Ecology, 82(1), 2001, pp. 290–297
© 2001 by the Ecological Society of America

FITTING MULTIVARIATE MODELS TO COMMUNITY DATA: A COMMENT ON DISTANCE-BASED REDUNDANCY ANALYSIS

BRIAN H. MCARDLE AND MARTI J. ANDERSON[1]

Department of Statistics, University of Auckland, Private Bag 92019, Auckland, New Zealand

Abstract. Nonparametric multivariate analysis of ecological data using permutation tests has two main challenges: (1) to partition the variability in the data according to a complex design or model, as is often required in ecological experiments, and (2) to base the analysis on a multivariate distance measure (such as the semimetric Bray-Curtis measure) that is reasonable for ecological data sets. Previous nonparametric methods have succeeded in one or other of these areas, but not in both. A recent contribution to *Ecological Monographs* by Legendre and Anderson, called distance-based redundancy analysis (db-RDA), does achieve both. It does this by calculating principal coordinates and subsequently correcting for negative eigenvalues, if they are present, by adding a constant to squared distances. We show here that such a correction is not necessary. Partitioning can be achieved directly from the distance matrix itself, with no corrections and no eigenanalysis, even if the distance measure used is semimetric. An ecological example is given to show the differences in these statistical methods. Empirical simulations, based on parameters estimated from real ecological species abundance data, showed that db-RDA done on multifactorial designs (using the correction) does not have type 1 error consistent with the significance level chosen for the analysis (i.e., does not provide an exact test), whereas the direct method described and advocated here does.

Key words: distance measures; distance-based redundancy analysis (db-RDA); hypothesis testing; linear models; MANOVA; multivariate analysis; nonparametric; partitioning; principal coordinate analysis; redundancy analysis; semimetric measures; statistical method.

INTRODUCTION

Many ecologists are faced with the task of analyzing the simultaneous responses of many species to several factors in some experimental design. This requires a multivariate analysis, where each species is considered a variable. The traditional approach is to use parametric MANOVA. For ecological applications, however, nonparametric approaches may be preferred for three reasons. First, the assumption that counts of abundances of species conform to a multivariate normal distribution (required by MANOVA) is not generally, or even likely, to be true. Distributions of abundances of species are often highly aggregated or skewed, and there are also usually rare species that contribute many zeros to ecological data sets. Second, partitioning in traditional MANOVA implicitly uses Euclidean distances among sampling units. By partitioning, we mean attributing additive proportions of the total variability to individual factors in an experimental design. It is generally agreed that the Euclidean distance measure is not appropriate for use with ecological data of species abundances (e.g., Faith et al. 1987, Clarke 1993, Legendre and Legendre 1998). Finally, there are often more variables (species) in the system than there are sampling units (or degrees of freedom), which makes the traditional MANOVA statistics impossible to calculate.

Several nonparametric multivariate methods for use in biology, ecology, and the social sciences have been proposed (Mantel 1967, Mantel and Valand 1970, Hubert and Schultz 1976, Mielke et al. 1976, Smith et al. 1990, McArdle 1991, Clarke 1993, Pillar and Orlóci 1996). For these, a test statistic is obtained directly from distances calculated among sampling units. Thus, a distance measure other than the Euclidean distance may be used as the basis of the analysis. Also, the *P* value associated with these tests is calculated by permutation (i.e., shuffling of the sampling units across treatments and recalculating the test statistic to obtain its distribution under a true null hypothesis), thus avoiding any need to comply with the assumption of multivariate normality.

A sharp dichotomy exists among the methods proposed. First, there are those that can be based on any distance measure of choice, including semimetric measures such as the Bray-Curtis measure (Mantel 1967, Hubert and Schultz 1976, Smith et al. 1990, Clarke 1993). For these, the variability is not partitioned according to an experimental design, because it has previously been unclear how to partition a semimetric measure such as the Bray-Curtis measure. Second, there are those that can partition the total variation according to any linear analysis of variance model, but these must use metric distance measures, such as χ^2 or Euclidean distances (Mielke et al. 1976, Pillar and Orlóci 1996). In the latter category, one may include the

Manuscript received 28 June 1999; revised 24 January 2000; accepted 28 February 2000.

[1] Corresponding author. Email: mja@stat.auckland.ac.nz

traditional MANOVA statistics, such as Pillai's trace (1955), but where permutations are used instead of tabled values for obtaining probabilities.

Recently published in *Ecological Monographs,* Legendre and Anderson (1999, hereafter referred to as LA) have proposed a method called distance-based redundancy analysis (db-RDA). It has been presented as advantageous over previous methods and as especially appropriate for use in ecology for two important reasons: (1) it can be based on any distance measure of choice (including the semimetric Bray-Curtis measure), and (2) it can provide a multivariate partitioning to test any individual term in a multifactorial ANOVA experimental design. This is a significant development, because it is precisely such designs that are most often used in ecological studies, due to the inclusion of several interacting factors and/or spatial and temporal replication.

