Assessment of Sampling Stability in Ecological Applications of Discriminant Analysis
Author(s): Byron K. Williams and Kimberly Titus
Source: *Ecology*, Vol. 69, No. 4 (Aug., 1988), pp. 1275-1285
Published by: Ecological Society of America
Stable URL: http://www.jstor.org/stable/1941283
Accessed: 13/03/2014 09:09

Ecology, 69(4), 1988, pp. 1275-1285
© 1988 by the Ecological Society of America

ASSESSMENT OF SAMPLING STABILITY IN ECOLOGICAL APPLICATIONS OF DISCRIMINANT ANALYSIS[1]

BYRON K. WILLIAMS[2]
United States Fish and Wildlife Service, Patuxent Wildlife Research Center, Laurel, Maryland 20708 USA

AND

KIMBERLY TITUS[3]
Virginia Cooperative Fish and Wildlife Research Unit, Department of Fisheries and Wildlife Sciences, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061 USA

Abstract. A simulation study was undertaken to assess the sampling stability of the variable loadings in linear discriminant function analysis. A factorial design was used for the factors of multivariate dimensionality, dispersion structure, configuration of group means, and sample size. A total of 32 400 discriminant analyses were conducted, based on data from simulated populations with appropriate underlying statistical distributions.

Results from the simulations suggest that minimum sample sizes must exceed multivariate dimensionality by at least a factor of three to achieve reasonable levels of stability in discriminant function loadings. However, the requisite sample size would vary with respect to each of the design factors and, especially, with the overall amount of system variation.

A review of 60 published studies and 142 individual analyses indicated that sample sizes in ecological studies often have met that requirement. However, individual group sample sizes frequently were very unequal, and checks of assumptions usually were not reported. We recommend that ecologists obtain group sample sizes that are at least three times as large as the number of variables measured.

Key words: canonical variates; classification; discriminant function analysis; sample size; stability in discriminant function loadings.

INTRODUCTION

Discriminant analysis is applicable to a wide range of ecological problems in which multiple measurements are made on samples of observations possessing an identifiable group structure. For example, an ecological application of discriminant analysis would focus on the structure of plant or animal communities indexed by geographically distinct habitats. Replicated samples in each habitat would consist of the abundances of species, and the objective would be to highlight differences in community structure (e.g., Matthews 1979, Tonn and Magnuson 1982, Culver and Beattie 1983). Conversely, discriminant analysis also could be used to highlight habitat differences separating different animal species. In this application replicated samples corresponding to each species would consist of multiple habitat measurements, and the objective would be to highlight differences in habitat use (e.g., Titus and Mosher 1981, Thompson and Gates 1982, Munro and Rounds 1985, Seagle 1985). The basic statistical features necessary for discriminant analysis are illustrated by these examples: that samples are separable into distinct groups, and each sample consists of the measurement of several attributes. Williams (1983) characterized a number of ecological applications of discriminant analysis in terms of their grouping indices and multivariate attributes.

Discrimination methods include both classification ("predictive discriminant analysis") and separatory approaches ("descriptive discriminant analysis") (Geisser 1977), with the linear combinations of descriptive discrimination known as linear discriminant functions or, more formally, canonical variates. Though predictive and separatory discrimination methods differ theoretically and operationally, they are nonetheless closely related (Williams 1982, 1983). Under assumptions described below, both approaches yield mathematically equivalent classification procedures (Kshirsager and Arseven 1975, Williams 1982).

Most ecological studies have used a descriptive approach (see, however, Rice et al. 1983 and Verner et al. 1986 for applications of predictive methods). The structure of the canonical variates is often of primary concern to ecologists. A stepwise procedure frequently is used to select variables that are useful in separating groups, and then canonical transformations of these

[1] Manuscript received 17 February 1987; revised 17 September 1987; accepted 4 November 1987.

[2] Present address: Office of Migratory Bird Management, United States Fish and Wildlife Service, 18th and C St. NW, Washington, D.C. 20240 USA.

[3] Present address: Patuxent Wildlife Research Center, United States Fish and Wildlife Service, Laurel, Maryland 20708 USA.

variables are determined. The canonical transforms are interpreted through the signs and magnitudes of the associated canonical coefficients (Green 1971, 1979, Campbell and Atchley 1981, Williams 1981, 1983) and by means of their correlations with the original variables (e.g., Anderson and Shugart 1974, Reinert 1984*a*, *b*). The observations usually are plotted on the corresponding canonical axes, and the resulting display is analyzed for structure (Tatsuoka 1970, Green 1979, Campbell and Atchley 1981).

In practical applications the canonical coefficients must be estimated from available data, the amount of which may be "small" relative to multivariate dimensionality. Thus the canonical variates are characterized by substantial, although largely unstudied, variability (Neff and Marcus 1980). Few investigations have addressed directly the effects of sampling variability. In a Monte Carlo study, Carnes and Slade (1982) assessed the effect of group sample sizes on the group positions in canonical space. Sample sizes were found to influence the relative positions of groups in canonical space, but statistical variability of the canonical variates themselves was not examined. Morrison (1984) subsampled a set of data collected as part of a field study to examine the influence of sample sizes on discriminant functions. However, his approach did not lend itself readily to any generalizations about stability in ecological applications. Other sample size problems in the interpretation of discriminant analysis have been highlighted by Van Horne and Ford (1982) and Titus et al. (1984). Though the assessment of stability has been identified as an important problem (Harner and Whitmore 1977, Neff and Marcus 1980), we were unable to find any studies that specifically addressed the problem of sampling variability in canonical variates analysis.

In this article we present the results of a simulation study of these issues. Our objectives were to identify sources of variability in canonical variates and to determine minimum sample sizes necessary to insure adequate estimation of them. In the sections below we outline the mathematics of canonical variates analysis and discuss its application in the ecological literature. We then describe the simulation procedure used to assess stability of the canonical variates and provide a description of results from the simulations. Finally, we offer some guidelines for determining adequate sample sizes.

Canonical Variates Analysis

Data for a canonical variates analysis consist of samples of observations from two or more groups. Each observation consists of a vector $\boldsymbol{x}$ of measurements, and each has associated with it a grouping index that identifies its group membership. We assume here that within-group distributions of the measurement vectors are specified by group means $\boldsymbol{\mu}_i$, $i = 1, \ldots, g$ and common dispersion Σ, where Σ is nondegenerate. Canonical variates analysis is essentially a linear transformation of these multidimensional data, consisting of a set of canonical variates that are chosen to exhibit optimal separation of groups, as described in the Appendix. They are obtained from the solution of

$$[\mathrm{A} - \lambda\Sigma]\boldsymbol{u} = \boldsymbol{0}, \tag{1}$$

with

$$\mathrm{A} = \Sigma_i q_i(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})'$$

and $\boldsymbol{u}$ scaled so that

$$\boldsymbol{u}'\Sigma\boldsymbol{u} = 1. \tag{2}$$

For simplicity it is assumed that the means $\boldsymbol{\mu}_i$ are linearly independent. Then A has rank $k = \min[p, g - 1]$, and k canonical varieties are defined. They are expressed by the equation

$$\boldsymbol{z} = \mathrm{U}\boldsymbol{x},$$

where each row of U is a transpose solution of Eq. 1. The k-dimension vector $\boldsymbol{z}$, which has mean $\boldsymbol{\eta}_i = \mathrm{U}\boldsymbol{\mu}_i$ and identity dispersion (Williams 1982), is the canonical transform.

