Ecological Monographs, 83(4), 2013, pp. 557–574

PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: What null hypothesis are you testing?

Marti J. Anderson[1,3] and Daniel C. I. Walsh[2]

[1]*New Zealand Institute for Advanced Study (NZIAS), Massey University, Albany Campus, Private Bag 102 904, Auckland 0745 New Zealand*
[2]*Institute of Natural and Mathematical Sciences (INMS), Massey University, Albany Campus, Private Bag 102 904, Auckland 0745 New Zealand*

Abstract. ANOSIM, PERMANOVA, and the Mantel test are all resemblance-based permutation methods widely used in ecology. Here, we report the results of the first simulation study, to our knowledge, specifically designed to examine the effects of heterogeneity of multivariate dispersions on the rejection rates of these tests and on a classical MANOVA test (Pillai's trace). Increasing differences in dispersion among groups were simulated under scenarios of changing sample sizes, correlation structures, error distributions, numbers of variables, and numbers of groups for balanced and unbalanced one-way designs. The power of these tests to detect environmental impacts or natural large-scale biogeographic gradients was also compared empirically under simulations based on parameters derived from real ecological data sets.

Overall, ANOSIM and the Mantel test were very sensitive to heterogeneity in dispersions, with ANOSIM generally being more sensitive than the Mantel test. In contrast, PERMANOVA and Pillai's trace were largely unaffected by heterogeneity for balanced designs. PERMANOVA was also unaffected by differences in correlation structure, unlike Pillai's trace. For unbalanced designs, however, all of the tests were (1) too liberal when the smaller group had greater dispersion and (2) overly conservative when the larger group had greater dispersion, especially ANOSIM and the Mantel test. For simulations based on real ecological data sets, PERMANOVA was generally, but not always, more powerful than the others to detect changes in community structure, and the Mantel test was usually more powerful than ANOSIM. Both the error distributions and the resemblance measure affected results concerning power.

Differences in the underlying construction of these test statistics result in important differences in the nature of the null hypothesis they are testing, their sensitivity to heterogeneity, and their power to detect important changes in ecological communities. For balanced designs, PERMANOVA and PERMDISP can be used to rigorously identify location vs. dispersion effects, respectively, in the space of the chosen resemblance measure. ANOSIM and the Mantel test can be used as more "omnibus" tests, being sensitive to differences in location, dispersion or correlation structure among groups. Unfortunately, none of the tests (PERMANOVA, Mantel, or ANOSIM) behaved reliably for unbalanced designs in the face of heterogeneity.

Key words: ANOSIM; Bray-Curtis; community composition; dispersion; dissimilarities; homogeneity; multivariate analysis; null hypothesis; PERMANOVA; PERMDISP; permutation test; species abundances.

Introduction

The validity of classical multivariate analysis of variance (MANOVA) relies on certain assumptions, including the independence of the sample units (e.g., row vectors), the multivariate normality of errors, and the homogeneity of variance–covariance matrices among the groups (e.g., Mardia et al. 1979, Seber 1984, Rencher 1998). The classical MANOVA test statistics (i.e., Wilks' lambda, the Hotelling-Lawley trace, Pillai's trace, and Roy's largest root criterion) are designed specifically to test the null hypothesis (H_0) of no differences in the multivariate centroids (the central location, or vector of mean parameters for all variables) among the groups. These tests also require the total number of sample units (N, say) to be large relative to the number of variables (p, say), and cannot be calculated when $p > N$.

In many biological, ecological, and environmental data sets, the assumptions of MANOVA are not likely to be met (e.g., Clarke 1993, McArdle and Anderson 2001). A number of more robust methods to compare groups of multivariate sample units have been proposed and several of these have now become very widely used in ecology. They include the analysis of similarities (ANOSIM; Clarke 1993, with >3700 citations according

Manuscript received 15 November 2012; revised 15 April 2013; accepted 25 April 2013. Corresponding Editor: M. Fortin.
[3] E-mail: m.j.anderson@massey.ac.nz

to ISI's Web of Science), permutational multivariate analysis of variance (PERMANOVA; Anderson 2001, with >1600 citations; see also Pillar and Orlóci 1996, Gower and Krzanowski 1999, Legendre and Anderson 1999, McArdle and Anderson 2001), and the Mantel test (Mantel 1967, with >5200 citations; see also Mantel and Valand 1970). Although the Mantel test is usually used to compare two distance matrices, it can be used to compare groups of samples by coding a contrast of between- vs. within-group distances in a model matrix (Appendix A). The particular form of this test that will be examined in what follows is where the specific values chosen for the model matrix yield an ANOSIM test on the basis of the dissimilarities themselves, rather than on their ranks (see Appendix A).

These methods all construct ANOVA-like test statistics from a matrix of resemblances (distances, dissimilarities, or similarities) calculated among the sample units, and obtain P values using random permutations of observations among the groups, thereby assuming only exchangeability for the one-way case. Any resemblance measure may be chosen as the basis of the analysis (optionally after first transforming the data, e.g., Clarke and Green 1988, Clarke 1993) to reflect whatever qualities among the samples may be of greatest interest (e.g., Legendre and Legendre 1998, Clarke et al. 2006, Anderson et al. 2011).

The test statistics inherent in resemblance-based permutation tests were modeled to varying degrees on Fisher's F statistic used in univariate ANOVA (Snedecor 1934), specifically by contrasting some function of the between-group vs. the within-group resemblances (e.g., Mantel), their squares (e.g., PERMANOVA), or their ranks (e.g., ANOSIM). They are therefore generally used and interpreted by practitioners for detection of differences in the locations (centroids) of multivariate groups. What is not widely appreciated, however, is that they are actually testing different null hypotheses.

The null hypothesis tested by PERMANOVA is that, under the assumption of exchangeability of the sample units among the groups, H_0: "the centroids of the groups, as defined in the space of the chosen resemblance measure, are equivalent for all groups." Thus, if H_0 were true, any observed differences among the centroids in a given set of data will be similar in size to what would be obtained under random allocation of individual sample units to the groups (i.e., under permutation). In contrast, the null hypothesis for the Mantel test, as originally described, is: H_0: "there is no relationship between the inter-point distances in one distance matrix and the inter-point distances in a second distance matrix." However, when the second distance matrix contains codes that contrast between- vs. within-group distances (see Appendix A for details), then the null hypothesis (once again, under the assumption of exchangeability) becomes H_0: "the average of the within-group distances is greater than or equal to the average of the between-group distances." This is deliberately phrased as a one-tailed test here; the alternative hypothesis being that the within-group distances are smaller, on average, than the between-group distances. The null hypothesis for the ANOSIM test is closely related to this, namely H_0: "the average of the *ranks* of within-group distances is greater than or equal to the average of the *ranks* of between-group distances," where a single ranking has been done across all inter-point distances in the distance matrix and the smallest distance (highest similarity) has a rank value of 1. For both the ANOSIM test and the Mantel test (in this form), the essence of what is being tested is the degree to which there is greater clumping (smaller distances) among samples within the same group compared to that observed among samples in different groups. The null hypotheses for ANOSIM or the Mantel test are therefore more general (less specific) than the null hypothesis tested by PERMANOVA. Thus, a significant result using ANOSIM or the Mantel test could indicate that the groups differ in their location, their dispersion, or some other distributional quality, such as their degree of skewness, non-sphericity (correlation structure), or some combination of these things, any of which can make the distribution of samples within a given group distinguishable from the rest.

The relative sensitivity of resemblance-based permutation tests to detect heterogeneity of multivariate dispersions is unknown. One might expect that non-parametric tests, especially those based on ranks, would be more robust than their classical counterparts to heterogeneity. In univariate analysis, however, a permutation test for differences in means using the t statistic is not necessarily more robust to heterogeneity of variances among the groups than the classical normal-theory t test (Boik 1987, Romano 1990, Hayes 1996, Manly and Francis 2002).

The comparative robustness of the classical MANOVA test statistics to heterogeneity has been studied to some extent and, generally, Pillai's trace was found to be the most robust (Olson 1974, 1979, Stevens 1979). More recently, Torres et al. (2010) measured the size and power of PERMANOVA and a related test described by Pillar and Orlóci (1996) by simulating multivariate normal, lognormal, and uniform data in one-way and two-way crossed ANOVA designs. Their main purpose, however, was to investigate the properties of different permutation methods for multivariate data. Thus, the empirical behavior of resemblance-based permutation tests in the presence of heterogeneity of multivariate dispersions among groups, either by comparison to the classical MANOVA tests or to one another, remains virtually unexplored.

Here, we describe a simulation study done to investigate the empirical rates of rejection of H_0 (at an a priori chosen significance level of $\alpha = 0.05$) for ANOSIM, the Mantel test, PERMANOVA, and (wherever possible) classical MANOVA (Pillai's trace; see Pillai 1955), with a special focus on the effects of

heterogeneity of multivariate dispersions. More complete descriptions of the test statistics examined here and their relationships with several other related methods, such as multi-response permutation procedures (MRPP; Mielke et al. 1981), are given in detail in Appendix A. Comparisons were also done with PERMDISP, a resemblance-based permutation test focused strictly on the null hypothesis of homogeneity of multivariate dispersions (Anderson 2006). Even if centroids differ, PERMDISP explicitly tests only H_0: "the average within-group dispersion (measured by the average distance to group centroid and as defined in the space of the chosen resemblance measure), is equivalent among the groups."

Increasing differences in dispersion among groups were simulated under scenarios of changing sample sizes, correlation structures, error distributions, numbers of variables, and numbers of groups for balanced and unbalanced one-way designs. The power of these tests to detect real potential changes due to environmental impact or natural large-scale biogeographic gradients was also compared empirically under simulations based on parameters derived from real ecological data sets.

Methods

Rationale

We began with simple scenarios (multivariate normal data with small numbers of variables and analyses based on Euclidean distances) and progressed to more complex and realistic scenarios (data simulated from truncated multivariate lognormal or Poisson or negative binomial distributions based on parameters for hundreds of species estimated from real data sets). Initially, simulation scenarios were done for normal and non-normal data in Euclidean space (i.e., based on the Euclidean distance measure and without any transformation) so that known and deliberately set differences in the parameters of the original variables would correspond directly to known changes in the relative location, dispersion, and correlation structure of points in the multivariate space. These are essential first steps to understanding the behavior of the test statistics with respect to null hypotheses. Unfortunately, ecological resemblance measures, such as Bray-Curtis or Jaccard, do not maintain location and dispersion effects that may be present in the Euclidean space of the original variables (see Appendix B and Warton et al. 2012). It was also particularly important to simulate scenarios initially where the multivariate dispersions could be altered independently of the centroids (e.g., using multivariate normal distributions, or negative binomial distributions altering only the dispersion parameter).