To achieve this end, db-RDA uses principal coordinate analysis (Gower 1966). Gower (1966) has shown how any distance matrix can be written as a linear form of Euclidean coordinates. Now, when a semimetric measure such as the Bray-Curtis distance measure is used (Bray and Curtis 1957), such an analysis produces both real and imaginary Euclidean coordinates (vectors), corresponding to positive and negative eigenvalues, respectively. The big dilemma faced by LA and previous workers was this: what does one do with the coordinates corresponding to negative eigenvalues, i.e., the imaginary (or complex) portion of information in the semimetric distance measure?

Others have suggested simply leaving the imaginary portion out of the analysis and using the coordinates corresponding to positive eigenvalues only (e.g., Pillar and Orlóci 1996). It is generally thought that, proportionally, not much information will be tied up in these imaginary axes, and, in any event, no ecologically meaningful interpretation can necessarily be found for them, as separated from the real axes. Thus, although ecologists generally agree that a semimetric index, namely the Bray-Curtis measure, seems to provide the most meaningful intuitive measure of dissimilarity in ecological community structure (Odum 1950, Hajdu 1981, Faith et al. 1987, Clarke 1993, Legendre and Legendre 1998), the mathematically complex portion of the information inherent in the measure has generally been ignored.

The LA approach to this dilemma was to ‘‘correct’’ for negative eigenvalues by adding a constant to the squared distances in the manner of Lingoes (1971) (called correction Method 1 in LA, see also Gower and Legendre 1986). It is not altogether clear, however, what the effect of adding such a constant might be on the test statistics and corresponding P values for the ensuing multifactorial MANOVA on corrected coordinates, although LA did provide some simulation results for constants added to Euclidean distances.

Here we give a direct method of partitioning a symmetric distance matrix according to any linear model. Our purpose is to show that multivariate models (including MANOVA) based on semimetric distances can be tested without using any correction to distances. This is so because the negative eigenvalues simply correspond to negative sums of squares. We show that db-RDA inflates the total sum of squares in the analysis. The approach of using only the axes corresponding to positive eigenvalues also inflates the total sum of squares. While the correction advocated by LA, being monotonic with respect to squared distances, does not affect P values obtained by permutation in the case of one-way ANOVA (as shown by LA for Euclidean distances, this is also true for the semimetric Bray-Curtis distances), it does affect P values in more complex models, e.g., multifactorial designs.

First we provide the necessary theory for our approach. We then reanalyze a data set presented in LA as an example to demonstrate the difference between our approach and db-RDA. Finally, we provide some simulations for two-way factorial designs to investigate how db-RDA with the correction advocated by LA, or the use of real axes only, may affect rates of rejection of a true null hypothesis (type 1 error).

Theory

The sums of squares associated with any term in any linear model (i.e., for MANOVA, MANCOVA, or multivariate regression) can be calculated directly from a distance matrix. This is because, for any centered data matrix $\mathbf{Y}_{(n\times p)}$ (of n sample units for each of p variables), the relevant information contained in the ($p \times p$) inner product matrix $\mathbf{Y}'\mathbf{Y}$ (used in classical multivariate analysis) is also contained in the ($n \times n$) outer product matrix $\mathbf{YY}'$. In addition, an outer product matrix can be obtained from any ($n \times n$) distance matrix (Gower 1966), thus allowing the analysis to be based on a distance measure of choice, including semimetric measures like Bray-Curtis. (For earlier references and discussion, see Seber [1984:238]).

Let $\mathbf{X}_{(n\times m)}$ be a model (aka design or regression) matrix, with m the number of parameters. Traditional multivariate analysis proceeds through partitioning of the ($p \times p$) total sum of squares and cross products (SSCP) matrix, which is the inner product $\mathbf{Y}'\mathbf{Y}$. The total sum of squares (S_T) is the trace, or sum of diagonal elements (sums of squares for each variable) in this matrix, which we will symbolize by tr($\mathbf{Y}'\mathbf{Y}$). Partitioning can be done according to the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\beta}$ is the matrix of model parameters, $\boldsymbol{\epsilon}$ is the matrix of errors, and we wish to test $\mathrm{H}_0 : \boldsymbol{\beta} = \mathbf{0}$. The least-squares solution for $\boldsymbol{\beta}$ is $\mathbf{B} = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{Y}$. The matrix of fitted values is $\hat{\mathbf{Y}} = \mathbf{XB} = \mathbf{HY}$, where $\mathbf{H}$ is the idempotent ‘‘hat’’ matrix $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'$ (e.g., Johnson and Wichern 1992). So the matrix of residuals is $\mathbf{R} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$. The total SSCP matrix is thus partitioned into predicted (model) and residual SSCP matrices in the following manner: $\mathbf{Y}'\mathbf{Y} = \hat{\mathbf{Y}}'\hat{\mathbf{Y}}$

