Testing for homogeneity of multivariate dispersions using dissimilarity measures

Irène Gijbels

Department of Mathematics and Leuven Statistics Research Center (LStat),

Katholieke Universiteit Leuven, Celestijnenlaan 200B, B-3001 Leuven (Heverlee), Belgium

email: Irene.Gijbels@wis.kuleuven.be

and

Marek Omelka

Department of Probability and Mathematical Statistics, Faculty of Mathematics and Physics,

Charles University Prague, Sokolovská 83, 186 75 Praha 8, Czech Republic.

email: omelka@karlin.mff.cuni.cz

Summary: Testing homogeneity of dispersions may be of its own scientific interest as well as an important auxiliary step verifying assumptions of a main analysis. The problem is that many biological and ecological data are highly skewed and zero-inflated. Also the number of variables often exceeds the sample size. Thus data analysts often do not rely on parametric assumptions, but use a particular dissimilarity measure to calculate a matrix of pairwise differences. This matrix is then the basis for further statistical inference. Anderson (2006) proposed a distance-based test of homogeneity of multivariate dispersions for a one-way ANOVA design, for which a matrix of pairwise dissimilarities can be taken as an input. The key idea, like in Levene's test, is to replace each observation with its distance to an estimated group centre. In this paper we suggest an alternative approach that is based on the means of within-group distances and does not require group centre calculations to obtain the test statistic. We show that this approach can have theoretical as well as practical advantages. A permutation procedure that gives type I error close to the prescribed value even in small samples is described.

Key words: ANOVA, dissimilarity measure, Levene's test, multivariate analysis, permutation tests, principal coordinates, spatial median, U-statistics.

1. Introduction

When analysing multivariate data it is often of crucial importance to know if groups of observations (e.g. as defined by different treatments) differ in their relative dispersions. This question may be of interest on its own (e.g. to find the treatments having stabilising/destabilising effects) or an auxiliary step when verifying assumptions or interpreting results of a main analysis.

Data coming from many biological applications, in particular ecological community data, are often very skewed, possibly containing many zeroes, which makes the assumption of normality (as well as any other parametric assumptions) very difficult to justify. Also the number of variables is usually not small in comparison to the sample size. That is why an analyst often uses an appropriate dissimilarity (distance) measure to calculate the matrix of pairwise dissimilarities (distances) and the statistical inference is based on that matrix.

In this paper we will consider the simple one-way ANOVA design, where each observation belongs to exactly one treatment. The usual question of scientific interest is to find out which treatments have effects on outcome. To answer this question several tests based on a dissimilarity matrix have been proposed, e.g. Mantel and Valand (1970), Mielke et al. (1981), Smith et al. (1990), Excoffier et al. (1992), Clarke (1993), Pillar and Orlóci (1996), Gower and Krzanowski (1999), Legendre and Anderson (1999) and McArdle and Anderson (2001). Roughly speaking, all the above tests are based on comparing the within-group against between-group dissimilarities and in particular they aim at finding the differences in centres of the multivariate distributions underlying the observed data. A significant result is usually

interpreted as the tendency of the observations belonging to the same treatment to 'cluster together' around different group centres. But this interpretation may not be correct as all the tests are to some extent also omnibus tests and a significant result may be purely the effect of differences in group dispersions.

Anderson (2006) developed a distance-based test for homogeneity of multivariate dispersions which is inspired by the popular Levene's test in univariate ANOVA. The key idea is to replace each observation with its distance to an estimated group centre.

The aim of this paper is to suggest an alternative test, which is directly based on the means of distances within the same group and does not require group centre calculations to obtain the test statistic. This test overcomes several difficulties of the test of Anderson (2006). First, it can be easily and explicitly stated in terms of the distance matrix which feature of the groups is compared. Second, a large sample version of this test can be calculated for a general dissimilarity matrix without the need for calculating the principal coordinate representation. Third, the permutation version of the test does not require re-calculation of the centre of the data with each permutation. Last but not least, our simulation experiences show that the proposed tests control slightly better the type I error for comparisons of small samples (up to 20 observations per group).

Note that in ecology the concept of dispersion can be very useful when analysing species diversity. If one observation represents the composition of species for a given site and observations are coming from several locations, then testing the homogeneity of dispersions of the groups is useful when testing for differences between locations in terms of species composition, because differences between locations (beta diversity) depend on dispersions within locations (alpha diversity). Observations can also be compared between years instead of locations. On the other hand, if an observation stands for characteristics of an individual (animals, plants, ...) and the observations are coming from different species (or even well

defined groups within a given species), then the presented testing problem can be used in the analysis of genetic or phenotypic differences between species' populations.

With no loss of generality, we talk only about dissimilarity matrices in this paper, with the understanding that similarity matrices can be handled in a similar fashion, after applying a suitable transformation to dissimilarities.

The paper is organized as follows. Section 2 introduces the tests. Permutation procedures to estimate p-values of the tests are discussed in Section 3. In Section 4 the suggested tests are compared with the tests of Anderson (2006) in a simulation study. Section 5 summarizes and discusses the results. Detailed simulation results are provided in a Web Appendix.

2. Description of the test

In a simple one-way ANOVA design each observation is associated with exactly one of the K treatments and the sample sizes of the corresponding groups are $n_1, \ldots, n_K$ with the total sample size $n = n_1 + \ldots + n_K$. Let $\mathbf{Y}_i^{(k)} = (Y_{i1}^{(k)}, \ldots, Y_{ip}^{(k)})^\mathsf{T}$ be the vector of length p corresponding to the i-th observation in the k-th group for each p variables and suppose that all the observation vectors are independent and the p-length observation vectors in the same group follow the same multivariate distribution.

To describe the tests we use two of the data sets discussed in Anderson (2006). We refer to that paper and the references therein for a more detailed description of the data.

2.1 *Using the Euclidean distance*

The Bumpus' sparrow data set consists of five morphological characteristics of sparrows measured in Rhode Island after a severe storm. The sparrows are divided into two groups, those that survived the storm and those that died. The data are a subset of the original data recorded by Bumpus (1899) and can be found in Manly (2005).

As the general theory of stabilizing selection (Campbell et al., 2008) suggests that the char-

acteristics of non-survivors should be more dispersed than the characteristics of survivors, one of the questions of interest is to test for a difference in the multivariate dispersions of these two groups. Thus, we require a test for homogeneity of multivariate dispersions.

The first option is the traditional likelihood ratio test described e.g. in Rencher (1998), pp. 138–140. This test assumes multivariate normality of the observations and tests the null hypothesis

$$H_0 : \mathbf{\Sigma}_1 = \ldots = \mathbf{\Sigma}_K, \tag{1}$$

where $\mathbf{\Sigma}_k$ is the variance-covariance matrix of random variables underlying the observations in the k-th group (note that $K = 2$ in our example). As the resulting test is rather sensitive to the assumption of multivariate normality, more robust procedures have been developed, see e.g. Tiku and Balakrishnan (1985) and O'Brien (1992). In the latter paper it was also suggested that, instead of trying to test the very specific hypothesis (1), it is often reasonable to concentrate on the overall level of dispersion. This basically means that one constructs a simple measure that aims at summarizing the dispersion of each group and then tests for the equality of these measures of dispersion among groups.

The test suggested in Anderson (2006) follows the idea of concentrating on the overall level of dispersion. Although the test can be based on any dissimilarity measure, it is instructive to illustrate with the Euclidean distance d_E. The core idea is that if group A is more dispersed than another group B, then the distances of the observation to the centre in group A tend to be larger than those in group B. To be more precise, let $\hat{\mathbf{t}}_k$ stand for an estimated centre of the k-th group and define

$$\mathbf{X} = \left(X_1^{(1)}, \ldots, X_{n_1}^{(1)}, X_1^{(2)}, \ldots, X_{n_2}^{(2)}, \ldots, X_1^{(K)}, \ldots, X_{n_K}^{(K)}\right)^{\mathsf{T}}, \text{ where } X_i^{(k)} = d_E(\mathbf{Y}_i^{(k)}, \hat{\mathbf{t}}_k).$$

Now, the vector $\mathbf{X}$ is treated as K independent random samples and the traditional ANOVA F-statistic is used to compare the means across K groups.

A p-value for this F-statistic is obtained either by using an F-distribution or by using the

following permutation procedure: (i) Permute the 'residuals' $\mathbf{r}_i^{(k)} = \mathbf{Y}_i^{(k)} - \hat{\mathbf{t}}_k$; (ii) Calculate the new group centres $\hat{\mathbf{t}}_k^*$ based on the permuted residuals; (iii) Take the distances of the permuted data from the new centres and recalculate the F-statistic.