Ultimately, however, interest lies in comparing the methods under more realistic scenarios; e.g., with non-normal high-dimensional count data, and with tests based on measures that focus on community composition, such as Bray-Curtis, chi-square, or Jaccard. The list of potential alternative hypotheses regarding changes in community structure that might be constructed to examine power is too vast to be undertaken exhaustively here. However, a suite of simulations were done to provide meaningful comparisons for certain ecological questions of very broad potential interest: namely, the effects of pollution surrounding an oilfield (Gray et al. 1990) and large-scale latitudinal changes in beta diversity (Ellingsen and Gray 2002).

Simulation methods

We used purpose-built code written in R for this study (R Development Core Team 2012). All code is provided in Supplement 1 and was checked with independent software (e.g., PRIMER v6; Clarke and Gorley 2006, Anderson et al. 2008). Simulations were designed to investigate and highlight the essential empirical characteristics of the tests and to compare and contrast their sensitivity to heterogeneity. Each set of scenarios defined in the sections below is designated as "Sim1", "Sim2", and so on, to identify their corresponding R code files in Supplement 1. A detailed outline and relevant parameters for simulation scenarios are provided in Table 1, for reference. For each individual scenario within each set, 1000 simulated data sets were generated under known parameters. For each simulated data set, the test statistic and associated P value was calculated for each of ANOSIM, Mantel, and PERMANOVA using 999 random permutations on the basis of Euclidean distances. For Pillai's trace, a P value was calculated using the classical F distribution approximation (Appendix A). The PERMDISP test for homogeneity of dispersions (Anderson 2006) was also done for each simulated data set, using distances to centroids, and with P values obtained using 999 permutations of residuals under a reduced model. The significance level to reject the null hypothesis was set a priori at $\alpha = 0.05$ in all cases, and the rejection rate of each test was calculated as the proportion of P values (out of the 1000 simulated data sets) that were less than or equal to α. Note that we deliberately do not refer to rejection rates as "Type I error" for any of these scenarios, because the tests being examined here do differ fundamentally in their underlying null hypothesis. The aim here was to uncover the relative sensitivity of the tests to specific known distributional differences between groups. Standard errors on each empirical rejection rate were calculated under the binomial distribution, using the function prop.test in R.

Simulation scenarios

Balanced designs (Sim1).—The first set of simulations examined the effect of increasing sample size and increasing degree of heterogeneity for balanced designs in Euclidean space for either multivariate normal or Poisson/negative binomial data. Multivariate normal (MVN) data for $g = 2$ groups and $p = 2$ independent variables were generated where the means in both

TABLE 1. Detailed outline of simulation scenarios conducted for the study, indicated as Sim1–Sim4.

Name	Distr.	No. variables, p	No. groups, g	Sample sizes	Variances/correlation structure
Balanced designs, $n_1 = n_2$, uncorrelated data ($\rho_1 = \rho_2 = 0$)					
Sim1a–1d	Norm	{2, 3, 5, 10}	2	$n_1 = n_2 = \{4, 6, 9, 12, 18, 24\}$	$m = \{1, 2, 5, 10\}$
Sim1e	Pois/NB	2	2	$n_1 = n_2 = 12$	$\theta_1 = 0$, $\theta_2 = \{0, 0.1, 0.4, 0.9\}$
Sim1f–h	Pois/NB	{3, 5, 10}	2	$n_1 = n_2 = 12$	$\theta_1 = 0$, $\theta_2 = \{0, 0.1\}$
Unbalanced designs, $n_2 \geq n_1$, uncorrelated data ($\rho_1 = \rho_2 = 0$)					
Sim2a–d	Norm	{2, 3, 5, 10}	2	$n_1 = \{3, 4, 6, 8, 12, 16\}$, $n_r = 2$; $n_1 = \{3, 4, 5, 6, 9, 12\}$, $n_r = 3$; $n_1 = \{3, 4, 5, 6, 8\}$, $n_r = 5$	$m = \{1, 2, 5, 10\}$; $m = \{0.5, 0.2, 0.1\}$
Sim2e	Pois/NB	2	2	$n_1 = 8$, $n_2 = 16$, $n_r = 2$	$\theta_1 = \{0.1, 0.4, 0.9\}$, $\theta_2 = 0$; $\theta_1 = 0$, $\theta_2 = \{0, 0.1, 0.4, 0.9\}$
Sim2f–h	Pois/NB	{3, 5, 10}	2	$n_1 = 8$, $n_2 = 16$, $n_r = 2$	$\theta_1 = 0.1$, $\theta_2 = 0$; $\theta_1 = 0$, $\theta_2 = 0.1$
Changes in correlation structure, constant variances ($m = 1$)					
Sim3	Norm	2	2	$n_1 = n_2 = \{4, 6, 9, 12, 18, 24\}$; $n_1 = \{3, 4, 6, 8, 12, 16\}$, $n_r = 2$; $n_1 = \{3, 4, 5, 6, 9, 12\}$, $n_r = 3$; $n_1 = \{3, 4, 5, 6, 8\}$, $n_r = 5$	$\rho_1 = 0$, $\rho_2 = \{0, 0.6, 0.9\}$; $\rho_1 = \rho_2 = \{0.6, 0.9\}$; ρ_1:$\rho_2 = \{-0.6$:$0.6, -0.9$:$0.9\}$
Changes in numbers of groups, constant total sample size (N)					
Sim4a	Norm	2	2	$n_i = 30$, $N = 60$	$\sigma_1^2 = \sigma_2^2 = 1$; $\sigma_1^2 = 1$, $\sigma_2^2 = 5$
Sim4b	Norm	2	4	$n_i = 15$, $N = 60$	$\sigma_{1,2,3,4}^2 = 1$, all small; $\sigma_{1,2,3}^2 = 1$, $\sigma_4^2 = 5$, one large; $\sigma_{1,2}^2 = 1$, $\sigma_{3,4}^2 = 5$, half large; $\sigma_1^2 = 1$, $\sigma_{2,3,4}^2 = 5$, one small
Sim4c	Norm	2	6	$n_i = 10$, $N = 60$	$\sigma_{1-6}^2 = 1$, all small; $\sigma_{1-5}^2 = 1$, $\sigma_6^2 = 5$, one large; $\sigma_{1-3}^2 = 1$, $\sigma_{4-6}^2 = 5$, half large; $\sigma_1^2 = 1$, $\sigma_{2-6}^2 = 5$, one small
Sim4d	Norm	2	10	$n_i = 6$, $N = 60$	$\sigma_{1-10}^2 = 1$, all small; $\sigma_{1-9}^2 = 1$, $\sigma_{10}^2 = 5$, one large; $\sigma_{1-7}^2 = 1$, $\sigma_{8-10}^2 = 5$; $\sigma_{1-4}^2 = 1$, $\sigma_{5-10}^2 = 5$, half large; $\sigma_{1-3}^2 = 1$, $\sigma_{4-10}^2 = 5$; $\sigma_1^2 = 1$, $\sigma_{2-10}^2 = 5$, one small
Changes in numbers of groups, constant group sample size (n_i)					
Sim4e	Norm	2	2	$n_i = 6$, $N = 12$	as in Sim4a
Sim4f	Norm	2	4	$n_i = 6$, $N = 24$	as in Sim4b
Sim4g	Norm	2	6	$n_i = 6$, $N = 36$	as in Sim4c
Sim4d	Norm	2	10	$n_i = 6$, $N = 60$	as in Sim4d

Notes: For each scenario, 1000 data sets were simulated, and P values for ANOSIM, Mantel, and PERMANOVA were obtained using 999 permutations on the basis of Euclidean distances. Distributions (Distr.) were Normal (Norm), or Poisson/Negative Binomial (Pois/NB), as specified. Variables are: p, the number of variables; g, the number of groups; N, the total number of samples, while, for the ith individual group, n_i is the the sample size; ρ_i, the correlation between variables; σ_i^2, the variance; and θ_i, the dispersion parameter for all variables in group i. For two groups, m equals the ratio of two variances (σ_2^2/σ_1^2), and n_r equals the ratio of sample sizes (n_2/n_1).

groups had a common value of $\mu = 10$ and the covariance matrix in both groups was initially

$$\Sigma_1 = \Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Heterogeneity was then introduced and gradually increased by introducing scalar multipliers, m_1 for Σ_1 and m_2 for Σ_2. For clarity in what follows, we define the ratio of these multipliers simply as $m = m_2/m_1$. Thus, as Σ_1 and Σ_2 are identity matrices, m is simply the ratio of the variances for any variable in group 2 vs. group 1 (i.e., σ_2^2/σ_1^2). A set of simulations was done for each of $m = \{1, 2, 5, 10\}$. The above set was repeated for each of several sample sizes (gradually increasing) in a balanced design for two groups: $n_1 = n_2 = \{4, 6, 9, 12, 18, 24\}$. To investigate the effect of increasing numbers of variables, all of the above was done for $p = \{2, 3, 5, 10\}$, while always maintaining a value of $\mu = 10$ for all variables, covariance values of 0 (thus, independence) among variables, and the identity matrix (variances = 1) for Σ_1, while increasing n and m, as described.

Simulations were also done on the basis of the Poisson and negative binomial (NB) distributions. For these discrete distributions, often used to model count data,

the variance is known to increase with the mean. More particularly, for the NB distribution having mean μ, the variance is $\sigma^2 = \mu + \theta\mu^2$, where θ is the aggregation parameter (also referred to as the dispersion parameter). The NB distribution reduces to the Poisson distribution when $\theta = 0$ and $\sigma^2 = \mu$. We simulated non-normal negative binomial count data for $p = \{2, 3, 5, 10\}$, and $\mu = 10$ and $\theta_1 = 0$ for all variables, while gradually introducing heterogeneity in group 2 by changing the value of θ_2 for all variables in that group to $\theta_2 = \{0, 0.1, 0.4, 0.9\}$. McArdle and Anderson (2004) have discussed how the degree of aggregation (as measured by θ) can shift in different habitats for organisms of the same species in natural systems.

Unbalanced designs (Sim2).—The second set of simulations was designed to examine the effects of heterogeneity for unbalanced designs. All of the scenarios and parameters remained as described above for Sim1, except for the sample sizes, which could differ in the two groups. We wished to distinguish the effect of a simple increase in the total sample size ($N = n_1 + n_2$) vs. a change in the degree of imbalance in the sample sizes between groups. The sample sizes were altered such that $n_2 \geq n_1$ and the ratio was $n_2/n_1 = \{1, 2, 3, 5\}$. We generated data, in each case, where the group with the larger sample size (n_2) also had the larger variance, putting $m = \{1, 2, 5, 10\}$ as in Sim1. We also simulated data where the group with the smaller sample size (n_1) had the larger variance, so putting $m = \{0.5, 0.2, 0.1\}$.