$+ \mathbf{R}'\mathbf{R}$. Also, S_{T} is partitioned into hypothesis sums of squares $S_{\mathrm{H}} = \mathrm{tr}(\hat{\mathbf{Y}}'\hat{\mathbf{Y}})$ and residual sums of squares $S_{\mathrm{R}} = \mathrm{tr}(\mathbf{R}'\mathbf{R})$, as follows:

$$\mathrm{tr}(\mathbf{Y}'\mathbf{Y}) = \mathrm{tr}(\hat{\mathbf{Y}}'\hat{\mathbf{Y}}) + \mathrm{tr}(\mathbf{R}'\mathbf{R}). \tag{1}$$

An appropriate statistic to test the null hypothesis of no effect of the model parameters is a pseudo F statistic:

$$F = \frac{\mathrm{tr}(\hat{\mathbf{Y}}'\hat{\mathbf{Y}})/(m - 1)}{\mathrm{tr}(\mathbf{R}'\mathbf{R})/(n - m)}. \tag{2}$$

Note that if there is only one variable, then Eq. 2 reduces to Fisher's univariate F statistic. For a nonparametric test, the P value may be obtained as $P = P(F^{\pi} \geq F)$ where F^{π} is the value of F obtained by a random equiprobable permutation across the n units. The degrees of freedom $(m - 1)$ and $(n - m)$ are not necessary in Eq. 2 for the test by permutation, as they remain constant.

Now, the same partitioning can be achieved using outer product matrices. This is so because, for any two matrices $\mathbf{A}_{(n\times p)}$ and $\mathbf{B}_{(p\times n)}$, $\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA})$. Thus, $\mathrm{tr}(\mathbf{Y}'\mathbf{Y}) = \mathrm{tr}(\mathbf{YY}')$. The $(n \times n)$ outer product matrix contains the same information that the $(p \times p)$ inner product matrix contains, in an exact duality, so far as the trace is concerned. So, exactly the same partitioning can be achieved using outer product matrices as

$$\mathrm{tr}(\mathbf{YY}') = \mathrm{tr}(\hat{\mathbf{Y}}\hat{\mathbf{Y}}') + \mathrm{tr}(\mathbf{RR}'). \tag{3}$$

Indeed, even if all that is available is an outer product $(n \times n)$ matrix $(\mathbf{YY}')$ and one does not know $\mathbf{Y}$, partitioning is still achievable because $\hat{\mathbf{Y}}\hat{\mathbf{Y}}' = \mathbf{H}(\mathbf{YY}')\mathbf{H}$ and $\mathbf{RR}' = (\mathbf{I} - \mathbf{H})(\mathbf{YY}')(\mathbf{I} - \mathbf{H})$ (McArdle 1991).

The reason this duality is important is that an $(n \times n)$ outer product matrix, ready for partitioning, can be obtained from any $(n \times n)$ symmetric matrix of distances or dissimilarities (Gower 1966). Namely, let $\mathbf{D} = (d_{ij})$ be an $(n \times n)$ distance matrix. Let $\mathbf{A} = (a_{ij}) = (-\frac{1}{2}\, d_{ij}^2)$, then calculate Gower's centered matrix $(\mathbf{G})$ by centering the elements of $\mathbf{A}$, i.e.,

$$\mathbf{G} = \left(\mathbf{I} - \frac{1}{n}\mathbf{11}'\right)\mathbf{A}\left(\mathbf{I} - \frac{1}{n}\mathbf{11}'\right)$$

where $\mathbf{1}$ is a column of 1's of length n. Matrix $\mathbf{G}$ is then an outer product matrix that can be partitioned directly in the manner we have described. Thus, replacing $(\mathbf{YY}')$ with $\mathbf{G}$, we have $S_{\mathrm{T}} = \mathrm{tr}(\mathbf{G})$ and the pseudo F statistic is

$$F = \frac{\mathrm{tr}(\mathbf{HGH})/(m - 1)}{\mathrm{tr}[(\mathbf{I} - \mathbf{H})\mathbf{G}(\mathbf{I} - \mathbf{H})]/(n - m)}. \tag{4}$$

This can be tested using permutation. Once again, the constants $(m - 1)$ and $(n - m)$ can be dropped from Eq. 4 for the permutation test. We have included them to maintain the relationship of this statistic with Fisher's F ratio: namely, for one variable and Euclidean distances, Eq. 4 is the traditional univariate F. This is a type III statistic (sensu Shaw and Mitchell-Olds 1993) and is therefore suitable for regression, MANCOVA, MANOVA, and unbalanced experimental designs. Partitioning for multifactorial designs is an easy extension of this procedure. An appropriate pseudo F ratio can be constructed as in Eq. 4 for terms in mixed or nested models, using relevant traces of outer product matrices.

If $\mathbf{D}$ is a matrix of Euclidean distances, then $\mathbf{G} = (\mathbf{YY}')$ and Eq. 4 is exactly equal to Eq. 2. This is how classical methods are based implicitly on the Euclidean distance measure. The great advantage of the approach using outer product matrices is that one is not restricted to using Euclidean distances as the basis of the analysis. $\mathbf{G}$ can be calculated from any symmetric distance matrix, and the test statistic in Eq. 4 can then be calculated directly. Thus, one can fit any linear model to metric or semimetric distance matrices, without the use of any corrections or the loss of possibly relevant information. Individual model parameters (or sets of them) can then be tested by permutation. (Note that we chose to consider the centered data matrix $\mathbf{Y}$ in order to avoid complexity in notation, but without any loss of generality. It is not necessary to center the raw data before calculating distances for this analysis).

