Australian & New Zealand Journal of Statistics

*Aust. N. Z. J. Stat.* **59**(1), 2017, 57–79 doi: 10.1111/anzs.12176

# Some solutions to the multivariate Behrens–Fisher problem for dissimilarity-based analyses

**Marti J. Anderson**[1]*, **Daniel C. I. Walsh**[2], **K. Robert Clarke**[3], **Ray N. Gorley**[4] **and Edlin Guerra-Castro**[5]

*Massey University and PRIMER-E Limited and Universidad Nacional Autónoma de México*

## Summary

The essence of the generalised multivariate Behrens–Fisher problem (BFP) is how to test the null hypothesis of equality of mean vectors for two or more populations when their dispersion matrices differ. Solutions to the BFP usually assume variables are multivariate normal and do not handle high-dimensional data. In ecology, species' count data are often high-dimensional, non-normal and heterogeneous. Also, interest lies in analysing compositional dissimilarities among whole communities in non-Euclidean (semi-metric or non-metric) multivariate space. Hence, dissimilarity-based tests by permutation (e.g., PERMANOVA, ANOSIM) are used to detect differences among groups of multivariate samples. Such tests are not robust, however, to heterogeneity of dispersions in the space of the chosen dissimilarity measure, most conspicuously for unbalanced designs. Here, we propose a modification to the PERMANOVA test statistic, coupled with either permutation or bootstrap resampling methods, as a solution to the BFP for dissimilarity-based tests. Empirical simulations demonstrate that the type I error remains close to nominal significance levels under classical scenarios known to cause problems for the un-modified test. Furthermore, the permutation approach is found to be more powerful than the (more conservative) bootstrap for detecting changes in community structure for real ecological datasets. The utility of the approach is shown through analysis of 809 species of benthic soft-sediment invertebrates from 101 sites in five areas spanning 1960 km along the Norwegian continental shelf, based on the Jaccard dissimilarity measure.

*Key words*: Behrens–Fisher problem; bootstrap; dissimilarity matrix; ecological community data; PERMANOVA; permutation test.

## 1. Introduction

The Behrens–Fisher problem (BFP) is one of the oldest puzzles in problems (Behrens 1929; Fisher 1935; Welch 1938). The essence of this problem is how validly to compare the means (or multivariate mean vectors) of two or more populations when their variances (or multivariate dispersions) differ. Solutions to the BFP for univariate data generally assume

*Author to whom correspondence should be addressed.
[1]New Zealand Institute for Advanced Study (NZIAS), Massey University, Albany campus, Private Bag 102 904, North Shore, Auckland 0745, New Zealand e-mail: m.j.anderson@massey.ac.nz
[2]Institute of Natural & Mathematical Sciences (INMS), Massey University, Albany Campus, Private Bag 102 904, North Shore, Auckland, 0745, New Zealand.
[3]Plymouth Marine Laboratory, Prospect Place, The Hoe, Plymouth PL1 3DH, UK.
[4]PRIMER-E Limited, c/o Plymouth Marine Laboratory, Prospect Place, The Hoe, Plymouth PL1 3DH, UK.
[5]CONACYT, Unidad Multidisciplinaria de Docencia e Investigacón Sisal, Facultad de Ciencias, Universidad Nacional Autónoma de México, Puerto de Sisal, Yucatán, México
*Acknowledgement.* This work was supported by a Royal Society of New Zealand Marsden Grant.

variables to be normally distributed (e.g., Wang 1971; Brown & Forsythe 1974; Clinch & Keselman 1982; Weerhandi 1993; Ghosh & Kim 2001). Similarly, the majority of solutions to the multivariate BFP (e.g., Johnson & Weerhandi 1988; Coombs & Algina 1996; Christensen & Rencher 1997; Gamage, Mathew & Weerhandi 2004; Belloni & Didier 2008; Krishnamoorthy & Lu 2010) assume variables to be multivariate normal (MVN), and are not amenable to high-dimensional problems where the number of variables may equal or exceed sample sizes.

High-dimensional multivariate data are very frequently encountered, however, in genetics, bioinformatics, ecology and environmental science. In ecology, high-dimensional multivariate data often consist of counts of species' abundances in a community. Such data show intrinsic mean-variance relationships (Taylor 1961) and tend to be long-tailed and overdispersed (Aitchison & Ho 1989), with each species having a unique degree of aggregation (McArdle & Anderson 2004) and potential zero inflation (McArdle, Gaston & Lawton 1990; Fletcher, Mackenzie & Villouta 2005). Promising new approaches to test the equality of mean vectors for high-dimensional heteroscedastic data on the basis of $U$-statistics have recently been proposed by Ahmad, von Rosen & Singull (2012) and Ahmad (2014). In ecology, however, interest more often lies in models of the associations (similarities, or dissimilarities) between pairs (or sets) of whole communities of species, rather than in models of individual variables.

Over the past 20+ years, several non- or semi-parametric methods have been developed to analyse multivariate ecological data. Specifically, the analysis of similarities (ANOSIM, Clarke) and permutational multivariate analysis of variance (PERMANOVA, McArdle & Anderson 2001; Anderson 2001) are very widely used. These methods are based on dissimilarities among sampling units which need not be Euclidean, so they flexibly allow tests of hypotheses regarding changes in the structure of communities represented by ecological measures, such as Jaccard, Bray-Curtis or Modified Gower (e.g., Anderson, Ellingsen & McArdle 2006; Clarke, Somerfield & Chapman 2006). These dissimilarity-based tests hence posit their hypotheses in (potentially) non-metric or semi-metric spaces, and use permutations (random re-assignment of sampling units to groups) to calculate $p$-values, so assume only exchangeability of the sampling units under the general null hypothesis of no differences among groups (Clarke 1993; Manly 2006, pp. 162–163).

Permutation tests are, however, sensitive to differences in dispersion; groups with different dispersions are not strictly exchangeable (Boik 1987; Clarke 1993; Hayes 1996). A recent study of the effects of heterogeneity of dispersions on dissimilarity-based tests (Anderson & Walsh 2013) found that ANOSIM was very strongly affected: rejection of the null hypothesis could easily be caused by differences in location, differences in dispersion, or both. Although PERMANOVA was robust to heterogeneity for balanced designs (its null hypothesis, by virtue of the construction of its test statistic, being much more focused on detecting differences in the location of the groups in the space of the chosen dissimilarity measure – see Anderson & Walsh (2013) for more details), it nevertheless demonstrated measurable effects on error rates for unbalanced cases. Specifically, the test is conservative when high dispersion occurs in larger groups and liberal when high dispersion occurs in smaller groups (Anderson & Walsh 2013), reflecting directly what has been observed for classical univariate ANOVA (Welch 1938; Horsnell 1953; Box 1954; Glass, Peckham & Sanders 1972). A separate test for homogeneity of multivariate dispersions based on dissimilarities is available (Anderson 2006), but the presence of significant heterogeneity

may still obscure inferences regarding potential differences in location. Currently, there is no test of the null hypothesis of equality in the location of groups of multivariate sampling units in the space of a chosen dissimilarity measure that simultaneously allows for heterogeneity of dispersions in that space among those groups.

Here, we propose some solutions to the multivariate BFP for dissimilarity-based analyses. Specifically, we propose a modification to the PERMANOVA pseudo $F$ test statistic that differs from the original for unbalanced designs, using a direct analogue to the modified $F$-statistic proposed by Brown & Forsythe (1974) for the generalised univariate BFP. A $p$-value can then be obtained using either the usual permutation algorithm or by using a bootstrap approach (e.g., Davison & Hinkley 1997, p. 161; Manly 2006, p. 71). A parametric bootstrap has been used successfully to tackle the univariate BFP (Krishnamoorthy, Lu & Mathew 2007). Here, we used a separate-sample bootstrap of residuals (groups must be centred on a common centroid under a true null hypothesis), thereby explicitly conditioning on unequal dispersions (Efron & Tibshirani 1993, p. 222; Manly 2006, p. 117). The bootstrap is known to be biased (e.g., Efron 1982, p. 27), hence some form of bias-correction can therefore also optionally be applied (e.g., Hall 1992, p. 36; Efron & Tibshirani 1993, p. 342; Davison & Hinkley 1997, p. 103).

This paper is structured as follows. Section 2 provides a description of the new proposed test statistic. The reader is also referred to Appendix A in which other relevant test statistics are briefly described. Section 3 describes a simulation study used to compare the performance of the new method with existing approaches, including ANOSIM, unmodified PERMANOVA, Pillai's trace (classical MANOVA) and a modification to Pillai's trace that is a direct multivariate analogue to Brown & Forsythe's (1974) approach (Coombs & Algina 1996). The method described by Coombs & Algina (1996) relies on multivariate normality, but in other respects provides a multivariate solution to the BFP in Euclidean space that is highly comparable to what we propose here more generally for dissimilarity-based analysis. This is the rationale for including it in our simulations as opposed to other potential solutions to the multivariate BFP — for purposes of comparison. An application of the new method, demonstrating its use in providing meaningful statistical inferences for high-dimensional multivariate ecological community data, is given in Section 4. Section 5 then describes a simulation study based on real high-dimensional ecological datasets, with special emphasis on the investigation of power. The paper concludes with a general discussion and directions for future research (Section 6).

## 2. Description of methods

### 2.1. PERMANOVA

Let $\boldsymbol{Y}$ be an $N \times p$ matrix of multivariate row vectors $\boldsymbol{y}_{ij}$ of length $p$, each belonging to one of $i = 1,\ldots,g$ groups, with $j = 1,\ldots,n_i$ sample rows in the $i$th group and $N = \sum_{i=1}^{g} n_i$. In ecology, these are commonly counts (or biomass or cover) of each of $p$ species in each of $N$ transects, quadrats, cores or other standardised sampling units. Thus $\boldsymbol{Y}$ describes the positions of $N$ sample points in a space of $p$ dimensions. Let $\boldsymbol{D}$ be an $N \times N$ symmetric matrix of distances or dissimilarities $\{d_{ij,i'j'}\}$ calculated between every pair of points. Any appropriate measure may be used here, depending on the hypotheses of interest, including those used to compare communities in ecology, such as Bray-Curtis, Jaccard, or Euclidean on $\log(y+1)$-transformed or proportional abundance data.