Correlation structure (Sim3).—The third set of simulations was designed to study the effects of differences in correlation structure among the variables across different groups. Here, the variances were kept constant across groups (equal to one for all variables), but the degree of difference in the correlation between variables was altered. First, we considered scenarios of two groups where the first group had no correlation between the variables ($\rho_1 = 0$ for all pairs of variables), so was spherical in shape, while the second group had an increasing degree of correlation among all variables, namely, $\rho_2 = \{0, 0.6, 0.9\}$. We then considered the situation where the two groups had a common nonzero correlation structure (i.e., $\rho_1 = \rho_2 = \{0.6, 0.9\}$), but then increased the differences in the degree and direction of correlation between the two groups, that is, the correlations between the variables for group 1 vs. group 2, respectively, were $\rho_1{:}\rho_2 = \{0{:}0, -0.6{:}0.6, -0.9{:}0.9\}$. These were done as for Sim1 and Sim2 for balanced and unbalanced designs and $p = 2$ variables.

Numbers of groups (Sim4).—The fourth set of simulations examined the effect of heterogeneity in the context of increasing numbers of a priori groups (as Sim1, Sim2, and Sim3 treated only the case of $g = 2$). Two different kinds of situations of increasing group numbers were considered. First, a fixed total number of samples (N) can be partitioned into more and more groups (increasing g), but with smaller and smaller numbers of samples per group (n). For this, we set $N = 60$ and for $g = \{2, 4, 6, 10\}$; this yielded $n = \{30, 15, 10, 6\}$, respectively. Second, the sample size per group (n) can be held constant while more groups (having the same sample size) are added, increasing both g and N. For this, we used $n = 6$ and $g = \{2, 4, 6, 10\}$, which yielded $N = \{12, 24, 36, 60\}$, respectively. Next, for each of these situations, four different kinds of heterogeneity were simulated. All of the variables within a given group either had variances of 1 (small) or 5 (large). For a given number of groups and sample size, we did simulations where: (1) all groups had small variances (equal dispersions, a baseline reference); (2) one group had a large variance and the others were small; (3) one group had a small variance and the others were large; and (4) half of the groups had large variances and half of them had small variances.

Simulations based on real data

We simulated data from two ecological data sets (referred to as "Ekofisk" and "Norwegian continental shelf"; discussed in the following sections), available as examples in the PRIMER v6 computer package (Clarke and Gorley 2006) with the PERMANOVA+ add-on (Anderson et al. 2008). A full description of the methods used for simulating data and calculating power from these data sets is given in Appendix C. Source data files and R code for both the estimation of parameters and the simulations are provided in Supplement 2. For all simulations based on real data sets, we consider the rejection rates to be empirical measures of the relative power of these tests to detect genuine differences between groups whenever any of the underlying parameters differed between those groups. We recognize that there is an infinite number of ways that simulations could be done to measure power and these simulations are not intended to be exhaustive. They do, however, allow some preliminary insights regarding the behavior of these tests with more realistic data structures and dissimilarity measures.

Ekofisk.—The first data set, exemplifying changes in species' abundances in response to pollution, comes from a study of marine soft-sediment benthic communities (173 taxa) surrounding the Ekofisk oil platform in the North Sea (Gray et al. 1990). There were 39 sites classified into four groups (A, B, C, and D) that occurred along a gradient of increasing proximity to the oil platform (Gray et al. 1990, Clarke and Gorley 2006). For each pairwise comparison of groups along the gradient (A vs. B, B vs. C, and C vs. D), power curves for each of the resemblance-based tests were generated on the basis of each of three different distance measures: Euclidean distances on $\log(y + 1)$-transformed abundances, chi-square distances, and Bray-Curtis distances on fourth-root transformed abundances. Three different distributional approaches were used to simulate abundance data, using parameters estimated from the real data sets: (1) species' values were drawn from a multivariate lognormal distribution (MVLN), with

TABLE 2. Rejection rates (out of 1000 simulations) for each of five different multivariate tests for data generated under either a multivariate normal (MVN) or Poisson/negative binomial (NB) distribution in Euclidean space for $g = 2$ groups and $p = 2$ or 10 variables, as indicated.

Test	MVN		Poisson/NB	
	$p = 2$	$p = 10$	$p = 2$	$p = 10$
a) Balanced and homogeneous ($m = 1$, $n_1 = n_2 = 12$)				
ANOSIM	0.051	0.063	0.052	0.053
Mantel	0.047	0.060	0.050	0.050
PERMANOVA	0.050	0.056	0.049	0.055
Pillai	0.053	0.042	0.046	0.044
PERMDISP	0.046	0.046	0.064	0.050
b) Balanced and heterogeneous ($m = 2$, $n_1 = n_2 = 12$)				
ANOSIM	0.129	0.335	0.104	0.306
Mantel	0.082	0.164	0.074	0.152
PERMANOVA	0.047	0.053	0.055	0.059
Pillai	0.057	0.058	0.054	0.071
PERMDISP	0.291	0.919	0.266	0.875
c) Unbalanced and homogeneous ($m = 1$, $n_1 = 8$, $n_2 = 16$)				
ANOSIM	0.040	0.044	0.041	0.057
Mantel	0.041	0.045	0.045	0.055
PERMANOVA	0.053	0.059	0.041	0.044
Pillai	0.041	0.052	0.052	0.051
PERMDISP	0.049	0.062	0.048	0.066
d) Unbalanced and heterogeneous ($m = 2$, $n_1 = 8$, $n_2 = 16$)				
ANOSIM	0.004	0.000	0.013	0.000
Mantel	0.003	0.000	0.012	0.000
PERMANOVA	0.019	0.009	0.033	0.009
Pillai	0.027	0.028	0.027	0.026
PERMDISP	0.275	0.912	0.232	0.868
e) Unbalanced and heterogeneous ($m = 0.5$, $n_1 = 8$, $n_2 = 16$)				
ANOSIM	0.348	0.913	0.352	0.889
Mantel	0.358	0.908	0.359	0.879
PERMANOVA	0.090	0.150	0.101	0.128
Pillai	0.088	0.133	0.092	0.117
PERMDISP	0.273	0.826	0.262	0.760

Notes: Five different simulation scenarios are shown here: (a) equal sample sizes and homogeneity; (b) equal sample sizes and heterogeneity; (c) unequal sample sizes and homogeneity; (d) unequal sample sizes and heterogeneity with greater dispersion in the group with more samples; and (e) unequal sample sizes and heterogeneity with greater dispersion in the group with fewer samples.

values truncated to integers; (2) species' values were drawn independently from either a Poisson or a negative binomial distribution (Poisson/NB) depending on their degree of aggregation (dispersion parameter θ), which for a given species was held constant across the groups; or (3) the same approach as (2) was used, but the value of θ was estimated separately for each group, so individual species' dispersions varied among the groups.

Norwegian continental shelf.—The second data set exemplifies changes in species' composition and beta diversity of benthic soft-sediment macrofauna (809 taxa) along a large-scale biogeographic gradient from 101 sites sampled across five areas along the Norwegian continental shelf, spanning 15° of latitude from the North Sea into the Arctic (Ellingsen and Gray 2002). We simulated presence/absence data by randomly drawing each species as a Bernoulli binary (0, 1) random variable with probability of occurrence set equal to parameters estimated from the data. Power curves for each of the resemblance-based tests were generated for pairwise comparisons between each of the areas along the continuum from south to north (i.e., 1 vs. 2, 2 vs. 3, 3 vs. 4, and 4 vs. 5) on the basis of the Jaccard resemblance measure.

RESULTS

The full set of simulation results obtained under all scenarios (Sim1–Sim4) is provided in Supplement 3. The full set of simulation results obtained on the basis of real data sets (Ekofisk and Norwegian continental shelf) is provided in Supplement 4. Key findings are summarized in the following.

Simulation scenarios

Balanced designs (Sim1).—With uncorrelated bivariate normal balanced data and two groups in Euclidean space, increasing heterogeneity of dispersions yielded substantial increases in the rejection rates for both ANOSIM and the Mantel test (Table 2, Fig. 1a). Rejection rates for both of these tests increased with

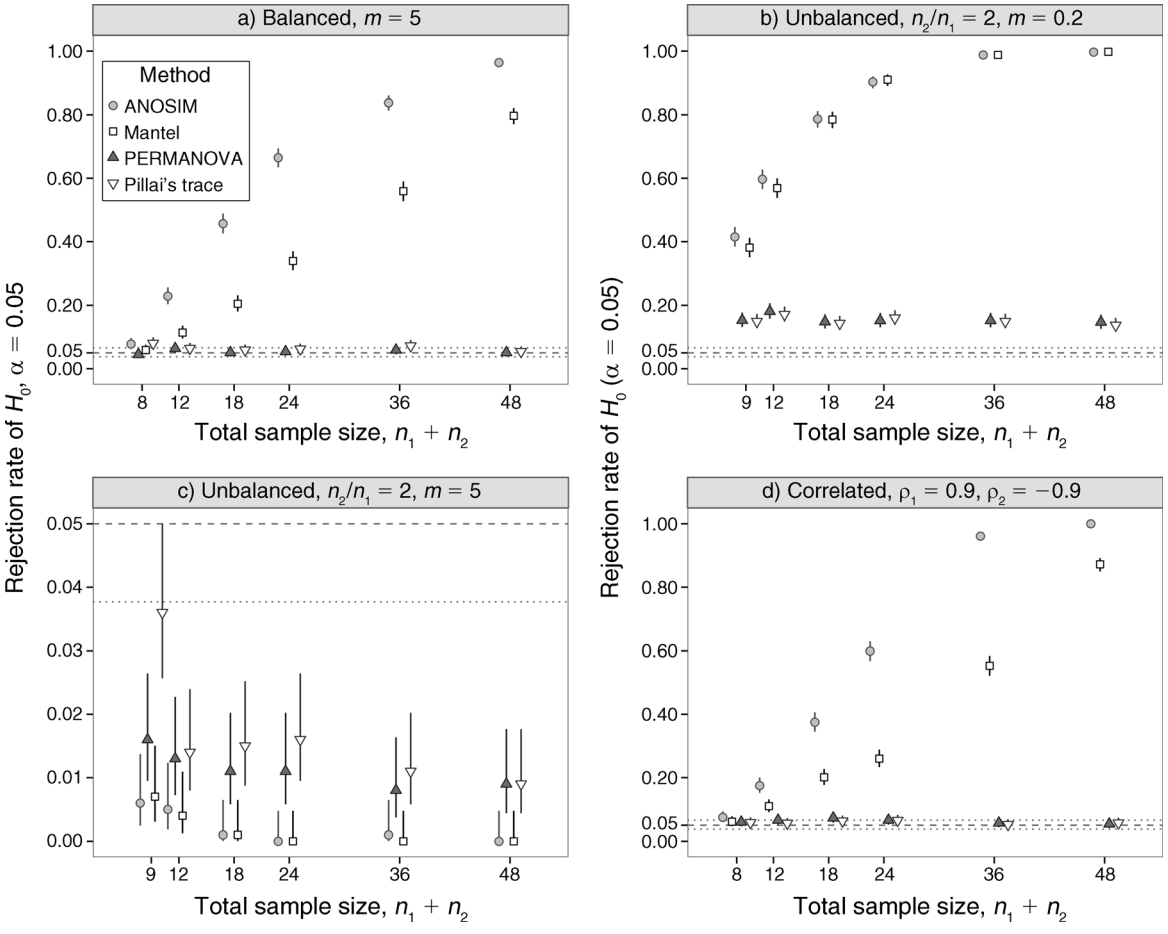

FIG. 1. Empirical rejection rates (±SE, $\alpha = 0.05$) for each of four different test statistics with increasing sample size under heterogeneity for 1000 simulated data sets from a multivariate normal distribution with $p = 2$ variables, $g = 2$ groups, and (a) equal sample sizes, $m = 5$; (b) unequal sample sizes ($n_2/n_1 = 2$), with greater variance in the smaller group ($m = 0.2$); (c) unequal sample sizes ($n_2/n_1 = 2$), with greater variance in the larger group ($m = 5$); and (d) equal sample sizes and variances ($m = 1$), but groups have different correlation structures, with $\rho_1 = 0.9$ and $\rho_2 = -0.9$. The significance level of $\alpha = 0.05$ is shown as a dashed line, and the 95% confidence interval for a test whose true rejection rate is equal to α is shown with dotted lines. Also, for the ith individual group, n_i is the sample size, ρ_i is the correlation between pairs of variables for all variables in group i, and m is the ratio of the variances in the two groups.