Three properties of the canonical variates are key in applications to ecology. First, on condition that within-group distributions are multivariate normal and within-group dispersions are equal, the canonical variates maintain posterior probabilities (Kshirsager and Arseven 1975, Williams 1982). Therefore relative distances among group means are maintained with the canonical variates, and the statistical attributes relevant to both optimal classification and separation of groups are carried over into canonical space. This is an important result for ecologists, because ecological interpretations are almost always based on patterns that are recognizable in canonical space.

Second, in almost all applications the number of groups g is substantially smaller than the number of measurement variables p. In this case the number of canonical variates is one less than the number of groups, so that the canonical variates provide a significant reduction of dimensionality. In addition, one or a few of the variates often are sufficient to exhibit group separation, so that dimensionality can be reduced yet further. The feature of dimension reduction is an important characteristic of canonical variates analysis, since it allows ecologists to fruitfully interpret patterns in canonical space.

Third, the canonical variates are "scale-invariant." Thus, if measurement x_j is standardized by x_j/σ_j, the relationship between coefficients for standardized and unstandardized data is simply

$$u_{ij}^* = \sigma_j u_{ij},$$

where u_{ij} is the canonical coefficient for (unscaled) variable x_j in the i^{th} canonical variate. This property is useful for interpreting the canonical variates in ecological applications, because ecological data frequently are scaled by standard deviations prior to a discriminant analysis (e.g., Busdosh et al. 1982).

In most practical problems the key parameters A and Σ in Eq. 1 are unknown. The usual practice is to use sample proportions and within-group sample means and dispersions to estimate A and Σ by

$$\hat{\mathrm{A}} = \Sigma_i \hat{q}_i (\hat{\boldsymbol{\mu}}_i - \hat{\boldsymbol{\mu}})(\hat{\boldsymbol{\mu}}_i - \hat{\boldsymbol{\mu}})', \qquad (3)$$

and

$$\hat{\Sigma} = \Sigma_i \Sigma_j (\boldsymbol{x}_{ij} - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{x}_{ij} - \boldsymbol{\mu}_i)'/(N - g), \qquad (4)$$

where

$$\begin{aligned} N &= \Sigma_i n_i, \\ \hat{q}_i &= n_i/N, \\ \hat{\mu}_i &= \Sigma_j \boldsymbol{x}_{ij}/n_i, \end{aligned}$$

and

$$\hat{\mu} = \sum\nolimits_i \hat{q}_i \hat{\boldsymbol{\mu}}_i.$$

Substitution of Â and $\hat{\Sigma}$ in Eq. 1 yields the standard computing procedures for discriminant analysis (Seal 1964, Jennrich 1977, Williams 1983). The resulting canonical variates are therefore random, inheriting distributions from the sample on which they are based. Though some work has been done on the distribution of principal components, left largely unanswered are questions concerning the statistical properties of sample-based canonical coefficients. In particular, the important question of small sample behaviors has not been adequately examined.

Literature Review

We conducted a literature search for discriminant analysis in eight journals (*American Journal of Botany, Auk, Copeia, Ecology, Journal of Ecology, Journal of Mammalogy, Journal of Wildlife Management, Oikos*). The *American Journal of Botany* was searched for the years 1978–1983, *Oikos* was searched for 1980–1985, and the other six journals were searched for 1980–1984. Articles from other sources were included as they were encountered during the review. We believe this search provided a representative overview of discriminant anlaysis in ecological studies.

We reviewed a total of 60 papers and summarized 142 discriminant analyses. For each discriminant analysis we recorded the number of variables, total sample size, number of groups, ratio of total sample size to the number of variables, and whether classification results were given. These attributes could not be determined for all 142 analyses.

For 126 analyses, sample sizes varied from 18 to >3000, with a median of 104.5. Seventy-four of 142 analyses were conducted with two or three groups (mode = 2, maximum = 32 groups). The mean number of variables at the beginning of the discriminant analyses was 17.8 (SD = 17.1, n = 140), although it was not always clear how many variables might have been aggregated or eliminated prior to analysis. Seventy of 140 analyses used between 8 and 20 variables.

The ratio of the total sample size to the number of variables varied from 0.78:1 (Tonn and Magnuson 1982) to more than 50:1 (Rakocinski 1980, Ryan et al. 1984, Munro and Rounds 1985, Niemi 1985). The median ratio was 7.9:1. Seventeen of the 125 analyses had ratios less than 3:1 while 54 of these analyses had sample sizes >10 times the number of variables.

At least some aspect of classification was mentioned in 95 of 142 analyses. Studies such as Baltz and Moyle (1981), Gilmore and Gates (1985), Peterson and Gauthier (1985), and Troy (1985) mentioned the use of the BMDP jackknife classification procedure (Dixon 1983), which is useful for small sample sizes. The classification of independent data sets was reported by Holbrook 1982, Conners 1983, Howard and Larson 1985, and Parren and Capen 1985.

Stepwise procedures were utilized in 63 of 105 analyses (e.g., Thompson and Gates 1982, Squibb and Hunt 1983), and in 42 they were not used (e.g., Gotfryd and Hansell 1985). For 37 analyses we were unable to determine whether stepwise or direct discriminant analyses were conducted.

Methods

Since our literature review revealed the need to address issues concerning sample sizes, parameter effects, and stability of the canonical variates, we conducted a simulation study of these issues. The simulation consisted of the replicated generation of groups of multivariate observations, followed by the determination of sample-based canonical variates. The resulting output was examined for bias and stability in both the canonical variate coefficient and the corresponding correct classification rates. The overall procedure consisted of three main parts.

Parameterization.—First, the means, dispersions, and sample sizes were input for each group. Group-specific means

$$\boldsymbol{\mu}_i = [\mu_{i1} \ldots \mu_{ip}]$$

consisted of p distinct parameters, and within-group dispersions were assumed to be of the form

$$\Sigma = (1 - \rho)\sigma \mathrm{I} + \rho\sigma \mathbf{11}', \qquad (5)$$

where $\mathbf{1}' = [1 \ldots 1]$. The inclusion of both variance and correlation parameters allowed for the investigation of dispersion structure in both the variance and covariance terms. Sample sizes n_i also were specified, to enable us to examine the statistical effects of sample size, dispersion, system dimensionality, and configuration of means.

