Journal of Modern Applied Statistical Methods

Volume 10 | Issue 2 Article 2

11-1-2011

Robustness, Power and Interpretability of Pairwise Tests of Discriminant Functions in MANOVA

Philip H. Ramsey
Queens College of CUNY, Flushing, Philip.Ramsey@qc.cuny.edu

Patricia P. Ramsey
Fordham University

Priscila Hachimine
Graduate Center of the City University of New York, phachimine@gmail.com

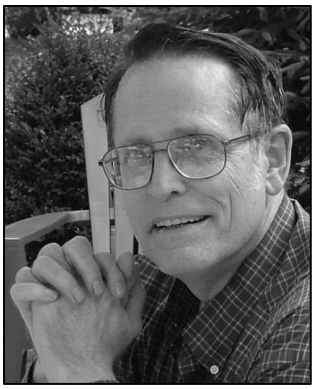
Nancy Andiloro
Graduate Center of the City University of New York, nancy6183@gmail.com

Journal of Modern Applied Statistical Methods
November 2011, Vol. 10, No. 2, 403-417

1538 – 9472/11/$95.00

Invited Article

Robustness, Power and Interpretability of Pairwise Tests of Discriminant Functions in MANOVA




Philip H. Ramsey
Queens College of the City University of New York, Flushing, NY

Patricia P. Ramsey
Fordham University, New York, NY

Priscila Hachimine
Nancy Andiloro
Graduate Center of the City University of New York, New York, NY

Limiting follow-up hypotheses to be tested can reduce problems relating to the control of Type I and Type II errors in multivariate analysis of variance (MANOVA). Such limitations can also improve the interpretability of results. The importance of sample size, shape of population distribution, within-group correlations and heterogeneity of variances are demonstrated. The protected greatest characteristic root (GCR) procedure is shown to work well for small, group size, N (≤ 10). The unprotected GCR is shown to work well for larger N.

Key words: Any-pair power, discriminant functions, MANOVA, pair-wise test.

Introduction

Testing for the significance of differences in means of k groups on p variables can be accomplished with multivariate analysis of variance (MANOVA). The full, null hypothesis is

$$H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \ldots = \boldsymbol{\mu}_k,$$

where $\boldsymbol{\mu}_i$ ($i = 1, \ldots, k$) is the vector of population means for group i on the p variables. The hypothesis degrees of freedom is $df_h = k - 1$. In the general case, the parameter, $s = \min(p, df_h)$. In MANOVA a variety of test statistics for the null hypothesis are possible. Taking p x p matrices, **H** and **E**, of the sum-of-products for hypotheses and error respectively as

$$\mathbf{H} = \sum_{i=1}^{k} \boldsymbol{n}_i (\bar{\mathbf{X}}_i - \bar{\mathbf{X}})(\bar{\mathbf{X}}_i - \bar{\mathbf{X}})', \quad (1)$$

Philip Ramsey is a Professor of Psychology. His research interests include psychometrics, applied statistics, and multiple comparisons. Email: Philip.Ramsey@qc.cuny.edu. Patricia Ramsey passed away in 2011. She was a Professor in the Graduate School of Business at Fordham University. Priscila Hachimine is a graduate student in psychology at CUNY. Email: phachimine@gmail.com. Nancy Andiloro is a graduate student in educational psychology at CUNY. Email: nancy6183@gmail.com.

and

$$\mathbf{E} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\mathbf{X}_{ij} - \bar{\mathbf{X}}_{\mathbf{j}})(\mathbf{X}_{ij} - \bar{\mathbf{X}}_{\mathbf{j}})', \tag{2}$$

where $\mathbf{X}_{ij}$ is the j^{th} of n_i observation vectors in group i, $\bar{\mathbf{X}}_{\mathbf{i}}$ is the mean vector for the i^{th} group and $\bar{\mathbf{X}}$ is the grand mean vector. The s, nonzero eigenvalues of $\mathbf{HE}^{-1}$ can be designated as λ_1, …, λ_s in order from largest to smallest. Equivalently, the s, nonzero eigenvalues (also called characteristic roots) of $\mathbf{H(H+E)^{-1}}$ can be designated as θ_1, … θ_w in order from largest to smallest. Each corresponding member of the respective sets of eigenvalues can be related by $\theta = \lambda/(1 + \lambda)$.

Multivariate Test Procedures

The four, most common MANOVA test statistics are:

1. The Pillai-Bartlett trace, $V = \sum_{i=1}^{s} \theta_i$;
2. Wilks' likelihood ratio, $W = \prod_{i=1}^{s} (1 - \theta_i)$;
3. The Hotelling-Lawley trace, $T = \sum_{i=1}^{s} \lambda_i$; and
4. Roy's greatest characteristic root (GCR), $R = \theta_1$.

Computer packages such as SPSS and SAS typically provide approximate and sometimes exact p values for each of these four test statistics.

Routines for Testing

In each of the following routines s is defined as shown above and $df_E = \Sigma(N_i - 1)$. One method of evaluating V for a group of k means is with an F test (Pillai, 1955; Seber, 1984, p. 564) defined by

$$F = \frac{cV}{b(s - V)},$$

where $c = df_E - p + s$, and $b = \max(p, k - 1)$. To test at level α requires critical value, $CV = F_{1-\alpha}$ (sb, sc). This method is designated here as VPB.