Consider a multi-response permutation test formulated as follows. First, as in Gower (1966), let matrix $\boldsymbol{A} = \{(-1/2)d_{ij,i'j'}^2\}$. Centring $\boldsymbol{A}$ on its rows and columns gives $\boldsymbol{G} = [\boldsymbol{I} - (1/N)\boldsymbol{J}_N]\boldsymbol{A}[\boldsymbol{I} - (1/N)\boldsymbol{J}_N]$, where $\boldsymbol{J}_N$ denotes an $N \times N$ matrix of 1s and $\boldsymbol{I}$ denotes an $N \times N$ identity matrix. Next, construct a projection matrix of the design of rank $r = g - 1$ as $\boldsymbol{H} = \text{diag}[(1/n_1)\boldsymbol{J}_{n_1}, \ldots, (1/n_g)\boldsymbol{J}_{n_g}] - (1/N)\boldsymbol{J}_N$. The PERMANOVA pseudo $F$ statistic (Anderson 2001; McArdle & Anderson 2001) for comparing the centroids among the $g$ groups is:

$$F_1 = \frac{\text{tr}(\boldsymbol{HG})/(g-1)}{\text{tr}[(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{G}]/(N-g)}, \tag{1}$$

where 'tr(·)' denotes the trace of a matrix. A $p$-value is then calculated as $P_1^\pi = \Pr(F_1^\pi \geqslant F_1)$, where $F_1^\pi$ is the value of $F_1$ obtained by a random equiprobable permutation $\pi$ (re-ordering) of the $1, \ldots, N$ rows of observations. The $F_1^\pi$ are realisations of $F_1$ under a true null hypothesis ($H_0$) of equality of centroids in the space of the chosen dissimilarity measure with the row units being exchangeable among the $g$ groups. Note that $F_1$ is equivalent in value to Fisher's univariate $F$-ratio if $\boldsymbol{D}$ contains Euclidean distances and $p = 1$. The denominator of $F_1$ is a pooled estimate of within-group dispersion, and exchangeability implies independence of row vectors and also homogeneity. In passing, we also note that, for balanced designs (equal sample sizes per group), the random re-ordering of observation rows results in there being an equal probability that any observation will fall into any particular group. This property does not hold, however, for unbalanced designs, for which we assert under $H_0$ merely that any ordering of the existing observations relative to the (fixed) grouping structure of specified sample sizes is equally likely.

## 2.2. Modified PERMANOVA

Following Brown & Forsythe (1974), we propose a modified pseudo $F$ statistic to account for heterogeneity:

$$F_2 = \frac{\text{tr}(\boldsymbol{HG})}{\sum_{i=1}^{g}(1 - \frac{n_i}{N})V_i}, \tag{2}$$

where $V_i = \sum_{j=1}^{(n-1)} \sum_{j'=(j+1)}^{n} d_{ij,ij'}^2/[n_i(n_i-1)]$ is the within-group dispersion for group $i$. The construction of $F_2$ explicitly acknowledges potential heterogeneity *via* estimation of these separate individual dispersions for each group. Hence, the null hypothesis for $F_2$ is equality of centroids in the space of a chosen dissimilarity measure *given* potential diffferences in dispersions. Note that $F_2$ is equivalent to $F_1$ for equal sample sizes and shares with $F_1$ the property that the expectations of numerator and denominator are equal under a true null hypothesis of no difference in the location of group centroids. Note also that for $p = 1$ variable and if the entries of $\boldsymbol{D}$ are Euclidean distances, then $V_i$ is the usual classical univariate unbiased measure of the sample variance for group $i$.

Separate measures of within-group dispersion for each group can also be obtained efficiently by considering the projection matrix for the residuals of the PERMANOVA model in the space of the dissimilarity measure. Specifically, the projection matrix for the residuals is $\boldsymbol{H}_{Res} = (\boldsymbol{I} - \boldsymbol{H})$ and a residualised $\boldsymbol{G}$ matrix is given by $\boldsymbol{R} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{G}(\boldsymbol{I} - \boldsymbol{H})$. Let $\boldsymbol{u}_i$ be an indicator vector of length $N$ for group $i$ that contains 1s in all positions $k \in \{1, \ldots, N\}$

where the $k$th sample unit belongs to group $i$ and zeros elsewhere. If we let $\boldsymbol{U}_i = \mathrm{diag}(\boldsymbol{u}_i)$ then a separate projection matrix for the residuals associated with group $i$ only is given by:

$$\boldsymbol{H}_{Res}^{(i)} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{U}_i(\boldsymbol{I} - \boldsymbol{H}). \tag{3}$$

Note also that $\boldsymbol{H}_{Res} = \sum_{i=1}^{g} \boldsymbol{H}_{Res}^{(i)}$. The within-group dispersions are then given by

$$V_i = \mathrm{tr}(\boldsymbol{H}_{Res}^{(i)}\boldsymbol{G})/(n_i - 1). \tag{4}$$

### 2.3. Calculating *p*-values

We shall consider here three potential methods for calculating a $p$-value for the modified PERMANOVA test statistic, $F_2$. First, a permutation test can be done in the usual way, as for $F_1$; namely, a $p$-value is calculated as $P_2^{\pi} = \Pr(F_2^{\pi} \geqslant F_2)$, where $F_2^{\pi}$ is the value of $F_2$ obtained by a random equiprobable permutation of the $N$ rows of observations. Intuitively, this may not seem satisfactory, as the act of exchanging observations appears to violate the key idea that heterogeneity of dispersions is already acknowledged among these groups under $H_0$ for the Behrens–Fisher case. However, permutation of raw data has been shown to provide satisfactory approximate $p$-values when coupled with pivotal test statistics (such as the $F$-statistic for balanced designs) in other contexts where conditioning on ancillary quantities would be desirable; e.g., for tests of partial regression coefficients in multiple regression models (Anderson & Robinson 2001) or for tests of individual factors in multi-way ANOVA designs (Anderson & ter Braak 2003). See also Manly (1995), Manly & Francis (1999, 2002) and Francis & Manly (2001) for some proposed solutions to the univariate BFP using approximate randomisation methods.

Second, we shall consider a separate-sample bootstrap of residuals (Efron & Tibshirani 1993, p. 222; Manly 2006, p. 117). Here, the basic idea is to condition on the known differences among groups and obtain realisations of outcomes for the test statistic under a true null hypothesis of equal centroids given unequal dispersions. This is achieved by: (i) centring the data from each group onto a common centroid (*i.e.*, by calculating residuals); and (ii) taking a random bootstrap sample of $n_i$ observation vectors (with replacement) from the $i$th group. As the bootstrap resampling is done separately within each group, any original differences in the within-group dispersions among those groups are maintained.

Specifically, for the bootstrap, consider the indexing vector $\boldsymbol{\tau} = (\boldsymbol{\tau}_1^{\top}, \boldsymbol{\tau}_2^{\top}, \ldots, \boldsymbol{\tau}_g^{\top})^{\top}$, where $\boldsymbol{\tau}_i$ is of length $n_i$ for the $i$th group and $\boldsymbol{\tau}$ simply contains integer values $\boldsymbol{\tau} = (1, \ldots, N)^{\top}$ corresponding to the original ordering of the observation units. Let $\boldsymbol{\tau}_i^{\beta}$ be a bootstrap sample with replacement of the $n_i$ indices from group $i$ alone. The complete index vector $\boldsymbol{\beta} = \{\beta_k\}$, $k = 1, \ldots, N$ for the separate-sample bootstrap is then given by $\boldsymbol{\beta} = [(\boldsymbol{\tau}_1^{\beta})^{\top}, (\boldsymbol{\tau}_2^{\beta})^{\top}, \ldots, (\boldsymbol{\tau}_g^{\beta})^{\top}]^{\top}$. Centred residuals of the full model in the space of the dissimilarity measure are represented by the residualised $\boldsymbol{G}$ matrix, $\boldsymbol{R} = \{r_{kk'}\}$; hence, a separate-sample bootstrap of these residuals is given by $\boldsymbol{R}^{\beta} = \{r_{\beta_k, \beta_{k'}}\}$. The modified PERMANOVA test statistic obtained by a separate-sample bootstrap of residuals in the space of the dissimilarity measure is then given by

$$F_2^{\beta} = \frac{\mathrm{tr}(\boldsymbol{H}\boldsymbol{R}^{\beta})}{\sum_{i=1}^{g}(1 - \frac{n_i}{N})V_i^{\beta}}, \tag{5}$$

where the within-group dispersions under bootstrapping are obtained as

$$V_i^{\beta} = \mathrm{tr}(\boldsymbol{H}_{Res}^{(i)}\boldsymbol{R}^{\beta})/(n_i - 1). \tag{6}$$

Following Fisher & Hall (1990), the $p$-value based on the bootstrap for hypothesis-testing is then calculated as $P_2^{\beta} = \Pr(F_2^{\beta} \geqslant F_2)$.

Bootstrap estimation of dispersion is known to be biased (see Appendix B and Efron 1982, p. 27). The third method we shall consider here is an empirically bias-adjusted bootstrap approach. Suppose a total of $B$ empirical bootstrap re-samples are done, yielding $\ell = 1,\ldots,B$ bootstrap values of within-group dispersion $V_{i\ell}^{\beta}$ for each of the $i = 1,\ldots,g$ groups. Let $\bar{V}_i^{\beta} = \sum_{\ell=1}^{B} V_{i\ell}^{\beta}/B$ be the mean of the empirical bootstrap distribution of dispersions for group $i$. One possible estimate of the bias for group $i$ is:

$$b_i = \bar{V}_i^{\beta} - V_i. \tag{7}$$

Hence, one possible empirical bias-adjustment to the individual dispersions under bootstrapping is $V_i^{\beta(ba)} = V_i^{\beta} - b_i$, and a possible bias-adjusted test statistic for a given bootstrap sample may then be constructed as

$$F_2^{\beta(ba)} = \frac{\mathrm{tr}(\boldsymbol{H}\boldsymbol{R}^{\beta})}{\sum_{i=1}^{g}(1 - \frac{n_i}{N})V_i^{\beta(ba)}}, \tag{8}$$

and its associated $p$-value is $P_2^{\beta(ba)} = \Pr(F_2^{\beta(ba)} \geqslant F_2)$.

We note in passing that no bias-adjustment has been performed for the numerator of equation (8) above, as the construction of residuals (*i.e.*, centring of groups onto a common centroid) for the full set of data already asserts that the "true" among-group variation is equal to zero under $H_0$. Hence, there is no obvious empirical estimate of potential bias in the numerator under bootstrap re-sampling. Note also that in the special case where the entries of **D** are Euclidean distances based on $p$ independent random variables, then we can use the direct result: $b_i = (-1/n)(V_i)$ (Appendix B).