Data generation.—Second, n_i samples of multivariate observations were obtained for group i. Observations were based on computer-generated random samples from a standard normal distribution, with subsequent adjustment by the transformation

$$\boldsymbol{y} = \Sigma^{1/2}\boldsymbol{x} + \boldsymbol{u}_i.$$

This transformation produced samples from a multivariate normal distribution with mean $\boldsymbol{\mu}_i$ and disper-

TABLE 1. Canonical coefficients corresponding to four dispersion structures and three group means. Dispersion structures are specified by parameters for variance (σ^2) and correlation (ρ). Group means are given by $S_1 = \{\boldsymbol{u}_1, 2\boldsymbol{u}_2, 3\boldsymbol{u}_3\}$, $S_2 = \{2\boldsymbol{u}_1, 4\boldsymbol{u}_2, 6\boldsymbol{u}_3\}$, and $S_3 = \{\boldsymbol{u}_1, 4\boldsymbol{u}_2, 9\boldsymbol{u}_3\}$. Coefficients are given for three levels of system dimension (ρ).

	$\sigma^2\rho = 0$						$\sigma^2\rho = 0.5$		
	$\sigma^2 = 1$			$\sigma^2 = 2$			$\sigma^2 = 1$		
	S_1	S_2	S_3	S_1	S_2	S_3	S_1	S_2	S_3
$p = 10$	−0.101	−0.101	−0.05	−0.071	−0.071	−0.032	−0.187	−0.187	−0.149
	−0.404	−0.404	−0.25	−0.286	−0.286	−0.175	−0.640	−0.640	−0.459
	0.909	−0.909	−0.97	0.643	0.643	0.684	1.238	1.238	1.306
	0	0	0	0	0	0	−0.05	−0.05	−0.09
	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.
	0	0	0	0	0	0	−0.05	−0.05	−0.09
$p = 20$	−0.101	−0.101	−0.05	−0.071	−0.071	−0.032	−0.166	−0.166	−0.109
	−0.404	−0.404	−0.25	−0.286	−0.286	−0.175	−0.607	−0.607	−0.407
	0.909	−0.909	−0.97	0.643	0.643	0.684	1.26	1.26	1.336
	0	0	0	0	0	0	−0.027	−0.027	−0.045
	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.
	0	0	0	0	0	0	−0.027	−0.027	−0.045
$p = 30$	−0.101	−0.101	−0.05	−0.071	−0.071	−0.032	−0.159	−0.159	−0.094
	−0.404	−0.404	−0.25	−0.286	−0.286	−0.175	−0.596	−0.596	−0.389
	0.909	−0.909	−0.97	0.643	0.643	0.684	1.269	1.269	1.347
	0	0	0	0	0	0	−0.019	−0.019	−0.031
	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.
	0	0	0	0	0	0	−0.019	−0.019	−0.031

sion Σ (Graybill 1976), which subsequently were used to estimate parameters in Eqs. 3 and 4, for use in Eq. 1.

Analysis.—Third, the sample data were analyzed by means discriminant analysis program BMDP7M (Dixon 1983). This process of analyzing randomly generated samples was repeated 100 times for each parameterization and sample size specification. Preliminary runs indicated that 100 replications of the procedure were sufficient for precise specification of statistical properties, in that additional replications beyond 100 affected neither means nor variances to a recognizable degree.

A factorial structure was used for the simulations, involving multivariate dimensionality, dispersion structure, configuration of means, and sample size as design factors. Three groups were assumed throughout, and multivariate dimensions were varied over the range of 10, 20, and 30 variates. Based on field sampling procedures reported in the biology literature, a range of system dimensionality with the largest dimension three times that of the smallest was expected to express any statistical effects of dimensionality.

Three configurations of group means were used, the first group having means along three of the multivariate axes: $S_1 = \{\boldsymbol{u}_1, 2\boldsymbol{u}_2, 3\boldsymbol{u}_3\}$, where $\boldsymbol{u}_i$ is a unit vector along the i^{th} axis. Thus the means in S_1 were chosen to increase in length arithmetically along the first three axes. A second configuration simply doubled the separation of group means along these axes while retaining the same geometric configuration: $S_2 = \{2\boldsymbol{u}_1, 4\boldsymbol{u}_2, 6\boldsymbol{u}_3\}$. A third configuration increased the degree of asymmetry among means by squaring the lead coefficients in S_1: $S_3 = \{\boldsymbol{u}_1, 4\boldsymbol{u}_2, 9\boldsymbol{u}_3\}$.

The effect of these configurations on variation in the discriminant loadings was not completely predictable. For a given dispersion structure, increasing the separation among means for a given geometric configuration was expected generally to decrease variation in the loadings. An increase in asymmetry in the locations of means, for a given amount of separation, also was expected to decrease variation. However, the precise degree and pattern of these changes could not be anticipated.

The effect of dispersion was analyzed by varying both the variance and covariance terms. In each simulation all variances were identical, with values of either 1 or 2. Covariances also were all identical, with a value of 0 or 0.5. Wilks' generalized variance (Wilks 1962) for patterned matrices of the form of Eq. 5 is given by

$$\det(\Sigma) = [(1 - \rho)\sigma^2]^{p-1}[1 + (p - 1)\rho]\sigma^2 \qquad (6)$$

(Graybill 1969), with a local maximum at $\rho = 0$. Thus the generalized variance decreases with positive covariance, indicating a decrease in overall system variability and suggesting an increase in the stability of the discriminant loadings.

Finally, group sample sizes were varied from 10 to 90 observations per group, with sample sizes identical among groups for each simulation. The ratio of sample size to multivariate dimensionality therefore was var-

Table 1. Continued.

$\sigma^2\rho = 0.5$		
$\sigma^2 = 2$		
S_1	S_2	S_3
−0.144	−0.144	−0.078
−0.363	−0.363	−0.255
0.719	0.719	0.760
−0.025	−0.025	−0.043
⋮	⋮	⋮
−0.025	−0.025	−0.043
−0.095	−0.095	−0.061
−0.349	−0.349	−0.232
0.729	0.729	0.773
−0.014	−0.014	−0.024
⋮	⋮	⋮
−0.014	−0.014	−0.024
−0.091	−0.091	−0.053
−0.343	−0.343	−0.223
0.733	0.733	0.778
−0.01	−0.01	−0.017
⋮	⋮	⋮
−0.01	−0.01	−0.017

ied from about 1:1 (10 observations per group and system dimension of size 30) to 27:1 (90 observations per group and system dimension of size 10).

One hundred discriminant analyses were conducted for each combination of dimensionality, dispersion, sample size, and configuration of group means. Thus $3 \times 3 \times 4 \times 9 \times 100 = 32\,400$ analyses were conducted, involving ≈1500 h of computer time on a Callan STATCAT supermicrocomputer.