Anderson (2006) suggested that either centroids (component-wise means) or spatial medians can be taken as the group centres ($\hat{\mathbf{t}}_k$'s). As there are many definitions of a multivariate median in the literature, by a spatial median of data-points in this paper we understand a point (say $\hat{\mathbf{t}}$) that minimizes the sum of the Euclidean distances of data-points from $\hat{\mathbf{t}}$ (Haldane, 1948).

The test F_{And} is very appealing as it seems to be a very natural multivariate analogue to Levene's test (see Van Valen (1978) and Manly (2005) for similar suggestions of a multivariate Levene's test). The test is also intuitive and simple to understand.

Despite these nice properties of the test F_{And}, one may feel uncomfortable that the test statistic depends on the choice of the centre of the groups. This could be of particular concern if a centroid and a spatial median are not close to each other which is often the case if the data are asymmetric or contain outliers. A straightforward alternative to calculating distances from group centres is to consider inter-point distances within groups. If for example the group of non-survivors is more dispersed than the group of survivors, then one would expect that the inter-point distances within the group of non-survivors are, on average, larger than those within the group of survivors.

Let us make these ideas more precise. Let $d_{ij}^{(k)} = d(\mathbf{Y}_i^{(k)}, \mathbf{Y}_j^{(k)})$ where d stands for the used dissimilarity measure (the Euclidean distance in our example) be the dissimilarity between observations i and j. Means of the within-group distances can be calculated as

$$\bar{d}_k = \frac{1}{\binom{n_k}{2}} \sum_{i=1}^{n_k-1} \sum_{j=i+1}^{n_k} d_{ij}^{(k)}, \qquad k = 1, \ldots, K. \tag{2}$$

Note that if the dispersions in the groups are the same, the quantities $\bar{d}_1, \ldots, \bar{d}_K$ are expected to be close to each other. Thus, as in a standard ANOVA, we want to test the equality of these

K means. The only thing we have to be careful about is that each of the $\bar{d}_k$ is not a mean of independent random variables, but rather is a U-statistic of degree two (see Chapter 5 of Serfling, 1980). The asymptotic variance of $\bar{d}_k$ is usually estimated with the jackknife estimator (formula (9) of Callaert and Veraverbeke, 1981)

$$\hat{\sigma}_k^2 = \frac{S_k^2}{n_k}, \qquad \text{where} \quad S_k^2 = \frac{4\,(n_k-1)}{(n_k-2)^2} \sum_{i=1}^{n_k} (\hat{D}_i^{(k)} - \bar{d}_k)^2, \quad k = 1, \ldots, K, \tag{3}$$

where the average distance from observation i to every other observation within its group is

$$\hat{D}_i^{(k)} = \frac{1}{n_k - 1} \sum_{j=1, j\neq i}^{n_k} d_{ij}^{(k)}, \quad i = 1, \ldots, n_k,\ k = 1, \ldots, K. \tag{4}$$

Finally, the test statistic is given by

$$F_{\bar{d}} = \frac{\sum_{k=1}^{K} n_k\,(\bar{d}_k - \bar{d})^2}{(K-1)\,\hat{\sigma}^2}, \quad \text{where} \quad \bar{d} = \frac{1}{K}\sum_{k=1}^{K} n_k\,\bar{d}_k,\ \hat{\sigma}^2 = \frac{\sum_{k=1}^{K}(n_k-1)\,S_k^2}{n-K}. \tag{5}$$

The null hypothesis is rejected when $F_{\bar{d}}$ exceeds the $(1-\alpha)$-quantile of an F-distribution with $K-1$ and $n-K$ degrees of freedom. The asymptotic validity of this test follows from independence and the asymptotic normality of each of the quantities $\bar{d}_1, \ldots, \bar{d}_k$. In Section 3 we describe a permutation procedure that improves the small sample properties of this test.

Remark 1: As pointed out by one of the referees, the test statistic $F_{\bar{d}}$ can be easily calculated by the standard ANOVA F-test applied to jackknifed 'pseudo-values' (Callaert and Veraverbeke, 1981). In our situation the jackknifed 'pseudo-values' are given by

$$P_i^{(k)} = \frac{2}{n_k - 2} \sum_{j=1, j\neq i}^{n_k} d_{ij}^{(k)} - \frac{n_k}{n_k - 2}\,\bar{d}_k, \quad i = 1, \ldots, n_k,\ k = 1, \ldots, K.$$

The fact that one does not need to specify the group centres has some methodological advantages. The main advantage is that one can easily specify the null hypothesis in terms of the pairwise distances. Let $\mathcal{L}_k$ stand for the distribution of the within sample distances in the k-th group. Although in the sequel we will be interested in the following null hypothesis

$$H_0 : \mathcal{L}_1 = \mathcal{L}_2 = \ldots = \mathcal{L}_K, \tag{6}$$

different null hypotheses may be of interest. For instance a researcher may be interested

in a very broad null hypothesis stating that only the mean values of the distributions $\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_K$ coincide. The corresponding test could be constructed by a modification of an ANOVA test for unequal variances (see e.g Volaufová, 2009, and references therein).

REMARK 2: Note that the test statistic $F_{\bar{d}}$ defined in (5) is not the only way to test the null hypothesis (6). When using $F_{\bar{d}}$ one hopes that a difference in distributions $\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_K$ can be discovered as a difference in their mean values. But generally speaking one can take any K-sample test and modify it for the within sample distances.

REMARK 3: As in fact we test for equality of dispersions, it seems reasonable to expect that the variances of the distributions $\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_K$ increase with the means. Thus one can use a logarithmic transformation of $\bar{d}_k$ to stabilize the variances. By the delta method (see e.g. Chapter 3.1 of Serfling, 1980) it follows that the resulting test statistic $F_{\bar{d}}^{\log}$ is given by (5) with $\bar{d}_k$ replaced with $\log(\bar{d}_k)$ and S_k^2 with $S_k^2/(\bar{d}_k)^2$.

REMARK 4: The test procedure can be visualized with the help of $\hat{D}_i^{(k)}$ defined in (4). This quantity gives an average distance from the i-th observation to the other data points in the k-th group. A small/large value of $\hat{D}_i^{(k)}$ means that the i-th observation is close/far from the 'centre' of the k-group. Further note that the mean within-group distance $\bar{d}_k$ defined in (2) can also be calculated as $\bar{d}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \hat{D}_i^{(k)}$. Figure 1 plots $\hat{D}_i^{(k)}$ together with their means $\bar{d}_k$ and an approximate 95%-confidence interval for the mean μ_k of the distributions $\mathcal{L}_k$. The confidence interval is based on the asymptotic normality of $\bar{d}_k$ and is given by $(\bar{d}_k - u_{0.975}\,\hat{\sigma}_k,\ \bar{d}_k + u_{0.975}\,\hat{\sigma}_k)$, where u_p is the p-quantile of the standard normal distribution and $\hat{\sigma}_k$ is the estimate of the standard derivation of $\bar{d}_k$ given in (3). Note that non-overlapping confidence intervals in Figure 1 would already indicate that the null hypothesis (6) does not hold. As for our data there is a degree of overlap in the intervals, a formal test is necessary to decide about the null hypothesis.

[Table 1 about here.]

[Figure 1 about here.]

Table 1 gives the values of the test statistics and p-values for the sparrow data set (after standardization for each variable). $F_{\bar{d}}$ stands for the test (5) and $F_{\bar{d}}^{\log}$ for the 'log-transformed' test described in Remark 3. The test statistics of Anderson (2006) are denoted by F_{And}^{c} when centred by a centroid or F_{And}^{m} when centred by a spatial median. Further, 'p-value (as.)' stands for the p-value given by a standard F-distribution and 'p-value (perm.)' for the p-value given by the permutation test with 99 999 permutations. The permutation procedure used to get the p-values $F_{\bar{d}}(p)$ and $F_{\bar{d}}^{\log}(p)$ is described in Section 3.

Note that the p-values of all the tests are 'borderline', i.e. close to 0.05 indicating a possible difference between the two groups of sparrows in the dispersions of these measured morphological characteristics.

2.2 *Using a general dissimilarity measure*

The Tikus Island coral data set consists of the percentage cover of each 75 coral species along each of 10 replicate transects in six different years from 1981 to 1988. Differences among the coral assemblages in different years are expected as an El Niño event occurred in 1982-1983. The data are given as a data set called `tikus` in the R-package `mvabund` (Wang et al., 2012). As these data include many zeroes and are highly skewed, the approach that uses the Euclidean distance does not seem to be appropriate here. Following the analysis of Warwick et al. (1990) the matrix of pair-wise Bray-Curtis dissimilarities (given by (8)) calculated from square-root transformed data was used as a starting point of the analysis.