## 3. Simulation study – Type I error

We used simulations to measure and compare empirical type I errors under a range of scenarios for the following seven methods: (i) classical MANOVA using Pillai's trace ('Pillai'); (ii) Pillai's trace modified to accommodate heterogeneous dispersions (Coombs & Algina 1996) ('Mod.Pillai'); (iii) ANOSIM (Clarke 1993); (iv) PERMANOVA based on $F_1$ and with $p$-values obtained by random permutations of raw observation vectors ('$F_1$(perm)'); modified PERMANOVA based on $F_2$ with $p$-values obtained by: (v) random permutations of observation vectors ('$F_2$(perm)'), (vi) separate-sample bootstraps ('$F_2$(boot)'), or (vii) separate-sample bootstraps with bias-adjustment ('$F_2$(ba.boot)'). Simulation scenarios were specifically designed to compare these methods under circumstances that have already been shown to be problematic for PERMANOVA and ANOSIM by Anderson & Walsh (2013). These circumstances consist of unbalanced designs having heterogeneous dispersions where either the greater dispersion occurs in the group with a smaller sample size (leading to overly liberal tests) or the greater dispersion occurs in the group with a larger sample size (leading to overly conservative tests). For relevance and simplicity in interpretation, as well as to allow comparisons with more classical techniques, we limited this part of our investigation of type I error to $g = 2$ groups from MVN distributions and to

analyses based on Euclidean distances. More complex and realistic scenarios (non-normal count data, with $p > n_i$ for all $i$ and with analyses based on ecological dissimilarity measures) are described in Section 5. We also calculated the type I error for $F_2$ where $p$-values were obtained using permutation of residuals ('$F_2$(perm.res)'), so as to differentiate the effect of calculating residuals from the effect of bootstrapping in the comparison of $F_2$(perm) with $F_2$(boot). All simulations described in Sections 3 and 5 were performed using R (R Core Team 2015).

### 3.1. Methods

Simulated datasets under a true null hypothesis were drawn randomly from each of $g = 2$ $p$-variate MVN populations having equal mean vectors (*i.e.*, $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$); every mean in each group had a value of 10. All covariances were set equal to zero and the $p \times p$ dispersion matrices for the two groups ($\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, respectively) differed by a constant scalar multiplier; *i.e.*, $\boldsymbol{\Sigma}_1 = m_1 \boldsymbol{I}_p$ and $\boldsymbol{\Sigma}_2 = m_2 \boldsymbol{I}_p$, where $\boldsymbol{I}_p$ is a $p \times p$ identity matrix. Under any particular scenario, $n_1 \leqslant n_2$ and the ratio of sample sizes is denoted by $r_n = n_2/n_1$. Note that these simulations targetted cases where heterogeneity occurs as differences only along the diagonal of dispersion matrices. This was our logical approach, as Anderson & Walsh (2013) have demonstrated that the usual PERMANOVA test ($F_1$) is unaffected by differences in correlation structure among groups (although the ANOSIM test is not). See, for example, the results of simulations labeled 'Sim3' in Anderson & Walsh (2013).

The full set of simulation scenarios is identifiable by all combinations of the following variables: $p = \{2, 5, 10\}$, $n_1 = \{10, 20, 40\}$, and $r_n = \{1, 2, 3\}$, first for the conservative case ($m_1 = 1$ and $m_2 = \{5, 10\}$) and then for the liberal case ($m_1 = \{5, 10\}$ and $m_2 = 1$). Under each scenario, 1000 datasets were simulated and the resampling methods (iii–vii above) each used 999 re-samples (either permutations or separate-sample bootstraps) to calculate a $p$-value for each simulated dataset. The $p$-values for methods (i) and (ii) above were obtained by referring to appropriate classical $F$-distributions (Appendix A). The *a priori* significance level was set at $\alpha = 0.05$ and the empirical type I error for each method was calculated as the proportion of $p$-values (out of 1000 simulated datasets) satisfying $P \leqslant \alpha$. A useful test is not only expected to have a type I error that matches $\alpha$, but is also expected to have a uniform distribution of $p$-values under a true null hypothesis; hence, we also formally compared the distribution of $p$-values obtained by each test under each simulation scenario with a uniform distribution, using the Anderson–Darling test (Anderson & Darling 1954; Marsaglia & Marsaglia 2004).

### 3.2. Results

The results of all simulation scenarios are provided as Supporting Information. Table S1 contains empirical type I errors and Table S2 the $p$-values of the Anderson–Darling tests for all methods. As the results for $F_2$(perm.res) mirrored those obtained for $F_2$(perm) (see Tables S1 and S2), $F_2$(perm.res) is not discussed further here, and any differences in the results for $F_2$(boot) versus $F_2$(perm) are interpreted as being due to the contrast of resampling 'with' versus 'without' replacement (and not due to the centring that is required prior to implementing the bootstrap).

The proposed modification of the PERMANOVA test statistic makes a world of difference to the performance of the test under heterogeneity for unbalanced designs. The type

Table 1. Empirical type I error for methods (i) - (vii) obtained from 1000 simulated MVN datasets with $p=5$ and $g=2$ based on Euclidean distances and with $p$-values obtained using 999 re-samples for methods (iii)–(vii). Scenarios were designed to examine balanced designs ($n_1=n_2$) and unbalanced designs ($n_1 \leqslant n_2$) in situations where existing methods are known to be either conservative (top-half, higher dispersion in larger-sized group) or liberal (bottom-half, higher dispersion in smaller-sized group).

| $n_1$ | $n_2$ | $m_1$ | $m_2$ | (i) Pillai | (ii) Mod.Pillai | (iii) ANOSIM | (iv) $F_1$(perm) | (v) $F_2$(perm) | (vi) $F_2$(boot) | (vii) $F_2$(ba.boot) |
|---|---|---|---|---|---|---|---|---|---|---|
| 20 | 20 | 1 | 5 | 0.077 | 0.060 | 1.000 | 0.066 | 0.066 | 0.044 | 0.056 |
| 20 | 40 | 1 | 5 | 0.007 | 0.048 | 0.000 | 0.007 | 0.055 | 0.043 | 0.053 |
| 20 | 60 | 1 | 5 | 0.003 | 0.053 | 0.000 | 0.000 | 0.050 | 0.047 | 0.054 |
| 20 | 20 | 1 | 10 | 0.082 | 0.047 | 1.000 | 0.046 | 0.046 | 0.038 | 0.045 |
| 20 | 40 | 1 | 10 | 0.001 | 0.047 | 0.000 | 0.002 | 0.041 | 0.040 | 0.045 |
| 20 | 60 | 1 | 10 | 0.000 | 0.062 | 0.000 | 0.000 | 0.068 | 0.059 | 0.067 |
| 20 | 20 | 5 | 1 | 0.057 | 0.040 | 1.000 | 0.045 | 0.045 | 0.031 | 0.038 |
| 20 | 40 | 5 | 1 | 0.232 | 0.051 | 1.000 | 0.232 | 0.052 | 0.030 | 0.045 |
| 20 | 60 | 5 | 1 | 0.356 | 0.040 | 1.000 | 0.342 | 0.052 | 0.034 | 0.040 |
| 20 | 20 | 10 | 1 | 0.086 | 0.043 | 1.000 | 0.054 | 0.054 | 0.034 | 0.043 |
| 20 | 40 | 10 | 1 | 0.311 | 0.039 | 1.000 | 0.258 | 0.049 | 0.027 | 0.035 |
| 20 | 60 | 10 | 1 | 0.502 | 0.060 | 1.000 | 0.446 | 0.060 | 0.033 | 0.046 |

I error was much closer to the *a priori* significance level of $\alpha=0.05$ when using $F_2$ rather than $F_1$, regardless of whether permutation or bootstrapping was used to obtain a $p$-value (Table 1, Table S1). This improvement in the dissimilarity-based PERMANOVA test when using $F_2$ versus $F_1$ mirrored that observed for Mod.Pillai versus Pillai for scenarios which can be modeled in Euclidean space on MVN data and where $p$ is substantially less than all $n_i$. The distribution of $p$-values was not significantly different from uniform when using Mod.Pillai or $F_2$(perm) in such cases (e.g., Fig. 1).

The results obtained using $F_2$(perm) tended to be closer to nominal $\alpha$ than those obtained using $F_2$(boot), which almost always yielded more conservative tests, particularly when higher dispersion occurred in the smaller-sized group (e.g., see the lower half of Table 1). The bias-adjusted bootstrap ($F_2$(ba.boot)), although still conservative, did tend to improve the performance of the unadjusted bootstrap, yielding type I errors closer to nominal $\alpha$ (Table 1, Table S1). More generally, conservativism for $F_2$(boot) and $F_2$(ba.boot) tended to increase with increasing dimensionality, increasing heterogeneity and increasing differences in sample size (Table S1). The fact that ANOSIM has high rejection rates in the presence of most types of heterogeneity (Table 1) is a direct reflection of it being a more general 'portmanteau' test for differences of any sort among groups of samples (Clarke 1993; Anderson & Walsh 2013).