Results

Since the number of groups was limited to three, only two discriminant functions were identified in each simulation. Table 1 summarizes the canonical coefficients corresponding to each combination of parameters. Differences in canonical coefficients are seen to be associated with both structure of Σ as well as the dispersion among group means. However, these factors affect the coefficients in opposite directions. As argued in the Appendix, low levels of stochastic variation generally correspond to large values in Σ^{-1} and, through Eqs. A.1 and A.2, to large values for the canonical coefficients. But reduced amounts of stochastic variation can be produced either by small variances or by large amounts of multicolinearity among individual variables. In Table 1, for example, larger values for the canonical coefficients correspond to $\sigma^2 = 1$ than to $\sigma^2 = 2$, and larger coefficients also occur when $\rho \neq 0$. Larger values for the canonical coefficients also occur when there is greater variation in distances between means. It is argued in the Appendix that asymmetric patterns in these distances correspond both to increased dominance of the lead canonical variate and to increases in the magnitudes of their coefficients. It is also shown that the canonical variates are invariant to simple scale changes. Thus, for example, A and 2A correspond to the same canonical variates. This is seen in Table 1, wherein the structure for means for group S_2 is simply a rescaling of those in group S_1.

If one considers the objectives of discriminant analysis and the scalings involved, these patterns have an intuitive appeal. The canonical variates are optimally chosen to represent differences among groups, relative to within-group variation. Their ability to do this should increase as dispersion among group means increases and as the amount of stochastic variation decreases. That is, the canonical variates should have greater discriminating power for groups that are "far apart" relative to the underlying stochastic variation. In general, this discriminating power corresponds to the magnitude of the coefficients, which increases with separation among groups and decreases with stochastic variation.

The patterns of variation in the canonical coefficients that arose from the simulations are characterized in Figs. 1–4, which display means and standard deviations of the dominant coefficient of the lead canonical variate. Means and standard deviations are based on 100 simulations, with sample sizes for each simulation ranging from 10 samples per group to 90 samples per group. Note that the means and standard deviations corresponding to dimension 30 and sample size 10 are not shown in the figures. In this case Σ is singular, and the computing procedures could not produce the corresponding estimates.

Effect of sample size

The effect of the sample size can be seen by comparison of the plots for a given configuration of means, dispersion structure, and system dimension. As shown in each of the figures, the effect of sample size is fairly uniform across the other factors in the study. For each configuration of group means and each dispersion pattern the estimates of the dominant canonical coefficient are quite unstable for small sample sizes. However, variation in the estimates decreases rapidly with increases of sample size. The point beyond which gains in precision become marginal is specific to dispersion structure and dimensionality, but not to the configuration of means.

Multivariate dimensionality

The effect of dimensionality on coefficient stability is indicated by comparison of plots within each part of the figures. As expected, increases in dimensionality have a destabilizing effect on the coefficient estimates. Thus the standard deviations for coefficient estimates with dimension 30 are generally higher than for di-

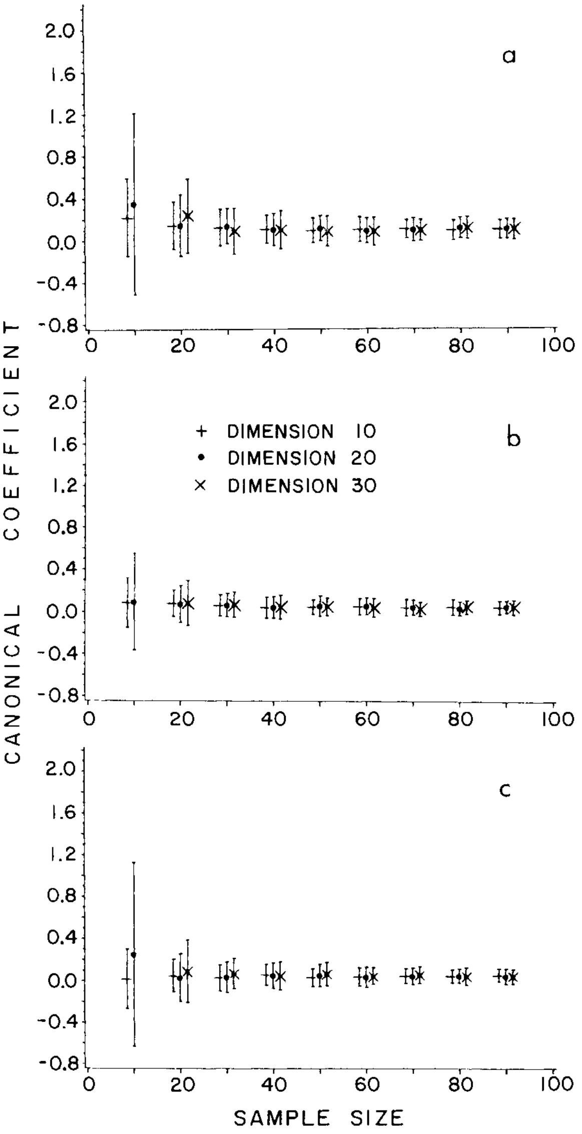

FIG. 1. Means and standard deviations for the lead canonical coefficient of the dominant canonical variate, for dispersion with $\sigma^2 = 1$ and $\rho = 0$. Part a: set S_1 of group means. Part b: set S_2 of group means. Part c: set S_3 of group means. Simulation results are shown for multivariate dimensions of 10, 20, and 30, for nine sample sizes.

mension 20, and variation for dimension 20 is greater than for dimension 10. This pattern holds for all dispersions and all configurations of means.

The effect is most clearly seen for small sample sizes, where, for example, with 10 samples per group there are large differences in variation for dimension 10 and 20. As sample sizes become large relative to dimensionality, however, coefficient variation becomes indistinguishable among the three dimensionalities.

Configuration of group means

The effect of geometric configuration of means on coefficient stability can be seen by comparison of parts a–c in each of these figures. Though the general pattern of response to changes in sample sizes is little affected by configuration, there are differences in stability for a given sample size. Configuration S_1 engenders the largest variation in estimates, whereas configuration S_3 corresponds to the least variation. However, the differences in variability are relatively minor, and suggest that neither absolute distance among group means nor the geometric relationship are significant determinants of coefficient stability.

Effect of dispersion structure

Comparison among the four figures indicates that variation in coefficients is highest for $\sigma^2 = 1$, $\rho = 0.5$,

