 0·05 except in the left tail and, in Setting 4, also in the right tail; see the Supplementary Material.

4·2. *Power calculations*

We conclude this section with a comparison of the empirical detection probability (2) of Roy's test to the theoretical formulas. We first consider the multivariate analysis of variance setting.

Table 3. *Power* (%) *of Roy's test for multivariate analysis of variance*

Dim. m	Groups p	Samples per group, n_k	Noncentrality ω	P_D sim. $\alpha = 1\%$	P_D theory	P_D sim. $\alpha = 5\%$	P_D theory
3	3	10	10	28·3	27·1	54·4	53·3
3	3	10	20	67·8	67·9	88·2	88·4
3	3	10	40	97·5	97·7	99·7	99·7
6	3	10	10	15·0	13·8	36·9	33·9
6	3	10	20	44·1	42·8	71·8	70·4
6	3	10	40	87·5	87·9	97·5	97·5
6	6	10	10	10·4	6·4	27·4	18·6
6	6	10	20	35·7	30·8	61·3	55·4
6	6	10	40	85·0	83·9	95·6	95·1
10	6	20	10	8·3	5·4	22·9	14·3
10	6	20	20	31·2	25·4	55·1	45·6
10	6	20	40	82·8	79·5	94·0	91·7

Table 4. *Power* (%) *of Roy's test for canonical correlation analysis*

p	q	n	ρ	P_D sim. $\alpha = 1\%$	P_D theory	P_D sim. $\alpha = 5\%$	P_D theory
2	5	40	0·50	34·4	33·6	59·6	57·8
2	5	40	0·60	65·3	64·9	84·9	84·2
2	5	40	0·70	91·8	91·7	97·8	97·7
3	7	50	0·50	31·3	27·8	56·5	51·4
3	7	50	0·60	64·3	61·8	84·2	82·2
3	7	50	0·70	92·5	92·1	98·0	97·9
5	10	50	0·50	13·5	8·5	32·7	22·2
5	10	50	0·60	35·1	28·9	60·3	52·3
5	10	50	0·70	72·3	68·9	88·9	86·6

Table 3 compares the theoretical power, which follows from our Proposition 4, to the results of simulations. Each entry in the table is the result of 2 000 000 independent random realizations of matrices H and E, with $\Sigma = I$ and different noncentrality values ω. The parameters in the table are a subset of those studied by Olson (1974). Two features are apparent from the table: first, our approximations are quite accurate for small sample size and dimension, and become less accurate as the dimension increases. This is to be expected given that the leading error terms in our expansion are of the form $O(m^{1/2})$. Second, the approximation is relatively more accurate at high powers, say larger than 80%, which fortunately are those most relevant to the design of studies in practice. This too is expected, as our approximation is based on a high signal-to-noise ratio, and is valid when no eigenvalue crossover has occurred, meaning that the largest eigenvalue is not due to large fluctuations in the noise. At the other extreme, when the signal strength is weak, our approximation of power is usually conservative since we do not model the case where the largest eigenvalue may arise due to large deviations of the noise.

Finally, we consider setting 5 of canonical correlation analysis. The corresponding comparison of simulations to theory is reported in Table 4, with similar behaviour to Table 3. For simulation results for the case of detection of signals in noise, we refer to Nadler & Johnstone (2011).

5. Discussion

The typical approach in classical statistics studies the asymptotics of the random variable of interest as the sample size $n_H \to \infty$. Propositions 1–5, in contrast, keep n_H, n_E, m fixed but let $\lambda_H \to \infty$, or equivalently $\sigma \to 0$. If the signal strength is sufficiently large, by their construction and as verified in the simulations, Propositions 1–5 are quite accurate for small dimension and sample size values. However, the error in these approximations increases with the dimensionality m, and so may not be suitable in high-dimensional small-sample settings.

Next, we mention some directions for future research. The study of the distribution of Roy's largest root test under higher-dimensional alternatives is a natural extension. However, depending on the particular alternative, it may be less powerful than other common tests in that situation. It should be possible to study the resulting distribution under, say, two strong signals, or perhaps one strong signal and several weak ones. We briefly elaborate on this in the Supplementary Material. Sensitivity of the distributions to departures from normality is important. Finally, our approach can be applied to study other test statistics, such as the Hotelling–Lawley trace. In addition, the approach can also provide information about eigenvector fluctuations, which are important in a variety of applications.

Supplementary material

Supplementary material available at *Biometrika* online includes further discussion and examples, further simulation results, proofs of Propositions 1 to 5, supporting lemmas, discussion of error terms, and the rank-r case. Matlab code for power calculations and scripts that reproduce the figures and tables are also provided.

References

Anderson, T. W. (1977). Asymptotic expansions of the distributions of estimates in simultaneous equations for alternative parameter sequences. *Econometrica* **45**, 509–18.