To be more precise, let $\mathbf{Y}_i^{(k)}$ stand for already square-root transformed observations. Denote the joint sample as

$$(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)^{\mathsf{T}} = \left(\mathbf{Y}_1^{(1)}, \ldots, \mathbf{Y}_{n_1}^{(1)}, \mathbf{Y}_1^{(2)}, \ldots, \mathbf{Y}_{n_2}^{(2)}, \ldots, \mathbf{Y}_1^{(K)}, \ldots, \mathbf{Y}_{n_K}^{(K)}\right)^{\mathsf{T}}. \tag{7}$$

The dissimilarity matrix $\mathbb{D}$ is a matrix of pair-wise distances with the elements $d_{ij} = d(\mathbf{Z}_i, \mathbf{Z}_j)$, where d now stands for the Bray-Curtis dissimilarity given by

$$d_{BC}(\mathbf{Z}_i, \mathbf{Z}_j) = \frac{\sum_{q=1}^{p} |Z_{iq} - Z_{jq}|}{\sum_{q=1}^{p} |Z_{iq} + Z_{jq}|}. \tag{8}$$

For the situation when the analyst starts with a matrix $\mathbb{D}$ of pairwise-dissimilarities Anderson (2006) suggests the following way of calculating the test F_{And}:

Step 1. Using principal coordinate analysis (PCoA, see e.g. Legendre and Legendre, 1998, pp. 424–438) find a representation (say $\mathbf{U}$) of the dissimilarity matrix $\mathbb{D}$. The observation $\mathbf{Z}_i$ is now represented by the vector $\mathbf{u}_i = (\mathbf{u}_i^+, \mathbf{u}_i^-)^\mathsf{T}$ (the i-th row of the matrix $\mathbf{U}$), where $\mathbf{u}_i^+$ ($\mathbf{u}_i^-$) stands for the coordinates of the vectors that corresponds to real (imaginary) axes of the representation $\mathbf{U}$.

Step 2. If $\mathbf{U}$ contains only real axes, the analysis of Section 2.1 can be directly used with the original data being replaced with the representation $\mathbf{U}$. If there are also some imaginary axes, then define $X_i^{(k)} = \sqrt{d_E^2(\mathbf{u}_i^+, \mathbf{t}_k^+) - d_E^2(\mathbf{u}_i^-, \mathbf{t}_k^-)}$, where $\hat{\mathbf{t}}_k = (\hat{\mathbf{t}}_k^+, \hat{\mathbf{t}}_k^-)$ stands for the corresponding group centre (either a centroid or a spatial median) with $\hat{\mathbf{t}}_k^+$ ($\hat{\mathbf{t}}_k^-$) being the coordinates corresponding to real (imaginary) axes.

Step 3. Analogously as in Section 2.1 an F-statistic is computed and its significance is assessed either with the help of an F-distribution or via a permutation method where $\mathbf{r}_i = \mathbf{u}_i - \hat{\mathbf{t}}_k$ are permuted.

Although the procedure seems to give reasonable results in practice, it is not without difficulties. Although one intuitively feels that the distances from the group centres in the PCoA representation should reflect the within-group variability, it is not at all straightforward to write down rigorously what feature of the original data is tested and how to formulate the null hypothesis. This is particularly true when the representation $\mathbf{U}$ also includes imaginary axes.

The test statistics based directly on pairwise distances, suggested in this paper, enable

an analyst to specify the null hypothesis by reference to the inter-point distances directly, and decide which feature of within-group distances to compare. For instance when using the statistic $F_{\bar{d}}$ defined in (5) the analyst concentrates on comparing the means of within-sample dissimilarities.

Figure 2 presents the visualization of the test procedure and the p-values for the coral data set are given in Table 1. To estimate the p-values of the permutation tests, 99 999 random permutations were used. The results suggest that there is a statistically significant difference in the means of group dissimilarities. Already the visual inspection of Figure 2 reveals that multivariate dispersion, as measured by the Bray-Curtis dissimilarity, is significantly higher in 1983, which corresponds to the year of the El Niño event. Analogously as in Anderson (2006) one can now proceed and try to find which pairs of years are significantly different. The findings (not presented here) when using either $F_{\bar{d}}$ or $F_{\bar{d}}^{\log}$ are similar to the findings based on F_{And}.

[Figure 2 about here.]

2.3 *Interpreting results*

It is worth mentioning that, as in O'Brien (1992) and Anderson (2006), the test statistics proposed in this paper concentrate on the overall level of dispersions as measured by the mean within-group dissimilarities. Thus, it might happen that the null hypothesis (6) is retained even if (1) is not true. Thus, for example, differences in rotations of the groups cannot be detected by the suggested tests when a rotation invariant dissimilarity measure is used. But this can be viewed also as a desirable feature of the test when analysing beta diversity (see Section 5).

Generally speaking, when using distance-based tests one should be careful when interpreting the results. In our case, rejecting the null hypothesis (6) says that the distributions $\mathcal{L}_1, \ldots, \mathcal{L}_K$ differ in their means. This can be safely interpreted as the difference in multi-

variate dispersions if a location invariant dissimilarity measure is used (e.g. Euclidean or Manhattan) and the underlying distribution of the data can be considered to belong to a multivariate location-scale family. This seems to be reasonable for Bumpus' sparrow data set, but not for the coral data set. When one is dealing with data sets that represent a percentage coverage (such as the coral data set) or with abundance data, the interpretation of any distance-based test is difficult because of the intrinsic mean-variance relationships inherent in the distribution of counts or species abundances (Warton et al., 2012).

Further, one should be aware that the results of tests are usually strongly influenced by the choice of a dissimilarity measure and by transformation/normalization of the data. This can be viewed also as an advantage though. As argued and illustrated in Anderson et al. (2006), by using different dissimilarity measures together with different transformations of the data an analyst can explore various aspects of data. Rejecting the null hypothesis (6) however always indicates that there is a difference among groups. Differences in dispersion may or may not be detected by the traditional ANOSIM-type procedures (see e.g. Clarke, 1993; Anderson, 2001) as these techniques target, in particular, location differences.

3. Resampling procedures

So far, the new proposed tests have been described where inferences rely only on asymptotic distributions of the test statistics. It is well known that finite-sample properties of these tests can often be improved through resampling procedures. Several resampling algorithms can be proven to be asymptotically valid for our problem. In this section we describe a modified permutation procedure that works very well in all situations we have encountered so far and that we recommend for general use.

3.1 *The standard permutation procedure*

The standard permutation approach consists of permuting the original observations $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$. Note that this is equivalent to permuting both the rows and the columns of the matrix $\mathbb{D}$. But this method gives an exact permutation test only if the distribution of the original observations is the same in all the groups. But as the null hypothesis (6) allows (among others) for different locations of the groups, this method cannot be generally recommended. Our experience is that this standard permutation approach does not hold the type I error if the between-group dissimilarities are much bigger than the within group distances. That is why we recommend the 'centred' permutation procedure described below.

3.2 *The 'centred' permutation procedure*

Improvement of the properties of the standard resampling procedure can be obtained by reducing the distances among groups. To achieve this we make use of the principal coordinate representation (PCoA) of the distance matrix $\mathbb{D}$. In order to prevent the imaginary axes of the representation that raises difficulties in interpretation, we use the *Correction method 2* of Legendre and Legendre (1998) (pp. 434–435). This methods adds the smallest positive constant c to all non-diagonal elements of the matrix $\mathbb{D}$, such that the new matrix ($\mathbb{D}^c$) has PCoA representation with only real axes. As proved by Cailliez (1983) the constant c can be found as the largest positive eigenvalue of the matrix

$$\begin{bmatrix} \mathbf{0} & 2\,\mathbf{\Delta}_1 \\ -\mathbf{I} & -4\,\mathbf{\Delta}_2 \end{bmatrix}$$

where $\mathbf{0}$ is a $n \times n$ null matrix and $\mathbf{I}$ is a $n \times n$ identity matrix. Let matrix $\mathbf{A}$ of elements $\{a_{ij}\}$ be defined element-wise as $a_{ij} = -0.5\, d_{ij}^2$. Matrix $\mathbf{\Delta}_1$ with elements $\{\delta_{ij}\}$ is defined as: $\delta_{ij} = a_{ij} - \bar{a}_{i\cdot} - \bar{a}_{\cdot j} + \bar{a}_{\cdot\cdot}$, where $\bar{a}_{i\cdot}$, $\bar{a}_{\cdot j}$ and $\bar{a}_{\cdot\cdot}$ are the means for row i, column j and the overall mean, respectively, from matrix $\mathbf{A}$. Matrix $\mathbf{\Delta}_2$ is defined in precisely the same way, but where matrix $\mathbf{A}$ contains, instead, the elements $a_{ij} = -0.5\, d_{ij}$.