As an alternative to the direct calculations that we have described, one can consider the eigenvalues and eigenvectors of $\mathbf{G}$ in a principal coordinate analysis, as in LA. Here, the sum of the eigenvalues of $\mathbf{G}$ is equal to the total sum of squares. The essential point is that this relationship holds, even if some of the eigenvalues are negative. That is, if λ_l for $l = 1, \ldots, r$ denote the r nonzero ordered eigenvalues of $\mathbf{G}$, and if there are p positive and q negative eigenvalues $(p + q = r)$, then

$$S_T = \mathrm{tr}(\mathbf{G}) = \sum_{l=1}^{r} \lambda_l = \sum_{l=1}^{p} \lambda_l - \left|\sum_{l=p+1}^{p+q} \lambda_l\right|. \tag{5}$$

When coordinates are scaled to $\sqrt{\lambda_l}$ (i.e., so that their sum of squares equals their corresponding eigenvalue, as is customary in principal coordinate analysis), then negative eigenvalues correspond to complex (imaginary) axes (Gower 1966, Legendre and Legendre 1998). So, if λ_l have corresponding scaled centered orthogonal coordinates $\boldsymbol{u}_{lj}$, $j = 1, \ldots, n$, we can let $\boldsymbol{u}_{lj} = iv_{lj}$ wherever λ_l is negative, using $i = \sqrt{-1}$ to indicate the imaginary axes. One can then separately calculate the sums of squares corresponding to real and imaginary portions of information as positive $(S_{\mathrm{T}(+)})$ and negative $(S_{\mathrm{T}(-)})$ sums of squares, respectively:

$$S_{T(+)} = \sum_{l=1}^{p} \sum_{j=1}^{n} u_{lj}^2 \quad \text{and} \quad S_{T(-)} = \sum_{l=p+1}^{p+q} \sum_{j=1}^{n} (iv_{lj})^2.$$

Then $S_{\mathrm{T}} = S_{\mathrm{T}(+)} + S_{\mathrm{T}(-)}$, or, perhaps more transparently, since $i^2 = -1$

$$S_{\mathrm{T}} = S_{\mathrm{T}(+)} - |S_{\mathrm{T}(-)}|. \tag{6}$$

Thus, the total sum of squares in the system of n points, as dictated by the distance measure, is equal to the positive sum of squares (corresponding to the real axes) minus the absolute value of the negative sum of squares (corresponding to the imaginary axes). It is not necessary, therefore, to correct for negative eigenvalues. The correction advocated by LA (i.e., adding a constant to each of the squared distances) inflates the total sum of squares. Similarly, ignoring the imaginary axes also inflates the total sum of squares.

Methods of permutation for multifactorial ANOVA and multiple regression are discussed elsewhere (Edgington 1995, Manly 1997, Gonzalez and Manly 1998, Anderson and Legendre 1999). Permutation of raw data (i.e., randomly shuffling individual sampling units across treatments) provides an exact unbiased test for the one-way case; for more complex designs, either the permutation of raw data or of residuals will provide an asymptotically unbiased test (Anderson and Legendre 1999). Permutation of raw data can be achieved by simultaneously permuting rows and columns of matrix **G**, as in a Mantel's test. Permutation of residuals under either the reduced or full model (Freedman and Lane 1983, ter Braak 1992) can be achieved by simultaneously permuting rows and columns of the ($n \times n$) matrix of residuals $(\mathbf{I} - \mathbf{H})\mathbf{G}(\mathbf{I} - \mathbf{H})$, where **H** contains the hat matrix corresponding to either the reduced or full model, respectively.

Pillar and Orlóci (1996) suggested using their Q statistic (equivalent to a sum of squares, described by them for Euclidean metric measures only) as the test statistic for a test by permutation of raw data. This is fine for the one-way case, but sums of squares cannot be used as test statistics in the case of multifactorial designs or partial tests in regression. In multifactorial ANOVA, unrestricted permutation of raw data (Manly 1997) or permutation of residuals (Freedman and Lane 1983, ter Braak 1992) will only give a correct test when the test statistic used is pivotal, like a t or F statistic (Manly 1997, Gonzalez and Manly 1998, Anderson and Legendre 1999).

