

RESEARCH ARTICLE

Use of multivariate distance measures for high-dimensional data in tests for difference, superiority, equivalence and non-inferiority

Siegfried Kropf¹ | Kai Antweiler² | Ekkehard Glimm^{1,3} 

¹ Institute of Biometry and Medical Informatics, Otto-von-Guericke University Magdeburg, Magdeburg, Germany

² Department of Medical Statistics, Universitätsmedizin Göttingen, Germany

³ Novartis Pharma AG, Basel, Switzerland

Correspondence

Ekkehard Glimm, Novartis Pharma AG, Basel, Switzerland.

Email: ekkehard.glimm@novartis.com

Funding information

German Federal Ministry of Education and Research, Grant/Award Number: 05M10NMA



This article has earned an open data badge “**Reproducible Research**” for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially due to their computational complexity.

Abstract

Tests based on pairwise distance measures for multivariate sample vectors are common in ecological studies but are usually restricted to two-sided tests for differences. In this paper, we investigate extensions to tests for superiority, equivalence and non-inferiority.

KEYWORDS

equivalence test, multivariate pairwise distance measure, non-inferiority test, superiority test

1 | INTRODUCTION

Multivariate tests with high-dimensional data are a challenge, particularly as sample sizes in many applications are small. One attempt to overcome this difficulty is the use of one- or low-dimensional scores derived from the high-dimensional data. Starting from asymptotic test versions such as the proposals of O’Brien (1984) or Tang et al. (1993), Läuter (1996), and Läuter et al. (1996, 1998) showed that exact parametric tests are possible with linear scores or score vectors with weights derived from the original data. The most common of these tests is the principal component test. A modification for very large number of variables is given by Ding et al. (2012). Other proposals for asymptotic test versions avoiding the inversion of the covariance matrix of the sample vectors had previously been given by Box (1954) and Dempster (1958, 1960). More recent papers are from Srivastava and von Rosen (2004), Srivastava and Fujikoshi (2006), and Bathke et al. (2009). Most of these multivariate tests are two-sided, but directed tests are available, too, such as the ALR test of Tang, Geller and Pocock (1993) or one-sided tests with one score, where special problems with monotonicity may occur as reported by Glimm and Läuter (2010).

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2021 The Authors. *Biometrical Journal* published by Wiley-VCH GmbH.

In this paper, we consider a class of tests based on pairwise multivariate distance measures between the sample vectors. The basic idea behind these tests is that distances between sample vectors from different groups must be stochastically larger than those between sample vectors from the same group if the groups differ in their distribution. Such tests are rather common in the field of ecology (Anderson, 2001). In medical biometry, they are getting more relevant with the increasing interest in the microbiome of patients but also with high-resolution imaging techniques. As an example, Marozzi (2015) used interpoint distance-based tests for the analysis of cardiovascular magnetic resonance imaging data.

Usually, these tests are applied to detect (undirected) differences between populations. As a starting point for further discussion, we review this situation. Then we present proposals for extensions that can be used to investigate directed differences, equivalence, or noninferiority. In contrast to multiple test procedures, no statements about single variables under familywise error control are intended here but rather a characterization of the variables in their entirety. This can be of special interest in the case of a large number of variables with similar importance. For example, consider the composition of microbial communities in biological material. The microbiome consists of a large number of different species of bacteria or fungi, many of which may not even be concretely characterized genetically. In such a situation, it is more meaningful to compare the total microbial composition rather than to try to identify and quantify the abundance of single species. As another example, in clinical trials, usually a large number of adverse event types is monitored. In this situation, multiple test procedures with strong control of the familywise error rate might have only low power—in contrast to a global comparison of the whole set of adverse event types.

Following a short specification of the model and of the used distance measures in Section 2, we will consider the different testing problems. Section 3 reviews tests for undirected difference. Section 4 focuses on one-sided tests for superiority. Tests for equivalence and noninferiority are proposed in Sections 5 and 6, respectively. Section 7 considers the robustness of the procedures under various conditions. Section 8 presents two applications. A short discussion concludes the paper.

2 | POPULATION MODEL AND DISTANCE MEASURES

Consider the simple situation of two independent samples of p -dimensional random vectors of sizes n_1 and n_2 , respectively. Let $y_i = (y_{i1}, \dots, y_{ip})'$ ($i = 1, \dots, n, n = n_1 + n_2$) be the n sample units ordered by sample

$$y_1, \dots, y_{n_1} \sim F_1(y), \quad y_{n_1+1}, \dots, y_{n_1+n_2} \sim F_2(y) \quad (2.1)$$

with multivariate distribution functions F_1 and F_2 in the two populations.