The distribution of $p$-values for $F_2$(boot) and $F_2$(ba.boot) deviated most from a uniform distribution under scenarios where the number of variables became large relative to sample sizes (Table 2, Table S2). In such cases, bootstrapping yielded unimodal distributions of $p$-values under simulation (Fig. 2). Mod.Pillai also suffered under these scenarios, yielding predominantly large $p$-values (Fig. 2). These issues disappeared for Mod.Pillai once the individual sample sizes were about four times the number of variables (Table 2). Although the type I error for $F_2$(boot) and $F_2$(ba.boot) similarly approached more acceptable levels (closer to $\alpha$) with increasing $n_i$, their distributions of $p$-values still deviated significantly from a uniform distribution (Table 2). In contrast, $F_2$(perm) maintained type I error close

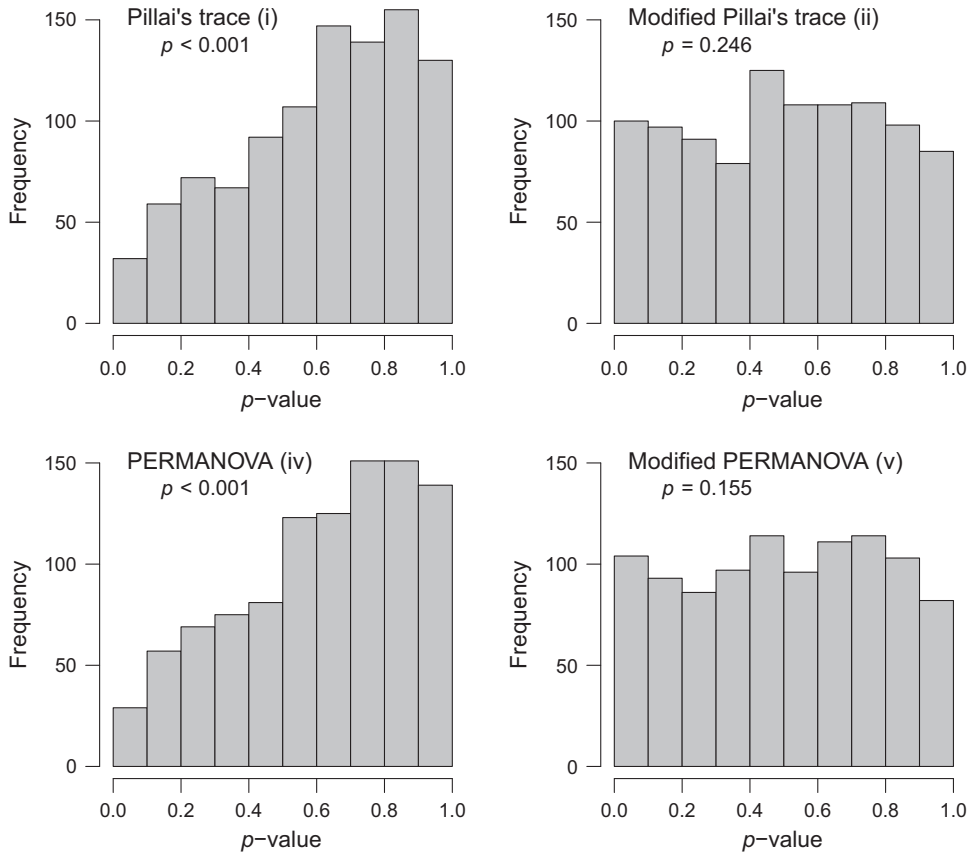


Figure 1. Frequency distributions of 1000 $p$-values obtained from MVN simulated data for each of four different methods: (i) Pillai, (ii) ModPillai, (iv) $F_1$(perm) and (v) $F_2$(perm) under a scenario of an unbalanced design producing conservativism for methods (i) and (iv), specifically: $p = 5$, $g = 2$, $n_1 = 20$, $n_2 = 40$, $m_1 = 1$ and $m_2 = 5$. Calculations for methods (iv) and (v) were based on Euclidean distances and each $p$-value was obtained using 999 permutations. The $p$-value associated with the Anderson–Darling test for uniformity is also shown on the plot for each method.

to the nominal $\alpha$ and also yielded uniform distributions of $p$-values for virtually all of the scenarios we examined (Tables 1 and 2; Tables S1 and S2).

## 4. Ecological application

Ellingsen & Gray (2002) described spatial variation in marine biodiversity along the Norwegian continental shelf, examining data consisting of counts of $p = 809$ species of benthic soft-sediment macrofauna in $N = 101$ sites sampled from five areas (Areas 1–5, from south to north) spanning 15° of latitude from the North Sea to the Arctic. Sample sizes (the number of sites) differ across the areas: $n_1 = 16$, $n_2 = 21$, $n_3 = 25$, $n_4 = 19$ and $n_5 = 20$. A non-metric multi-dimensional scaling (MDS) ordination plot based on the Jaccard dissimilarities among sites, using only presence/absence (1,0) information (Fig. 3), shows not only the gradient in community structure from south to north (along MDS axis 1), but

Table 2. Results of empirical simulations for methods obtained from 1000 simulated MVN datasets with $p = 10$ and $g = 2$ based on Euclidean distances, with $m_1 = 10$ and $m_2 = 1$ (*i.e.*, larger dispersion in the smaller-sized group) showing (a) empirical type I error and (b) $p$-value associated with the Anderson–Darling test to compare the distribution of $p$-values obtained under simulation with a uniform distribution.

| $n_1$ | $n_2$ | (ii) Mod.Pillai | (v) $F_2$(perm) | (vi) $F_2$(boot) | (vii) $F_2$(ba.boot) |
|---|---|---|---|---|---|
| *(a) Empirical Type I error* | | | | | |
| 10 | 10 | 0.000 | 0.063 | 0.009 | 0.027 |
| 20 | 20 | 0.037 | 0.059 | 0.027 | 0.041 |
| 40 | 40 | 0.053 | 0.054 | 0.036 | 0.040 |
| 10 | 20 | 0.000 | 0.061 | 0.017 | 0.035 |
| 20 | 40 | 0.037 | 0.056 | 0.020 | 0.027 |
| 40 | 80 | 0.046 | 0.037 | 0.016 | 0.024 |
| 10 | 30 | 0.000 | 0.057 | 0.013 | 0.026 |
| 20 | 60 | 0.031 | 0.052 | 0.019 | 0.026 |
| 40 | 120 | 0.056 | 0.060 | 0.037 | 0.046 |
| *(b) Anderson–Darling p-value (deviation from uniformity)* | | | | | |
| 10 | 10 | 0.000 | 0.061 | 0.000 | 0.000 |
| 20 | 20 | 0.037 | 0.235 | 0.000 | 0.000 |
| 40 | 40 | 0.816 | 0.626 | 0.007 | 0.000 |
| 10 | 20 | 0.000 | 0.214 | 0.000 | 0.000 |
| 20 | 40 | 0.000 | 0.034 | 0.002 | 0.000 |
| 40 | 80 | 0.564 | 0.417 | 0.000 | 0.000 |
| 10 | 30 | 0.000 | 0.131 | 0.000 | 0.000 |
| 20 | 60 | 0.001 | 0.364 | 0.000 | 0.000 |
| 40 | 120 | 0.376 | 0.726 | 0.003 | 0.000 |

also clear heterogeneity in dispersions for sites from different areas. Area 3 has markedly greater dispersion, while Area 1 has markedly less dispersion, than the other three areas (2, 4 or 5). Recalling that Jaccard dissimilarity can be directly interpreted as the proportion of unshared species, multivariate dispersion here provides a direct measure of ecological beta diversity (variation in the identities of species, see Anderson, Ellingsen & McArdle 2006).

An overall test comparing the groups for changes in community structure in Jaccard space is statistically significant, using either ANOSIM ($R = 0.729$, $P = 0.0001$, 9999 permutations) or PERMANOVA ($F_{4,96} = 12.9$, $P = 0.0001$, 9999 permutations). However, the test for heterogeneity is also statistically significant (PERMDISP, $F_{4,96} = 49.8$, $P = 0.0001$, 9999 permutations of residuals). Given the unbalanced design, one may ask whether the differences among groups detected by either ANOSIM or PERMANOVA are caused by differences in centroids, differences in within-group dispersions or both. The visual pattern of separation of the sites from different areas observed in the MDS plot provides some support for the notion of differences in centroids (Fig. 3, top), but does not yield a probabilistic statement for direct statistical inference or interpretation. However, focussing on centroids only and explicitly conditioning on the known differences in within-group dispersions among groups in Jaccard space, the modified PERMANOVA test we propose here indeed yields a statistically significant result. This occurs whether the $p$-value is obtained using permutations ($F_{4,96} = 12.9$, $P = 0.0001$), bootstraps ($P = 0.0001$), or bias-adjusted bootstraps ($P = 0.0001$). Thus, our new test provides unequivocal evidence against the null

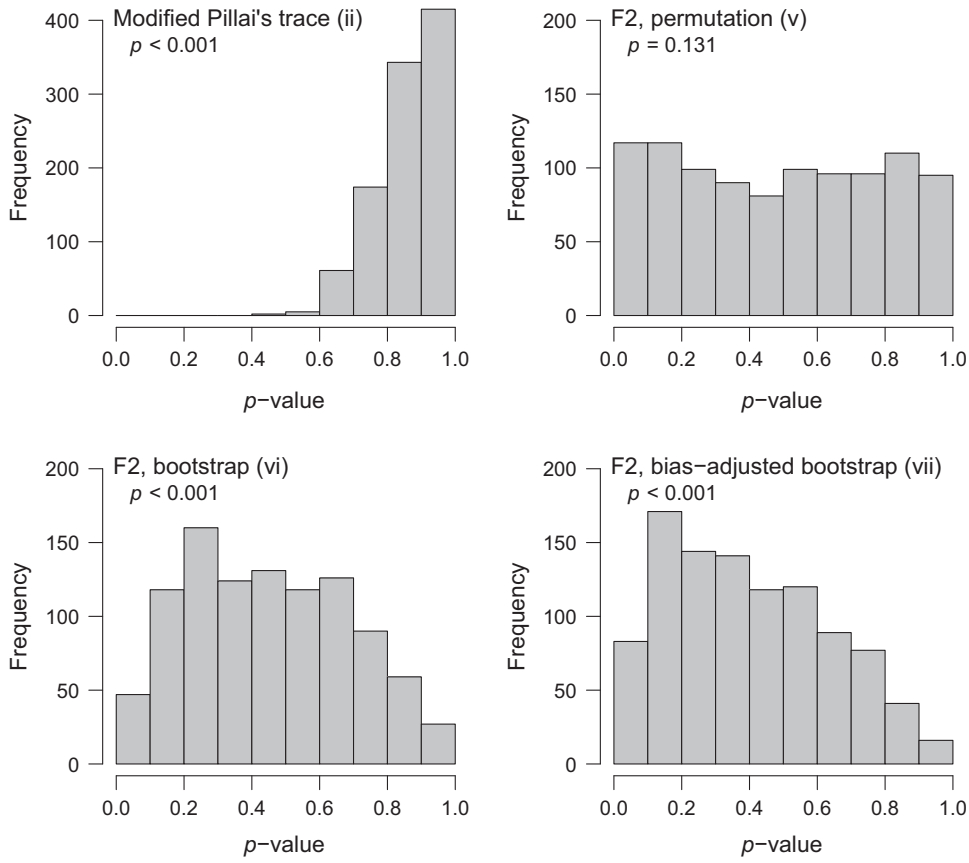


Figure 2. Frequency distributions of 1000 $p$-values obtained from MVN simulated data for each of four different methods: (ii) ModPillai, (v) $F_2$(perm), (vi) $F_2$(boot) and (vi) $F_2$(ba.boot) under a scenario where $p = n_1$, producing conservativism for methods (ii), (vi) and (vii), specifically: $p = 10$, $g = 2$, $n_1 = 10$, $n_2 = 30$, $m_1 = 10$ and $m_2 = 1$. Calculations for methods (v), (vi) and (vii) were based on Euclidean distances and each $p$-value was obtained using 999 re-samples. The $p$-value associated with the Anderson–Darling test for uniformity is also shown on the plot for each method.

hypothesis of no differences in centroids among the groups, and in the clear presence of heterogeneous dispersions.