The PCoA applied on the matrix $\mathbb{D}^c$ gives a matrix $\mathbf{U}$ such that for each $i, j \in \{1, \ldots, n\}$

$$d_{ij} + c = d_E(\mathbf{u}_i, \mathbf{u}_j), \quad \text{for} \quad i \neq j,$$

where d_E stands for the Euclidean distance and $\mathbf{u}_k$ is the k-th row of the matrix $\mathbf{U}$.

Put $\mathbf{U}_1$ for the first n_1 rows of the matrix $\mathbf{U}$, $\mathbf{U}_2$ for the next n_2 rows of the matrix $\mathbf{U}$ and so on. Thus, one can write

$$\mathbf{U} = [\mathbf{U}_1^\mathsf{T}, \mathbf{U}_2^\mathsf{T}, \ldots, \mathbf{U}_K^\mathsf{T}]^\mathsf{T}, \tag{9}$$

with $\mathbf{U}_k$ being the representation for the k-th sample. Now, put $\widetilde{\mathbf{U}}_k$ for $\mathbf{U}_k$ centred with either a centroid or a group spatial median defined as

$$\tilde{\mathbf{t}}^{(k)} = \arg\min_{\mathbf{t}} \sum_{i=1}^{n_k} d_E(\mathbf{u}_i^{(k)}, \mathbf{t}), \tag{10}$$

where $\mathbf{u}_i^{(k)}$ is the i-th row of the matrix $\mathbf{U}_k$. Replacing $\mathbf{U}_k$ with $\widetilde{\mathbf{U}}_k$ in (9) one gets the 'centred' representation $\widetilde{\mathbf{U}}$, that is used to construct the 'centred' distance matrix $\widetilde{\mathbb{D}}^c$ with the elements

$$\widetilde{d}_{ij}^c = d_E(\widetilde{\mathbf{u}}_i, \widetilde{\mathbf{u}}_j).$$

Finally, remove the constant c from all the non-diagonal elements of the matrix $\widetilde{\mathbb{D}}^c$ and use the resulting matrix ($\widetilde{\mathbb{D}}$) for the standard permutation approach (instead of the matrix of the original distances $\mathbb{D}$) of Section 3.1.

Remark 5: Note that

$$\widetilde{d}_{ij} = d_{ij} \qquad \text{if } \mathbf{Z}_i \text{ and } \mathbf{Z}_j \text{ belong to the same group.}$$

Thus switching from $\mathbb{D}$ to $\widetilde{\mathbb{D}}$ modifies only the dissimilarities between observations from different groups.

Further note for the test statistic $F_{\bar{d}}$ defined in (5) one can already resample the matrix $\widetilde{\mathbb{D}}^c$ as adding a constant to all distances does not affect the statistic $F_{\bar{d}}$.

Remark 6: Note that PCoA is used only as a vehicle to estimate a sampling distribution

of the test statistic $F_{\bar{d}}$ ($F_{\bar{d}}^{\log}$) under the null hypothesis. In contrast with the test F_{And}, PCoA is not necessary to calculate the values of the test statistics. The user thus has the possibility to use the method of adding a constant to the original dissimilarity matrix to have a PCoA with only real axes. But for F_{And} the dissimilarity matrix $\mathbb{D}^c$ gives generally a different value of the test statistic than when using the original distance matrix $\mathbb{D}$.

3.3 *Centring the original data*

The construction of the matrix $\widetilde{\mathbb{D}}$ described above may be summarized as

$$\text{DATA} \to \mathbb{D} \to \mathbb{D}^c \to \mathbf{U} \to \widetilde{\mathbf{U}} \to \widetilde{\mathbb{D}}^c \to \widetilde{\mathbb{D}}.$$

Sometimes it seems even more natural to calculate the matrix $\widetilde{\mathbb{D}}$ from the original observations that are appropriately centred rather than with the help of PCoA.

This approach has basically two requirements. First, the centring of the observations has to be a reasonable operation. Note that this is true for the sparrow data set of Section 2.1, but not for the abundance-type of data of Section 2.2. Second, the chosen dissimilarity measure must be location-invariant, so that centring of the observations does not affect the within sample distances. In this situation one can centre data either with a centroid or with a spatial median, where the latter is recommended if outliers are present in the data.

A general class of dissimilarity measures that are location invariant is generated by the *Minkowski distance* defined as

$$d_r(\mathbf{Z}_i, \mathbf{Z}_j) = \left(\sum_{q=1}^{p} |Z_{iq} - Z_{jq}|^r \right)^{1/r}, \qquad \text{where } r > 0.$$

Note that for $r = 2$ the Minkowski distance d_2 reduces to the *Euclidean distance* d_E, and for $r = 1$ the Minkowski distance results in the *Manhattan distance*.

Further, a number of dissimilarity measures are 'Euclidean-transformable', that is they coincide with the Euclidean distance after appropriate transformations of the data. Examples of such dissimilarity measures are the *Chord dissimilarity measure*, the *Chi-square distance*,

the *Distance between species profiles* and the *Hellinger distance* (Legendre and Gallagher, 2001).

On the other hand, the distances that are usually used for abundance-type of data are not location invariant. Among others let us mention the *Canberra metric*, the *Coefficient of divergence*, the *Bray-Curtis dissimilarity* (Legendre and Legendre, 1998) or the *scale-invariant binomial deviance* (Anderson and Millar, 2004)

$$d_{bin}(\mathbf{Z}_i, \mathbf{Z}_j) = \sum_{q=1}^{p} \frac{1}{S_q} \left[Z_{iq} \log\left(\frac{Z_{iq}}{S_q}\right) + Z_{jq} \log\left(\frac{Z_{jq}}{S_q}\right) - S_q \log\left(\frac{1}{2}\right) \right], \tag{11}$$

where $S_q = Z_{iq} + Z_{jq}$. For these dissimilarity measures, centring must be done using the PCoA representation, as described in Section 3.2.

4. Simulation study – findings

In the simulation study we investigated the type I error and power properties of both asymptotic as well as permutation versions of the tests suggested in Section 2.1 and compared them with the performances of the F_{And}-tests.

The following test procedures were considered:

1. $F_{\bar{d}}(as)$ – $F_{\bar{d}}$ given by (5) + F-distribution;
2. $F_{\bar{d}}(p_{med})$ – $F_{\bar{d}}$ + the 'centred' (by median) permutation procedures of Section 3.2 or Section 3.3 (if the Euclidean distance is employed);
3. $F_{\bar{d}}(p_{centr})$ – $F_{\bar{d}}$ + the 'centred' permutation procedure as for $F_{\bar{d}}(p_{centr})$ but with centring by a centroid instead of a spatial median
4. $F_{\bar{d}}^{\log}(as)$ – $F_{\bar{d}}^{\log}$ of Remark 3 + F-distribution;
5. $F_{\bar{d}}^{\log}(p_{med})$ – $F_{\bar{d}}^{\log}$ + the 'centred' permutation procedure as for $F_{\bar{d}}(p_{med})$
6. $F_{\bar{d}}^{\log}(p_{centr})$ – $F_{\bar{d}}^{\log}$ + the 'centred' permutation procedure as for $F_{\bar{d}}(p_{centr})$
7. $F_{And}(as_{centr})$ – test of Anderson (2006) + centroid + F-distribution;
8. $F_{And}(p_{centr})$ – test of Anderson (2006) + centroid + permutation;
9. $F_{And}(as_{med})$ – test of Anderson (2006) + spatial median + F-distribution;
10. $F_{And}(p_{med})$ – test of Anderson (2006) + spatial median + permutation.

4.1 *The models used to generate data*

We considered four types of data generation processes. Detailed descriptions of these data generations are given in the Web Appendix.

1. *Sparrows type data* – This data generation process is inspired by the data set of Section 2.1. The simulated data came from a five-dimensional normal distribution and the Euclidean distance d_E was used.
2. *Fish type data* – This data generation example is inspired by the data coming from a study on spatial variation in temperate reef fish assemblages (Anderson and Millar, 2004). In this data set each observation records abundance of 57 fish species. In this model the samples were simulated from a multivariate Poisson-lognormal distribution (Aitchison and Ho, 1989). As a dissimilarity measure we used the scale-invariant binomial deviance d_{bin} (11).
3. *Corals type data* – This data generation process is inspired by the Tikus Islands coral

data discussed in Section 2.2. In this data set each observation records the percentage cover of 75 coral species. In our simulation model these covers were generated as a mixture of independent normal distributions and zeroes. Here the Bray-Curtis dissimilarity measure d_{BC} (8) was used.

4. *Gaussian data with outliers* – Here the data were generated from a bivariate normal distribution with 10 % of outliers. The Euclidean distance d_E was used.

The detailed results of the simulation study are given in the Web Appendix. Here we limit ourselves to describing and discussing simulation findings.