Two, more accurate F tests for V are available (Muller, 1998). Method 1 is

$$F = \frac{df_2}{df_1} \frac{V}{d - V}, \tag{4}$$

where $df_1 = p(k - 1)$,

$$df_2 = \frac{[p(k-1)+2]df_E}{df_E(k+p)+(k+1)(k-2)}(df_E + k - 1 - p),$$

and

$$d = \frac{p(k-1)+df_2}{df_2 + k - 1}.$$

To test at level α requires $CV = F_{1-\alpha}(df_1, df_2)$. This method is designated here as VM1.

For Method 2 (Muller, 1988) the F test is

$$F = \frac{df_2}{df_1} \frac{V}{s - V}, \tag{5}$$

where

$$K = \frac{1}{s(df_E + k - 1)} \left[\frac{s(df_E + s - p)(df_E + k + 1)(df_E + k - 2)}{df_E(df_E + k - 1 - p)} \right] - 2$$

$df_1 = p(k - 1)K$, $c = df_E - p + s$, and $df_2 = scK$. To test at level α requires $CV = F_{1-\alpha}(df_1, df_2)$. This method is designated here as VM2.

One method of evaluating W for a group of k mean vectors is with an F test (Rao, 1951; Seber, 1984, p. 41) defined by

$$F = \frac{1 - U}{U} \frac{df_2}{df_1}, \tag{6}$$

where

$$t = \sqrt{\frac{p^2(k-1)^2 - 4}{p^2 + (k-1)^2 - 5}},$$

$$f = \frac{\mathrm{df_E} - (p - k + 2)}{2}$$

$$g = \frac{\mathrm{p(k\text{-}1)\text{-}2}}{2},$$

$$\mathrm{df_1} = p(k - 1),$$

$$\mathrm{df_2} = \mathrm{ft} - \mathrm{g},$$

and

$$U = W^{1/t}.$$

To test at level α requires CV = $F_{1-\alpha}$ (df_1, df_2). This method is designated here as WLR. It can be shown that (6) provides an exact *F* test for $p = 1, 2$ or $k = 2, 3$ (Seber, 1984, pp. 40-41).

One method of evaluating *T* for a group of *k* mean vectors is with an *F* test (McKeon, 1974; Seber, 1984, p. 39) defined by

$$F = \frac{\mathrm{T}}{\mathrm{c}} \qquad (7)$$

where

$$B = \frac{(\mathrm{df_E} + \mathrm{k\text{-}p\text{-}2})(\mathrm{df_E}\text{-}1)}{(\mathrm{df_E}\text{-}\mathrm{p}\text{-}3)(\mathrm{df_E}\text{-}\mathrm{p})},$$

$$b = 4 + \frac{\mathrm{a+2}}{\mathrm{B\text{-}1}},$$

and

$$c = \frac{\mathrm{a(b\text{-}2)}}{\mathrm{b}(\mathrm{df_E}\text{-}\mathrm{p}\text{-}1)}.$$

To test at level α requires CV = $F_{1-\alpha}$ (a, b). This method is designated here as THL.

Routines for computing *p* values for Roy's *R* are either quite complex or rather crude. The versions used by statistical packages are not very accurate. For example, SAS prints a footnote on output warning that the corresponding *F* ratio for *R* is an upper bound. Consequently, the *p* value is a lower bound. Therefore, a *p* value of .04 would only tell the user that the exact *p* value is no less than .04. It would be more helpful to know that the exact *p* value was no greater than some value. Tables of critical values for *R* are available (Harris, 2001, pp. 518-531; Sever, 1984, pp. 593-598). Routines described by Harris (2001) were used to determine *p* values and critical values in the present study; the method is designated here as GCR.

Pairwise testing on a discriminant function can be performed as described by Harris (2001, p. 222). The *F* test for the difference between a given pair of means on the discriminant function is compared to a critical value, F_{CRIT}. The value of F_{CRIT} is found from $df_E(\theta_{CRIT})/(1 - \theta_{CRIT})$ where θ_{CRIT} is the critical value for *R*.

Noncentrality

In the non-null case, the *p* x *p* matrix Φ can be defined as

$$\Phi = \sum_{i=1}^{k} n_i(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})', \qquad (8)$$

where $\boldsymbol{\mu}$ is the grand mean vector of the population.

Take the *p* x *p* matrix Γ as

$$\Gamma = \Phi\Sigma^{-1},$$

where Σ is the population covariance matrix. The *p* eigenvalues of Γ are $\gamma_1, \ldots, \gamma_p$. The noncentrality parameter, δ^2, is

$$\delta^2 = \sum_{i=1}^{p} \gamma_i. \qquad (10)$$

Populations vary along a continuum from a concentrated structure where γ_1 is the only nonzero eigenvalue of Γ to a diffuse structure where *s* eigenvalues of Γ are nonzero. When the usual MANOVA assumptions are satisfied the most powerful tests of the four listed above for evaluating a concentrated structure would be *R*. For the diffuse structure the most powerful of the four tests would be *V* (Olson, 1974).

Robustness

Investigations of various testing procedures have shown marked differences in robustness (Olson, 1974). All test procedures in MANOVA have reduced control of Type I and Type II errors in the presence assumption

failure. The most extreme problems occur for *R* and the least for *V*.

Follow-Up Tests

Roy's *R* has been found to be more useful than *V*, *W*, or *T* for finding specific differences between groups (Bird & Hadzi-Pavlovic, 1983). In order to improve the robustness and interpretability of significant group differences, Bird and Hadzi-Pavlovic, (1983) proposed limiting the testing of group contrasts in two ways. First, they proposed the examination of group differences on single dependent variables, sums of dependent variables, differences between dependent variables, or combinations of these. That is, complex weightings of dependent variables used to form discriminant functions were avoided.