Example

We provide a reanalysis of the data set concerning effects of grazing by gastropods in Australian intertidal estuarine communities used in Legendre and Anderson (1999) (LA; for further details of the study, see Anderson and Underwood [1997]). The experimental design was a two-way factorial mixed model ANOVA including the fixed effect of gastropod grazers (three treatments: grazers excluded by cages, grazers not excluded, and a control for the presence of a cage) and the random effect of three repeated trials of the experiment, with $n = 8$ experimental units per treatment combination (Table 1). These data (18 taxa, after excluding gastropods and algal species) were transformed to their fourth roots, and the Bray–Curtis distance measure was calculated between all pairs of units. Nonparametric MANOVA was done for the entire data set using one of three methods: (1) a direct analysis using sums of squared Bray–Curtis distances and the pseudo F ratio, as we have outlined, (2) distance-based redundancy analysis (db-RDA) advocated in LA, including the correction for negative eigenvalues, and (3) analysis of the data using real Euclidean axes only (corresponding to positive eigenvalues). For completeness and to demonstrate the relationship among the methods, we also provide (4) the analysis of the data using the imaginary axes only (corresponding to negative eigenvalues). Permutation of residuals under the full model was used to obtain P values in all cases (999 permutations). Results using permutation of raw data gave similar results (not shown here).

TABLE 1. Nonparametric multivariate analyses (sum of squares [SS], F statistic, and P value) of data from an experiment on the effects of grazers.

Source	SS	F	P
Method 1: Direct analysis from Bray-Curtis distances			
Grazers	1.09567	9.4491	0.008
Time	0.89766	31.6590	0.001
G × T	0.23191	4.0895	0.001
Residual	0.89315	. . .	. . .
Total	3.11839	. . .	. . .
Method 2: db-RDA with correction for negative eigenvalues			
Grazers	1.39092	3.3825	0.001
Time	1.19292	3.6863	0.001
G × T	0.82242	1.2707	0.001
Residual	10.19377	. . .	. . .
Total	13.60003	. . .	. . .
Method 3: Positive sums of squares only			
Grazers	1.11147	6.1489	0.001
Time	0.91682	17.3014	0.001
G × T	0.36152	3.4111	0.001
Residual	1.66922	. . .	. . .
Total	4.05903	. . .	. . .
Method 4: Negative sums of squares only			
Grazers	0.01580	0.2438	1.000
Time	0.01916	0.7778	0.903
G × T	0.12961	2.6303	0.001
Residual	0.77607	. . .	. . .
Total	0.94064	. . .	. . .

Note: The methods used were (1) partitioning of the sums of squares of the Bray-Curtis distances, calculated directly from the distance matrix, (2) distance-based redundancy analysis (db-RDA) with correction for negative eigenvalues, (3) partitioning using the principal coordinates associated with positive eigenvalues only (real axes), and (4) partitioning using the principal coordinates associated with negative eigenvalues only (imaginary axes).

Table 1 shows that the total sum of squares, as calculated directly from Bray–Curtis distances, can indeed be partitioned into additive components corresponding to factors in the experimental design. Second, notice that the sum of squares for each term in the model obtained directly from the distance matrix in nonparametric MANOVA method 1 is equivalent to the positive sum of squares in method 3, minus the negative sum of squares in method 4, providing direct evidence of the validity of Eq. 6. Also, the analysis using db-RDA has inflated the total sum of squares, as has the analysis based only on real axes.

The inflation of total sum of squares in db-RDA results in the pseudo F ratios for each term being very different from those obtained using the direct method (method 1). The P values obtained under permutation for db-RDA are not wildly different, however, from the P values obtained for the direct test; for these analyses, there is no change to the interpretation of results using the direct method as opposed to db-RDA (method 2).

If one chooses to consider the redundancy (sensu Gittins [1985], redundancy is the proportion of multivariate variability explained by particular factors, analogous to R^2 in multiple regression), then db-RDA will give, in general, quite different results to the direct approach. The proportion of explained variability for the full model in the direct analysis is 71% (including grazers at 35%, time at 29%, and their interaction [G × T] at 7%); whereas, for db-RDA, it is a mere 25% (with grazers at only 10%, time at 9%, and their interaction at 6%). Thus, the inflation of the total sum of squares using db-RDA can have important effects on the interpretation of results concerning variance partitioning (as in Borcard et al. [1992] or Anderson and Gribble [1998]).

It is difficult, however, to determine from such an individual example how the correction for negative eigenvalues might generally affect rates of type 1 error in multifactorial designs. We provide some simulations to investigate this.

Empirical Simulations

The proof in Appendix B of Legendre and Anderson (1999) (LA) shows that adding a constant to squared Euclidean distances results in a monotonic transformation of the pseudo F statistic. Thus, for a one-way analysis, db-RDA and the direct method (method 2 vs. method 1; see *Example*) will give the same P value under permutation, because this proof also holds for semimetric distance measures. However, db-RDA does not necessarily give the same P values as the direct method in the multifactorial case.