The utility of the proposed method is even more striking when we consider the pair-wise comparison of Area 2 versus Area 3. Here, the MDS plot is dominated by high variability in community structure for Area 3 (Fig. 3, bottom), which clearly represents an important transitional area of high biotic turnover in species' identities. Area 2 had an average distance-to-centroid of 42.5% in Jaccard space, while Area 3 had an average distance-to-centroid of 56.4%, a significant difference by PERMDISP ($F_{1,44} = 76.1$, $P = 0.0001$). The pair-wise test is also highly statistically significant using either ANOSIM ($R = 0.313$, $P = 0.0002$) or PERMANOVA ($F_{1,44} = 6.53$, $P = 0.0001$), but the MDS plot offers little additional help to unravel the potential confounding of inferences regarding a difference in dispersions versus a difference in centroids between these two groups. One could rely on the expectation that greater dispersion in the group having a larger sample size (Area 3 in this case) should yield,

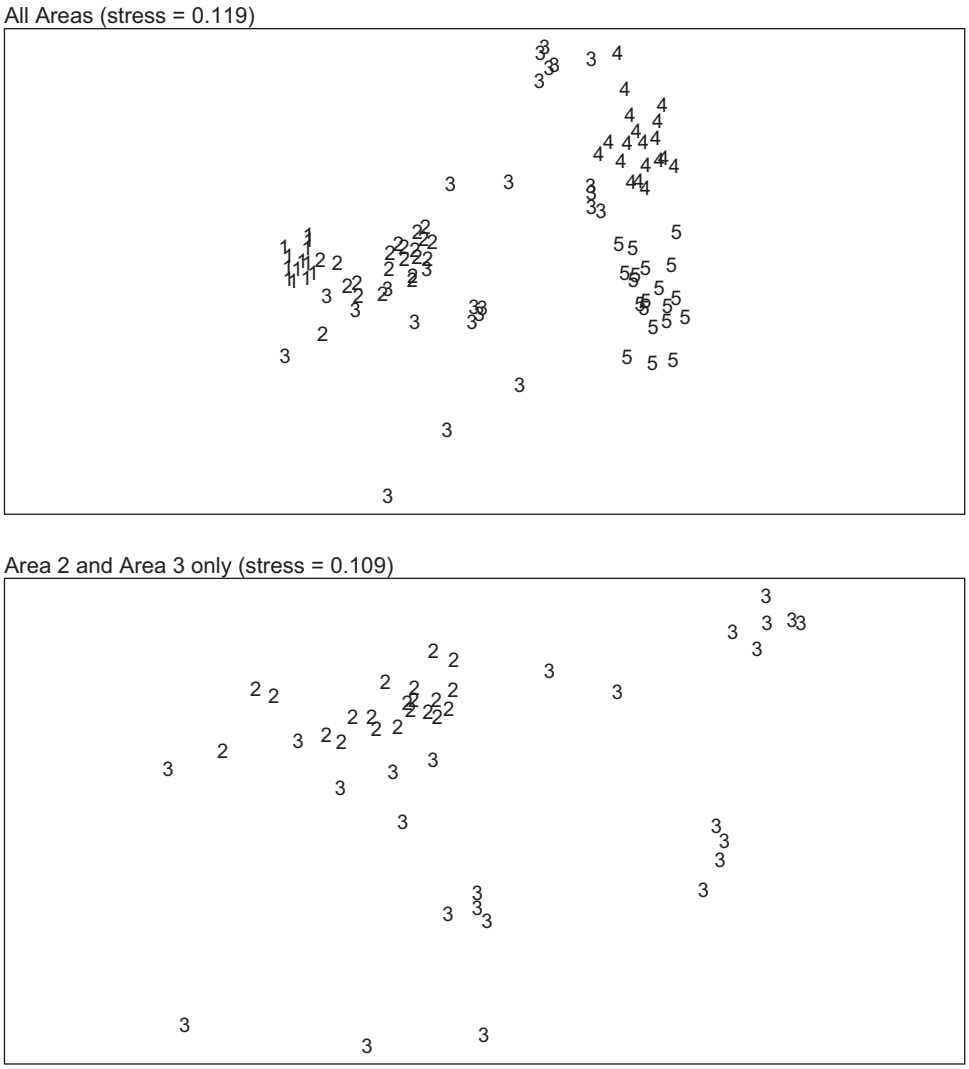


Figure 3. Non-metric multi-dimensional scaling (MDS) ordination plots of benthic soft-sediment assemblages based on Jaccard dissimilarities calculated on the presence/absence (0,1) of each of 809 species variables in 101 sites across five areas spanning 1960 km off the coast of Norway (labeled 1–5 from south to north, see Ellingsen & Gray (2002) for a map and further details).

if anything, a more conservative test *via* PERMANOVA. Thus, although the very small $p$-value obtained certainly lends support to the notion of differences in centroids, unfortunately such an inference is still, at best, indirect. In contrast, the modified PERMANOVA analysis specifically homes in on the null hypothesis of no difference in centroids, *given* differences in dispersion, and rejects this null hypothesis resoundingly ($F_{1,44} = 6.84$, $P = 0.0001$), regardless of the re-sampling method used. (An equivalent $p$-value was obtained using permutations, bootstraps or bias-adjusted bootstraps of residuals).

## 5. Simulations based on real datasets – Power

Simulations to examine type I error and power of the newly proposed methods for analysing changes in community structure based on dissimilarity measures for real ecological data were based on two datasets. These were also used by Anderson & Walsh (2013) to compare rejection rates for several existing multivariate tests. A full description of the simulation methods and associated R code for implementation have been provided by Anderson & Walsh (2013); brief descriptions are provided below.

### 5.1. Methods

The first dataset of interest was the benthic marine soft-sediment macrofaunal dataset from Norway described by Ellingsen & Gray (2002) and analysed in Section 4 above. These data are also provided here as Supporting information (Data S1). For simulations, we focused on the pair-wise comparison of Area 2 versus Area 3 for presence/absence data only. These areas had different sample sizes and differed significantly in their multivariate dispersions (Section 4, Fig. 3).

The probability of occurrence of each species in each Area was estimated directly using the method of moments. Presence/absence data were then simulated by taking a Bernoulli random draw (separately for each species) with probabilities set equal to these estimated parameters. First, the parameters for both Areas in the pair-wise comparison matched those for Area 2 (*i.e.*, $H_0$ was true). Then, to generate a power curve, the parameters for Area 3 were gradually changed in 10 equal steps, eventually to match those originally estimated from the real data for Area 3. At each step (and also under $H_0$), 1000 simulated datasets were produced from which Jaccard dissimilarity matrices were calculated. The empirical proportion of rejections of the null hypothesis (using $\alpha = 0.05$) was recorded for each of the following tests: (iii) ANOSIM; (iv) $F_1$(perm), (v) $F_2$(perm), (vi) $F_2$(boot), and (vii) $F_2$(ba.boot), with $p$-values estimated using 999 re-samples (permutations or bootstraps).

The second dataset consisted of counts of abundances of $p = 173$ taxa of benthic macrofauna sampled from $N = 39$ sites in a five-spoke radial design at increasing distances from the Ekofisk oil platform in the North Sea (Gray et al. 1990). The sites were classified into groups (A, B, C or D) to indicate increasing proximity to the oil platform. We focused here on the pair-wise comparison of group B (1–3.5 km from the platform, $n_B = 12$) versus group C (250 m–1 km from the platform, $n_C = 10$). Individual response variables (taxa) were first classified as being well-modeled using either a Poisson or negative binomial (NB) distribution. Means and dispersion parameters for each taxon in each group were estimated using the method of moments. Power curves were generated as described for the Norway data, in ten equal steps between the two sets of parameters for all taxa across the two groups, but where counts of individual taxon abundances were simulated by a random draw from their respective distribution (either Poisson or NB). This approach generally follows the spirit of simulations of multivariate ecological data along gradients for power analyses done by Somerfield, Clarke & Olsgard (2002). Empirical power was estimated as described above, but the procedure was repeated three times, based on each of the following: Bray-Curtis dissimilarities on fourth-root-transformed abundances, Chi-squared distances, and Euclidean distances on $\log(y + 1)$-transformed abundances. For more details, see Anderson & Walsh (2013). Of course, dispersion in dissimilarity space is often more about differences in species turnover within different groups than it is about differing variation in individual

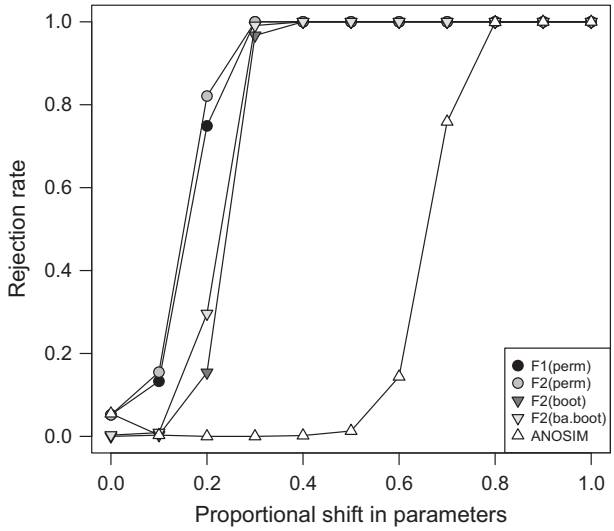


Figure 4. Proportion of rejections of the null hypothesis (at $\alpha = 0.05$) for each of five methods based on Jaccard dissimilarities for 1000 presence/absence datasets simulated as random draws for each species from a Bernoulli distribution with probabilities based on occurrences of Norwegian soft-sediment benthic macrofauna ($p = 809$ species). The $x$-axis depicts the proportional shift in the parameters along a continuum for the comparison of Area 2 versus Area 3. $P$-Values for the tests were calculated using 999 re-samples (permutations or bootstraps).

species variables, inevitably so when the dissimilarity measure (such as Jaccard) is based solely on presence/absence information, or when a strong transformation such as fourth-root or log has been used on quantitative data.