total sample size (Fig. 1a) and with increases in the number of variables, and ANOSIM also had greater rejection rates than the Mantel test (Fig. 1a, Table 2). In contrast, PERMANOVA and Pillai's trace remained unaffected by heterogeneity for balanced designs (Fig. 1a, Table 2). Highly similar results were obtained when the negative binomial distribution was used, although Pillai's trace showed modest inflated type I error under severe heterogeneity (e.g., $m = 10$; Table 3a, c).

Unbalanced designs (Sim2).—For unbalanced designs, all tests were liberal when the large dispersion occurred in the group with the smaller sample size (Fig. 1b, Tables 2 and 3), but especially ANOSIM and the Mantel test. Rejection rates for PERMANOVA and Pillai's trace were constant for a given ratio of sample sizes in the two groups (e.g., leveling off at <0.20 in Fig. 1b; see also Supplement 3), whereas for ANOSIM and the Mantel test, rejection rates increased with increases in the total sample size.

In contrast, all tests, and especially ANOSIM and the Mantel test, were overly conservative (with rejection rates <0.05) when the group with the larger sample size also had the greater dispersion (Fig. 1c, Tables 2 and 3). This conservatism became worse with increasing numbers of variables (Tables 2 and 3), increasing disparity in the sample sizes, or increasing heterogeneity (Supplement 3). Regardless of which group had the greater dispersion, however, PERMDISP was able to detect this heterogeneity equally reliably (compare, e.g., PERMDISP vs. either ANOSIM or the Mantel test when $m = 0.1$, then when $m = 10$ in Table 3).

Correlation structure (Sim 3).—Both ANOSIM and the Mantel test (albeit to a lesser extent) were sensitive to differences in correlation structure between groups and

TABLE 3. Rejection rates (out of 1000 simulations) for each of five different multivariate tests, including one designed specifically to detect differences in dispersion among groups (PERMDISP), for data generated under a multivariate normal or negative binomial distribution in Euclidean space with $g = 2$ groups and $p = 2$ variables.

Test and m value	PERMDISP	ANOSIM	Mantel	PERMANOVA	Pillai
a) MVN, balanced					
1.0	0.046	0.051	0.047	0.050	0.053
2.0	0.291	0.129	0.082	0.047	0.057
5.0	0.921	0.665	0.339	0.054	0.062
10.0	0.994	0.948	0.744	0.064	0.075
b) MVN, unbalanced					
0.1	0.986	0.995	0.996	0.171	0.187
0.2	0.871	0.903	0.910	0.152	0.160
0.5	0.273	0.348	0.358	0.090	0.088
1.0	0.049	0.040	0.041	0.053	0.041
2.0	0.275	0.004	0.003	0.019	0.027
5.0	0.835	0.000	0.000	0.011	0.016
10.0	0.980	0.002	0.000	0.008	0.012
c) Negative binomial, balanced					
1.0	0.064	0.052	0.050	0.049	0.046
2.0	0.266	0.104	0.074	0.055	0.054
5.0	0.808	0.621	0.320	0.049	0.067
10.0	0.968	0.931	0.681	0.070	0.117
d) Negative binomial, unbalanced					
0.1	0.923	0.989	0.991	0.221	0.276
0.2	0.780	0.892	0.892	0.177	0.181
0.5	0.262	0.359	0.359	0.101	0.092
1.0	0.048	0.045	0.045	0.041	0.052
2.0	0.232	0.012	0.012	0.033	0.027
5.0	0.637	0.001	0.001	0.018	0.026
10.0	0.767	0.005	0.005	0.015	0.036

Notes: Two different simulation scenarios are shown here: (a) equal sample sizes ($n_1 = n_2 = 12$); and (b) unequal sample sizes ($n_1 = 8$, $n_2 = 16$). Values of m indicate when dispersions were homogeneous ($m = 1$); when heterogeneity occurred ($m \neq 1$) and, for unbalanced designs, whether there was greater dispersion in the group with more samples ($m > 1$); or greater dispersion in the group with fewer samples ($m < 1$).

their rejection rates increased with increasing sample sizes (Fig. 1d). Rejection rates were as high as 100% for ANOSIM when one group had strong negative correlation structure, while the other had strong positive correlation structure (Fig. 1d). Pillai's trace was unaffected by heterogeneity in correlation structure for balanced designs, but showed increased rejection rates (up to about 30%) for unbalanced designs (Supplement 3). PERMANOVA, however, remained completely unaffected by differences in correlation structure (Fig. 1d; Supplement 3).

Numbers of groups (Sim 4).—For ANOSIM, any form of heterogeneity among multiple groups increased rejection rates (Fig. 2). Having half of the groups with large dispersions and half with small dispersions resulted in the highest rejection rates (dark triangles, left-hand panels in Fig. 2), followed first by the situation where one group had large dispersion relative to the others (white squares), and then by the situation where one group had small dispersion relative to the others (white triangles). Under the "half large" scenario for ANOSIM, when the total sample size remained constant (constant N and decreasing n; top left-hand panel of Fig. 2), rejection rates decreased with increasing group number, whereas when the sample sizes per group remained constant (increasing N and constant n; bottom left-hand panel of Fig. 2), rejection rates increased as more groups were added. Very similar results (but with lower average rejection rates) were demonstrated by the Mantel test (Supplement 3). The power of PERMDISP to detect heterogeneity under these scenarios was greater than that of ANOSIM, but in other ways, patterns were similar (Fig. 2).

In contrast, PERMANOVA was not sensitive to heterogeneity under any of these scenarios, but slightly increased rejection rates (between 0.05 and 0.10) were obtained when one group had large dispersion relative to the others (Fig. 2). The results for Pillai's trace mirrored those for PERMANOVA (see Supplement 3).

Simulations based on real data

Ekofisk.—There were only slight differences in power among methods in analyses of simulations based on the Ekofisk data set (Fig. 3; Appendix D). The method having the greatest empirical power also depended on the distribution upon which the simulations were based. When data were generated using the MVLN distribution, then the Mantel test tended to have the greatest

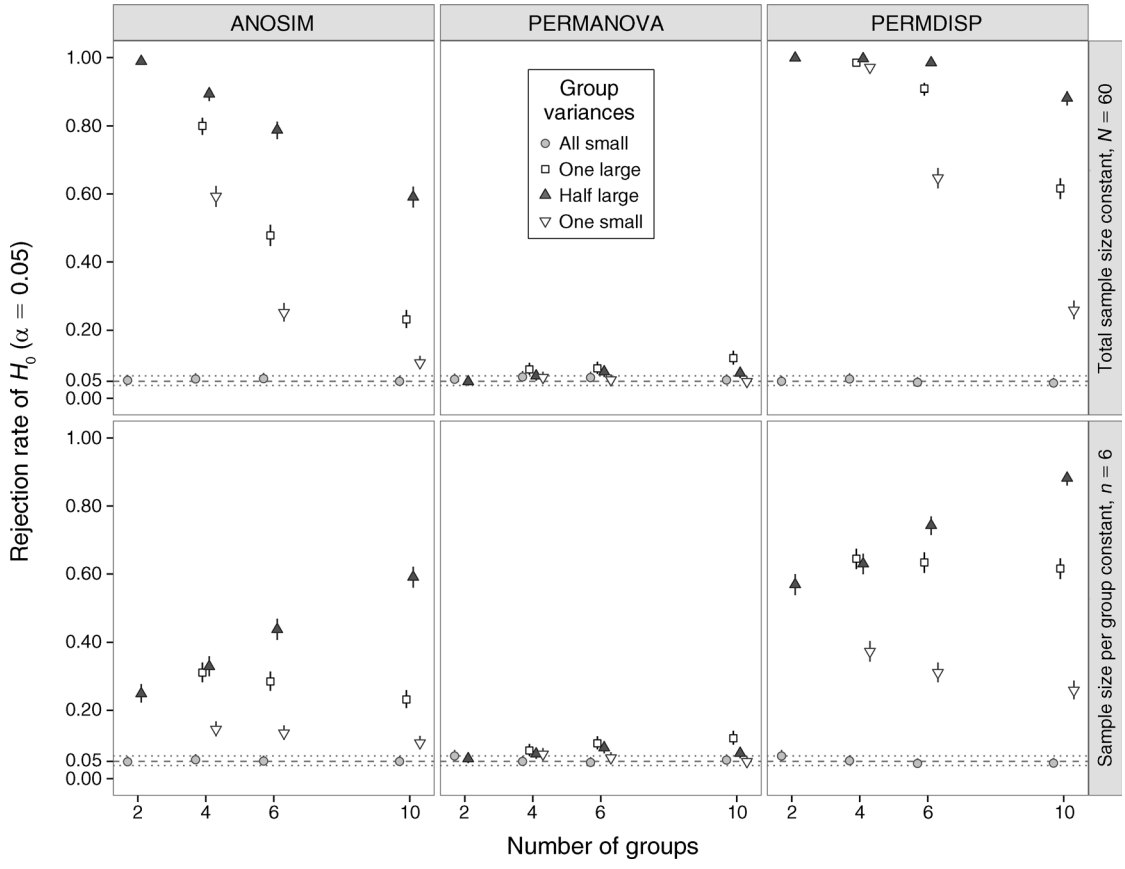

FIG. 2. Empirical rejection rates (±SE, $\alpha = 0.05$) for each of three different tests with increasing numbers of groups (g), for balanced data under four different dispersion scenarios (all small, one large, half large/half small ["half large"], or one small), and where either the total sample size (N) remained constant (top panels), or the number of samples per group (n) remained constant (bottom panels), as calculated from 1000 simulated data sets from a multivariate normal distribution with $p = 2$ variables. Dashed and dotted lines are as described for Fig. 1.

power, followed by ANOSIM, then PERMANOVA (Appendix D). In contrast, when data were generated using a suite of negative binomial variables and Poisson variables, either with or without changes in the dispersion parameters among groups, then PERMANOVA generally had the greatest power, followed by the Mantel test, then ANOSIM (Fig. 3; Appendix D).

