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.