### 5.2. Results

Under all scenarios, $F_2$(perm) had the greatest power, matching or exceeding that obtained by $F_1$(perm) (Figs 4 and 5). This was true even for simulations based on the Ekofisk dataset, which did not show significant heterogeneity of dispersions for the two groups being compared (PERMDISP: $P_{BC} = 0.458$, $P_{Chi} = 0.242$, and $P_{Euc} = 0.936$, 9999 permutations). The use of bootstrapping ($F_2$(boot)) led to more conservative and hence less powerful tests, and although the use of a bias-adjustment ($F_2$(ba.boot)) improved power markedly, it did not match the power obtained using permutations (Figs 4 and 5). For the Norway data, ANOSIM had very low power, lagging far behind the other tests, presumably due to the greater dispersion occuring in the group with the larger sample size (Fig. 4). The oddly non-monotonic shape of the power curve for ANOSIM in this case was also previously noted by Anderson & Walsh (2013). In contrast to the Norway example, ANOSIM had more power than the bootstrapping methods for simulations based on the Ekofisk data, and came close to that obtained by $F_2$(perm) when the analysis was based on either Euclidean distances of $\log(y+1)$-transformed data or Bray-Curtis dissimilarities on fourth-root-transformed data (Fig. 5). In summary, although these simulations are by no means exhaustive, the use of $F_2$(perm) clearly had the best overall performance in terms of both type I error and power across all of the scenarios investigated here.

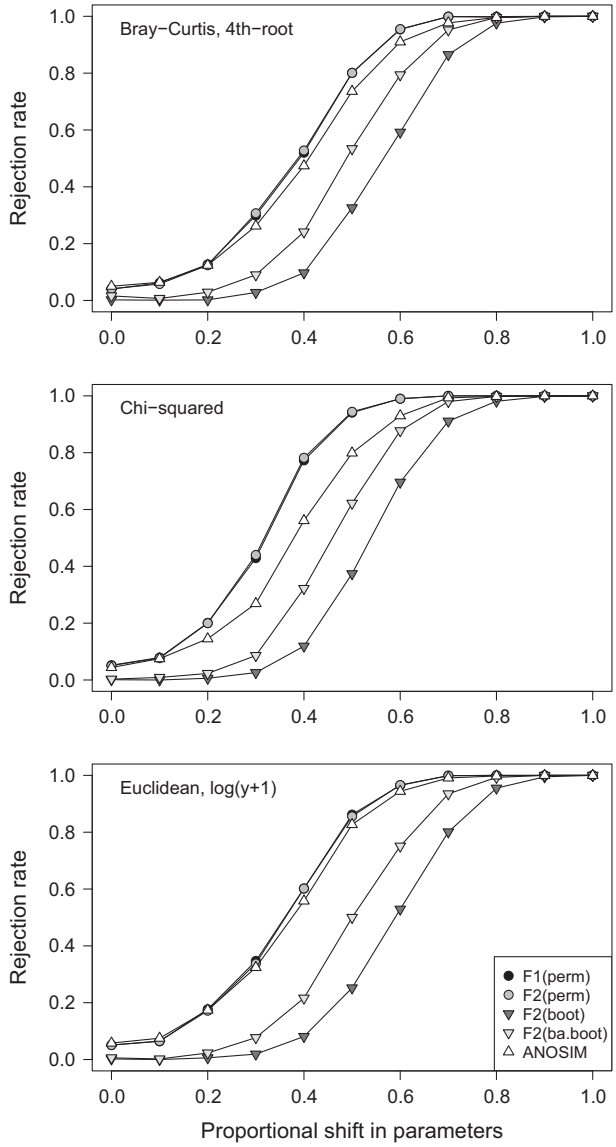


Figure 5. Proportion of rejections of the null hypothesis (at $\alpha = 0.05$) for each of five methods based on three different resemblance measures (Bray-Curtis on fourth-root-transformed counts, Chi-squared distances and Euclidean distances on $\log(y + 1)$-transformed counts) for 1000 species abundance datasets simulated as random draws for each species from either Poisson or negative binomial distributions, with parameters estimated from soft-sediment benthic macrofauna surrounding the Ekofisk oil platform ($p = 173$ taxa). The $x$-axis depicts the proportional shift in the parameters along a continuum for the comparison of samples from group B versus group C. $P$-Values for the tests were calculated using 999 re-samples (permutations or bootstraps).

## 6. Discussion

The proposed Behrens–Fisher modification to the PERMANOVA test statistic ($F_2$), coupled with a simple permutation algorithm ($F_2$(perm)) provides a test that maintains both level accuracy and uniform distributions of $p$-values under a true null hypothesis in the face of heterogeneity in multivariate dispersions for balanced or unbalanced designs. It matches the robustness of Mod.Pillai for MVN data, and out-performs this classical multivariate BFP approach for cases where one or more $n_i$ are small relative to $p$. The key motivation for the development of $F_2$, however, is to analyse ecological datasets, where interest lies in modelling community structure in the space of a chosen dissimilarity measure. This cannot be achieved using any other existing approaches to the multivariate BFP. Our simulation studies provide evidence that $F_2$(perm) is more powerful than the separate-sample bootstrap ($F_2$(boot) or $F_2$(ba.boot)). Indeed, in our simulations its power matched or exceeded the unmodified PERMANOVA test for alternative hypotheses simulated from real ecological data, even in the absence of heterogeneity.

How does $F_2$(perm) maintain the empirical probability of a type I error at the nominal significance level ($\alpha$)? It is likely that the construction of the $F_2$ test statistic ensures that a consistent under- or over-estimate of the pooled within-group dispersion that would arise under permutation (which mixes the original unequal dispersions among all groups) simply cancels out, as the numerator and denominator would still have equivalent expectations under a true null hypothesis. Although the permutation method performed the best overall, the separate-sample bootstrap is more satisfying conceptually, as it conditions explicitly on known differences in dispersion. We therefore consider that there remains a role for the bootstrap approach. The known conservatism of the bootstrap test can be used to advantage. In the face of clear differences in within-group dispersions, a statistically significant difference obtained using either $F_2$(boot) or $F_2$(ba.boot) must be viewed as very solid (and not spurious) evidence against the null hypothesis of equality of centroids in the space of a chosen dissimilarity measure.

Further work should be done to compare the performance of the methods proposed here with other proposed solutions to the multivariate BFP that cater to non-normal high-dimensional data, such as methods based on $U$-statistics (Ahmad, von Rosen & Singull 2012; Ahmad 2014). Although the latter are based on analyses purely in the Euclidean space of the original variables, their potential utility in non-Euclidean spaces may be explored through the use of either principal coordinates (Gower 1966) or metric MDS of dissimilarity matrices. Another potentially useful approach would be to generalise the approximate randomisation test proposed by Manly (1995) for the univariate BFP to multivariate situations. The essential idea here is to perform a permutation test after first standardising the data to have a distribution with a common variance (the appropriate linear transformation to apply must be discovered iteratively; see Manly (1995)). Further research would be required to uncover how this might be achieved in the space of a chosen dissimilarity measure and also to clarify how covariance structures would be treated in this context. With the development of additional potential methodologies, a broader simulation study of these and other methods, across a wider array of ecological datasets, as well as work investigating performance for other types of high-dimensional data, would clearly be desirable.

Finally, we note that an additional important and useful property of the methods we have proposed here is that they can be readily extended to multi-factor hierarchical ANOVA designs, where one might wish to allow for differences in both the numbers of levels and

in the dispersions of the centroids associated with those levels, either within or between factors that occur at different positions in the hierarchy. Indeed one may wish to allocate greater resources and increase the replication of sites within certain regions in which there is known to be greater variability, generating *a priori* unbalanced designs (Cochran 1977, pp. 96–99). For example, consider an asymmetrical hierarchical design, with varying numbers of sites nested within each of several regions, and with varying numbers of replicate sample units within each site. The $F_2$ test statistic can be extended to allow for differences in dispersions of replicates within sites, and differences in dispersions of site centroids within regions, for relevant tests of individual factors at each spatial scale. Similar extensions can be formulated and derived for tests of individual terms in fixed, random or mixed multi-way ANOVA models including interactions. We shall leave these extensions (beyond the scope of the current contribution) for a future endeavour.

## Appendix A. Description of other comparative test statistics

Three other test statistics were compared for their performance in simulation studies alongside the newly proposed techniques: analysis of similarities (ANOSIM; Clarke 1993), Pillai's trace from classical MANOVA (Bartlett 1939; Pillai 1955) and a modification to Pillai's trace that provides a direct multivariate analogue to Brown & Forsythe's (1974) solution to the univariate BFP (Coombs & Algina 1996).

### A.1. ANOSIM

Let the $M = N(N-1)/2$ pair-wise dissimilarities (or distances) in the sub-diagonal of matrix $\boldsymbol{D}$ be replaced by their ranks, with the lowest dissimilarity being given a rank of 1, and let the average of the ranked values between sample units that are in the same group be denoted by $\bar{r}_W$ and the average of the ranked values between sample units that are in different groups be denoted by $\bar{r}_B$. The ANOSIM $R$-statistic is a measure of the distinctiveness of the groups, ranging from $-1$ to $+1$, and is defined as:

$$R = \frac{(\bar{r}_B - \bar{r}_W)}{M/2},$$

A $p$-value associated with the general null hypothesis of no differences among the groups is obtained by calculating values of the test statistic $R$ under random equiprobable permutation $\pi$ of the original $i = 1,\ldots,N$ observation vectors to yield $R^\pi$; then, $P_R^\pi = \Pr(R^\pi \geqslant R)$. For more details regarding the null hypothesis and a comparison with PERMANOVA, see Anderson & Walsh (2013).