These general patterns were broadly consistent for each pair of groups being compared (A vs. B, B vs. C, and C vs. D; see Appendix D and Supplement 4). Simultaneous estimation of empirical power using PERMDISP showed that for some pairwise comparisons, the change in location was the dominant feature of group differences (e.g., B vs. C), whereas for others, substantial differences in dispersion occurred as well (e.g., C vs. D). Interestingly, when simulated differences very clearly had dispersion as well as location effects (e.g., see the PERMDISP results for the comparison of C vs. D using Poisson/NB distributions with varying dispersion parameter), PERMANOVA had greater power than either Mantel or ANOSIM (Appendix D, Supplement 4).

The choice of transformation and resemblance measure also affected power. For some of the groups being compared from the Ekofisk data (e.g., A vs. B), analyses based on Euclidean distances of $\log(y + 1)$-transformed values had the greatest power, while for others (e.g., B vs. C), analyses based on chi-square distances had the greatest power (Fig. 3; Appendix D, Supplement 4). The rank-order differences in power between the different resemblance measures investigated here tended to remain consistent, however, for a given pair of groups being compared, regardless of which distributions were used to simulate the underlying variables (Appendix D, Supplement 4).

Norwegian continental shelf.—PERMANOVA had much greater power than either Mantel or ANOSIM to detect changes in composition in comparisons of area 1 vs. 2 and also area 2 vs. 3 (Fig. 4; Appendix E). For both of these comparisons, the MDS plot and PERMDISP revealed a substantial change in dispersion, as well

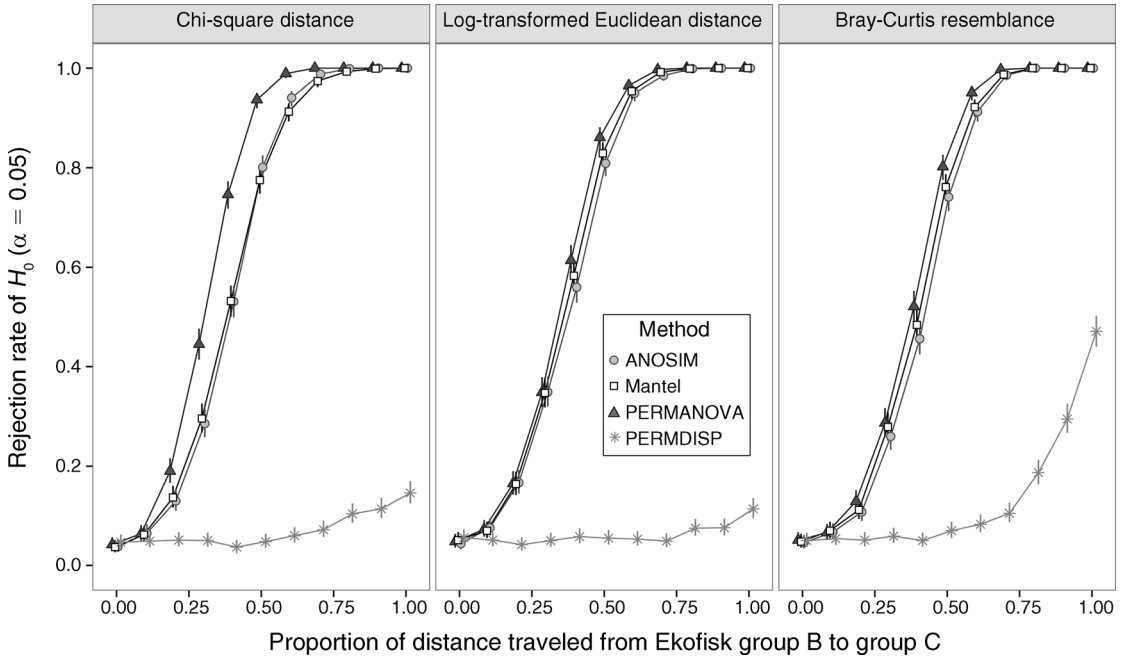

FIG. 3. Empirical power of four multivariate tests to detect changes between group B (1–3.5 km from the center of the oil-drilling activity, $n_B = 12$) and group C (0.25–1 km from the center of the oil-drilling activity, $n_C = 10$) for data simulated using parameters for individual species of soft-sediment benthic macrofauna estimated from the Ekofisk oilfield data sets ($p = 173$ species). The distributions used to simulate the data were a mixture of independent variables having either Poisson or negative binomial distributions with dispersion parameters that varied between the two groups (see Appendix C for more details). The resulting counts were analyzed using chi-square distances (left), Euclidean distances of $\log(y + 1)$-transformed values (middle), or Bray-Curtis resemblances on fourth-root transformed values (right). Results comparing other pairs of groups for the Ekofisk data set and based on other distributions for underlying variables are provided in Appendix D.

as location, of the multivariate data cloud in Jaccard space (see Anderson et al. 2006). The Mantel test, in turn, was more powerful than ANOSIM to detect these compositional differences. PERMANOVA was also more powerful to detect differences for area 4 vs. 5 (Appendix E), while, in contrast, Mantel and ANOSIM had slightly better power than PERMANOVA for the comparison of area 3 vs. 4, although the disparity in

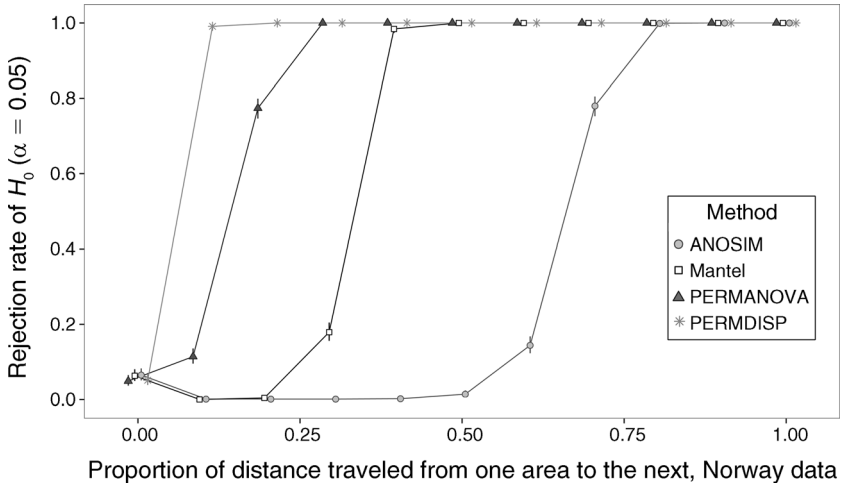

FIG. 4. Empirical power of four multivariate tests to detect changes in community structure and beta diversity of benthic soft-sediment macrofauna (809 species) between area 2 and area 3 ($n_2 = 21$, $n_3 = 25$), along a biogeographic transition on the Norwegian continental shelf. Analyses were based on Jaccard resemblances of presence/absence data from Bernoulli(0, 1) random draws, where probabilities were estimated from proportional occurrences of species in the data set for each group (see Appendix C for further simulation details). Results comparing other pairs of areas for the Norwegian continental shelf data set are given in Appendix E.

power among methods was not so large for these two sets of comparisons (Fig. 4, Appendix E).

In some cases, ANOSIM and the Mantel test demonstrated the unusual behavior of being initially highly conservative, generating power curves that were not monotonically increasing. More specifically, their power to detect initial (smaller) differences in frequencies of occurrences of species between groups actually sunk down to substantially below the 0.05 significance level, even effectively dropping down to zero (Fig. 4; Appendix E). This odd behavior was more pronounced for ANOSIM than for Mantel, and occurred primarily under scenarios where the increase in dispersion was accompanied by an increase in the sample size (e.g., $n_1 = 16$, $n_2 = 21$). PERMANOVA, in contrast, demonstrated no such effect, generating power curves that were both monotonic and quickly responsive to quite small changes in compositional frequencies between groups (Fig. 4; Appendix E).

Discussion

After the original description of the F statistic (Fisher 1925, Snedecor 1934), it was some time before the behavior of the univariate F test in ANOVA for real data became better understood by reference to its performance under potential violations of its assumptions (Pearson 1931, Cochran 1947, Box 1953, 1954) or by comparison with nonparametric alternatives developed later (e.g., Feir-Walsh and Toothaker 1974, Tomarken and Serlin 1986). We are in a similar situation with respect to our current knowledge of the multivariate resemblance-based permutation tests. Although ANOSIM, PERMANOVA, and the Mantel test are now very widely used in ecology and other disciplines, this is the first study, to our knowledge, which focuses on the effects of heterogeneity of dispersions on these multivariate resemblance-based tests.

Effects of heterogeneity for balanced designs

Given that the ANOSIM R statistic is described as a test of the very general null hypothesis of H_0: "no differences among the groups," and the fact that Clarke's (1993: 131) original description states that the test "will have some power to detect" this kind of change, referring to differences in dispersion, researchers have so far been very wise to interpret significant R statistics simply as providing evidence for "a difference" among groups, but without honing their inferences down any further in terms of differences in locations, dispersions, the shape of the data cloud, or perhaps all of these things, within a given context. Indeed, the fundamental idea that the true underlying null hypothesis in a statistical comparison of two (or more) sampled groups has (at least) two parts (equality of means and equality of variances) can be traced back to the work of Fisher (1939).

One might expect the ANOSIM test to be more robust to heterogeneity than either PERMANOVA or the Mantel test, as it not only uses a permutation algorithm, but also reduces the distance matrix down to ranks. Surprisingly, this did not occur. In fact, it is clear that the construction of the test statistic itself makes it a kind of "omnibus" test, being much more sensitive to heterogeneity of dispersions and differences in correlation structure among groups than was PERMANOVA. The Mantel test has a broadly similar construction, and although it was not as severely affected by heterogeneity as the ANOSIM test under any scenario examined here, it did follow all of the general patterns observed for ANOSIM in its essential behavior.

Why does heterogeneity lead to small P values for ANOSIM?

The value of the R statistic in ANOSIM measures directly the degree of distinctiveness of groups, regardless of sample size (Clarke 1993). Under a scenario of one group having larger dispersions than another, when centroids are equal, the value of R does not necessarily get very large. For example, in Fig. 5, a plot is shown of a single set of simulated bivariate normal data, where the population variances for the two variables in group 2 are twice those in group 1. The value of the PERMANOVA pseudo-F and ANOSIM R statistics for this particular set of data are each shown to the right of this, placed within the context of their distributions under permutation. Note that even though the value of R is quite small for the simulated data ($R = 0.0963$), the distribution of the R statistic under permutation has been shifted to the left, thus resulting in a small P value for the ANOSIM test. Under repeated simulation of such data sets, the distribution of the P values for ANOSIM is therefore not uniform. Instead, many are small, and the percentage of P values less than $\alpha = 0.05$ (the rejection rate) is around 24% (Fig. 5). In contrast, the pseudo-F statistic is quite robust to this heterogeneity, showing a large P value for this particular simulated data set ($P = 0.725$), as well as a quite uniform distribution of P values for the full set of simulations and a rejection rate of ~4% (Fig. 5).