Simulations of multivariate ecological data sets were done. Data from a study by Connell and Anderson (1999) were used as the basis of simulations. This was a study of the effects of predation by fish on assemblages of invertebrates and algae colonizing wooden surfaces in the intertidal zone of the Port Stephens Estuary in New South Wales (Australia; see Connell and Anderson [1999] for details). For each of 21 taxa, a mean and variance were estimated from the real data, and correlations among all pairs of taxa were also estimated. New data were then generated by drawing randomly from a multivariate normal distribution with parameters set to the estimates from the real data. The values were rounded to the nearest integer, because counts of the abundances of individual taxa cannot occur as noninteger values. Also, any negative values obtained were set at zero. Thus, rare species with very small means occurred in simulated data infrequently, and data sets realistically contained many zeros. In addition, data were simulated in the same manner, but using a multivariate lognormal distribution. For this, means, variances, and correlations were estimated for the original data after a transformation of $x' = \ln(x + 1)$ was applied. Nonmetric multidimensional scaling ordinations of simulated data with real data showed that data generated in this way from either the multivariate normal or lognormal distribution gave reasonable models for the joint multivariate distribution of these taxa (not shown).

The data were generated for a two-way factorial design (factor A with two levels and factor B with two levels), where the null hypotheses of no significant main effects or interaction were true. We examined the following situations: the fixed effects model (A and B fixed), the mixed model (A fixed, B random), and the random effects model (A and B random), where the number of units per treatment combination was either $n = 5$ or 10, and where data were either left untransformed or transformed to fourth root. For each situation, we simulated 1000 data sets. For each data set, nonparametric MANOVA was done using (1) the direct method, (2) db-RDA, and (3) axes from positive eigenvalues only, with probabilities obtained from 999 random permutations under the full model (ter Braak 1992). Type 1 error was recorded as the proportion of the number of rejections of the null hypothesis out of each of the 1000 simulations at a significance level of 0.05. The entire set of simulations was also done using permutation of raw data, instead of permutation of residuals (Manly 1997).

Similar results were obtained using either method of permutation. Results using permutation of residuals are shown in Table 2. As the rejection rate has a binomial distribution, type 1 error rates falling outside the range {0.036–0.064} were considered to differ significantly from the expected value of 0.05. In general, type 1 error associated with the tests of significance of main effects or interaction was not affected for the fixed-effects model. However, with a mixed model or random-effects model and $n = 5$, db-RDA generally gave results that were too conservative for tests of main effects. This conservatism was most notable for data simulated using the multivariate lognormal distribution, which is also a more realistic model of species data (Table 2). With a larger sample size ($n = 10$), the type 1 error was, however, comparable for the three methods and did not differ significantly from 0.05. Importantly, type 1 error is not inflated by db-RDA, which is good news for the studies that have used this approach. As there are many kinds of alternatives possible in multivariate analyses, no explicit calculations of power are provided here. We suggest that, if anything,

TABLE 2. Empirical type 1 errors (proportion of rejections of a true null hypothesis) for tests of each of two main effects (A and B) and their interaction (A × B) using 1000 simulated data sets, with P values calculated from 999 permutations under the full model for each data set and an a priori significance level of 0.05.

			Factor A			Factor B			Interaction A × B		
n	Transform	Model	Dir	LA	Pos	Dir	LA	Pos	Dir	LA	Pos
Multivariate normal simulations											
5	none	fixed	0.038	0.037	0.038	0.058	0.056	0.062	0.054	0.054	0.055
		mixed	0.042	**0.031**	0.037	0.044	0.044	0.044	0.047	0.045	0.050
		random	0.043	0.040	0.043	0.044	0.040	0.041	0.042	0.043	0.042
	4th root	fixed	0.040	0.038	0.040	0.047	0.044	0.048	0.036	0.036	0.036
		mixed	0.046	**0.030**	0.039	0.045	0.043	0.044	0.048	0.047	0.051
		random	0.055	**0.029**	0.045	0.051	0.037	0.044	0.042	0.040	0.045
10	none	fixed	0.042	0.042	0.043	0.052	0.054	0.053	0.062	0.062	0.063
		mixed	0.056	0.047	0.047	0.057	0.060	0.060	0.041	0.041	0.041
		random	0.046	0.041	0.044	0.046	**0.030**	**0.034**	0.041	0.041	0.042
	4th root	fixed	0.047	0.048	0.047	0.060	0.061	0.061	0.045	0.045	0.049
		mixed	0.046	0.045	0.051	0.050	0.048	0.050	0.056	0.056	0.055
		random	0.043	0.040	0.042	0.048	0.047	0.047	0.048	0.049	0.047
Multivariate lognormal simulations											
5	none	fixed	0.045	0.044	0.043	0.043	0.043	0.044	0.043	0.044	0.043
		mixed	0.042	**0.032**	0.043	0.043	0.043	0.047	0.052	0.050	0.049
		random	0.044	**0.033**	0.041	0.038	0.038	0.041	0.046	0.046	0.046
	4th root	fixed	0.050	0.050	0.053	0.049	0.048	0.050	0.052	0.051	0.054
		mixed	0.041	**0.023**	**0.029**	0.045	0.046	0.045	0.049	0.048	0.048
		random	0.045	**0.034**	**0.032**	0.047	**0.033**	0.039	0.053	0.050	0.053
10	none	fixed	0.048	0.048	0.050	0.059	0.057	0.060	0.051	0.053	0.051
		mixed	0.048	0.044	0.046	0.055	0.057	0.055	0.052	0.052	0.052
		random	0.048	0.041	0.046	0.044	0.039	0.046	0.045	0.046	0.046
	4th root	fixed	0.050	0.050	0.053	0.062	0.064	**0.065**	0.052	0.051	0.054
		mixed	0.042	0.052	0.048	0.039	0.038	0.037	0.050	0.051	0.053
		random	0.049	0.039	0.043	0.055	**0.034**	0.039	0.048	0.045	0.048