### A.2. Pillai's trace

We wish to test the equality of $g$ population mean vectors when independent random samples are drawn from populations that are distributed as multivariate normal (MVN) with equal dispersion matrices, using multivariate analysis of variance (MANOVA). Let $\boldsymbol{y}_{ij}$ denote the $j$th multivariate sample unit (a vector of length $p$) from the $i$th group, let $\bar{\boldsymbol{y}}_i$ denote the sample mean vector for each of $i = 1,\ldots,g$ groups and let $\bar{\boldsymbol{y}}$ denote the overall sample mean vector. Between-group variation is quantified by

$$\boldsymbol{B} = \sum_{i=1}^{g} n_i(\bar{\boldsymbol{y}}_i - \bar{\boldsymbol{y}})(\bar{\boldsymbol{y}}_i - \bar{\boldsymbol{y}})^\top,$$

and within-group variation is quantified by $\boldsymbol{W}=\sum_{i=1}^{g}(n_i-1)\boldsymbol{S}_i$, where

$$\boldsymbol{S}_i=\frac{1}{(n_i-1)}\sum_{j=1}^{n_i}(\boldsymbol{y}_{ij}-\bar{\boldsymbol{y}}_i)(\boldsymbol{y}_{ij}-\bar{\boldsymbol{y}}_i)^{\top}.$$

The Pillai-Bartlett trace criterion (Bartlett 1939; Pillai 1955) is then

$$V_P=\mathrm{tr}[\boldsymbol{B}(\boldsymbol{B}+\boldsymbol{W})^{-1}].$$

For MVN samples having equal dispersion matrices and under a true null hypothesis of equality of the population mean vectors, the transformation

$$F_{V_P}=\frac{(2t+s+1)V_P}{(2q+s+1)(s-V_P)}$$

is approximately distributed as an $F$ distribution with $s(2q+s+1)$ and $s(2t+s+1)$ degrees of freedom, where $v_1=(g-1)$, $v_2=(N-g)$, $s=\min(v_1,p)$, $q=(|v_1-p|-1)/2$ and $t=(v_2-p-1)/2$, and we must have $v_2\geqslant p$.

### A.3. Modified Pillai's trace

Following Coombs & Algina (1996), consider a modification to the matrix $\boldsymbol{W}$ above in the spirit of the Brown & Forsythe (1974) univariate solution to the BFP, as

$$\boldsymbol{M}=\frac{f}{(g-1)}\sum_{i=1}^{g}\left(1-\frac{n_i}{N}\right)\boldsymbol{S}_i,$$

where $f$ is the degrees of freedom for $\boldsymbol{M}$, and is defined in practice (where true but unknown dispersion matrices for individual groups are replaced by $\boldsymbol{S}_i$) as

$$f=\frac{[\mathrm{tr}(\sum_{i=1}^{g}c_i\boldsymbol{S}_i)]^2+\mathrm{tr}[(\sum_{i=1}^{g}c_i\boldsymbol{S}_i)^2]}{\sum_{i=1}^{g}\{[\mathrm{tr}(\sum_{i=1}^{g}c_i\boldsymbol{S}_i)]^2+\mathrm{tr}[(\sum_{i=1}^{g}c_i\boldsymbol{S}_i)^2]\}/(n_i-1)}$$

where $c_i=1-n_i/N$ for $i=1,\ldots,g$.

The modified Pillai's trace statistic is then defined as:

$$V_P^*=\mathrm{tr}[\boldsymbol{B}(\boldsymbol{B}+\boldsymbol{M})^{-1}].$$

Let $t^*=(f-p-1)/2$. An approximate $p$-value is obtained by using the transformed variable

$$F_{V_P^*}=\frac{(2t^*+s+1)V_P^*}{(2q+s+1)(s-V_P^*)},$$

which is distributed approximately according to an $F$ distribution with $s(2q+s+1)$ and $s(2t^*+s+1)$ degrees of freedom.

## Appendix B. Bias in the bootstrap

### B.1. Univariate

We show here that the exact downward bias in the bootstrap estimate of the variance of the mean of a univariate random variable is $(1-1/n)$. Consider a univariate random variable (r.v.) $Y$ with mean $\mu=0$ and variance $\sigma^2=1$, so $n$ samples from this distribution yield r.v.s $\{Y_j;j=1,\ldots,n\}$ and a single set of realised sample values $\{y_j;j=1,\ldots,n\}$, with sample

average $\bar{y} = \sum_{j=1}^{n} y_j/n$. Next, obtain $B$ random bootstrap samples with replacement of size $n$ from the sample values and let the r.v.s $\{Z_\ell; \ell = 1,\ldots,B\}$ denote $B$ bootstrap averages. If the bootstrap were unbiased, then we would have $\text{var}(Z_\ell) = 1/n$.

Conditioning on the sample values $\{Y_j = y_j; j = 1,\ldots,n\}$, $\{Z_\ell\}$ are independent and identically distributed for all $\ell$. Without loss of generality (w.l.o.g.), consider only the first bootstrap sample and its mean, $Z$. Let $T_j$ be the r.v. denoting the number of times $y_j$ is picked ($j = 1,\ldots,n$), so $\sum_{j=1}^{n} T_j = n$ and $\{T_j; j = 1,\ldots,n\}$ are multinomial with

$$\Pr(T_1 = t_1, T_2 = t_2,\ldots, T_n = t_n) = \frac{n!}{t_1!t_2!\cdots t_n!}\left(\frac{1}{n}\right)^{t_1}\left(\frac{1}{n}\right)^{t_2}\cdots\left(\frac{1}{n}\right)^{t_n}.$$

Then $Z = \sum_{j=1}^{n} T_j y_j/n$. The marginal distribution of $T_j$ is binomial$(n,p)$, where $p = 1/n$; so $\text{E}(T_j) = np = 1$ and $\text{var}(T_j) = npq = (1 - 1/n)$. Thus, $\text{E}(Z) = (\sum_{j=1}^{n} y_j/n)\cdot\text{E}(T_j) = \bar{y}$, showing the bootstrap average is an unbiased estimator for the sample average $\bar{y}$ if the latter is regarded as fixed. Unconditionally, $\text{E}(Z) = \text{E}_{Y_1,Y_2,\ldots,Y_n}[\text{E}(Z|Y_1, Y_2,\ldots, Y_n]$ (from the formula for conditional expection) so $\text{E}(Z) = \text{E}(\bar{Y}) = 0$ and the bootstrap average is an unbiased estimator of the mean of the underlying distribution. Next,

$$\text{var}(Z|\{Y_j = y_j\}) = \text{var}\left(\left[\sum_{j=1}^{n} T_j y_j\right]/n\right) = \frac{1}{n^2}\left[\sum_{j=1}^{n}\text{var}(T_j y_j) + \sum_{j}\sum_{\substack{j' \\ j\neq j'}} y_j y_{j'}\text{cov}(T_j, T_{j'})\right].$$

Now, the joint distribution of $(T_j, T_{j'})$ is trinomial; w.l.o.g., consider just $(T_1, T_2)$:

$$\Pr(T_1{=}t_1, T_2{=}t_2) = \frac{n!}{t_1!t_2!(n{-}t_1{-}t_2)!}\left(\frac{1}{n}\right)^{t_1}\left(\frac{1}{n}\right)^{t_2}\left(1-\frac{2}{n}\right)^{n-t_1-t_2}$$

So

$$\text{E}(T_1T_2) = \sum_{t_1}\sum_{t_2}\frac{n(n{-}1)(n{-}2)!}{(t_1{-}1)!(t_2{-}1)!(n{-}t_1{-}t_2)!}\left(\frac{1}{n}\right)^{2}\left(\frac{1}{n}\right)^{t_1-1}\left(\frac{1}{n}\right)^{t_2-1}\left(1-\frac{2}{n}\right)^{n-t_1-t_2}$$

and the right-hand side can be written as $[n(n-1)]/(n^2)$ multiplied by

$$\sum_{t_1}\sum_{t_2}\frac{(n{-}2)!}{(t_1{-}1)!(t_2{-}1)!\,[(n{-}2){-}(t_1{-}1){-}(t_2{-}1)]!}\left(\frac{1}{n}\right)^{t_1-1}\left(\frac{1}{n}\right)^{t_2-1}\left(1-\frac{2}{n}\right)^{(n-2)-(t_1-1)-(t_2-1)}$$

hence $\text{E}(T_1T_2) = [n(n-1)]/(n^2) = 1 - 1/n$, so $\text{cov}(T_1, T_2) = \text{E}(T_1T_2) - \text{E}(T_1)\text{E}(T_2) = -1/n$. Thus,

$$\begin{aligned}\text{var}(Z|\{Y_j = y_j\}) &= \frac{1}{n^2}\left[\sum_{j} y_j^2\left(1-\frac{1}{n}\right) - \sum_{j}\sum_{j',j\neq j'} y_j y_{j'}\frac{1}{n}\right] \\ &= \frac{1}{n^2}\left[\sum_{j} y_j^2 - \frac{1}{n}\left(\sum_{j} y_j\right)^2\right]\end{aligned}$$

$$
\begin{aligned}
&= \frac{(n-1)}{n^2}\left[\frac{1}{(n-1)}\sum_j (y_j - \bar{y})^2\right] \\
&= \left(1 - \frac{1}{n}\right)\cdot s_y^2/n,
\end{aligned}
$$

where $s_y^2$ is the unbiased form of sample variance of the $\{y_j\}$. Unconditionally,

$$
\begin{aligned}
\mathrm{E}[\mathrm{var}(Z)] &= \mathrm{E}_{Y_1,\ldots,Y_n}[\mathrm{E}(\mathrm{var}(Z)|Y_1,\ldots,Y_n)] \\
&= \mathrm{E}_{Y_1,\ldots,Y_n}\left[\left(1 - \frac{1}{n}\right)s_y^2/n\right] \\
&= \left(1 - \frac{1}{n}\right)\cdot\frac{1}{n}\mathrm{E}(s_y^2) \\
&= \left(1 - \frac{1}{n}\right)\cdot \mathrm{var}(Y)/n \\
&= \left(1 - \frac{1}{n}\right)\cdot\frac{1}{n}
\end{aligned}
$$

We are expecting a variance of $1/n$, so there is an exact downward bias in the bootstrap estimate of the variance of the mean of $(1 - 1/n)$. Note that no normality assumptions are needed for this; the result holds for any r.v. $Y$.