The reason that the ANOSIM test yields a significant result (a small P value) is not because the observed value of R is large, but rather because the distribution of $R^{(\pi)}$ under permutation is shifted to the left, including a large number of negative values (Fig. 5). Negative values of R indicate that the average of the ranks of within-group dissimilarities is greater than the average of the ranks of between-group dissimilarities (e.g., Chapman and Underwood 1999). This arises under permutation for groups with heterogeneous dispersions because, if there is greater clumping of samples in (say) one or more of the groups in the original data, then under permutation the within-group dissimilarity values will look larger (on average) and the between-group dissimilarities will look smaller (on

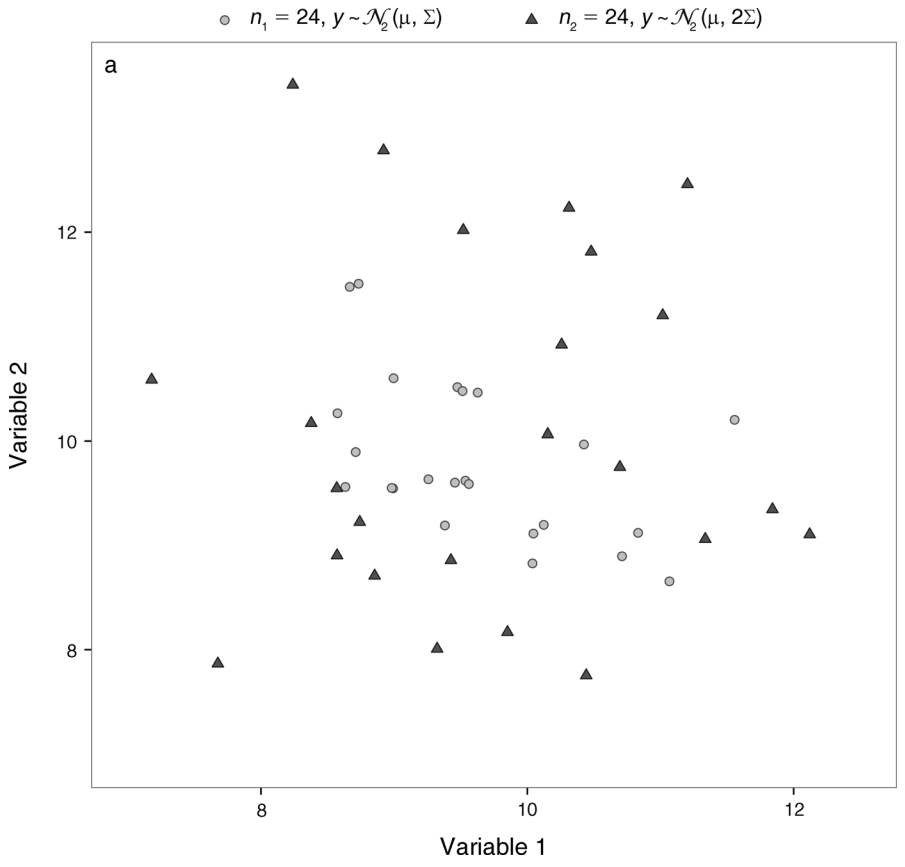

FIG. 5. (a) Scatterplot of a single set of simulated data from a bivariate normal distribution ($p = 2$) for each of $g = 2$ groups having equal sample sizes ($n_1 = n_2 = 24$), but different dispersions ($m = 2$), along with the test statistics and permutation distributions for (b) the PERMANOVA pseudo-F statistic and (c) the ANOSIM R statistic. Also shown are the distributions of P values and associated rejection rates obtained for (d) PERMANOVA and (e) ANOSIM for 1000 such data sets.

average) compared to what was originally observed, so the permutation distribution of $R^{(\pi)}$ shifts to the left relative to the original R value.

If a statistically significant result is accompanied by a value of R that is not very large, this could be a signal that the difference is primarily a difference in dispersion. For example, under the scenario depicted in Fig. 5, the median value of R obtained under simulation was 0.021 (with 0.025 and 0.975 quantile values for R of −0.023 and 0.115, respectively). A statistically significant, yet small, value of R is no guarantee, however, that differences are indeed differences in dispersion and not (small) differences in location, whatever pattern might be evidenced on an accompanying MDS (or other) ordination plot. In practice, unfortunately, there is no way to unravel location vs. dispersion effects using ANOSIM (or, for that matter, Mantel).

The increased rejection rates of ANOSIM (or Mantel) caused by differences in the degree or direction of correlation structure among groups was unexpected. The detection of "distinctiveness" of groups purely on the basis that the groups have different shapes is something that is (once again) likely to be caused by the distribution of $R^{(\pi)}$ under permutation being tugged to the left, rather than the value of R being large, per se. In other words, strong correlations generate a "clumping" effect into distinct nonspherical shapes, which will be seen by ANOSIM or Mantel in much the same way as differences in dispersions. More specifically, under permutation, the within-group dispersions will look large relative to their values in the original data set, shifting $R^{(\pi)}$ to the left.

Robustness of PERMANOVA for balanced designs

Importantly, the response of either ANOSIM or Mantel to increases in sample size was to cause increasing rejection rates of the null hypothesis under a given scenario of heterogeneity. This is because more samples provide greater power to detect this type of change for these omnibus tests. An increase in rejection rates under heterogeneity with increasing sample size is also seen in the behavior of the univariate rank-based Kruskal-Wallis test (Feir-Walsh and Toothaker 1974). This effect did not occur for either Pillai's trace or

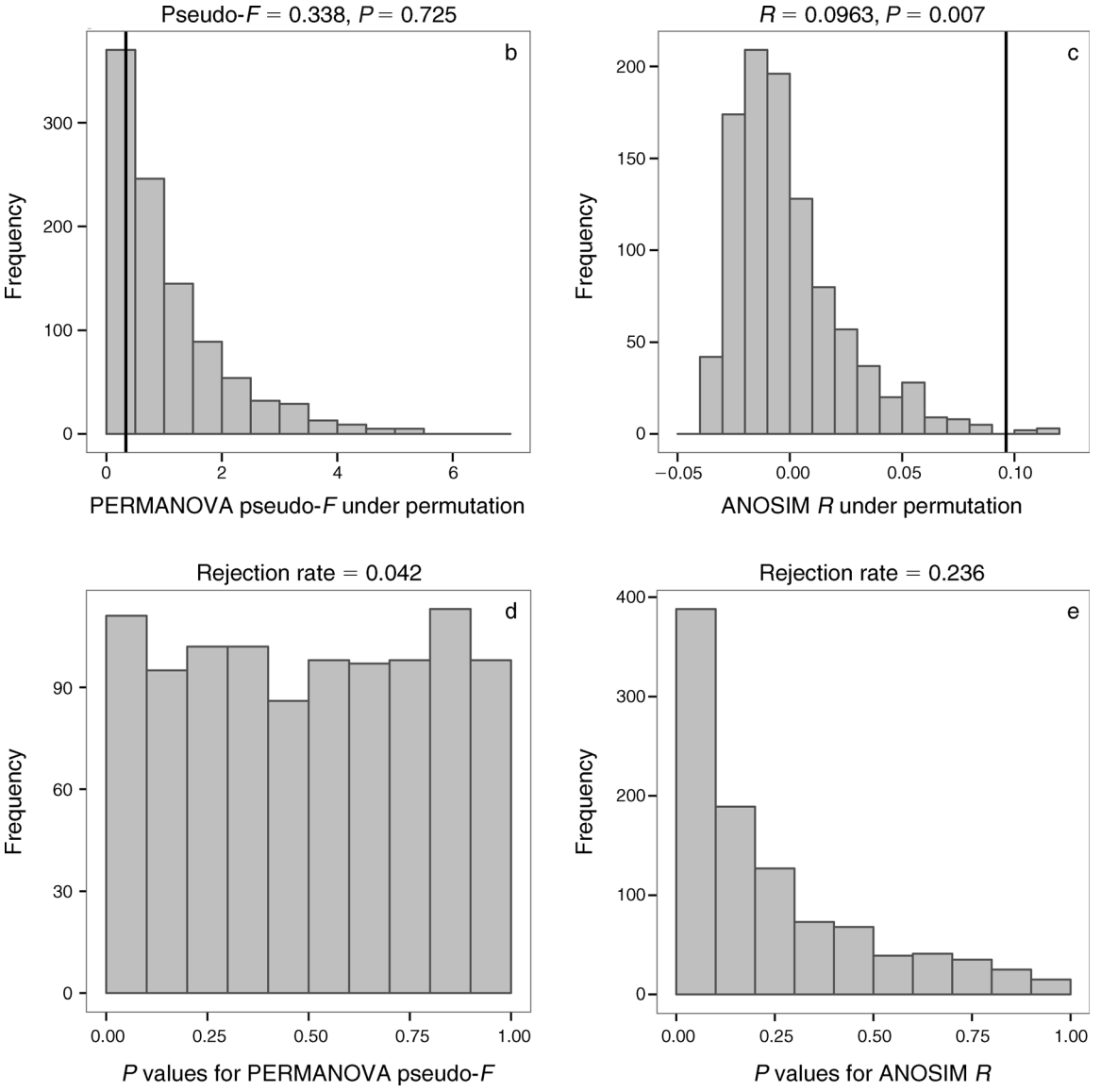

FIG. 5. Continued.

PERMANOVA. This is because, by design, PERMANOVA's pseudo-*F* and Pillai's trace statistics focus purely on measuring differences in locations (centroids). Thus, as in the famous reference to the univariate ANOVA *F* statistic being like an "ocean liner" (Box 1953), these approaches, too (for balanced designs), will not be easily rocked by differences in other ancillary quantities (like dispersions) that they are not designed to test.

When there were more than two groups, ANOSIM showed greater sensitivity to heterogeneity when half of the groups had large dispersions and half had small dispersions. In contrast, PERMANOVA and Pillai's trace were more sensitive to heterogeneity (albeit only mildly, as these were balanced designs) when it occurred in the form of one group being substantially more dispersed than the others. The latter result mirrors what has been found for univariate ANOVA (Box 1953, 1954). This led to the rationale behind the common use in ecology (and beyond) of Cochran's test for homogeneity (Cochran 1941, 1951), a statistic consisting of the ratio of the largest estimated within-group variance vs. the sum of the individual estimated within-group variances. Thus, Underwood (1981, 1997) recommended the use of Cochran's test for homogeneity of variances prior to implementing a univariate ANOVA test. A multivariate resemblance-based analogue to Cochran's test has not yet been developed, although this would be a useful topic for future research. Nevertheless, PERMANOVA was overall still quite robust to this form of

heterogeneity; for example, even with large numbers of groups ($g = 10$) and with one group having much larger dispersion than the others, the rejection rate only increased to around 0.10 (Fig. 2).