4.2 *An F-distribution or a permutation test?*

Generally we recommend to use permutation tests whenever it is computationally feasible. The tests $F_{\bar{d}}(as)$ and $F_{\bar{d}}^{\log}(as)$ tend to be conservative in particular for *fish data* and *coral data*. (Web Tables 2 and 6). Note that $F_{\bar{d}}^{\log}(as)$ is rather unreliable in terms of type I error when outliers were present and the Euclidean distance was used (Web Table 8). $F_{And}(as_{centr})$ tends to exceed the level slightly in balanced samples and considerably in unbalanced samples in all simulation settings. Finally, the test $F_{And}(as_{med})$ is usually conservative in balanced samples, but sometimes exceeds the level in unbalanced samples (Web Table 2).

The F-distribution approximation of critical values seems to work reasonably well when the group sample sizes exceed 50, but one should be aware that even for very large sample sizes $F_{And}(as_{centr})$ can exceed the level by about one percent, and $F_{And}(as_{med})$ by about a half percent (Web Table 4). The tests $F_{\bar{d}}(as)$ and $F_{\bar{d}}^{\log}(as)$ seem to work better in this aspect.

4.3 *Centring by a centroid or a spatial median?*

Not surprisingly, centring by a spatial median is preferred when there are outlying observations. Particularly in small samples even the permutation test $F_{And}(p_{centr})$ is not reliable

in terms of type I error. This can be seen from Web Table 8 as well as from Web Table 2; indeed the *fish type* data generation also produced values that are outlying.

When there is no evidence for outlying values, the differences in methods are subtle. $F_{And}(p_{centr})$ gives usually slightly higher power results than $F_{And}(p_{med})$, but in some models the former test exceeds the level for small and unbalanced samples (Web Tables 2, 3, 6 and 7).

The choice of centring seems to be of lesser importance for the permutation versions of the tests $F_{\bar{d}}$ and $F_{\bar{d}}^{\log}$ than for the F_{And}-tests. This can probably be explained by the fact that the type of centring does not affect the test statistics and it is used only as an adjustment for the data before resampling. The test $F_{\bar{d}}(p_{centr})$ ($F_{\bar{d}}^{\log}(p_{centr})$) usually achieves slightly higher power than its closest competitor $F_{\bar{d}}(p_{med})$ ($F_{\bar{d}}^{\log}(p_{med})$). The type I error properties of the tests $F_{\bar{d}}(p_{centr})$ and $F_{\bar{d}}^{\log}(p_{centr})$ are also very satisfactory and both the tests hold the level very closely for sample sizes bigger than 10. For very small sample sizes, the tests can exceed the level slightly (but usually not more than by a half percent). On the other hand $F_{\bar{d}}(p_{med})$ and $F_{\bar{d}}^{\log}(p_{med})$ are sometimes unnecessarily conservative for small samples.

4.4 *Distance to centres or within group distances?*

In most of the situations all the permutation tests hold the level quite satisfactorily. Only in small or unbalanced samples the F_{And}-tests slightly (usually with about a half or one percent) exceed the level (Web Tables 2 and 6). On the other hand all the permutation tests suggested in this paper hold the level very satisfactorily in all situations we encountered so far.

Regarding power, all the permutation tests give similar powers in balanced samples. This is not so surprising as the tests concentrate on very similar features of the data. For unbalanced samples, the suggested tests are more powerful when bigger groups tend to have smaller dispersions. If it is the other way around, then the tests of Anderson (2006) achieve higher

power. This seems to be a small sample feature of the tests that is diminishing with increasing sample size (Web Table 5).

4.5 *$F_{\bar{d}}$-test or $F_{\bar{d}}^{\log}$-test?*

In terms of type I error these two tests are comparable. Our experience is that also in terms of power performances these tests are very close when a scale-invariant dissimilarity measure is used, that is, a dissimilarity measure satisfying $d(c\,\mathbf{Z}_i, c\,\mathbf{Z}_j) = d(\mathbf{Z}_i, \mathbf{Z}_j)$ for each $c > 0$. Note that d_{BC} defined in (8) as well as d_{bin} defined in (11) are scale-invariant. On the other hand a noticeable difference in the power performances of the tests can be observed for dissimilarity measures that are scale-equivariant, that is $d(c\,\mathbf{Z}_i, c\,\mathbf{Z}_j) = c\,d(\mathbf{Z}_i, \mathbf{Z}_j)$. This is the case for the Euclidean measure used for the sparrow-type data (Web Table 1). As the test statistic $F_{\bar{d}}^{\log}$ transforms a scale effect into an additive effect and the F-statistic used in ANOVA is constructed to detect additive differences in mean values, we recommend to use $F_{\bar{d}}^{\log}$ for scale-equivariant dissimilarity measures. However, one should be aware that $F_{\bar{d}}^{\log}(as)$ can break down when outliers are present (Web Table 8).

4.6 *Computational aspects.*

Sometimes it might be advantageous that the test statistics $F_{\bar{d}}$ and $F_{\bar{d}}^{\log}$ can be computed without the necessity of calculating a PCoA representation. The permutation versions of the suggested tests also require a PCoA representation, thus from this point of view the suggested procedures are comparable with the tests of Anderson (2006). The amount of calculations needed to re-calculate the test statistic is for the suggested test procedure comparable with that for $F_{And}(p_{centr})$. The test $F_{And}(p_{med})$ is more computationally expensive when the total sample size n exceeds one hundred and not the Euclidean distance is used as a dissimilarity measure. The reason is that in this situation the PCoA representation has typically $(n-1)$ dimensions. Re-calculating a spatial median in such a high dimension, with each permutation and in each group, can take a substantial amount of computing time.

5. Conclusions

In this work we propose a testing approach that aims at detecting differences in the overall level of dispersion among independent groups of observations. The approach is based on pairwise distances of observations. We compare the suggested tests with the tests introduced in Anderson (2006) from the computational aspect as well as from aspects of performances in power and type I error.

While the tests of Anderson (2006) summarize the overall measure of dispersion by means of the distances from the group centres, the tests proposed in this paper used the mean within group distances to summarize the dispersion. While these concepts are close together when the Euclidean distance is used, the difference becomes more important for other dissimilarity measures. The reason is that in this case tests of Anderson (2006) are based on the distances from the group centres in the PCoA representation and it is less obvious what feature of the original distance matrix is tested. This is in particular true when also imaginary axes are present in the PCoA representation. On the other hand, the approach based on distances from the centroids can be more useful when one also wants to identify outlying observations visually.

As suggested in Anderson et al. (2006) for the mean distance-to centre, the mean within-group dissimilarity $\bar{d}_k$ can also be used as a direct measure of beta diversity (e.g., Whittaker, 1960, 1972; Vellend et al., 2007). For example, if one analyses species data like the Tikus Island coral data set, then $\bar{d}_k$ can be interpreted as a measure of beta diversity for a given year (e.g., Anderson et al., 2011).

The proposed new approach to test for homogeneity of overall dispersions among groups behaves just as well as the tests proposed by Anderson (2006) with respect to type I error and power, has reasonably good asymptotic properties, and is generally easier to compute and to interpret than the methods proposed by Anderson (2006).

6. Supplementary Materials

The data sets analysed in Section 2, an R-code implementing the considered tests and the Web Appendix referenced in Section 4 are available with this paper at the Biometrics website on Wiley Online Library.

Acknowledgement

The authors are grateful to the Editor, the Associate Editor and two referees for their very valuable comments which led to a considerable improvement of the manuscript. The authors are also very grateful to Prof. Marti J. Anderson for providing them with the details about the simulation study presented in Anderson (2006).

This research was supported by the Fellowship F+/11/028 from the Research fund KU Leuven. The first author gratefully acknowledges support from the projects GOA/07/04 and GOA/12/014 of the Research Fund KU Leuven and from the IAP research networks P6/03 and P7/13 of the Federal Science Policy, Belgium. The second author gratefully acknowledges support from the grant GACR P201/11/P290.

References

Aitchison, J. and 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, M. J., Crist, T. O., Chase, J. M., Vellend, M., Inouye, B. D., Freestone, A. L., Sanders, N. J., Cornell, H. V., Comita, L. S., Davies, K. F., Harrison, S. P., Kraft, N.

J. B., Stegen, J. C., and Swenson, N. G. (2011). Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist. *Ecology Letters* **14,** 19–28.

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

Anderson, M. J. and Millar, R. B. (2004). Spatial variation and effects of habitat on temperate reef fish assemblages in northeastern New Zealand. *Journal of Experimental Marine Biology and Ecology* **305,** 191–221.

Bumpus, H. C. (1899). The elimination of the unfit as illustrated by the introduced sparrow, passer domesticus. Marine Biology Laboratory, Woods Hole, Massachusetts, 11th Lecture, pp. 209-226.

Cailliez, F. (1983). The analytical solution of the additive constant problem. *Psychometrika* **48,** 305–308.