The second restriction was a limitation of the contrasts on group means to be tested. A moderate restriction on contrasts allows only one subset of means to be compared to another subset. With $k = 4$ there would be only 25 possible contrasts (6 pairwise, 3 pairs versus another pair, 12 pairs versus a single & 4 triples versus a single). With $p = 2$ dependent variables there would be four variables for testing (2 dependent variables, one sum, & one difference). That would allow only 100 contrasts to be tested. For $p = 6$ the total number of contrasts to be tested would be 9,100.

A strong restriction on the permissible contrasts for $k = 4$ would allow the 25 contrasts about the 4 groups to be applied only to each dependent variable. With $p = 2$, there would be only 50 tests performed. With $p = 6$ there would be 150. Bird and Hadzi-Pavlovic, (1983) reported considerable improvement in Type I error control under assumption failure with both moderate and strong restrictions. A univariate, Bonferroni- Scheffé (B-S) approach was also considered by testing contrasts on each dependent variable using the Scheffé (1953) procedure at level α/p. They also suggest the possibility of a so-called protected *R* test in which *R* is applied to testing contrasts only after a significant overall test such as *V*.

In an attempt to increase power, Sheehan-Holt (1998) considered a partially restricted condition. Sheehan-Holt placed no restriction on the variable thus allowing the testing of group contrasts on any discriminant function. For $k = 4$, the 25 contrasts would be tested on the first discriminant function. If the first discriminant function were limited to pairwise testing there would be only six tests of group differences on the discriminant function for $k = 4$.

A Monte Carlo Study

The present restriction on group contrasts to be tested is limited to pairwise testing. For $k = 4$, the six contrasts constitute fewer group contrasts than any considered by Bird and Hadzi-Pavlovic, (1983) or Sheehan-Holt (1998). However, the present investigation applies those group contrasts to all significant discriminant functions.

Seven procedures were used to test the full null hypothesis: VPB, VM1, VM2, THL, WLR, GCR, and the Bonferroni-Scheffé (B-S). The first five procedures follow a significant overall test with pairwise testing based on *R*. These five methods are examples of a protected *R* test. The GCR procedure also applies pairwise testing as an unprotected *R* test.

Conditions investigated included $k = 4$, common group sizes *N* of 10, 15 and 20, and $p = 4$. The population covariance matrix was varied to produce either uncorrelated variates ($\mathbf{\Sigma} = \mathbf{I}$) or $\mathbf{\Sigma}$ with all variables correlated by a common correlation ρ of either 0.71 or –0.2. For non-null conditions δ^2 was varied over a range of several values to produce power values in the neighborhood of 0.50.

Covariance Heterogeneity

Following Bird and Hadzi-Pavlovic (1983) and Olson (1974), heterogeneity was introduced by multiplying all variates in Group 1 by a constant chosen to produce a value of the coefficient of variation, *C*, (Box, 1954). If the variances in Group 1 are all initially set at $\sigma^2 = 1$ and a value *d* is the multiplicative value, C^2 can be calculated as

$$C^2 = \frac{1}{k\bar{\sigma}^4}\sum_{i=1}^{k}(\sigma_i^2 - \bar{\sigma}^2)^2, \qquad (11)$$

where $\sigma_i^2 = d$ for $i = 1$ and 1 for $i \neq 1$, and

$$\bar{\sigma}^2 = \frac{\sum_{i=1}^{k} \sigma_i^2}{k}.$$

Bird and Hadzi-Pavlovic (1983) used $C = 0.4$ as moderate covariance heterogeneity and $C = 0.8$ as substantial covariance heterogeneity. Thus, C^2 values would be .16 for moderate and 0.64 for substantial covariance heterogeneity. However, Olson (1974) investigated values as high as $C^2 = 2.4$. The present investigation examined values as high as $C^2 = 2.0$. Olson's (1974) results seem to suggest that error rates approach an upper limit for very high values of C^2.

Nonnormality

Previous studies have given little consideration to failure of the normality assumption. Some degree of kurtosis has been investigated showing relative little effect. However, the degree of kurtosis is not clear. For example, the fourth moment calibration was not reported.

Micceri (1989) reported many distributions that were clearly nonnormal. However, the data sets reported by Micceri are not as extreme as those used in many studies evaluating statistical robustness. Among skewed distributions, Micceri identified the most extreme distributions as being typified by the exponential distribution with standardized third and fourth moments as ($\sqrt{\beta_1} = 2.0$, $\beta_2 = 9.0$). Among symmetric, platykurtic distributions Micceri represented the shape as typical of the uniform distribution ($\sqrt{\beta_1} = 0.0$, $\beta_2 = 1.8$). Among symmetric, leptokurtic distributions Micceri identified the shape as double exponential ($\sqrt{\beta_1} = 0.0$, $\beta_2 = 6.0$).

To investigate the effects of distribution shape, four shapes were considered: the normal, uniform, exponential, and double exponential. The three nonnormal shapes represent the most extreme conditions reported by Micceri (1989). The uniform distribution was easily produced directly from the generated random numbers. The double exponential was approximated as a *t* distribution with *df* = 6. This *t* distribution has the same third and fourth moments as the double exponential distribution. The exponential distribution was approximated by Johnson's (1949) S_B method as described by Tadikamalla (1980) with $\sqrt{\beta_1} = 2.0$ and $\beta_2 = 9$.