Next, we define a distance measure between two sample vectors

$$\delta_{jk} = \delta(y_j, y_k), \quad j, k = 1, \dots, n. \quad (2.2)$$

The measure shall fulfil the conditions

$$\delta_{ij} \geq 0, \quad (2.3)$$

$$\delta_{ii} = 0. \quad (2.4)$$

For now, we will also assume that

$$\delta_{ij} = \delta_{ji}, \quad (2.5)$$

($i, j = 1, \dots, n$). Later on, in tests for directed alternatives, the symmetry condition (2.5) will be dropped.

Common examples are

- the squared Euclidean distance

$$\delta_{ij} = (y_j - y_i)'(y_j - y_i) = \|y_j - y_i\|^2, \quad (2.6)$$

- the maximum absolute distance

$$\delta_{ij} = \max_{k=1, \dots, p} |y_{jk} - y_{ik}|, \quad (2.7)$$

- the city block distance

$$\delta_{ij} = \sum_{k=1}^p |y_{jk} - y_{ik}|, \quad (2.8)$$

and

the minimum absolute distance

$$\delta_{ij} = \min_{k=1, \dots, p} |y_{jk} - y_{ik}|. \quad (2.9)$$

The Mahalanobis distance $(y_j - y_i)' \Sigma^{-1} (y_j - y_i)$ is another popular multivariate distance measure. However, it requires the inversion of a covariance matrix Σ . This can be a problem in a high-dimensional setup because of difficulties with estimating Σ from the data and because the inversion can be computationally costly.

Prominent applications of distance-based tests are comparisons of microbial populations, where the p variables indicate the relative frequency of different microbes (so-called operational taxonomic units, OTUs). In this case, the city block distance corresponds to two times the Bray–Curtis distance as a very common distance measure in ecologic studies. The relative frequencies are sometimes log or probit transformed (with an additive correction term to avoid zeros) in order to enhance the weight of rare OTUs in the analyses. Also in this context, the Pearson correlation between sample vectors is often used as a similarity measure instead of a distance measure. We will revisit this topic in the discussion.

The pairwise distances are averaged over those pairs of sample vectors that consist of sample vectors from different groups,

$$\delta_{\text{between}} = \frac{1}{n_1 n_2} \sum_{i \leq n_1, j > n_1} \delta_{ij}, \quad (2.10)$$

and over those pairs of vectors from the same group,

$$\delta_{\text{within}} = \frac{1}{n_1(n_1 - 1) + n_2(n_2 - 1)} \left(\sum_{i, j \leq n_1, i \neq j} \delta_{ij} + \sum_{i, j > n_1, i \neq j} \delta_{ij} \right). \quad (2.11)$$

Based on these two averages, we will consider the following three versions of a test statistic:

$$d = \delta_{\text{between}} - \delta_{\text{within}}, \quad (2.12)$$

$$d^* = \delta_{\text{between}}, \quad (2.13)$$

$$d^{**} = \frac{\delta_{\text{between}} - \delta_{\text{within}}}{\delta_{\text{within}}}. \quad (2.14)$$

In the definition of δ_{within} , δ_{ii} are excluded. Including them would amount to using $\frac{1}{n_1^2 + n_2^2}$ as the multiplicative factor in δ_{within} . If additionally $n_1 = n_2$, then d from (2.12) with the squared Euclidean distance (2.6) would be identical to $\|\bar{y}_{(1)} - \bar{y}_{(2)}\|_2$, where $\bar{y}_{(1)}$ and $\bar{y}_{(2)}$ are the sample mean vectors in the two groups. For the other distances, such intuitive reformulations of the difference are not available. In this paper, we will use the definition (2.11).

A more recent development are interpoint distance-based tests (Jurečková & Kalina, 2012; Marozzi et al., 2020). These start with the pairwise distance measures as well, but combine the within-group and between-group distances in a more sophisticated way, often using nonparametric elements. The resulting tests have a large power and are robust for a wide class of distributions. As our main focus here are the extensions to one-sided tests and equivalence/noninferiority tests, we restrict to the above statistics that can also be interpreted as effect measures for which tolerance limits can be given.