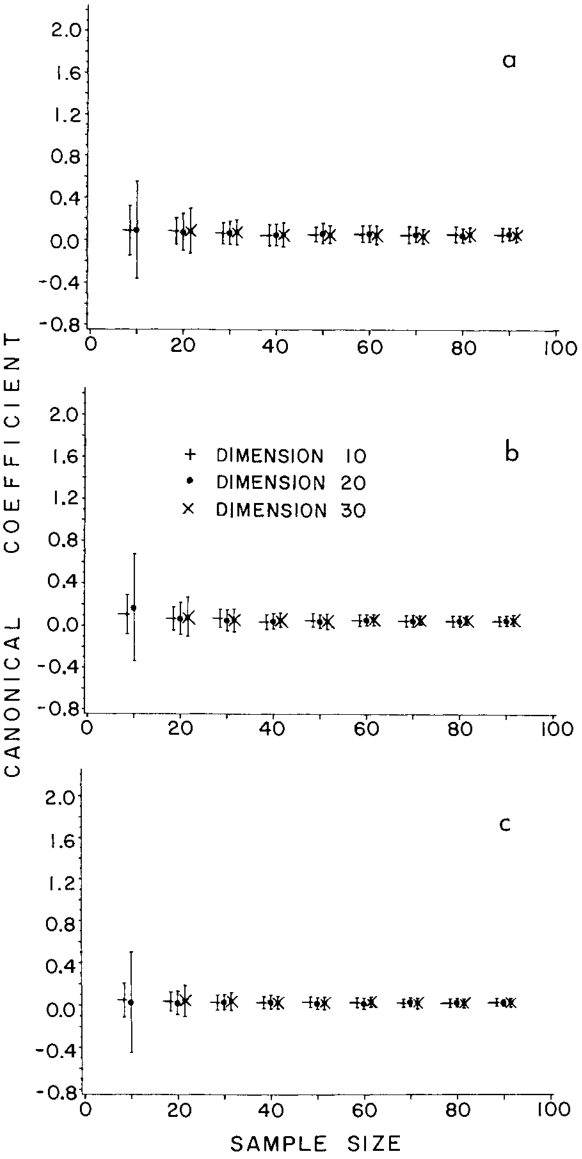

FIG. 2. Means and standard deviations for the lead canonical coefficient of the dominant canonical variate, for dispersion with $\sigma^2 = 2$ and $\rho = 0$. Parts a–c and display of simulation results as in Fig. 1.

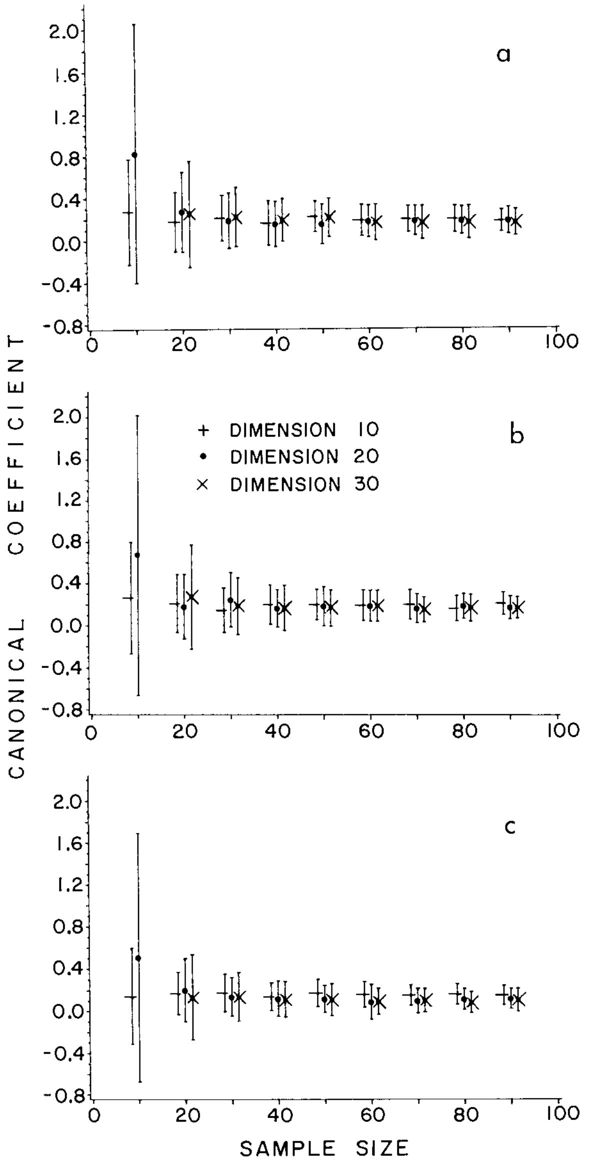

FIG. 3. Means and standard deviations for the lead canonical coefficient of the dominant canonical variate, for dispersion with $\sigma^2 = 1$ and $\rho = 0.5$. Parts a–c and display of simulation results as in Fig. 1.

lowest for $\sigma^2 = 2$, $\rho = 0$, and midrange for the other two cases. From Eq. 6 coefficient variation thus is largest when the generalized variance is smallest, and conversely, variation in the coefficients is smallest when the generalized variance is largest. This result is counterintuitive. We had anticipated that increases in coefficient variation would reflect increases in the stochastic variation of the system. Instead, stochastic variation influenced both the magnitude and the stability in the coefficients. This resulted in an association between means and variances, with high system variability corresponding to small coefficients and also to reduced levels of variation in them. Thus the coefficient of variation for any given configuration, dimension, and sample size was effectively constant for all four dispersion structures. On reflection this constancy makes sense. Since the canonical coefficients inherit their variation from sample-based estimates Σ and A in Eqs. 3 and 4, one would expect that increasing variation in the sample, and hence in Σ and A, would result in increasing variation in the coefficient. However, the coefficients are also scaled by Σ, as shown in Eq. 2. Since large variation results on average in large values of Σ, the effect of this scaling is to *reduce* the magnitude of the coefficients, and concomitantly, to reduce the amount of variation in them.

A further examination of patterns in the variability

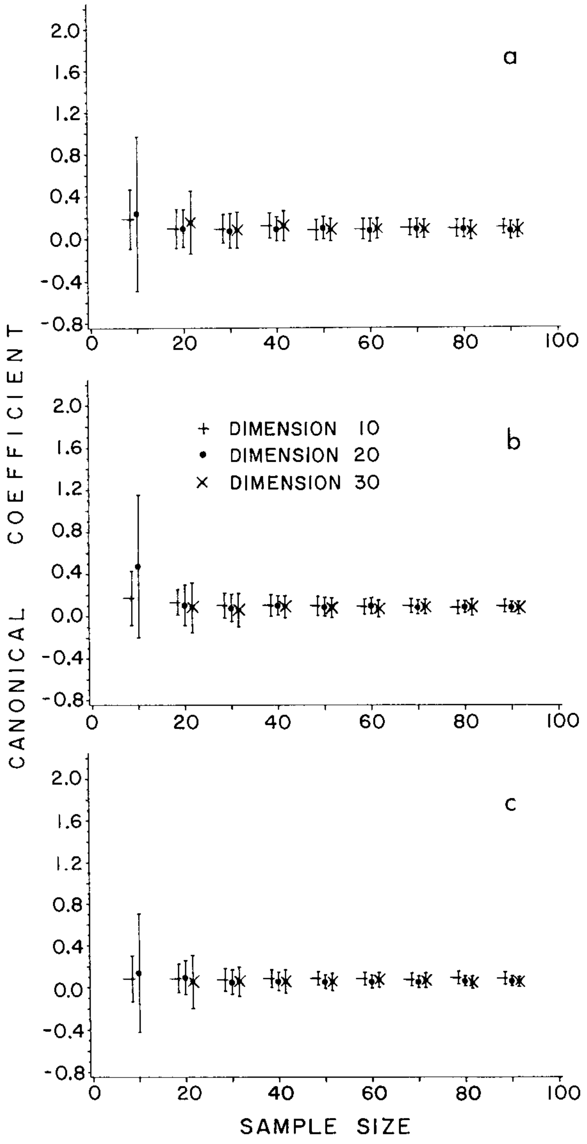

FIG. 4. Means and standard deviations for the lead canonical coefficient of the dominant canonical variate, for dispersion with $\sigma^2 = 2$ and $\sigma^2\rho = 0.5$. Parts a–c and display of simulation results as in Fig. 1.