The relative robustness of PERMANOVA and Pillai's trace to heterogeneity for balanced designs mirrors similar original results obtained for the F statistic in univariate ANOVA (Horsnell 1953, Box 1954, Glass et al. 1972), as well as the results obtained by Olson (1974) indicating that Pillai's trace was more robust to heterogeneity than the other classical MANOVA statistics. It might be tempting to consider using Pillai's trace routinely, or even to couple it with a permutation algorithm for calculating P values (as is done in the canonical analysis of principal coordinates, CAP, a method with which the classical MANOVA statistics have a clear kinship, see Anderson and Robinson [2003] and Anderson and Willis [2003] for details), but Pillai's trace will be sensitive to differences in correlation structure among groups, and, furthermore, it simply cannot be calculated when there are more variables than samples ($p > N$), nor can it be implemented on the basis of non-Euclidean distances (such as Bray-Curtis or Jaccard) as are commonly used in the analysis of ecological communities.

Effects of heterogeneity for unbalanced designs

All of the tests were sensitive to heterogeneity of dispersions for unbalanced designs. The direction of the effects depended on the direction of heterogeneity with respect to the differences in sample sizes. If a group with large dispersions also had a small number of samples, then rejection rates increased. Consider that the position of just a few points in a group drawn from a population that has large variation, relative to a large number of samples in a tightly clustered group, could well fall on one side of that cluster or the other, just by chance, even though the two groups have the same population centroid. In contrast, if the group with the large dispersions also had a large number of samples, then the tests all became quite conservative; it is very difficult to get a small tightly clustered group to fall outside of a large group that is widely dispersed, so differences in centroid (even if they were there), become very difficult to detect in such cases. These effects of heterogeneity for unbalanced designs mirror precisely what has been shown for univariate ANOVA: namely, conservatism when variances are positively related to group sample sizes and liberalism when the relationship is negative (e.g., Welch 1937, Glass et al. 1972).

The fact that effects of heterogeneity on rejection rates for unbalanced designs were constant for a given sample size ratio in the case of Pillai's trace and PERMANOVA was interesting and warrants further study. Although not pursued further here, it should be possible to demonstrate this result as an asymptotic property of these tests, especially by reference to the construction of these test statistics. In contrast, just as in the balanced-design case, ANOSIM and Mantel rejection rates always increased substantially with increases in the total sample size.

Power

PERMANOVA was more powerful than the other tests to detect changes in community structure for the majority of the scenarios simulated here based on real data sets. This aligns with previous work demonstrating greater power for canonical partitioning methods over the Mantel test (Legendre and Fortin 2010). Furthermore, the ANOSIM test was never found to be more powerful than the Mantel test. This occurred despite the fact that, under simpler idealized scenarios (Sim1–Sim4), ANOSIM rejection rates tended to be much higher than PERMANOVA or the Mantel test. This is likely due to the ranking of dissimilarities, intrinsic to the ANOSIM test, which will have different consequences for data sets where the underlying dissimilarities are distributed in different ways.

For simulations of count data (Ekofisk), differences in power were not large and the method with the greater power depended on the distributions used for the simulation. PERMANOVA generally had more power to detect changes when data were simulated using separate independent Poisson and negative binomial distributions, whereas Mantel had more power when the truncated MVLN distribution was used. One possible explanation for this result is that the Mantel test is sensitive to changes in correlation structure among variables (see Sim3; Fig. 1d), and such differences (as estimated from the real data) were indeed able to be built in to the truncated MVLN simulations. PERMANOVA, in contrast, is not sensitive to differences in correlation structure (shape). A resemblance-based test statistic that takes into account the correlation structure among variables (such as CAP; Anderson and Robinson 2003, Anderson and Willis 2003) would also be expected to have more power than PERMANOVA in the presence of high correlation structures in the data (Anderson and Robinson 2003), although this was not examined explicitly here.

Relative power also depended on the resemblance measure used as the basis of the analysis. Different resemblance measures emphasize fundamentally different aspects of the underlying multivariate data matrix (Clarke et al. 2006). Euclidean distance is more focused on differences in abundance per species and differences in total abundances per sample, chi-square distance emphasizes changes in proportional abundances, with heavier weights being given to rarer species (Legendre and Gallagher 2001), whereas Bray-Curtis is more focused on compositional changes in species' identities. Thus, for example, greater power obtained using Euclidean distances for a particular comparison may simply be a consequence of the differences in (log) abundance per species between groups being more pronounced in those cases than either the turnover in

species' identities or the changes in proportional abundances. These results serve to highlight that the resemblance measure used as the basis of the analysis should be chosen carefully by reference to the underlying ecological questions of greatest interest to the researcher within a given context.

Location vs. dispersion effects

Warton et al. (2012) stated that distance-based tests (such as ANOSIM or PERMANOVA) confound location and dispersion effects. However, as has been clearly demonstrated here, it is *not* the construction of the PERMANOVA test statistic itself that confounds location and dispersion effects, but rather the *underlying dissimilarity measure* that is used as the basis of the analysis which may do this (see Appendix B). PERMANOVA (in the case of balanced designs), as a test statistic, will focus on differences in location only, but this will be done *in the space of the resemblance measure chosen*. Thus, careful consideration of the meaning of the resemblance measure and what it actually measures by reference to the wealth of information in the underlying multivariate data set is clearly necessary. Measures such as Jaccard or Bray-Curtis, commonly used in ecology, do not retain the mean–variance properties of original abundances, but they do emphasize, instead, the similarity in composition of species' identities among samples, which the Euclidean distance measure does not.

In contrast to PERMANOVA, even for balanced designs, ANOSIM and the Mantel test really do confound location and dispersion effects in the sense that one cannot unravel which of these types of differences (in the space of the chosen resemblance measure) might be driving any reported significant results. Furthermore, their hyper-conservatism in the face of unbalanced designs where groups with larger sample sizes have greater dispersion suggests that they cannot, unfortunately, be relied upon more generally as "omnibus" tests for differences among groups. Importantly, PERMANOVA (and Pillai's trace) also lack the desired level of robustness to heterogeneity for unbalanced designs, pointing directly to the need for new methods to be developed that can be used to test for differences in location even in the presence of differences in dispersions among groups for cases where there are unequal sample sizes.

Other methods

Several other resemblance-based test statistics described to date will yield equivalent P values under permutation for the one-way case to the methods examined in detail here (see Appendix A and Warton and Hudson 2004). Specifically, the results of simulations obtained here for the Mantel test are equivalent to what would be obtained using either $\bar{d}_B/\bar{d}_W$, as proposed by Good (1982) and Smith et al. (1990), or the MRPP statistic calculated directly on dissimilarities (Mielke et al. 1981), provided the design is balanced and particular weights are used in the construction of the MRPP test statistic (see Appendix A). Similarly, the results of simulations obtained here for PERMANOVA are equivalent to what would be obtained using the statistics described by Pillar and Orlóci (1996) and Gower and Krzanowski (1999), or MRPP calculated on squared dissimilarities (for a particular choice of weights, but for both unbalanced and balanced designs; see Appendix A). Note, however, that the equivalence of these methods to PERMANOVA is only true for the one-way ANOVA design, and does not necessarily hold more generally for higher-way designs (e.g., Torres et al. 2010). In addition, if Euclidean distances are used as the basis of the analysis, then the results obtained using PERMANOVA are equivalent to what would be obtained using either a redundancy analysis (RDA) for an ANOVA factor by permutation (e.g., Verdonschot and ter Braak 1994) or the geometric F test by randomization proposed by Edgington (1995). Similarly, if the analysis is based on a chi-square distance matrix, then results obtained using PERMANOVA are expected to mirror results obtained using canonical correspondence analysis (CCA; ter Braak 1986, Legendre and Gallagher 2001), where the environmental predictor variables are orthogonal ANOVA codes for a factor.

Other approaches, not investigated here, include what might be called "stacked" test statistics or "variable-based" statistics, such as the sum of individual F ratios of Edgington (1995), the "LR-IND," or "sum-of-LR" tests (Warton and Hudson 2004, Warton 2011, Warton et al. 2012), all of which also use permutations to obtain P values for inference. These effectively treat the multivariate problem as a sum of individual univariate problems. It is well known, however, that high-dimensional information may not be manifest in the original individual variables, so individual variable-based approaches will reflect this limitation. We expect that the relative power of these approaches compared to PERMANOVA or to a MANOVA statistic like Pillai's trace, with P values obtained using permutations, or a dissimilarity-based approach that takes into account correlation structure, such as CAP, will depend heavily on the type of scenario being examined, the relative between- vs. within-group variation among different variables that show some effects, the extent to which rare species are responsible for turnover among groups, and the degree of correlation structure among the variables.

Future research directions

Although the present study was very broad in scope, there are clearly many avenues requiring further research. The initial focus here was on rejection rates when centroids were equal, and subsequent power comparisons were limited to examination of certain alternative hypotheses for specific data sets and distance

measures. Exploration of potential general principles regarding comparative power of these methods under different types of ecological scenarios and for a variety of distance measures is needed. More work is also needed to clarify the behavior of these and other tests for different shapes of distributions in underlying variables, such as those with extreme values or outliers, including to better understand the properties of different distance measures when used with variables having differently shaped distributions. In addition, despite their somewhat similar behavior overall when compared to PERMANOVA, ANOSIM and the Mantel test clearly are not equivalent and their rejection rates can be very different in size under a given scenario. Thus, identifying situations when ranking the distances either increases or decreases the relative power of ANOSIM compared to the Mantel test also requires more study.

Conclusions

For balanced designs, PERMANOVA was quite robust to heterogeneity, but ANOSIM and the Mantel test were not. These resemblance-based tests are clearly not testing the same null hypothesis. ANOSIM and the Mantel test examine the more general H_0: "samples in the same group are no more tightly clustered together than samples from different groups," whereas PERMANOVA focuses on the more specific H_0: "there are no differences in centroids among the groups." Note that in all cases, the "clumping of samples within groups" or the "differences in centroids" (a shift in the location of the multivariate data cloud) are defined *in the space of the resemblance measure chosen* for the analysis. As ANOSIM and the Mantel test are more general "omnibus" tests, rejection of the null hypothesis in either case will indicate only that some feature of the groups differ to make them distinct. This feature could be (1) locations, (2) dispersions, (3) the particular shape (correlation structure) of the data clouds being compared; or indeed, some combination of these things. Although reduced-space ordinations (such as nonmetric MDS) can assist in interpreting the potential nature of any differences detected, it is not possible with these tests or any associated plots to make more specific statistical inferences.