3 | TESTS FOR UNDIRECTED DIFFERENCE BETWEEN THE TWO POPULATIONS

In the simplest case, we want to test the null hypothesis of agreement of both distributions,

$$H_0 : F_1(y) = F_2(y) \quad \forall y \quad (3.1)$$

against differences in any direction

$$H_1 : F_1(y) \neq F_2(y) \text{ for some } y. \quad (3.2)$$

This case is easy to treat, because the null hypothesis ensures the basis for usual permutation tests for the case of two independent groups. Under the null hypothesis, all permutations of the n sample elements have the same distribution. Thus, the test statistic d according to (2.12), (2.13) or (2.14) is calculated once for the original sample and then repeatedly for the permuted samples. If d_0 denotes the original value and d_1, \dots, d_N the values obtained from permutations (all remaining permutations or a random selection of size N of all permutations), then the P -value of the permutation test can be calculated as the proportion of permutation test statistics that are larger than or equal to the original value d_0

$$P = \frac{\#\{i \in \{1, \dots, N\} : d_i \geq d_0\} + 1}{N + 1}. \quad (3.3)$$

Formula (3.3) yields a conservative estimate of the proportion of rejections under the null in the true underlying population from which the sample was obtained, since all permutations that yield d_0 are regarded as still supporting the null hypothesis. However, since in many fields of applications (e.g., clinical trials), the use of methods for “breaking the ties” are unpopular and the conservativeness of the approach is seen as a partly desirable feature of the method, we prefer not to address this topic here. Pesarin and Salmaso (2010) cover this issue in depth. The permutation test is already treated by Mantel (1967), and also described in Anderson (2001) or Kropf et al. (2004) among others.

In the special case of multivariate normal data with equal covariance matrix Σ in both samples, one can use rotation tests (Kropf et al., 2007; Langsrud, 2005) instead of permutation tests. The main advantage of rotation tests is that they are applicable in the case of very small samples sizes that lead to a low number of possible permutations and hence severe power loss. The number of possible rotations is always infinite, whereas the number of possible permutations is always finite.

Generally, permutation tests with the statistic d have a surprisingly good power even for moderate sample sizes in high-dimensional data and are well applicable for sample sizes of 4 or larger per group (Kropf et al., 2007). The arguments about the exactness of the permutation test remain true for the alternative test statistics d^* and d^{**} as well as for other test statistics mentioned in Section 1 that are derived from asymptotic arguments (Pesarin & Salmaso, 2010).

In these permutation tests for the undirected test of difference, the test statistics d , d^* and d^{**} give equivalent test results. This is easy to see by using the symmetry of the distance measures in this section and expressing the sum of all differences δ_{ij} as a decomposition into δ_{within} and δ_{between} . Since this sum is the same for all permutations of the treatment indicator, there is a fixed monotonous relation between δ_{within} and δ_{between} .

Unfortunately, this simple permutation approach does not yield a confidence interval for the expectation $E(d)$. If that is desired (the test statistic d can also be interpreted as an effect measure), then the approximate leave-one-out and bootstrap procedures of Section 5 can be used.

Figure 1 illustrates the power of the permutation tests based on the statistic d for the three distance measures (2.6)–(2.8). As comparator we used the maxT permutation test by Westfall and Young (1993). The simulation series considers two independent samples of size $n_1 = n_2 = 10$ with multivariate normal data vectors of dimension $p = 20$. The variables are grouped into two blocks of variables. They are uncorrelated between the blocks but have an equal pairwise correlation