Callaert, H. and Veraverbeke, N. (1981). The order of the normal approximation for a studentized U-statistics. *Annals of Statistics* **9,** 194–200.

Campbell, N. A., Reece, J. B., Urry, L. A., Cain, M. L., Wasserman, S. A., Minorsky, P. V., and Jackson, R. B. (2008). *Biology.* London: Persons International. 8th Edition.

Clarke, K. R. (1993). Non-parametric multivariate analysis of changes in community structure. *Australian Journal of Ecology* **18,** 117–143.

Excoffier, L., Smouse, P. E., and Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. *Genetics* **131,** 479–491.

Gower, J. C. and Krzanowski, W. J. (1999). Analysis of distance for structured multivariate data and extensions to multivariate analysis of variance. *Journal of the Royal Statistical Society. Series C.* **48,** 505–519.

Haldane, J. B. S. (1948). Note on the median of a multivariate distribution. *Biometrika* **35,**

414–415.

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

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

Legendre, P. and Legendre, L. (1998). *Numerical Ecology.* Elsevier Science, Amsterdam.

Manly, B. F. J. (2005). *Multivariate statistical methods. A Primer. 3rd ed.* Chapman & Hall/CRC, London.

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

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

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

O'Brien, P. C. (1992). Robust procedures for testing equality of covariance matrices. *Biometrics* **48,** 819–827.

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

Rencher, A. C. (1998). *Multivariate statistical inference and applications.* Wiley Series in Probability and Statistics. New York, NY: Wiley.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* Wiley, New York.

Smith, E. P., Pontasch., K. W., and Cairns Jr., J. (1990). Community similarity and

the analysis of multispecies environmental data: A unified statistical approach. *Water Research.* **24,** 507–514.

Tiku, M. L. and Balakrishnan, N. (1985). Testing the equality of variance-covariance matrices the robust way. *Communication in Statistics - Theory and Methods* **14,** 3033–3051.

Van Valen, L. (1978). The statistics of variation. *Evolutionary Theory* **4,** 33–43. (Erratum Evolutionary Theory 4, 202).

Vellend, M., Verheyen, K., Flinn, K., Jacquemyn, H., Kolb, A., Van Calster, H., Peterken, G., Graae, B. J., Bellemare, J., Honnay, O., Brunet, J., Wulf, M., Gerhardt, F., and Hermy, M. (2007). Homogenization of forest plant communities and weakening of species–environment relationships via agricultural land use. *Journal of Ecology* **95,** 565–573.

Volaufová, J. (2009). Heteroscedastic ANOVA: old p values, new views. *Statistical Papers* **50,** 943–962.

Wang, Y., Naumann, U., Wright, S. T., and Warton, D. I. (2012). mvabund– an R package for model-based analysis of multivariate abundance data. *Methods in Ecology and Evolution* **3,** 471–474.

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

Warwick, R. M., Clarke, K. R., and Suharsono (1990). A statistical analysis of coral community responses to the 1982-83 El Niño in the Thousand Islands, Indonesia. *Coral Reefs* **8,** 171–179.

Whittaker, R. H. (1960). Vegetation of the Siskiyou Mountains, Oregon and California. *Ecological Monographs* **30,** 279–338.

Whittaker, R. H. (1972). Evolution and measurement of species diversity. *Taxon* **21,** 213–251.

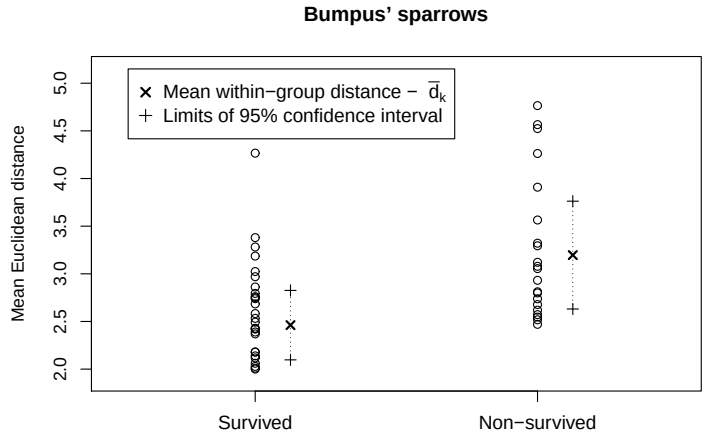

Figure 1. Plot of the pseudo-observations $\hat{D}_i^{(k)}$ for Bumpus' sparrow data set along with mean within-group distance and 95%-confidence interval.

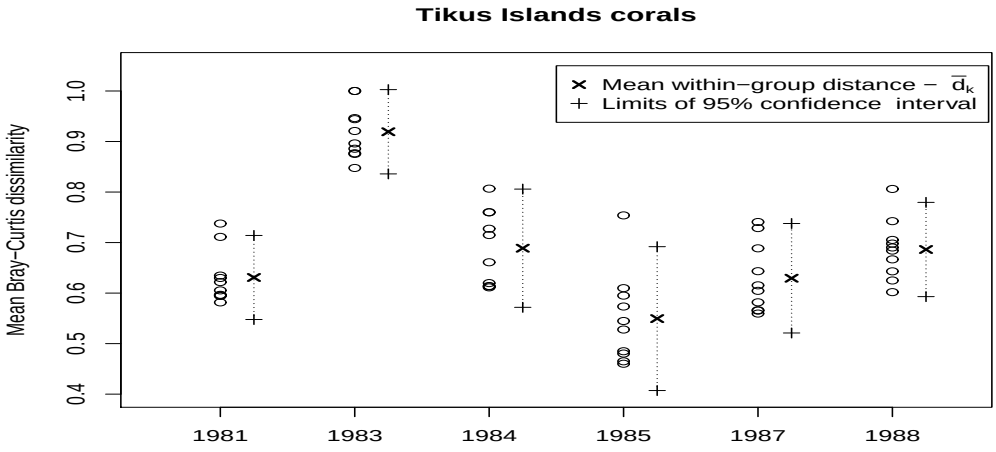

Figure 2. Plot of the pseudo-observations $\hat{D}_i^{(k)}$ for Tikus Island coral data set along with mean within-group distance and 95%-confidence interval.

Table 1
Test statistics and p-values for the Bumpus' sparrow and Tikus Islands coral data set.

		$F_{\bar{d}}$	$F_{\bar{d}}^{\log}$	F_{And}^{c}	F_{And}^{m}
Sparrows	test statistic	4.319	4.956	3.869	3.818
	p-value (as.)	0.043	0.031	0.055	0.057
	p-value (perm.)	0.041	0.034	0.055	0.048
Corals	test statistic	6.920	5.193	9.097	6.292
	p-value (as.)	< 0.0001	0.0006	< 0.0001	0.0001
	p-value (perm.)	< 0.0001	0.0004	< 0.0001	0.0001

Web based supplementary material for **"Testing for homogeneity of multivariate dispersions using dissimilarity measures"**

by Irène Gijbels and Marek Omelka

Web Appendix A: Details of the simulation study

This section provides detailed results of the simulations study. The considered tests have been described in Section 4 of the main manuscript.

The type I error is prescribed to be 0.05. A total of 10 000 samples (or 50 000 samples when at least one of the sample sizes was less or equal to ten) were generated to estimate the type I error of the tests and 5 000 samples were generated for assessing the power of the tests. A total of 999 random permutations were used to estimate p-values of permutation tests. We used the R-computing environment, version 2.10.1 (see R Development Core Team (2009)) to perform the simulations.

A1. **Sparrows.** This simulation study is inspired by the data set already introduced in Section 2.1 of the main manuscript. The simulated data came from a five-dimensional normal distribution with the parameters estimated from the Bumpus' sparrow data set. The means and the variances of the components were taken to be

$$\boldsymbol{\mu}_1 = (157.4, 241.0, 31.4, 18.5, 20.8)^\mathsf{T}, \qquad \boldsymbol{\sigma}_1^2 = (11.048, 17.500, 0.531, 0.176, 0.575)^\mathsf{T},$$

$$\boldsymbol{\mu}_2 = (158.4, 241.6, 31.5, 18.4, 20.8)^\mathsf{T}, \qquad \boldsymbol{\sigma}_2^2 = (15.069, 32.550, 0.728, 0.434, 1.321)^\mathsf{T}.$$

Note that the dispersion is 'larger' for the second group.

The correlation matrix was taken the same for both groups and estimated from the pooled sample as

$$\mathbf{Corr} = \begin{pmatrix} 1.00 & 0.73 & 0.66 & 0.65 & 0.61 \\ 0.73 & 1.00 & 0.67 & 0.77 & 0.53 \\ 0.66 & 0.67 & 1.00 & 0.76 & 0.53 \\ 0.65 & 0.77 & 0.76 & 1.00 & 0.61 \\ 0.61 & 0.53 & 0.53 & 0.61 & 1.00 \end{pmatrix}.$$

For estimating the empirical type I error the variance of the first group served as that for both groups.