Although the generality of these more omnibus tests can often be useful, in many ecological studies it may be quite important, however, to hone inferences further. For example, ecologists may want to distinguish—*has there been a fundamental shift in the community structure itself* (a change in location)? Or rather, *has the community structure become more (or less) variable* (a change in dispersion)? Or both? For balanced designs, PERMANOVA can be used effectively to make inferences about differences in centroids alone (i.e., shifts in the location of the multivariate cloud of sample units in the space of the resemblance measure), while PERMDISP can be used to make inferences about differences in multivariate dispersions alone.

Importantly, *none* of the tests examined here were robust to heterogeneity for unbalanced sampling designs, being either excessively liberal or extremely conservative under different scenarios, especially ANOSIM and the Mantel test, which became worse with increasing total numbers of samples. Thus, we do not recommend the routine use of these tests for unbalanced designs where heterogeneity of dispersions is known to occur, as interpreting results and drawing inferences in such cases can be problematic. Given the common occurrence of genuine heterogeneity in multivariate ecological data, the development of tests for differences in centroids that explicitly take into account heterogeneity of within-group dispersions is an important topic for future research, and will certainly be necessary to analyze unbalanced sampling designs rigorously in multivariate tests to compare groups.

Acknowledgments

This work was supported by a Royal Society of New Zealand Marsden Grant (MAU1005) to M. J. Anderson. This work benefited greatly from important insights and comments given by K. R. Clarke, R. N. Gorley, P. Legendre, and two anonymous reviewers. We also thank M. B. Jones, K. Parry, M. D. Pawley, O. Hannaford, and A. N. H. Smith for thought-provoking discussions of this work. M. J. Anderson has worked with PRIMER-e to develop commercial software to implement the method of PERMANOVA for one-way and multi-way designs, which is sold as an add-on package to the PRIMER v6 computer program. This may be perceived as posing a potential conflict of interest; however, the PERMANOVA add-on requires the base PRIMER package, in which ANOSIM is implemented, thus negating any perceived conflict of interest in presenting rigorous comparisons of PERMANOVA vs. ANOSIM. Furthermore, for one-way designs, all of the methods compared here can be implemented for free in R, and all of the results presented here are repeatable, using the R code provided in Supplements 1 and 2.

Literature Cited

Anderson, M. J. 2001. A new method for non-parametric multivariate analysis of variance. Austral Ecology 26:32–46.

Anderson, M. J. 2006. Distance-based tests for homogeneity of multivariate dispersions. Biometrics 62:245–253.

Anderson, M. J., et al. 2011. Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist. Ecology Letters 14:19–28.

Anderson, M. J., K. E. Ellingsen, and B. H. McArdle. 2006. Multivariate dispersion as a measure of beta diversity. Ecology Letters 9:683–693.

Anderson, M. J., R. N. Gorley, and K. R. Clarke. 2008. PERMANOVA+ for PRIMER: Guide to software and statistical methods. PRIMER-E, Plymouth, UK.

Anderson, M. J., and J. Robinson. 2003. Generalized discriminant analysis based on distances. Australian and New Zealand Journal of Statistics 45:301–318.

Anderson, M. J., and T. J. Willis. 2003. Canonical analysis of principal coordinates: a useful method of constrained ordination for ecology. Ecology 84:511–525.

Boik, R. J. 1987. The Fisher-Pitman permutation test: a non-robust alternative to the normal theory F test when variances are heterogeneous. British Journal of Mathematical and Statistical Psychology 40:26–42.

Box, G. E. P. 1953. Non-normality and tests on variances. Biometrika 40:318–335.

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

Chapman, M. G., and A. J. Underwood. 1999. Ecological patterns in multivariate assemblages: information and interpretation of negative values in ANOSIM tests. Marine Ecology Progress Series 180:257–265.

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

Clarke, K. R., and R. N. Gorley. 2006. PRIMER v6: User manual/tutorial. PRIMER-E, Plymouth, UK.

Clarke, K. R., and R. H. Green. 1988. Statistical design and analysis for a 'biological effects' study. Marine Ecology Progress Series 46:213–226.

Clarke, K. R., P. J. Somerfield, and M. G. Chapman. 2006. On resemblance measures for ecological studies, including taxonomic dissimilarities and a zero-adjusted Bray-Curtis coefficient for denuded assemblages. Journal of Experimental Marine Biology and Ecology 330:55–80.

Cochran, W. G. 1941. The distribution of the largest of a set of estimated variances as a fraction of their total. Annals of Eugenics 11:47–52.

Cochran, W. G. 1947. Some consequences when the assumptions for the analysis of variance are not satisfied. Biometrics 3:22–38.

Cochran, W. G. 1951. Testing a linear relation among variances. Biometrics 7:17–32.

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

Ellingsen, K. E., and J. S. Gray. 2002. Spatial patterns of benthic diversity: is there a latitudinal gradient along the Norwegian continental shelf? Journal of Animal Ecology 71: 373–389.

Feir-Walsh, B. J., and L. E. Toothaker. 1974. An empirical comparison of the ANOVA *F*-test, normal scores test and Kruskal-Wallis test under violation of assumptions. Educational and Psychological Measurement 34:789–799.

Fisher, R. A. 1925. Statistical methods for research workers. Oliver and Boyd, Edinburgh, UK.

Fisher, R. A. 1939. The comparison of samples with possibly unequal variances. Annals of Eugenics 9:174–180.

Glass, G. V., P. D. Peckham, and J. R. Sanders. 1972. Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. Review of Educational Research 42:237–288.

Good, I. J. 1982. An index of separateness of clusters and a permutation test for its significance. Journal of Statistical Computation and Simulation 15:261–275.

Gower, J. C., and W. J. Krzanowski. 1999. Analysis of distance for structured multivariate data and extensions to multivariate analysis of variance. Applied Statistics 48:505–519.

Gray, J. S., K. R. Clarke, R. M. Warwick, and G. Hobbs. 1990. Detection of initial effects of pollution on marine benthos: an example from the Ekofisk and Eldfisk oilfields, North Sea. Marine Ecology Progress Series 66:285–299.

Hayes, A. F. 1996. Permutation test is not distribution free. Psychological Methods 1:184–198.

Horsnell, G. 1953. The effect of unequal group variances on the *F*-test for the homogeneity of group means. Biometrika 40: 128–136.

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 M.-J. Fortin. 2010. Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. Molecular Ecology Resources 10:831–844.

Legendre, P., and E. D. Gallagher. 2001. Ecologically meaningful transformations for ordination of species data. Oecologia 129:271–280.

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

Manly, B. F. J., and R. I. C. C. Francis. 2002. Testing for mean and variance differences with samples from distributions that may be non-normal with unequal variances. Journal of Statistical Computation and Simulation 72:633–646.

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.

Mardia, K. V., J. T. Kent, and J. M. Bibby. 1979. Multivariate analysis. Academic Press, New York, New York, USA.

McArdle, B. H., and M. J. Anderson. 2001. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. Ecology 82:290–297.

McArdle, B. H., and M. J. Anderson. 2004. Variance heterogeneity, transformations and models of species abundance: a cautionary tale. Canadian Journal of Fisheries and Aquatic Sciences 61:1294–1302.

Mielke, P. W., K. J. Berry, P. J. Brockwell, and J. S. Williams. 1981. A class of nonparametric tests based on multiresponse permutation procedures. Biometrika 68:720–724.

Olson, C. L. 1974. Comparative robustness of six tests in multivariate analysis of variance. Journal of the American Statistical Association 69:894–908.

Olson, C. L. 1979. Practical considerations in choosing a MANOVA test statistic: a rejoinder to Stevens. Psychological Bulletin 86:1350–1352.

Pearson, E. S. 1931. The analysis of variance in cases of non-normal variation. Biometrika 23:114–133.

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 relevé groups. Journal of Vegetation Science 7:585–592.

R Development Core Team. 2012. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org

Rencher, A. C. 1998. Multivariate statistical inference and applications. John Wiley and Sons, New York, New York, USA.

Romano, J. P. 1990. On the behavior of randomization tests without group invariance assumption. Journal of the American Statistical Association 85:686–692.

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

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.

Snedecor, G. W. 1934. Calculation and interpretation of analysis of variance and covariance. Collegiate Press, Ames, Iowa, USA.

Stevens, J. 1979. Comment on Olson: Choosing a test statistic in multivariate analysis of variance. Psychological Bulletin 86:355–360.

ter Braak, C. J. F. 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology 67:1167–1179.

Tomarken, A. J., and R. C. Serlin. 1986. Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. Psychological Bulletin 99: 90–99.

Torres, P. S., M. B. Quaglio, and V. D. P. Pillar. 2010. Properties of a randomization test for multifactor comparisons of groups. Journal of Statistical Computation and Simulation 80:1131–1150.

Underwood, A. J. 1981. Techniques of analysis of variance in experimental marine biology and ecology. Oceanography and Marine Biology: An Annual Review 19:513–605.

Underwood, A. J. 1997. Experiments in ecology: their logical design and interpretation using analysis of variance. Cambridge University Press, Cambridge, UK.

Verdonschot, P. F. M., and C. J. F. ter Braak. 1994. An experimental manipulation of oligochaete communities in mesocosms treated with chlorpyrifos or nutrient additions: multivariate analyses with Monte Carlo permutation tests. Hydrobiologia 278:251–266.

Warton, D. I. 2011. Regularized sandwich estimators for analysis of high-dimensional data using generalised estimating equations. Biometrics 67:116–123.

Warton, D. I., and H. M. Hudson. 2004. A MANOVA statistic is just as powerful as distance-based statistics, for multivariate abundances. Ecology 85:858–874.

Warton, D. I., S. T. Wright, and Y. Wang. 2012. Distance-based multivariate analyses confound location and dispersion effects. Methods in Ecology and Evolution 3:89–101.

Welch, B. L. 1937. The significance of the difference between two means when the population variances are unequal. Biometrika 29:350–362.

Supplemental Material

Appendix A

Description of statistical tests and related methods (*Ecological Archives* M083-019-A1).

Appendix B

Example showing how data simulated from groups with equal centroids but different variances in Euclidean space can yield groups with unequal centroids in Bray-Curtis space (*Ecological Archives* M083-019-A2).

Appendix C

Description of data sets (Ekofisk and Norwegian continental shelf) and simulation methods used to compare the power of statistical tests to detect real changes in multivariate ecological assemblages (*Ecological Archives* M083-019-A3).

Appendix D

Additional figures showing empirical power of multivariate tests for comparisons of groups based on the Ekofisk data (*Ecological Archives* M083-019-A4).

Appendix E

Additional figure showing empirical power of multivariate tests for comparisons of groups based on the Norwegian continental shelf data (*Ecological Archives* M083-019-A5).

Supplement 1

R code and associated source files of parameters used to conduct simulations Sim1–Sim4 (*Ecological Archives* M083-019-S1).

Supplement 2

R code and associated source files used to generate parameters and conduct simulations based on real ecological data sets (*Ecological Archives* M083-019-S2).

Supplement 3

Full results of all simulation scenarios described in Sim1–Sim4 (*Ecological Archives* M083-019-S3).

Supplement 4

Full results of all simulations based on real ecological data sets (*Ecological Archives* M083-019-S4).