of the canonical coefficients was conducted with analysis of variance (ANOVA). We analyzed the standard deviations of the canonical coefficients with an ANOVA model that included factors for the levels of covariance, variance, system dimension, configuration of group means, and the number of samples per group. As expected, main effects for all design factors and most of the two-way interactions were highly significant ($P <$ .01). Two-way interactions that were not significant included interactions between group configuration and either dispersion structure or system dimension. With the exception of certain three-way interactions, most of the remaining interactions among design factors were not significant. The only three-way interactions of significance were between dispersion, system dimension, and sample size. These results thus confirm the importance of each of the design factors in influencing the statistical properties of the canonical variates. They also suggest that group configurations, and to a lesser extent the dispersion structure, influence stability of the discriminant functions more or less uniformly across the levels of the other factors. This can be seen by the overall similarities in pattern displayed in each of Figs. 1–4.

Again, these results concern patterns of variation only for the lead coefficient of the dominant canonical variate. However, the scaling of coefficients shown in Eq. 2 results in sampling correlations among the coefficients, resulting in similar patterns of variation for each of them.

Discussion

The simulation yielded a number of unanticipated results. We had expected the configuration of means to have considerable effect on the stability of estimates of the canonical coefficients. Lachenbruch (1968) found that error rates, and hence required group sample sizes, decreased with increased distance among means. However, our simulation results indicated that, at least within the range of values characterized by groups S_1, S_2, and S_3, mean configuration is only of marginal importance. We had also expected that the effect of dimensionality would be greater than was found. Though the effect of dimensionality was quite dramatic for small sample sizes, this effect was quickly damped as sample sizes increased. Finally, we had expected that increases in system variability, either through increases in sample variances or decreases in covariances, would result in less stable estimates of the canonical coefficients. Instead, the simulations indicated that increases in sample dispersion lead to decreases in variation of the coefficients. This decrease in variation corresponded to a reduction in magnitudes of the canonical coefficients for increasing sample dispersion.

The substantial variation corresponding to low sample sizes is quickly reduced as sample sizes increase. Furthermore, the point at which substantial reductions in variance cease to occur appears to be roughly a constant multiple of the system dimensionality. For example, in Fig. 2a reductions in variance occur for dimension 10, for sample sizes of 10 per group or less. For dimension 20, variance reductions occur up to sample sizes between 20 and 30 per group. For dimension 30 the corresponding point appears to be ≈ 30 samples per group. With some variation the relationship holds for each configuration of means and each dispersion structure. Since the simulations all involved three groups, these results correspond to sample sizes roughly three times as large as indicated. This suggests the following sampling rule:

> For discriminant analysis of ecological systems
> with homogeneous dispersions, choose the total
> number of samples per group to be at least
> three times the number of variables to be measured.

Two points should be made about this rule. First, it imposes demands on ecologists that may be difficult to meet in some field studies. Indeed, some applications have used sample sizes that were actually less than the system dimensionality. Our simulation results appear quite unambiguous about the inferences from such applications. They suggest that unless the statistical structure of the ecological system is very simple, sampling variability is likely to be so large that no confidence can be placed in the structure of the canonical variates. Similar reliability problems with small samples sizes were found by Rencher and Larson (1980) in a Monte Carlo study of stepwise procedures. There are of course many ecological situations in which the number of samples that can be obtained is limited by such factors as budgets, availability of personnel, sample availability, or other exigencies. Under such conditions a possible approach for assessing reliability would be to use quasireplication procedures such as bootstrap or jackknife sampling (Efron 1982, Efron and Gong 1983, Lanyon 1987). In any case some form of reliability testing is advisable (e.g., Frank et al. 1965, Stauffer et al. 1985). Failure to assess reliability casts serious doubt on both the analysis and interpretation of data in the study.

The second point to note is that the rule, though conservative by ecological standards, is nonetheless an improvement over conventional thinking about sample size requirements. Within the community of practioners of discriminant analysis it is generally believed (though poorly documented) that one needs at least five times as many samples as the system dimensionality. This is presumed to follow from the large number of parameters that must be effectively estimated in multivariate systems. Such a rule may indeed be appropriate for systems with completely general covariance structures. However, for systems with patterned covariance structures not dissimilar from those used here, our simulations suggest that fewer samples may suffice in some cases. It remains for ecologists to recognize, document, and take advantage of them in their sampling plans.

Finally, we suggest that researchers perform some tests on the covariances to better understand their data. This includes tests for homoscedasticity, following the rationale of Pimentel (1979:177), and tests for multivariate skewness and kurtosis (Mardia 1974). The results displayed above and the rule that is presented are contingent on appropriate assumptions about the nature of dispersions. The effect on the structure and performance of canonical variates if these assumptions are violated to any substantial degree cannot be predicted from the results of our study.

Acknowledgments

Helpful reviews by Donald W. Sparling, Dean F. Stauffer, and two anonymous reviewers greatly improved this manuscript. Technical support was provided by Grey W. Pendelton and Edward E. Burton.

Literature Cited

Anderson, S. H., and H. H. Shugart, Jr. 1974. Habitat selection of breeding birds in an east Tennessee deciduous forest. Ecology **55**:828–837.

Baltz, D. M., and P. B. Moyle. 1981. Morphometric analysis of the tule perch (*Hysterocarpus traski*) populations in three isolated drainages. Copeia 1981:305–311.

Busdosh, M., D. M. LaVigne, and G. A. Robilliard. 1982. Habitat separation by the amphipods *Pontoporeia affinis* and *P. femorata* near Prudhoe Bay, Alaska. Oikos **39**:77–82.

Campbell, N. A., and W. R. Atchley. 1981. The geometry of canonical variates analysis. Systematic Zoology **30**:268–280.

Carnes, B. A., and N. A. Slade. 1982. Some comments on niche analysis in canonical space. Ecology **63**:888–893.

Conners, P. G. 1983. Taxonomy, distribution and evolution of golden plovers (*Pluvialis dominica* and *Pluvialis fulva*). Auk **100**:607–620.

Culver, D. C., and A. J. Beattie. 1983. Effects of ant mounds on soil chemistry and vegetation patterns in a Colorado montane meadow. Ecology **64**:485–492.

Dixon, W. J., editor. 1983. BMDP statistical software. University of California Press, Berkeley, California, USA.