The generated data were first standardised (to have zero mean and unit variance) for each variable, and then the distance matrix based on the Euclidean distance measure was computed. Samples of various sizes were generated to cover the situation of balanced ($(n_1, n_2) = (10, 10)$, $(20, 20)$, $(40, 40)$) as well as unbalanced samples ($(n_1, n_2) = (10, 20)$, $(20, 10)$).

From Web Table 1 one can see that for this setting all the tests hold the type I error very satisfactory. Such type of data sets seem to be 'well-behaved' and the asymptotic and permutation tests give similar results. Note however that for the sample sizes $(10, 10)$, the asymptotic as well as the permutation procedures that rely on centring by centroids, slightly exceed the prescribed level 0.05.

When the sample sizes are equal, the powers of all tests are similar. When the larger sample size goes along with the less dispersed sample, both permutation and asymptotic versions of the test $F_{\bar{d}}$ achieve higher power than the F_{And}-tests. If the more dispersed

WEB TABLE 1. Sparrows type data – empirical type I error and power for the tests.

	Type I error				Power		
	(10,10)	(20,20)	(40,40)	(10,20)	(20,20)	(10,20)	(20,10)
$F_{\bar{d}}(as)$	0.051	0.044	0.048	0.046	0.423	0.237	0.319
$F_{\bar{d}}(p_{med})$	0.049	0.047	0.050	0.049	0.434	0.249	0.332
$F_{\bar{d}}(p_{centr})$	0.054	0.049	0.050	0.051	0.433	0.251	0.335
$F_{\bar{d}}^{\log}(as)$	0.051	0.051	0.051	0.054	0.440	0.303	0.303
$F_{\bar{d}}^{\log}(p_{med})$	0.049	0.048	0.050	0.049	0.421	0.281	0.281
$F_{\bar{d}}^{\log}(p_{centr})$	0.053	0.049	0.048	0.051	0.422	0.284	0.285
$F_{And}(as_{centr})$	0.060	0.053	0.051	0.057	0.443	0.307	0.308
$F_{And}(p_{centr})$	0.056	0.050	0.052	0.052	0.411	0.294	0.295
$F_{And}(as_{med})$	0.034	0.041	0.046	0.041	0.397	0.270	0.246
$F_{And}(p_{med})$	0.050	0.049	0.052	0.050	0.408	0.299	0.277

sample is larger in size, then it is the other way around. The powers of the $F_{\bar{d}}^{\log}$-tests are close to these of the F_{And}-tests.

A2. **Fish.** Here, the simulated data were inspired by the data coming from a study on spatial variation in temperate reef fish assemblages along the north-eastern coast of New Zealand. The observations were coming from four sites (Berghan Point, Home Point, Leigh and Hahei) and each observation recorded abundance of 57 fish species. One of the

interests of the original study was in spatial variation among the sites. More details about the original study can be found in Anderson and Millar (2004) and the references therein.

These data are highly skewed containing many zeros. Analogously as in Anderson (2006) the simulated samples were generated from a multivariate Poisson-lognormal distribution (Aitchison and Ho, 1989) by using the following three-steps process. First, we generated multivariate normal vectors with the parameters (means, variances for each of the group and a single pooled correlation matrix) that we were kindly provided by Prof. Marti J. Anderson. For generating the null hypothesis, only the parameters corresponding to these for Berghan Point were used for all four groups. For an alternative hypothesis, a mixture of the multivariate normal distributions was considered. With probability $\alpha = 0.75$ a vector was generated based on the parameters for Berghan Point and with probability $\alpha = 0.25$ a vector was generated using parameters of the actual group (e.g. Home Point when generating observations for this group). In a second step, the exponential function was applied to the generated vectors. The results from this step were then used as parameters of component-wise independent Poisson distributions to generate $(\mathbf{X}_1^{(1)}, \ldots, \mathbf{X}_{n_1}^{(1)}, \ldots, \mathbf{X}_1^{(4)}, \ldots, \mathbf{X}_{n_4}^{(4)})$.

Finally, for each of the four groups ($k = 1, \ldots, 4$), a random permutation $\boldsymbol{\pi}_k = (\pi_1^k, \ldots, \pi_{57}^k)$ of the numbers $\{1, \ldots, 57\}$ was generated and the coordinates of each of the observations were permuted with a permutation corresponding to its group, that is the observations were given by

$$\mathbf{Y}_i^{(k)} = (X_{i\pi_1^k}^{(k)}, \ldots, X_{i\pi_{57}^k}^{(k)}), \qquad i = 1, \ldots, 57, \quad k = 1, \ldots, 4,$$

where $X_{ij}^{(k)}$ is the j-th component of vector $\mathbf{X}_i^{(k)}$. Note that as a permutation of the coordinates does not affect the null hypothesis (6), the null hypothesis (6) continues to hold provided the vectors $\mathbf{X}_i^{(k)}$s were generated from the same distribution.

The reason for introducing the permutations of the coordinates was to have a situation for which that the null hypothesis (6) holds, but the distributions in different groups are not the same so that the observations are not identically distributed.

The scale-invariant binomial deviance dissimilarity was used to produce the distance matrix. In this setting there are four groups corresponding to different locations. The results for the various scenarios are to be found in Web Tables 2 and 3. The statement `4x10` means that all sample sizes are equal to 10. Similarly for `4x20`. Further, `2x10-2x20` means that the sample sizes of the first and second group are 10 and of the third and the fourth group are 20. Similarly for `2x20-2x40`.

Note that the asymptotic tests $F_{\bar{d}}(as)$, $F_{\bar{d}}^{\log}(as)$ and $F_{And}(as_{med})$ are conservative for balanced small sample sizes. For unbalanced sample sizes this still holds true for the suggested tests $F_{\bar{d}}(as)$ and $F_{\bar{d}}^{\log}(as)$, but $F_{And}(as_{med})$ exceeds the prescribed level considerably. The problem with holding the level is even bigger for $F_{And}(as_{centr})$.

The difficulties of the asymptotic tests are to a large extent overcome by the permutation versions of the tests. Note however that while the permutation versions of the suggested tests hold the level very closely, the $F_{And}(p_{centr})$ and $F_{And}(p_{med})$ tests slightly exceed the prescribed level.

WEB TABLE 2. Fish type data – empirical type I error

	4x10	4x20	4x40	2x7-2x15	2x10-2x20	2x20-2x40
$F_{\bar{d}}(as)$	0.020	0.031	0.037	0.039	0.030	0.042
$F_{\bar{d}}(p_{med})$	0.046	0.048	0.047	0.045	0.046	0.047
$F_{\bar{d}}(p_{centr})$	0.050	0.050	0.049	0.049	0.048	0.048
$F_{\bar{d}}^{\log}(as)$	0.019	0.031	0.036	0.026	0.030	0.037
$F_{\bar{d}}^{\log}(p_{med})$	0.045	0.048	0.048	0.046	0.045	0.047
$F_{\bar{d}}^{\log}(p_{centr})$	0.051	0.049	0.047	0.048	0.048	0.048
$F_{And}(as_{centr})$	0.049	0.053	0.055	0.138	0.106	0.084
$F_{And}(p_{centr})$	0.063	0.061	0.056	0.070	0.062	0.058
$F_{And}(as_{med})$	0.030	0.041	0.045	0.098	0.082	0.071
$F_{And}(p_{med})$	0.056	0.059	0.056	0.065	0.060	0.058

Power results are quite analogous to the results for the 'sparrows-type' data. The differences in power for balanced data correspond well with the way how the tests hold the level. For unbalanced samples the F_{And}-tests are more powerful than the tests based on $F_{\bar{d}}$ or $F_{\bar{d}}^{\log}$, when the larger sample size goes along with the more dispersed sample. If it is the other way around then the tests based on $F_{\bar{d}}$ or $F_{\bar{d}}^{\log}$ are more powerful. Note that for this type of data, the tests based on $F_{\bar{d}}$ or $F_{\bar{d}}^{\log}$ give very similar powers also in unbalanced samples.