Each simulated experiment was replicated 10,000 times. Significant differences in Type I error rates can be identified as deviating from an expected interval about the nominal rejection rates. For rejection rates between 0.0 and 1.0 the standard error (SE) depends on the value of the rate. If x is the proportion of replications exceeding a critical value, the SE is $[x(1 - x)/10000]^{1/2}$. For x = 0.5 the SE would be a maximum and have a value, SE = $\sqrt{.000025} = 0.005$ so a 50% rejection rate would be included in a 2SE interval from 0.49 to 0.51 in approximately 95% of the simulations. An x of 0.05 would have SE = $\sqrt{.00000475} = 0.002179$ and a 2SE interval from 0.045641 to 0.054358. Thus rates even as small as 5% will usually be estimated to differ from the correct value by no more than about 0.0044.

Even after Type I error rates are identified as significantly different from nominal levels and not due to chance, an additional question arises. How much deviation from the nominal level is acceptable to a given researcher? Bradley (1978) has suggested that a real error rate that differs from the intended nominal rate, α, by no more than 0.1α is negligibly non-robust. Thus, a rate of $\alpha = 0.05$ should not exceed 0.055 to be negligibly non-robust. Bradley (1978) also suggested that rates above $1.5\,\alpha$ (0.075 for $\alpha = 0.05$), should never be accepted as robust. All researchers must make their own decisions but an upper limit of 0.075 for the 0.05-level test seems a useful guideline.

Power rates require a different approach. To compare power rate for two statistical procedures requires that they have the same, or in some sense equivalent, control of Type I errors. If one procedure has true Type I error rates that never exceed the nominal level and a second procedure has true Type I error rates that never exceed one half the nominal level then both are limiting the Type I error rate to no more than the nominal level: Power rates can be expected to be higher for the first procedure but that may not always be the case.

Any uniformly, higher power rate for one of two such procedures justifies identifying it as more powerful. Higher power rates in specific conditions may guide a researcher to select a procedure based on conditions of the investigation. If power rates are uniformly higher but small then other factors such as ease of application may be considered. Einot and Gabriel (1975) used such an argument in the univariate case to support a slightly less powerful procedure. Power advantages less than 0.1 might be ignored, but advantages above 0.2 might be designated as substantial and override other considerations. Again, all researchers must make their own decisions.

McNemar's (1947) test of correlated proportions was used to test the significance of the difference between proportions as power rates in the non-null conditions. For greater efficiency the procedures were placed in order with consecutive procedures tested for power differences. The order is VPB, VM1, VM2, THL, WLR, GCR, and B-S.

Results

Type I Error Rates

Table 1 presents the Type I error rates for seven procedures with $k = 4$, equal N of 10, three values of ρ, four population distributions, and $C^2 = 1.6$. The overall maximum error rates are in bold print. Those are also the maximum error rates for the same conditions when C^2 has values 0.0, 0.8, and 1.2. Clearly, with C^2 values as high as 1.6, the error rates are well above the Bradley upper limit of 0.075. None of the procedures is robust by this criterion for that value of C^2.

The maximum error rates in Table 1 all occur for populations with an exponential distribution. This suggests that differences in skewness are more important than differences in kurtosis. Only differences in kurtosis were investigated in the previous studies (Bird & Hadzi-Pavlovic, 1983; Olson, 1974; Sheehan-Holt, 1998).

Table 2 presents summaries for N values of 10, 15 and 20 including the maximum rates for the results shown in Table 1. In every case the maximum error rate was found for the exponential population but could be for any one of the three values of ρ.

As shown in Table 2 (a) with N = 10, the $C^2 = 0.0$ condition shows all seven procedures to have a maximum Type I error rate below the nominal 0.05 level even when the maximum is taken over three values of ρ and four population distributions. When C^2 rises to 0.8, only VPB, the original testing formula for the V statistic is below the nominal level. However, VM1 and VM2 have maximum rates almost identical to the nominal level. Also, THL, WLR, and GCR satisfy the 0.075 limit to robustness. The Bonferroni-Scheffé is not robust for $C^2 \geq 0.8$.

If the $C^2 = 0.64$ definition of substantial covariance heterogeneity is accepted as suggested by Bird and Hadzi-Pavlovic (1983), the VPB combination of testing V and pairwise testing with R is robust for that condition. The same conclusion is probably justified for VM1 and VM2.

In all parts of Table 2 the Bonferroni-Scheffé, B-S, procedure has a simple, almost linear relationship between error rates and C^2. The greater the covariance heterogeneity the higher is the Type I error rate. The situation is quite different for the other six, multivariate procedures. Table 2 (b) presents results for $N = 15$. Even for $C^2 = 2.0$ the first five procedures have no more than negligible non-robustness (i.e. ≤ 0.055). GCR does exceed that limit but only for the most extreme case and is always robust (i.e. ≤ 0.075).

Table 2 (c) presents results for $N = 20$. All six multivariate procedures are conservative (i.e. rates ≤ 0.05). Even GCR is conservative and the protection of another procedure may not be needed. The greater control of Type I errors for all multivariate procedures as shown in Table 2(c) suggests that protected tests are not needed for sample sizes this large. The maximum Type I error rate for GCR is 0.0369 occurs for $C^2 = 0.8$.