Notes: The methods compared were (1) the direct analysis of Bray-Curtis distances (Dir), (2) the Legendre and Anderson (1999) method of distance-based redundancy analysis (db-RDA) with correction for negative eigenvalues (LA), and (3) the analysis of real axes, corresponding to positive eigenvalues (Pos). Numbers in bold indicate significant deviation from the expected value of 0.05.

both db-RDA (method 2) and the use of real axes alone (method 3) are generally more conservative tests and so may have less power to detect alternatives than the direct method (method 1).

DISCUSSION

The present results provide an improved method for the analysis of multivariate response data in complex designs (including MANOVA, MANCOVA, and multiple or partial regression) directly from a symmetric distance or dissimilarity matrix. A more complete description of the present method for nonparametric MANOVA in ecology, including reference to available software, is described elsewhere (Anderson, *in press*).

This method is an elegant and rigorous alternative to the partial Mantel test (Smouse et al. 1986), partial redundancy analysis (partial RDA, Davies and Tso 1982), or partial canonical-correspondence analysis (partial CCA, ter Braak 1988). Note also that, like CCA or RDA, an eigenvalue decomposition of matrix **HGH** gives orthogonal axes that can be readily used for constrained (i.e., canonical) ordination. Whereas partial RDA implicitly preserves Euclidean distances and partial CCA implicitly preserves χ^2 distances, the method given here provides partitioning on the basis of any symmetric distance or dissimilarity measure.

The method is a slight, but important, improvement to the method of distance-based redundancy analysis (db-RDA) given by Legendre and Anderson (1999) (LA). P values obtained using the direct method described here (Eq. 4, method 1) have correct type 1 error. The other advantages of this direct approach are that principal coordinate analysis (eigenanalysis of matrix **G**) and subsequent correction for negative eigenvalues are not necessary. Because the correction for negative eigenvalues advocated by LA is monotonic across the squared distances, db-RDA still gives a reasonable nonparametric test under permutation. Nevertheless, P values are affected in multifactorial designs using db-RDA with a correction. Simulations show that db-RDA (or the test using vectors of positive eigenvalues only) will generally be too conservative.

We do not make any statements concerning the ecological meaning per se of the negative eigenvalues. Indeed, the concept of negative sums of squares, like negative variance, is hard to grasp. We prefer to consider that an ecologist will choose a distance measure, such as the Bray-Curtis measure, because of its properties as a descriptor of multivariate ecological dissimilarities among assemblages of species. It is perhaps not surprising, after all, that a multivariate measure that intuitively encapsulates ecological information does not exactly conform to a straight-line distance, but instead contains some real and some complex information.

We also do not wish to presume that the Bray-Curtis measure is necessarily going to be the measure of choice for ecologists in all situations, even for data on abundances of species. Indeed, some metric measures, such as the χ^2 distance, may be the measure of choice for a multivariate analysis of ecological community composition. An excellent resource concerning the multitude of available coefficients of ecological similarity or dissimilarity, and their properties and uses, is found in Legendre and Legendre (1998:Chapter 7).

Provided an ecologist is satisfied that the distance measure chosen is reasonable, possessing all properties relevant for the kind of variables being investigated, then the direct analysis of these distances (method 1) follows as a logical nonparametric procedure. This contrasts with the rather complex series of analyses needed to do db-RDA. An appealing aspect of the approach outlined here is that, if only one response variable is measured and the Euclidean distance is used, then the pseudo F statistic in Eq. 4 is Fisher's univariate F statistic, which is well understood and widely used by practicing researchers. If a semimetric distance measure is to be used with multivariate data, our approach using the pseudo F statistic does not ignore complex portions of the information in the data, nor does it inflate the total sum of squares, and can be applied to any experimental design.

Acknowledgments

This research was supported by a U2000 Post-doctoral Fellowship to M. J. Anderson from the University of Sydney and funds from the Centre for Research on Ecological Impacts of Coastal Cities.

Literature Cited

Anderson, M. J. *In press.* A new method for non-parametric multivariate analysis of variance. Austral Ecology.

Anderson, M. J., and N. A. Gribble. 1998. Partitioning the variation among spatial, temporal and environmental components in a multivariate data set. Australian Journal of Ecology **23**:158–167.

Anderson, M. J., and P. Legendre. 1999. An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. Journal of Statistical Computation and Simulation **62**:271–303.