WEB TABLE 3. Fish type data – empirical power

	4x10	4x20	2x10-2x20	2x20-2x10	2x20-2x40	2x40-2x20
$F_{\bar{d}}(as)$	0.203	0.621	0.338	0.467	0.744	0.892
$F_{\bar{d}}(p_{med})$	0.270	0.622	0.369	0.508	0.756	0.896
$F_{\bar{d}}(p_{centr})$	0.288	0.628	0.375	0.518	0.759	0.900
$F_{\bar{d}}^{\log}(as)$	0.216	0.611	0.357	0.480	0.770	0.898
$F_{\bar{d}}^{\log}(p_{med})$	0.271	0.634	0.387	0.515	0.776	0.901
$F_{\bar{d}}^{\log}(p_{centr})$	0.293	0.638	0.386	0.522	0.774	0.905
$F_{And}(as_{centr})$	0.374	0.702	0.598	0.559	0.866	0.908
$F_{And}(p_{centr})$	0.340	0.672	0.480	0.461	0.819	0.876
$F_{And}(as_{med})$	0.249	0.614	0.480	0.445	0.812	0.870
$F_{And}(p_{med})$	0.283	0.621	0.427	0.415	0.781	0.850

A2.1. *Fish – asymptotic study.* To explore the large sample properties of the asymptotic tests, we concentrated on a comparison of two groups (Berghan Point and Leigh). Similarly as above, when generating under the null hypothesis only the parameters corresponding to Berghan Point were used.

The type I errors are to be found in Web Table 4. Note that while $F_{\bar{d}}(as)$ and $F_{\bar{d}}^{\log}(as)$ hold the level very closely, $F_{And}(as_{centr})$ has difficulties not to exceed the level even for large samples, and this is particularly the case for unbalanced samples. $F_{And}(as_{med})$ holds the level well for balanced samples, but slightly exceeds the level for unbalanced samples.

WEB TABLE 4. Fish type data – empirical type I error for large sample sizes

	(50,50)	(100,100)	(200,200)	(50,100)	(100,200)	(200,400)
$F_{\bar{d}}(as)$	0.047	0.047	0.048	0.049	0.050	0.050
$F_{\bar{d}}^{\log}(as)$	0.045	0.046	0.047	0.046	0.049	0.050
$F_{And}(as_{centr})$	0.055	0.056	0.057	0.064	0.062	0.060
$F_{And}(as_{med})$	0.049	0.051	0.051	0.058	0.056	0.054

When investigating the power of the tests, the probability α of generating from the alternative distribution was always chosen such that the power is around 0.5. The results can be found in Web Table 5. Note that with increasing sample size, it becomes less important if the larger sample size goes along with either the less or more dispersed sample. At the same time it becomes more important how well the test statistics are suited for detecting a particular type of deviation from the null hypothesis. This has been confirmed also for the other types of data generations for which, for brevity, the results are not included.

A3. **Corals.** This mechanism of data generation is inspired by the Tikus Islands coral data set discussed in Section 2.2. The data were generated in a two-steps process. First, independent normal random variables with the means and variances estimated from the original data were generated and truncated to the nearest integers. If there were any negative values, they were set to zero. Second, with probability equal to the proportion of non-zeroes values in the original data, the values generated in the first step were accepted

WEB TABLE 5. Fish type data – empirical power for large samples

	(100,100)	(200,200)	(50,100)	(100,50)	(200,400)	(400,200)
$F_{\bar{d}}(as)$	0.585	0.523	0.387	0.454	0.491	0.561
$F_{\bar{d}}^{\log}(as)$	0.585	0.523	0.390	0.447	0.493	0.560
$F_{And}(as_{centr})$	0.611	0.530	0.519	0.391	0.553	0.514
$F_{And}(as_{med})$	0.576	0.501	0.475	0.363	0.516	0.491

or otherwise set to zero. For the null hypothesis, only the parameters from the first group were used. For an alternative hypothesis, an average of the corresponding parameters of the first group and the actual group were used. Analogously as in the original analysis, the Bray-Curtis dissimilarity measure was used on the square root-transformed data. The results are given in Web Tables 6 and 7.

Note that the pattern of the results is similar to the pattern of results in the previous section. The asymptotic tests $F_{\bar{d}}(as)$, $F_{\bar{d}}^{\log}(as)$ and $F_{And}(as_{perm})$ are conservative for balanced samples, but for unbalanced samples $F_{And}(as_{perm})$ exceeds the given level. The test $F_{And}(as_{centr})$ slightly exceeds the level for balanced samples, but heavily for unbalanced samples. All the considered permutation tests do a good job in holding the type I error, but in unbalanced samples $F_{And}(p_{centr})$ slightly exceed the level.

A4. **Gaussian data with outliers.** In this simulation setup we considered two groups. As we were interested only in the type I error all observations were drawn from the bivariate

WEB TABLE 6. Corals type data – empirical type I error for large samples

	6x10	6x20	3x5-3x10	3x7-3x15	3x10-3x20
$F_{\bar{d}}(as)$	0.018	0.030	0.019	0.025	0.029
$F_{\bar{d}}(p_{med})$	0.042	0.047	0.042	0.043	0.047
$F_{\bar{d}}(p_{centr})$	0.046	0.048	0.048	0.046	0.049
$F_{\bar{d}}^{\log}(as)$	0.019	0.030	0.022	0.026	0.030
$F_{\bar{d}}^{\log}(p_{med})$	0.042	0.047	0.042	0.043	0.046
$F_{\bar{d}}^{\log}(p_{centr})$	0.047	0.048	0.048	0.046	0.049
$F_{And}(as_{centr})$	0.056	0.055	0.125	0.124	0.104
$F_{And}(p_{centr})$	0.049	0.049	0.057	0.055	0.056
$F_{And}(as_{med})$	0.026	0.036	0.054	0.070	0.069
$F_{And}(p_{med})$	0.041	0.047	0.050	0.052	0.053

distribution function

$$F(x_1, x_2) = 0.9\,\Phi_1(x_1, x_2) + 0.1\,\Phi_2(x_1, x_2), \qquad (x_1, x_2) \in \mathbb{R}^2,$$

where Φ_1 and Φ_2 are the distribution functions of centred bivariate Gaussian distributions with variance matrices $\mathbf{V_1}$ and $\mathbf{V_2}$ given by

$$\mathbf{V_1} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, \qquad \mathbf{V_2} = \begin{bmatrix} 100 & 1 \\ 1 & 100 \end{bmatrix}.$$

WEB TABLE 7. Corals type data – empirical power

	6x10	6x20	3x10-3x20	3x20-3x10	3x20-3x40	3x40-3x20
$F_{\bar{d}}(as)$	0.292	0.842	0.412	0.786	0.882	0.998
$F_{\bar{d}}(p_{med})$	0.414	0.885	0.479	0.830	0.901	0.998
$F_{\bar{d}}(p_{centr})$	0.433	0.888	0.485	0.838	0.900	0.999
$F_{\bar{d}}^{\log}(as)$	0.320	0.859	0.438	0.800	0.894	0.999
$F_{\bar{d}}^{\log}(p_{med})$	0.436	0.896	0.498	0.844	0.909	0.999
$F_{\bar{d}}^{\log}(p_{centr})$	0.454	0.900	0.507	0.847	0.909	0.999
$F_{And}(as_{centr})$	0.489	0.909	0.625	0.884	0.936	0.999
$F_{And}(p_{centr})$	0.439	0.892	0.488	0.799	0.896	0.998
$F_{And}(as_{med})$	0.354	0.862	0.533	0.828	0.910	0.998
$F_{And}(p_{med})$	0.413	0.884	0.477	0.789	0.895	0.998

The type I error results are to be found in Web Table 8. This table illustrates that in case of outliers it is better to centre with a spatial mean rather than with a centroid. This is particularly true for very small samples.

REFERENCES

Aitchison, J. and Ho, C. H. (1989). The multivariate Poisson-log normal distribution. *Biometrika*, 76(4):643–653.

WEB TABLE 8. Bivariate normal data with outliers – empirical type I error

	(10,10)	(20,20)	(40,40)	(10,20)	(20,40)
$F_{\bar{d}}(as)$	0.024	0.034	0.046	0.040	0.042
$F_{\bar{d}}(p_{med})$	0.046	0.049	0.049	0.053	0.050
$F_{\bar{d}}(p_{centr})$	0.050	0.050	0.050	0.053	0.051
$F_{\bar{d}}^{\log}(as)$	0.133	0.122	0.093	0.121	0.100
$F_{\bar{d}}^{\log}(p_{med})$	0.047	0.049	0.049	0.051	0.049
$F_{\bar{d}}^{\log}(p_{centr})$	0.067	0.055	0.050	0.059	0.051
$F_{And}(as_{centr})$	0.097	0.084	0.071	0.101	0.076
$F_{And}(p_{centr})$	0.114	0.068	0.054	0.078	0.057
$F_{And}(as_{med})$	0.017	0.030	0.043	0.031	0.037
$F_{And}(p_{med})$	0.047	0.049	0.049	0.051	0.050

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

Anderson, M. J. and Millar, R. B. (2004). Spatial variation and effects of habitat on temperate reef fish assemblages in northeastern New Zealand. *Journal of Experimental Marine Biology and Ecology*, 305:191–221.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.