Power Rates

For $N = 10$ the five protected procedures (VPB, VM1, VM2, THL, WLR) provide varying control of Type I errors for C^2 vales from 0.0 to about 0.8. The B-S procedure provides poor control in the same conditions of C^2. However, B-S represents a useful alternative provided it can be equated in Type I error control. Repeated testing of these six procedures (VPB, VM1,

Table 1: Type I Error Rates for Seven Pairwise Testing Procedures for k = 4, N = 10, $\alpha = .05$, $C^2 = 1.6$ and a True, Full-Null Hypothesis

ρ	Population	VPB	VM1	VM2	THL	WLR	GCR	B-S
0.00	Normal	.0240	.0242	.0241	.0256	.0253	.0262	.1046
	Uniform	.0327	.0333	.0333	.0347	.0340	.0348	.1193
	Exponential	.0788	.0814	.0810	.0893	.0875	.0921	**.1827**
	Double Exponential	.0206	.0208	.0207	.0225	.0219	.0229	.0786
0.71	Normal	.0279	.0284	.0284	.0300	.0296	.0311	.0844
	Uniform	.0329	.0335	.0335	.0346	.0345	.0349	.0864
	Exponential	.0792	.0814	.0814	.0904	.0886	.0927	.1142
	Double Exponential	.0236	.0238	.0237	.0253	.0248	.0256	.0745
–0.20	Normal	.0254	.0261	.0261	.0281	.0272	.0283	.1086
	Uniform	.0295	.0297	.0297	.0309	.0304	.0313	.1199
	Exponential	**.0823**	**.0855**	**.0852**	**.0943**	**.0914**	**.0977**	.1664
	Double Exponential	.0198	.0203	.0203	.0220	.0215	.0228	.0892

Notes: C^2 = measure of variance heterogeneity, ρ = correlation, VPB = *V* tested by Pillai, (1955) formula, VM1 = *V* tested by Muller (1988) Method 1, VM2 = *V* tested by Muller (1988) Method 2, THL = *T* tested by McKeon, (1974), WLR = *W* tested by Rao, (1951), GCR = *R* tested by Harris, (2001), B-S = Bonferroni-Scheffé. Pairwise testing of first six procedures done by ρ (see Harris, 2001, p. 222); Maximum value for each column in **bold**.

VM2, THL, WLR, B-S) showed that each would limit the Type I error rate to a maximum .05 in the conditions of Table 2(a) provided they were applied at the nominal rates of 0.0115, 0.0093, 0.0095, 0.0016, 0.0036 and 0.0024, respectively.

Any-pair power is defined as the probability of detecting one or more true differences between pairs of population means. Table 3 presents the any-pair power rates for the six procedures applied to the first discriminant function for data from four population distributions, $k = 4$, $N = 10$ and a diffuse noncentrality structure.

The most powerful procedure in all conditions is VM1 testing *V* with Muller's Method 1. McNemar's test showed each procedure to be significantly different from the one to the right provided the difference was at least 0.0006 or more. However, many differences are quite small. The power advantage of VM1 over the other protected *R* procedures can be seen in Table 3 to be modest. The power advantage of VM1 over VPB and

Table 2: Maximum Over Three ρ values, and Four Populations for Type I Error Rates for Seven Pairwise Testing Procedures for $k = 4$, $\alpha = .05$, C^2 = measure of variance heterogeneity, and a True, Full-Null Hypothesis

C^2	VPB	VM1	VM2	THL	WLR	GCR	B-S
(a) $N = 10$							
0.0	.0175	.0186	.0185	.0232	.0210	.0301	.0277
0.8	.0473	.0508	.0505	.0639	.0588	.0735	.1008
1.2	.0669	.0706	.0702	.0850	.0799	.0916	.1573
1.6	**.0823**	**.0855**	**.0852**	**.0943**	**.0914**	**.0977**	**.1827**
(b) $N = 15$							
0.0	.0172	.0179	.0178	.0208	.0196	.0265	.0241
0.8	.0421	.0425	.0425	.0467	.0452	.0507	.0936
1.2	.0509	.0511	.0510	**.0542**	.0532	**.0565**	.1427
1.6	.0524	.0525	.0525	.0533	.0531	.0537	.1736
2.0	**.0534**	**.0534**	**.0534**	.0534	**.0534**	.0535	**.2053**
(c) $N = 20$							
0.0	.0201	.0208	.0207	.0228	.0216	.0292	.0237
0.8	**.0328**	**.0334**	**.0333**	**.0349**	**.0342**	**.0369**	.0930
1.2	.0325	.0327	.0327	.0332	.0331	.0336	.1285
1.6	.0322	.0322	.0322	.0323	.0323	.0323	.1533
2.0	.0297	.0297	.0297	.0297	.0297	.0297	**.1882**

Notes: VPB = *V* tested by Pillai, (1955) formula, VM1 = *V* tested by Muller (1988) Method 1, VM2 = *V* tested by Muller (1988) Method 2, THL = *T* tested by McKeon, (1974), WLR = *W* tested by Rao, (1951), GCR = *R* tested by Harris, (2001), B-S = Bonferroni-Scheffé; Maximum value for each column in **bold**.

VM2 is always less than 0.01. The power advantage of VN1 over WLR is always less than 0.06. The greatest power advantage of VM1 over any protected *R* procedure is over THL but is always less than 0.15.

The power advantage of VM1 over B-S can be quite large. For normal populations the maximum is 0.4744 (= 0.6712 – 0.1968). For the other distributions the maximum power advantages are 0.4750 (= 0.6559 – 0.1809) for uniform distributions, 0.4453 (= 0.7514 – 0.3061) for exponential distributions, and 0.4652 (= 0.7062 – 0.2410) for double exponential distributions.

The maximum power advantages of VM1 over B-S for diffuse noncentrality structures and $C^2 = 0$ (i.e. homogeneous covariances) are shown in Table 4(a) for each of the four population distributions and three values of ρ. The power advantages vary from 0.4453 to 0.8896.