### B.2. Multivariate

Let $\mathscr{Y}$ denote a $p$-dimensional multivariate system with $n \times p$ independent random variables in matrix $\boldsymbol{Y} = \{Y_{jk}\}$, with each dimension $k = 1,\ldots,p$ having mean $\mu_k = 0$ and variance $\sigma_k^2 = 1$, and realised sample matrix $\boldsymbol{Y} = \{y_{jk}\}$ for $j = 1,\ldots,n$; $k = 1,\ldots,p$. For each dimension $k = 1,\ldots,p$, the sample mean is $\bar{y}_k = \sum_{j=1}^n y_{jk}/n$ and $s_k^2 = \sum_{j=1}^n (y_{jk} - \bar{y}_k)^2/(n-1)$ is an unbiased estimate of $\sigma_k^2$.

Suppose Euclidean distances are calculated among every pair of sample units based on all $p$ variables, yielding the $n \times n$ matrix $\boldsymbol{D} = \{d_{jj'}\}$. Apply Gower's (1966) transformation to obtain $\boldsymbol{G} = (\boldsymbol{I}_n - (1/n)\boldsymbol{J}_n)\boldsymbol{A}(\boldsymbol{I}_n - (1/n)\boldsymbol{J}_n)$ where $\mathbf{A} = \{(-1/2)d_{jj'}^2\}$. Following McArdle & Anderson (2001), we have

$$
\mathrm{tr}(\boldsymbol{G}) = (n-1)\sum_{k=1}^p s_k^2.
$$

Hence, as $Y_{jk}$ are independent for all $j, k$,

$$
\mathrm{E}\left\{\frac{1}{(n-1)}\mathrm{tr}(\boldsymbol{G})\right\} = \sum_{k=1}^p \sigma_k^2 = p.
$$

Consider a bootstrap sample of multivariate data $\{y_{\beta_j k}\}$, where $\boldsymbol{\beta} = \{\beta_j\}$ for $j = 1,\ldots,n$ is a vector of length $n$ containing a bootstrap sample of the integers from 1 to $n$. Gower's transformation of a Euclidean distance matrix from this bootstrap sample is then given by $\boldsymbol{G}^\beta$ and, from the univariate result above,

$$
\mathrm{E}\left\{\frac{1}{(n-1)}\mathrm{tr}(\boldsymbol{G}^\beta)\right\} = \left(1 - \frac{1}{n}\right)\sum_{k=1}^p \sigma_k^2 = \left(1 - \frac{1}{n}\right)p
$$

We are expecting a variance equal to $p$, so there is an exact downwards bias in the bootstrap estimate of the variance of $(1 - 1/n)$.

## Supporting information

Additional supporting information may be found in the online version of this article at http://wileyonlinelibrary.com/journal/anzs

**Data S1**. Norway macrofauna dataset.
**Tables S1 & S2**. Tables reporting results of simulations.

## References


AHMAD, M.R. (2014). A $U$-statistic approach for a high-dimensional two-sample mean testing problem under non-normality and Behrens–Fisher setting. *Annals of the Institute of Statistical Mathematics* **66**, 33–61.

AHMAD, M.R., VON ROSEN, D. & SINGULL, M. (2012). A note on mean testing for high dimensional multivariate data under non-normality. *Statistica Neerlandica* **67**, 88–99.

AITCHISON, J. & HO, C.H. (1989). The multivariate Poisson-log normal distribution. *Biometrika* **76**, 643–653.

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, T.W. & DARLING, D.A. (1954). A test of goodness-of-fit. *Journal of the American Statistical Association* **49**, 765–769.

ANDERSON, M.J., ELLINGSEN, K.E. & MCARDLE, B.H. (2006). Multivariate dispersion as a measure of beta diversity. *Ecology Letters* **9**, 683–693.

ANDERSON, M.J. & ROBINSON, J. (2001). Permutation tests for linear models. *Australian & New Zealand Journal of Statistics* **43**, 75–88.

ANDERSON, M.J. & TER BRAAK, C.J.F. (2003). Permutation tests for multi-factorial analysis of variance. *Journal of Statistical Computation and Simulation* **73**, 85–113.

ANDERSON, M.J. & WALSH, D.C.I. (2013). PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: what null hypothesis are you testing? *Ecological Monographs* **83**, 557–574.

BARTLETT, M.S. (1939). A note on tests of significance in multivariate analysis. *Proceedings of the Cambridge Philosophical Society* **35**, 180–185.

BEHRENS, W.V. (1929). Ein Beitrag zur Fehlerberechnung bei wenigen Beobachtungen. *Landwirtschaftliches Jaresbch* **68**, 807–837.

BELLONI, A. & DIDIER, G. (2008). On the Behrens–Fisher problem: a globally convergent algorithm and a finite sample study of the Wald, LR and LM tests. *Annals of Statistics* **36**, 2377–2408.

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. (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. *The Annals of Mathematical Statistics* **25**, 290–302.

BROWN, M.B. & FORSYTHE, A.B. (1974). The small sample behaviour of some statistics which test the equality of several means. *Technometrics* **16**, 129–132.

CHRISTENSEN, W.F. & RENCHER, A.C. (1997). A comparison of type I error rates and power levels for seven solutions to the multivariate Behrens–Fisher problem. *Communications in Statistics – Simulation and Computation* **26**, 1251–1273.

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

CLARKE, K.R., SOMERFIELD, P.J. & CHAPMAN, M.G. (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.

CLINCH, J.J. & KESELMAN, H.T. (1982). Parametric alternatives to the analysis of variance. *Journal of Educational Statistics* **7**, 207–214.

COCHRAN, W.G. (1977). *Sampling Techniques*, 3rd edn. New York: John Wiley & Sons.
COOMBS, W.T. & ALGINA, J. (1996). New test statistics for MANOVA/descriptive discriminant analysis. *Educational and Psychological Measurement* **56**, 382–402.
DAVISON, A.C. & HINKLEY, D.V. (1997). *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.
EFRON, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia, PA: Society for Industrial and Applied Mathematics.
EFRON, B. & TIBSHIRANI, R.J. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.
ELLINGSEN, K.E. & GRAY, J.S. (2002). Spatial patterns of benthic diversity: is there a latitudinal gradient along the Norwegian continental shelf? *Journal of Animal Ecology* **71**, 373–389.
FISHER, N.I. & HALL, P. (1990). On bootstrap hypothesis-testing. *Australian Journal of Statistics* **32**, 177–190.
FISHER, R.A. (1935). The fiducial argument in statistical inference. *Annals of Eugenics* **6**, 391–198.
FLETCHER, D., MACKENZIE, D. & VILLOUTA, E. (2005). Modelling skewed data with many zeros: a simple approach combining ordinary and logistic regression. *Environmental and Ecological Statistics* **12**, 45–54.
FRANCIS, R.I.C.C. & MANLY, B.F.J. (2001). Bootstrap calibration to improve the reliability of tests to compare sample means and variances. *Environmetrics* **12**, 713–729.
GAMAGE, J., MATHEW, T. & WEERHANDI, S. (2004). Generalized p-values and generalized confidence regions for the multivariate Behrens–Fisher problem and MANOVA. *Journal of Multivariate Analysis* **8**, 177–189.
GHOSH, M. & KIM, Y.-Y. (2001). The Behrens–Fisher problem revisited: a Bayes-frequentist synthesis. *Canadian Journal of Statistics* **29**, 5–17.
GLASS, G.V., PECKHAM, P.D. & SANDERS, J.R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research* **42**, 237–288.
GOWER, J.C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* **53**, 325–338.
GRAY, J.S., CLARKE, K.R., WARWICK, R.M. & HOBBS, G. (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.
HALL, P. (1992). *The Bootstrap and Edgeworth Expansion*. New York: Springer-Verlag.
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.
JOHNSON, R.A. & WEERHANDI, S. (1988). A Bayesian solution to the multivariate Behrens–Fisher problem. *Journal of the American Statistical Association* **83**, 145–149.
KRISHNAMOORTHY, K., LU, F. & MATHEW, T. (2007). A parametric bootstrap approach for ANOVA with unequal variances: fixed and random models. *Computational Statistics & Data Analysis* **51**, 5731–5742.
KRISHNAMOORTHY, K., & LU, F. (2010). A parametric bootstrap solution to the MANOVA under heteroscedasticity. *Journal of Statistical Computation and Simulation* **80**, 873–887.
MARSAGLIA, G. & MARSAGLIA, J. (2004). Evaluating the Anderson–Darling distribution. *Journal of Statistical Software* **9**(2), 1–5.
MANLY, B.F.J. (1995). Randomization tests to compare means with unequal variation. *Sankhya B* **57**, 200–222.
MANLY, B.F.J. (2006). *Randomization, Bootstrap and Monte Carlo Methods in Biology*, 3rd edn. London: Chapman and Hall.
MANLY, B.F.J. & FRANCIS, R.I.C.C. (1999). Analysis of variance by randomization when variances are unequal. *Australian & New Zealand Journal of Statistics* **41**, 411–429.
MANLY, B.F.J. & FRANCIS, R.I.C.C. (2002). Testing for mean and variance differences from samples from distributions that may be non-normal with unequal variances. *Journal of Statistical Computation and Simulation* **72**, 633–646.
MCARDLE, B.H. & ANDERSON, M.J. (2001). Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology* **82**, 290–297.
MCARDLE, B.H. & ANDERSON, M.J. (2004). Variance heterogeneity, transformations and models of species abundance: a cautionary tale. *Canadian Journal of Fisheries and Aquatic Science* **61**, 1294–1302.
MCARDLE, B.H., GASTON, K.J. & LAWTON, J.H. (1990). Variation in the size of animal populations: patterns, problems and artefacts. *Journal of Animal Ecology* **59**, 439–454.
PILLAI, K.C.S. (1955). Some new test criteria in multivariate analysis of variance. *The Annals of Mathematical Statistics* **26**, 117-121.

R Core Team (2015).*R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

SOMERFIELD, P.J., CLARKE, K.R. & OLSGARD, F. (2002). A comparison of the power of categorical and correlational tests applied to community ecology data from gradient studies. *Journal of Animal Ecology* **71**, 581–593.

TAYLOR, L.R. (1961). Aggregation, variance and the mean. *Nature* **189**, 732–735.

WANG, Y.Y. (1971). Probabilities of the type I errors of the Welch tests for the Behrens–Fisher problem. *Journal of the American Statistical Association* **66**, 605–608.

WEERHANDI, S. (1993). Generalized confidence intervals. *Journal of the American Statistical Association* **88**, 899–905.

WELCH, B.L. (1938). The significance of the difference between two means when population variances are unequal. *Biometrika* **29**, 350–362.