Anderson, M. J., and A. J. Underwood. 1997. Effects of gastropod grazers on recruitment and succession of an estuarine assemblage: a multivariate and univariate approach. Oecologia **109**:442–453.

Borcard, D., P. Legendre, and P. Drapeau. 1992. Partialling out the spatial component of ecological variation. Ecology **73**:1045–1055.

Bray, J. R., and J. T. Curtis. 1957. An ordination of the upland forest communities of southern Wisconsin. Ecological Monographs **27**:325–349.

Clarke, K. R. 1993. Nonparametric multivariate analyses of changes in community structure. Australian Journal of Ecology **18**:117–143.

Connell, S. D., and M. J. Anderson. 1999. Predation by fish on assemblages of intertidal epibiota: effects of predator size and patch size. Journal of Experimental Marine Biology and Ecology **241**:15–29.

Davies, P. T., and M. K.-S. Tso. 1982. Procedures for reduced-rank regression. Applied Statistics **31**:244–255.

Edgington, E. S. 1995. Randomization tests. Third edition. Marcel Dekker, New York, New York, USA.

Faith, D. P., P. R. Minchin, and L. Belbin. 1987. Compositional dissimilarity as a robust measure of ecological distance. Vegetatio **69**:57–68.

Freedman, D., and D. Lane. 1983. A nonstochastic interpretation of reported significance levels. Journal of Business and Economic Statistics **1**:292–298.

Gittins, R. 1985. Canonical analysis—a review with applications in ecology. Springer-Verlag, Berlin, Germany.

Gonzalez, L., and B. F. J. Manly. 1998. Analysis of variance by randomization with small data sets. Environmetrics **9**: 53–65.

Gower, J. C. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika **53**:325–338.

Gower, J. C., and P. Legendre. 1986. Metric and Euclidean properties of dissimilarity coefficients. Journal of Classification **3**:5–48.

Hajdu, L. J. 1981. Graphical comparison of resemblance measures in phytosociology. Vegetatio **48**:47–59.

Hubert, L., and J. Schultz. 1976. Quadratic assignment as a general data analysis strategy. British Journal of Mathematical and Statistical Psychology **29**:190–241.

Johnson, R. A., and D. W. Wichern. 1992. Applied multivariate statistical analysis. Third edition. Prentice–Hall, Englewood Cliffs, New Jersey, USA.

Legendre, P., and M. J. Anderson. 1999. Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. Ecological Monographs **69**:1–24.

Legendre, P., and L. Legendre. 1998. Numerical ecology. Second English edition. Elsevier, Amsterdam, The Netherlands.

Lingoes, J. C. 1971. Some boundary conditions for a monotone analysis of symmetric matrices. Psychometrika **36**: 195–203.

Manly, B. F. J. 1997. Randomization, bootstrap and Monte Carlo methods in biology. Second edition. Chapman & Hall, London, UK.

Mantel, N. 1967. The detection of disease clustering and a generalized regression approach. Cancer Research **27**:209–220.

Mantel, N., and R. S. Valand. 1970. A technique of nonparametric multivariate analysis. Biometrics **26**:547–558.

McArdle, B. H. 1991. Detecting and displaying impacts in biological monitoring: spatial problems and partial solutions. Pages 249–255 *in* Z. Harnos et al., editors. The XVth

International Biometrics Conference, Proceedings of Invited Papers. International Biometrics Society, Hungarian Region, Budapest.

Mielke, P. W., K. J. Berry, and E. S. Johnson. 1976. Multiresponse permutation procedures for a priori classifications. Communications in Statistics—Theory and Methods **A5**(14):1409–1424.

Odum, E. P. 1950. Bird populations of the Highlands (North Carolina) Plateau in relation to plant succession and avian invasion. Ecology **31**:587–605.

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

Pillar, V. D. P., and L. Orlóci. 1996. On randomization testing in vegetation science: multifactor comparisons of releve groups. Journal of Vegetation Science **7**:585–592.

Seber, G. A. F. 1984. Multivariate observations. John Wiley & Sons, New York, New York, USA.

Shaw, R. G., and T. Mitchell-Olds. 1993. ANOVA for unbalanced data: an overview. Ecology **74**:1638–1645.

Smith, E. P., K. W. Pontasch, and J. Cairns. 1990. Community similarity and the analysis of multispecies environmental data: a unified statistical approach. Water Research **24**:507–514.

Smouse, P. E., J. C. Long, and R. R. Sokal. 1986. Multiple regression and correlation extensions of the Mantel test of matrix correspondence. Systematic Zoology **35**:627–632.

ter Braak, C. J. F. 1988. Partial canonical correspondence analysis. Pages 551–558 *in* H. H. Bock, editor. Classification and related methods of data analysis. North-Holland, Amsterdam, The Netherlands.

ter Braak, C. J. F. 1992. Permutation versus bootstrap significance tests in multiple regression and ANOVA. Pages 79–85 *in* K.-H. Jöckel, G. Rothe, and W. Sendler, editors. Bootstrapping and related techniques. Springer-Verlag, Berlin, Germany.