The same conditions reported in Table 3 were investigated for a diffuse noncentrality structure but $C^2 = 1.6$. The maximum power advantages of VM1 over B-S for a diffuse

Table 3: Any-Pair Power of Five Procedures on the First Discriminant Function and B-S for N = 10, $\alpha = .05$, Four Distributions, A Diffuse Non-centrality Structure and Four Non-centrality Values and $C^2 = 0.0$

Population	δ^2	VPB	VM1	VM2	THL	WLR	B-S
Normal	30.0	.6679	**.6712**	.6694	.5478	.6366	.1968
	24.3	.5233	**.5303**	.5260	.3909	.4797	.1277
	19.2	.3760	**.3829**	.3775	.2425	.3275	.0733
	14.7	.2537	**.2610**	.2558	.1436	.2072	.0453
Uniform	30.0	.6526	**.6559**	.6542	.5275	.6172	.1809
	24.3	.4983	**.5038**	.5008	.3678	.4558	.1150
	19.2	.3536	**.3603**	.3560	.2271	.3073	.0672
	14.7	.2277	**.2354**	.2308	.1256	.1886	.0388
Exponential	30.0	.7479	**.7514**	.7486	.6434	.7214	.3061
	24.3	.6000	**.6048**	.6026	.4697	.5588	.1907
	19.2	.4580	**.4637**	.4602	.3143	.4054	.1169
	14.7	.3117	**.3196**	.3151	.1820	.2612	.0575
Double Exponential	30.0	.7028	**.7062**	.7044	.5970	.6767	.2410
	24.3	.5592	**.5650**	.5615	.4249	.5141	.1561
	19.2	.4072	**.4145**	.4093	.2704	.3594	.0874
	14.7	.2728	**.2788**	.2755	.1627	.2314	.0503

Notes: VPB = *V* tested by Pillai, (1955) formula, VM1 = *V* tested by Muller (1988) Method 1, VM2 = *V* tested by Muller (1988) Method 2, THL = *T* tested by McKeon, (1974), WLR = *W* tested by Rao, (1951), GCR = *R* tested by Harris, (2001), B-S = Bonferroni-Scheffé; Maximum value for each row in **bold**.

noncentrality structures are shown in Table 4(b) for each of the four population distributions and three values of ρ. The power advantages vary for 0.2238 to 0.7288.

The same conditions reported in Table 3 and Table 4(a) were investigated for a concentrated noncentrality structure where group differences existed only along a single dimension. The maximum power advantages of VM1 over B-S for a concentrated noncentrality structures are shown in Table 4(c) for each of the four population distributions and three values of ρ. The power advantages vary from –0.1454 to 0.5335. Of course, the negative advantage means that B-S has a power advantage over VM1 as high as 0.1454. This occurs only for ρ = 0.71 but for all four population distributions.

The same conditions reported in Table 4(b) were investigated for a concentrated noncentrality structure where group differences existed only along a single dimension. The maximum power advantages of VM1 over B-S for concentrated noncentrality structures are shown in Table 4(d) for each of the four population distributions and three values of ρ. The power advantages vary for –0.4019 to 0.4827. Again the negative advantage means

Table 4: Any-Pair Power Advantage of VM1 Over B-S for $k = 4$, $N = 10$, a = .05, and $C^2 = 0.0$ or 1.6

	ρ		
Population	0.0	0.71	–0.2
(a) Diffuse Noncentrality Structure with $C^2 = 0$			
Normal	.4744	.8748	.6418
Uniform	.4750	.8854	.6501
Exponential	.4453	.8696	.5997
Double Exponential	.4652	.8708	.6245
(b) Diffuse Noncentrality Structure with $C^2 = 1.6$			
Normal	.2975	.7288	.6217
Uniform	.3809	.7311	.6258
Exponential	.2238	.6543	.5781
Double Exponential.	.2920	.7259	.5974
(c) Concentrated Noncentrality Structure with $C^2 = 0$			
Normal.	.5133	-.1454	.8724
Uniform	.5335	-.1327	.8873
Exponential	.4579	-.1043	.8505
Double Exponential	.4780	-.1155	.8561
(d) Concentrated Noncentrality Structure with $C^2 = 1.6$			
Normal	.0487	-.3668	.4170
Uniform	.0484	-.4019	.4827
Exponential	.3826	-.1549	.3070
Double Exponential	.0553	-.3369	.4756

that B-S has a power advantage over VM1 as high as 0.4019. This occurs only for ρ = 0.71 and for all four population distributions.

As shown in Table 2(b), all six multivariate procedures, VPB, VM1, VM2, THL, WLR, and GCR, showed good control of Type I errors for $N = 15$. In the most extreme conditions each of these procedures has a Type I error rate slightly above the nominal level. Even GCR, with no additional multivariate test, had a maximum rate of only 0.0565. Although that exceeds Bradley's negligible nonrobustness limit of 0.055, it might be adequate for some researchers. The rates at which each of the seven procedures must be performed to limit the actual Type I error rate to the nominal 0.05 level are 0.044, 0.044, 0.044, 0.044, 0.044, 0.044, 0.0005 respectively for VPB, VM1, VM2, THL, WLR, GCR, and B-S.

Table 5 presents the power advantages of GCR over B-S for $N = 15$ just as did Table 4 for the power advantage of VM1 over B-S. In Table 5, the greater power for B-S over GCR for ρ = 0.71 with concentrated noncentrality structures occurs only for the heterogeneous covariance condition.

The power advantage of GCR over B-S for ρ = 0.0 in Table 5(d) is less than 0.1 for all populations and becomes slightly negative for exponential distributions.