Efron, B. 1982. The jackknife, the bootstrap, and other resampling plans. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, USA.

Efron, B., and G. Gong. 1983. A leisurely look at the bootstrap, the jackknife, and cross-classification. American Statistician **37**:36–48.

Frank, R. E., W. F. Massy, and D. G. Morrison. 1965. Bias in multiple discriminant analysis. Journal of Marketing Research **2**:250–258.

Geisser, S. 1977. Discrimination, allocatory and separatory, linear aspects. Pages 301–330 *in* J. Van Ryzin, editor. Classification and clustering. Academic Press, New York, New York, USA.

Gilmore, R. M., and J. E. Gates. 1985. Habitat use by the southern flying squirrel at a hemlock–northern hardwood ecotone. Journal of Wildlife Management **49**:703–710.

Gotfryd, A., and R. I. C. Hansell. 1985. The impact of observer bias on multivariate analyses of vegetation structure. Oikos **45**:223–234.

Graybill, F. A. 1969. Introduction to matrices with applications in statistics. Wadsworth, Belmont, California, USA.

———. 1976. Theory and application of the linear model. Duxbury, North Scituate, Massachusetts, USA.

Green, R. H. 1971. A multivariate statistical approach to the Hutchinsonian niche: bivalve molluscs in central Canada. Ecology **52**:543–556.

———. 1979. Sampling design and statistical principles for environmental biologists. John Wiley and Sons, New York, New York, USA.

Harner, E. J., and R. C. Whitmore. 1977. Multivariate measures of niche overlap using discriminant analysis. Theoretical Population Biology **12**:21–36.

Holbrook, S. J. 1982. Ecological inferences from mandibular morphology of *Peromyscus maniculatus*. Journal of Mammalogy **63**:399–408.

Howard, R. J., and J. S. Larson. 1985. A stream habitat classification system. Journal of Wildlife Management **49**: 19–25.

Hudlet, R., and R. Johnson. 1977. Linear discrimination and some further results on best lower dimensional representations. Pages 371–394 *in* J. Van Ryzin, editor. Classification and clustering. Academic Press, New York, New York, USA.

Jennrich, R. I. 1977. Stepwise discriminant analysis. Pages 76–95 *in* K. Enslein, A. Ralston, and H. S. Wilf, editors. Mathematical methods for digital computers. Volume 3. Statistical methods for digital computers. Wiley-Interscience, New York, New York, USA.

Kshirsager, A. M., and E. Arseven. 1975. A note on the equivalency of two discrimination procedures. American Statistician **29**:38–39.

Lachenbruch, P. A. 1968. On expected probabilities of misclassification in discriminant analysis, necessary sample size, and a relation with the multiple correlation coefficient. Biometrics **24**:823–834.

Lanyon, S. M. 1987. Jackknifing and bootstrapping: important "new" statistical techniques for ornithologists. Auk **104**:144–146.

Mardia, K. V. 1974. Applications of some measures of multivariate skewness and kurtosis in the testing of normality and robustness studies. Sankhya: The Indian Journal of Statistics **36**:115–128.

Matthews, J. A. 1979. A study of the variability of some successional and climax plant assemblage-types using multiple discriminant analysis. Journal of Ecology **67**:255–271.

Morrison, M. L. 1984. Influence of sample size on discriminant analysis of habitat use by birds. Journal of Field Ornithology **55**:330–335.

Munro, H. L.,and R. C. Rounds. 1985. Selection of artificial nest sites by five sympatric passerines. Journal of Wildlife Management **49**:264–276.

Neff, W. A., and L. F. Marcus. 1980. A survey of multivariate methods for systematics. American Museum of Natural History, New York, New York, USA.

Niemi, G. J. 1985. Patterns of morphological evolution in bird genera of new world and old world peatlands. Ecology **66**:1215–1228.

Parren, S. G., and D. E. Capen. 1985. Local distribution and coexistence of two species of *Peromyscus* in Vermont. Journal of Mammalogy **66**:36–44.

Peterson, B., and G. Gauthier. 1985. Nest site use by cavity-nesting birds of the Cariboo Parkland, British Columbia. Wilson Bulletin **97**:319–331.

Pimentel, R. A. 1979. Morphometrics: the multivariate analysis of biological data. Kendall/Hout, Dubuque, Iowa, USA.

Rakocinski, C. F. 1980. Hybridization and introgression between *Campostoma oligolepis* and *C. anomalum pullum* (Cypriniformes: Cyprinidae). Copeia 1980:584–594.

Reinert, H. K. 1984*a*. Habitat separation between sympatric snake populations. Ecology **65**:478–486.

———. 1984*b*. Habitat variation within sympatric snake populations. Ecology **65**:1673–1682.

Rencher, A. C., and S. F. Larson. 1980. Bias in Wilks' Λ in stepwise discriminant analysis. Technometrics **22**:349–356.

Rice, J., R. D. Ohmart, and B. W. Anderson. 1983. Habitat selection attributes of an avian community: a discriminant analysis investigation. Ecological Monographs **53**:263–290.

Ryan, M. R., R. B. Renken, and J. J. Dinsmore. 1984. Marbled godwit habitat selection in the northern prairie region. Journal of Wildlife Management **48**:1206–1218.

Seagle, S. W. 1985. Patterns of small mammal microhabitat utilization in cedar glade and deciduous forest habitats. Journal of Mammalogy **66**:22–35.

Seal, H. L. 1964. Multivariate statistical analysis for biologists. Methuen, London, England.

Squibb, R. C., and G. J. Hunt, Jr. 1983. A comparison of nesting-ledges used by seabirds on St. George Island. Ecology **64**:727–734.

Stauffer, D. F., E. O. Garton, and R. K. Steinhorst. 1985. A comparison of principal components from real and random data. Ecology **66**:1693–1698.

Tatsuoka, M. M. 1970. Selected topics in advanced statistics, an elementary approach. Number 6: discriminant analysis—the study of group differences. Institute for Personality and Ability Testing, Champaign, Illinois, USA.

Thompson, E. L., and J. E. Gates. 1982. Breeding pool segregation by the mole salamanders *Ambystoma jeffersonianum* and *A. maculatum* in a region of sympatry. Oikos **38**:273–279.

Titus, K., and J. A. Mosher. 1981. Nest site habitat selected by woodland hawks in the central Appalachians. Auk **98**: 270–281.

Titus, K., J. A. Mosher, and B. K. Williams. 1984. Chance-corrected classification for use in discriminant analysis: ecological applications. American Midland Naturalist **111**: 1–7.

Tonn, W. M., and J. J. Magnuson. 1982. Patterns in the species composition and richness of fish assemblages in northern Wisconsin lakes. Ecology **63**:1149–1166.

Troy, D. M. 1985. A phenetic analysis of the redpolls *Carduelis flammea flammea* and *C. hornemanni exilipes*. Auk **102**:82–96.

Van Horne, B., and R. G. Ford. 1982. Niche breadth calculation based on discriminant analysis. Ecology **63**:1172–1174.