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis* Hokoben, New Jersey: Wiley, 3rd ed.

Baik, J., Ben Arous, G. & Péché, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Prob.* **33**, 1643–97.

Butler, R. & Paige, R. (2010). Exact distributional computations for Roy's statistic and the largest eigenvalue of a Wishart distribution. *Statist. Comp.* **21**, 147–57.

Butler, R. W. & Wood, A. T. A. (2005). Laplace approximations to hypergeometric functions of two matrix arguments. *J. Mult. Anal.* **94**, 1–18.

Hand, D. J. & Taylor, C. C. (1987). *Multivariate Analysis of Variance and Repeated Measures: A Practical Approach for Behavioural Scientists*. London: CRC Press.

James, A. T. (1964). Distributions of matrix variates and latent roots derived from normal samples. *Ann. Math. Statist.* **35**, 475–501.

KADANE, J. B. (1970). Testing overidentifying restrictions when the disturbances are small. *J. Am. Statist. Assoc.* **65**, 182–5.

KADANE, J. B. (1971). Comparison of *k*-class estimators when the disturbances are small. *Econometrica* **39**, 723–37.

KAY, S. M. (1998). *Fundamentals of Statistical Signal Processing, Volume 2: Detection Theory*. Upper Saddle River, New Jersey: Prentice-Hall.

KOEV, P. & EDELMAN, A. (2006). The efficient evaluation of the hypergeometric function of a matrix argument. *Math. Comp.* **75**, 833–46.

KRITCHMAN, S. & NADLER, B. (2009). Non-parametric detection of the number of signals: Hypothesis testing and random matrix theory. *IEEE Trans. Sig. Proces.* **57**, 3930–41.

MARDIA, K. V., KENT, J. T. & BIBBY, J. M. (1979). *Multivariate Analysis*. London: Academic Press.

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

MULLER, K. E., LAVANGE, L. M., RAMEY, S. M. & RAMEY, C. T. (1992). Power calculations for general multivariate models including repeated measures applications. *J. Am. Statist. Assoc.* **87**, 1209–26.

MULLER, K. E. & PETERSON, B. L. (1984). Practical methods for computing power in testing the multivariate general linear hypotheiss. *Comp. Statist. Data Anal.* **2**, 143–58.

MULLER, K. E. & STEWART, P. W. (2006). *Linear Model Theory, Univariate, Multivariate and Mixed Models*. Hokoben, New Jersey: Wiley.

NADAKUDITI, R. R. & SILVERSTEIN, J. W. (2010). Fundamental limit of sample generalized eigenvalue based detection of signals in noise using relatively few signal-bearing and noise-only samples. *IEEE J. Selected Topics Sig. Proces.* **4**, 468–80.

NADLER, B. (2008). Finite sample approximation results for principal component analysis: A matrix perturbation approach. *Ann. Statist.* **36**, 2791–817.

NADLER, B. & COIFMAN, R. (2005). The prediction error in CLS and PLS: The importance of feature selection prior to multivariate calibration. *J. Chemometrics* **19**, 107–18.

NADLER, B. & JOHNSTONE, I. M. (2011). Detection performance of Roy's largest root test when the noise covariance matrix is arbitrary. In *2011 IEEE Statistical Signal Processing Workshop (SSP)*. Nice: Institute of Electrical and Electronics Engineers, pp. 681–4.

OLSON, C. L. (1974). Comparative robustness of six tests in multivariate analysis of variance. *J. Am. Statist. Assoc.* **69**, 894–908.

PAUL, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica* **17**, 1617–42.

ROY, S. N. (1957). *Some Aspects of Multivariate Analysis*. New York: Wiley.

SAS (1999). *SAS Manual*, release 8. Cary, North Carolina: SAS Institute Inc.

SCHATZOFF, M. (1966). Sensitivity comparisons among tests of the general linear hypothesis. *J. Am. Statist. Assoc.* **61**, 415–35.

SCHOTT, J. R. (1986). A note on the critical values used in stepwise tests for multiplicative components of interaction. *Commun. Statist.* A **15**, 1561–70.

STOICA, P. & CEDERVALL, M. (1997). Detection tests for array processing in unknown correlated noise fields. *IEEE Trans. Sig. Proces.* **45**, 2351–62.

WAX, M. & KAILATH, T. (1985). Detection of signals by information theoretic criteria. *IEEE Trans. Acoust. Speech Sig. Proces.* **33**, 387–92.

ZHAO, L., KRISHNAIAH, P. R. & BAI, Z. (1986). On detection of the number of signals when the noise covariance matrix is arbitrary. *J. Mult. Anal.* **20**, 26–49.

ZHU, Z., HAYKIN, S. & HUANG, X. (1991). Estimating the number of signals using reference noise samples. *IEEE Trans. Aerosp. Electron. Syst.* **27**, 575–9.