Table 5: Any-Pair Power Advantage of GCR Over B-S for $k = 4$, $N = 15$, $\alpha = .05$, and $C^2 = 0.0$ or 2.0

	ρ		
Population	0.0	0.71	–0.2
(a) Diffuse Noncentrality Structure with $C^2 = 0.0$			
Normal	.6516	.8984	.7528
Uniform	.6511	.8975	.7744
Exponential	.6030	.9049	.7346
Double Exponential	.6243	.9041	.7354
(b) Diffuse Noncentrality Structure with $C^2 = 2.0$			
Normal	.4737	.8081	.5719
Uniform	.5399	.8219	.6010
Exponential	.3207	.7342	.4180
Double Exponential.	.4110	.8002	.5380
(c) Concentrated Noncentrality Structure with $C^2 = 0.0$			
Normal.	.7970	.3290	.9241
Uniform	.8205	.3448	.9288
Exponential	.8159	.3556	.9187
Double Exponential	.7827	.3264	.9284
(d) Concentrated Noncentrality Structure with $C^2 = 2.0$			
Normal	.0607	-.3958	.5498
Uniform	.0618	-.4434	.5415
Exponential	-.0304	-.1979	.3274
Double Exponential	.0584	-.3700	.5469

Table 6 presents the power advantages of GCR over B-S for $N = 20$. The conservative Type I error rejection rate GCR implies that the procedure must be applied at a lenient rate of 0.099 to limit the rate to 0.05. In contrast B-S must be applied at a rate of 0.0008. The power advantages of GCR over B-S in Table6 are similar to those of Table 5.

Conclusion

The present investigation extends the previous work of Bird and Hadzi-Pavlovic (1983) and Sheehan-Holt (1998) on follow-up tests for MANOVA to pairwise testing on the discriminant functions. As shown in Tables 1 and 2, Type I error rates can be quite high depending upon ρ (the correlation between dependent variables), the population distribution, sample size N, and especially the covariance heterogeneity, C^2.

For samples of size, $N = 10$, and only moderate covariance heterogeneity (i.e. $C^2 = 0.8$), Three protected tests, VPR, VM1, and VM2, provide good control of Type I errors even for realistic nonnormality. Even for slightly higher covariance heterogeneity (i.e. $C^2 = 1.2$), these three protected R procedures are below Bradley's (1978) 1.5 α limit for robustness.

Power comparisons in the present investigation used adjusted alpha levels so that

Table 6: Any-Pair Power Advantage of GCR Over B-S for $k = 4$, $N = 20$, $\alpha = 0.05$ and $C^2 = .0$, 0.8 or 2.0

	ρ		
Population	0.0	0.71	–0.2
(a) Diffuse Noncentrality Structure with $C^2 = 0.0$			
Normal	.7159	.9364	.8048
Uniform	.7226	.9400	.8217
Exponential	.6883	.9343	.7683
Double Exponential	.7072	.9396	.7856
(b) Diffuse Noncentrality Structure with $C^2 = 2.0$			
Normal	.5582	.8397	.6291
Uniform	.5946	.8569	.6584
Exponential	.4314	.7616	.5010
Double Exponential.	.5059	.8214	.6094
(c) Concentrated Noncentrality Structure with $C^2 = 0.0$			
Normal.	.8386	.4161	.9649
Uniform	.8349	.4313	.9674
Exponential	.8159	.4446	.9517
Double Exponential	.8258	.4099	.9590
(d) Concentrated Noncentrality Structure with $C^2 = 0.8$			
Normal	.3355	-.2174	.7411
Uniform	.333	-.2292	.7420
Exponential	.2579	-.0916	.6048
Double Exponential	.3461	-.1822	.7541

power could be compared when all methods provided the same control of Type I errors. Table 3 shows a clear advantage in power over all procedures for homogeneous covariance and diffuses noncentrality condition for VM1. However, the power advantage over VPB and VM2 is only modest. The power advantage of VM1 over the Bonferroni-Scheffé (B-S) is shown in Tables 3 and 4 to be as high as 0.8854 but can be as low as –0.1454. On balance, the protected multivariate approach of VM1 is clearly superior to the univariate approach of B-S.

As shown in Table 2(b), a minimum sample size of about 15 is sufficient for GCR to provide adequate control of Type I errors even without the addition of the alternative protection of an additional multivariate test. Table 5 shows the power advantage of GCR over B-S to range from 0.9049 to –0.3958. As was true for Table 4 results, the power advantage of B-S is almost exclusively in conditions where $\rho = 0.71$. A univariate-based follow-up is most powerful when dependent variables are highly, positively correlated.

Table 6 provides power advantages for GCR over B-S for $N = 20$. These rates range from 0.94 to –0.2174 and are similar to those in Table 5. Although B-S can be powerful even when applied at a reduced alpha level to control

Type I errors, it would still not be practical in those conditions. Continually applying a test at different alpha levels is tedious and requires a large table of appropriate alpha levels.

Discriminant functions are more difficult to interpret than are simple combinations of dependent variables. However, MANOVA may profitably be considered not just as combined dependent variables but rather a blending of several ANOVAs and factor analysis. A discriminant function can be considered an approximation to a latent variable. The correlation between each dependent variable and the discriminant function could be used to identify the latent variable just as is done in factor analysis using factor loadings.

If a new statistical package is being developed, it might be desirable to replace the traditional VPB with VM1. However, the existing VPB reported by many statistical packages such as SAS and SPSS should provide adequate results in a protected *R* test for small *N*.