Verner, J., M. L. Morrison, and C. J. Ralph. 1986. Wildlife 2000—modeling habitat relationships of terrestrial vertebrates. University of Wisconsin Press, Madison, Wisconsin, USA.

Wilks, S. S. 1962. Mathematical statistics. John Wiley and Sons, New York, New York, USA.

Williams, B. K. 1981. Discriminant analysis in wildlife research: theory and applications. Pages 59–71 *in* D. E. Capen, editor. The use of multivariate statistics in studies of wildlife habitat. U.S. Forest Service General Technical Report RM-87. Rocky Mountain Forest and Range Experiment Station, Fort Collins, Colorado, USA.

———. 1982. A simple demonstration of the relationship between classification and canonical variates analysis. American Statistician **36**:363–365.

———. 1983. Some observations on the use of discriminant analysis in ecology. Ecology **64**:1283–1291.

APPENDIX

In this Appendix we explore the effects of stochastic dispersion and variation among group means on the canonical coefficients. The canonical variates are derived from group means and common dispersion, by transformation of data to eliminate covariances and standardize variances to unity. This is followed by projection of transformed means onto unit-length vectors $\boldsymbol{a}$ that are oriented to maximize the mean squared projection length

$$\Sigma_i q_i[\boldsymbol{\nu}_i - \boldsymbol{\nu})'\boldsymbol{a}]^2$$

Hudlet and Johnson 1977), where q_i is the prior probability associated with group i, $\boldsymbol{\nu}_i = \Sigma^{-1/2}\boldsymbol{\mu}_i$, and $\boldsymbol{\nu} = q_i\boldsymbol{\nu}_1 + \ldots + q_g\boldsymbol{\nu}_g$. This procedure is equivalent to determination of unit-length eigenvectors $\boldsymbol{a}$ of

$$\Sigma^{-1/2}\mathrm{A}\Sigma^{-1/2'}, \tag{A.1}$$

with the canonical variates given by

$$\mathrm{z} = \boldsymbol{a}'\Sigma^{-1/2}\boldsymbol{x} \tag{A.2}$$

(Williams 1982).

From A.1 and A.2 the effect of a scalar change in stochastic dispersion can be determined directly, by replacing Σ by $k\Sigma$. The effect on the eigenstructure of A.1 is simply to scale the eigenvalues of A.1 by $1/k$, while maintaining the directional cosines of eigenvectors $\boldsymbol{a}$. From A.2 the resulting canonical variate is

$$\begin{aligned} z^* &= \boldsymbol{a}'(k\Sigma)^{-1/2}\boldsymbol{x} \\ &= 1/\sqrt{k}(\boldsymbol{a}'\Sigma^{-1/2}\boldsymbol{x}) \\ &= 1/\sqrt{k}z, \end{aligned}$$

demonstrating that an increase in overall dispersion results in proportional decreases in the magnitudes of the canonical coefficients.

The effect of increasing multicolinearity is not as straightforward, because an increase in multicolinearity results in disproportionate changes in coefficient magnitudes. In this case direction cosines as well as magnitudes are subject to alteration. However, multicolinearity corresponds in general to a reduced level of stochastic variation and thus to smaller values of det(Σ). This results in overall (albeit disproportionate) increases in coefficient magnitudes, through the influence of $\Sigma^{-1/2}$ in A.2.

To see the effect of variation in group means, it is helpful to use a simple factorization

$$\mathrm{A}^{1/2} = [\sqrt{q_i}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}) \ldots \sqrt{q_g}(\boldsymbol{\mu}_g - \boldsymbol{\mu})] \tag{A.3}$$

of

$$\mathrm{A} = \Sigma_i q_i(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})'.$$

That A.3 is a factorization of A follows immediately from

$$\begin{aligned} \mathrm{A}^{1/2}\mathrm{A}^{1/2'} &= \Sigma_i[\sqrt{q_i}(\boldsymbol{\mu}_i - \boldsymbol{\mu})][\sqrt{q_i}(\boldsymbol{\mu}_i - \boldsymbol{\mu})]' \\ &= \Sigma_i q_i(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})' \\ &= \mathrm{A}. \end{aligned}$$

From A.3 the constant scaling of group means (i.e., replacement of $\boldsymbol{\mu}_i$ by $k\boldsymbol{\mu}_i$) can be seen simply to scale $\mathrm{A}^{1/2}$ by k, thereby scaling the eigenvalues of A.1 by k^{-2} while maintaining the direction cosines of $\boldsymbol{a}$. Since the canonical variate in A.2 is influenced by A only through these direction cosines, the effect is to leave its coefficients unchanged.

On the other hand, nonconstant scaling of group means can result in complicated changes in the canonical coefficients. Some general patterns can be deduced, however, by reformulating the optimization criterion Eq. 1 to include scaling parameters. The criterion can be expressed as

$$f(\boldsymbol{a}; \boldsymbol{k}) = \Sigma_i q_i \Sigma^{-1/2}[k_i\boldsymbol{\mu}_i - \boldsymbol{\mu}(\boldsymbol{k})'\boldsymbol{a}]^2, \tag{A.4}$$

where $\boldsymbol{k}' = [k_1 \ldots k_g]$ and $\boldsymbol{\mu}(\boldsymbol{k})' = \Sigma_i q_i k_i \boldsymbol{\mu}_i$. This is a direct extension of Eq. 1 to allow for replacement of $\boldsymbol{\mu}_i$ by $k_i\boldsymbol{\mu}_i$. The canonical variates corresponding to a vector $\boldsymbol{k}$ are given by

$$\max_{\boldsymbol{a}} f(\boldsymbol{a}; \boldsymbol{k})$$

subject to

$$\boldsymbol{a}'\boldsymbol{a} = 1.$$

The values of $\boldsymbol{k}$ producing extremes in this maximization are given by

$$\max_{\boldsymbol{k}}[\max_{\boldsymbol{a}} f(\boldsymbol{a}; \boldsymbol{k})]$$

and

$$\min_{\boldsymbol{k}}[\max_{\boldsymbol{a}} f(\boldsymbol{a}; \boldsymbol{k})]$$

subject to

$$\begin{gathered} \boldsymbol{a}'\boldsymbol{a} = 1 \\ \mathbf{1}'\boldsymbol{k} = 1, \end{gathered}$$

where the constraint $\mathbf{1}'\boldsymbol{k} = 1$ reflects the fact that the canonical variates are invariant to a constant rescaling of group means. It is tedious but straightforward to show that for equal prior probabilities the least-squares criterion is minimum when $\boldsymbol{k}$ is chosen such that the scaled group means are approximately equidistant from each other in canonical space. It is maximum when $\boldsymbol{k}$ is chosen so that equal numbers of group means are "clumped" at two distinct points in canonical space. In general, the latter condition leads to a dominance of the lead canonical variate and to substantial variability among its coefficients, whereas the former condition generally corresponds to equitability among eigenvalues and to smaller, less variable canonical coefficients.