Numerous additional conditions could be considered. Various patterns of correlations might have an effect. More powerful methods of pairwise testing then the Scheffé could be considered if one is willing to consider only pairwise testing. The higher rejection rates of such powerful pairwise tests are also likely to produce even higher Type I error rates. More extreme nonnormality than is considered there can be investigated.

Example

Baumann, Seifert-Kessell, and Jones (1992) report comparing three strategies for teaching reading comprehension to fourth-graders. One strategy was Think-Aloud (TA). A second strategy was Direct Reading Activity (DRA). The third was Direct Reading and Thinking Activity (DRTA). The two dependent variables were Error Detection Task (Y_1) and Degrees of Reading Power (Y_2). There were 21 students in each of the three groups. The means and standard deviations were:

TA

Y_1	Y_2
M = 7.7727	M = 43.4545
SD = 3.9271	SD = 7.8603

DRA

Y_1	Y_2
M = 6.6818	M = 42.0455
SD = 2.7669	SD = 6.6151

DRTA

Y_1	Y_2
M = 6.2273	M = 46.6364
SD = 2.0915	SD = 7.6441

Analysis in SAS produces:

Eigenvalues

	λ	θ
Root 1	.165844	.142252
Root 2	.019988	.019596

Eigenvectors

	Y_1	Y_2
Root 1	-.038037	.017307
Root 2	.027758	.008466

s = 2, m = –0.5, n = 30

Statistic	Value	P-Value
Wilks' Lambda	0.84093942	0.0286
Pillai's Trace	0.16184815	0.0284
Hotelling-Lawley Trace	0.18583147	0.0290
Roy's Greatest Root	0.16584380	0.0321

Dividing each eigenvector element by the square root of the sum of squared values for the eigenvector, convert each subjects' dependent variable scores to a score on the first discriminant function.

$$DF1 = 0.414159Y_2 - 0.910204Y_1$$

Group	1	2	3			
N	21	21	21	$MS_E = 9.0893$		
Mean	10.9223	11.3317	13.6468	Value	SS	F
Contrast 1	-1	0	1	2.7245	77.9405	8.58*
Contrast 2	-1	1	0	0.4094	1.7599	0.19
Contrast 3	0	-1	1	2.3151	56.2767	6.19

s	n	m	$\theta_{.95}$	$df_E(\theta_{.95})/(1-\theta_{.95})$	CV
2	30	–0.5	0.1287	30(0.1287)/(0.8713) =	6.73

Group 3 (DRTA) is significantly higher than Group 1 (TA) on the first discriminant function at $\alpha = 0.05$. The average, within-group correlation between Y_1 and DF1 is –0.50. The average, within-group correlation between Y_2 and DF1 is 0.54. The two, dependent variables have about the same size relationship to DF1, however, Y_1 is inversely related whereas Y_2 is directly related to DF1. Y_1 was measuring the number of errors to be detected so it is negatively related to Y_2, reading power. DF1 is a composite measure of error detection and reading power.

The three groups failed to differ significantly on either dependent variable even at $\alpha = 0.10$. A significant B-S would require group differences on at least one dependent variable to be significant at the 0.025 level.

References

Baumann, J. F., Seifert-Kessell, N., & Jones, L. A. (1992). Effect of think-aloud instruction on elementary students' comprehension monitoring abilities. *Journal of Reading*, *24*, 143-172.

Bird, K. D., & Hadzi-Pavlovic, D. (1983). Simultaneous test procedures and the choice of a test statistic in MANOVA. *Psychological Bulletin*, *93*, 167-178.

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems: Effects of inequality of variance in the one-way classification. *Annals of Mathematical Statistics*, *25*,290-302.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*, 141-152.

Einot, I., & Gabriel, K. R. (1975). A study of the powers of several methods of multiple comparisons. *Journal of the American Statistical Association*, *70*, 574-583.

Harris, R. J. (1985) Extending the GCR tables: n < 1 and n > 1000. *Multivariate Behavioral Research*, *20*, 475-481.

Harris, R. J. (2001). *A primer of multivariate statistics*. Mahwah, NJ: Lawrence Erlbaum.

McNemar, Q. (1947). Note on the sampling error of the differences between correlated proportions or percentages. *Psychometrika*, *12*, 153-157.

McKeon, J. (1974). F approximations to the distribution of Hotelling's T^2. *Biometrika*, *61*, 381-383.

Micceri, T. (1989).The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*,156-166.

Muller, K. E. (1998). A new F approximation of the Pillai-Bartlett trace under H_0, *Journal of Computational and Graphical Statistics*, *7*, 131-137.

Pillai, K. C. S. (1955). Some new test criteria in multivariate analysis. *Annals of Mathematical Statistics*, *26*, 117-121.

Rao, C. R. (1951). An asymptotic expansion of the distribution of Wilks' criterion. *Bulletin of the Institute of International Statistics*, *33*, 177-180.

Roy, S. N. (1966). Sensitivity comparisons among tests of the general linear hypotheses, *Journal of the American Statistical Association*, *61*, 415-435.

Olson, C. L. (1974). Comparative robustness of six tests in multivariate analysis of variance. *Journal of the American Statistical Association*, *69*, 894-908.

Seber, G. A. F. (1984). *Multivariate observations*. New York, NY: Wiley.

Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance (Corr: V56 p229) *Biometrika*, *40*, 87-104.

Sheehan-Holt, J. K. (1998). MANOVA simultaneous test procedures: The power and robustness of restricted multivariate contrasts. *Educational and Psychological Measurement*, *58*, 861-881.

Tadikamalla, P. R. (1980). On simulating non-normal distributions. *Psychometrika*, *45*, 273-279.