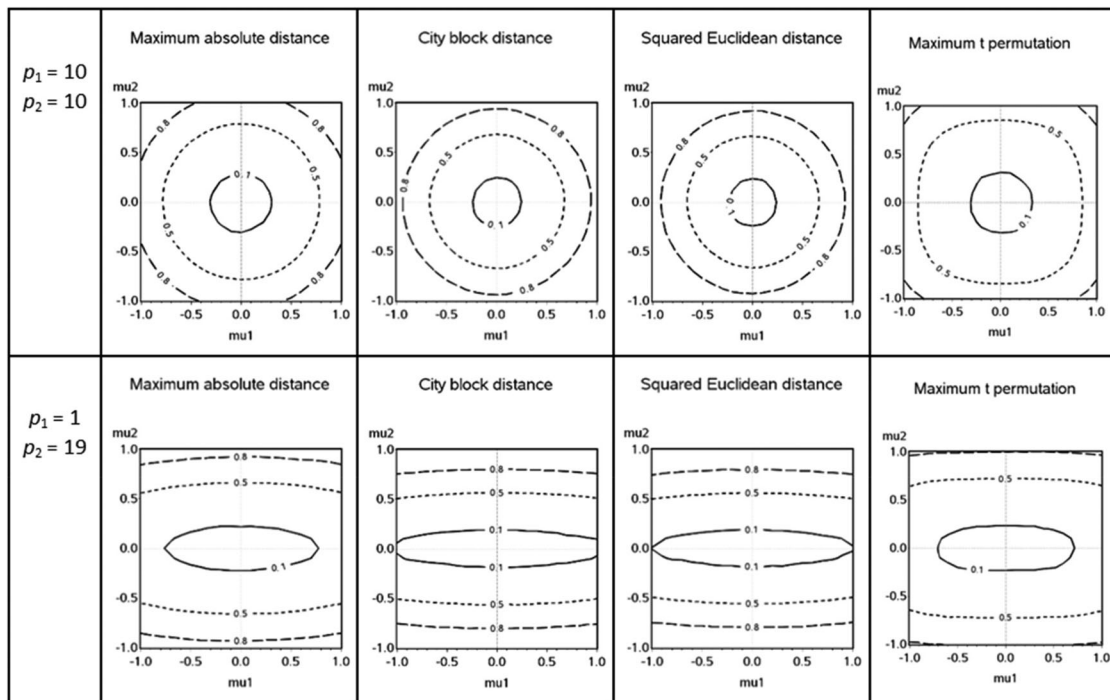


FIGURE 1 Contour plots for the power of distance based tests on two-sided difference between both groups (levels 0.1, 0.5, and 0.8). The parameters of the simulation runs are described in the text. Upper panel: equal number of variables in the two blocks of variables, lower panel: one variable belongs to block 1, the 19 others to block 2

coefficient of 0.3 within the blocks. All variables have variance 1 in both samples and expectation 0 in sample 1. Block 1 comprises p_1 variables with expectation μ_1 in sample 2, block 2 consists of the remaining p_2 variables with expectation μ_2 in sample 2. The two location parameters μ_1 and μ_2 vary in steps of 0.1 between -1 and $+1$. Each parameter constellation is investigated in 4000 independently simulated data sets. Permutation tests are carried out with 200 random permutations.

The upper panel of Figure 1 shows results for equal block sizes $p_1 = p_2 = 10$. Correspondingly, the contour plots are symmetric in the direction of both axes. For all three distance measures, the contour lines are approximately circles around the origin with the smallest diameter (largest power) for the squared Euclidean distance followed by the city block distance and the maximum absolute distance. The contour lines for the global test based on the maxT test have a more flattened shape around the axis and particularly much larger diameters reflecting a distinctly smaller power of this multiple test procedure when considered as a global test.

The plots in the lower panel show the results for $p_1 = 1$ and $p_2 = 19$, that is, the case where the group effect in the first variable differs from the common effect in all other 19. Correspondingly, the diameter of the contour lines is slightly smaller in μ_2 -direction (controlling the 19 variables) and much larger in μ_1 -direction. Deviations in μ_1 -direction are now better detected with the maximum absolute distance than with city block distance or squared Euclidean distance, whereas in μ_2 -direction again the maximum absolute distance has worst performance among the three distance measures. The maxT test is similar in performance to the maximum absolute distance and better than the other two distances for deviations in μ_1 -direction and still worse than all distance-based tests for deviations in μ_2 -direction.

Figure S2 in the Supporting Information shows additionally analogous results for the minimum absolute distance. In our parameter settings, this version yielded poor power and is thus omitted from subsequent consideration. According to Marozzi (2015), this distance has advantages in distributions with heavy-tailed and highly skewed distributions. In this paper, we will focus on moderate deviations from normal distributions (see discussion in Section 7).

4 | TESTS FOR SUPERIORITY

Next, we consider one-sided tests. The null hypothesis is still given by (3.1). However, we now attempt to design tests with a high power to reject H_0 when sample 2 is stochastically larger than sample 1.

As the symmetric distance measures (2.6)–(2.8) do not allow a distinction between “better” or “worse”, we use one-sided modifications here:

for the squared Euclidean distance

$$\delta_{ij}^{(+)} = \sum_{k: y_{jk} > y_{ik}} (y_{jk} - y_{ik})^2, \quad \delta_{ij}^{(-)} = \sum_{k: y_{jk} < y_{ik}} (y_{jk} - y_{ik})^2 \quad (4.1)$$

for the maximum absolute distance

$$\delta_{ij}^{(+)} = \max_{k: y_{jk} > y_{ik}} |y_{jk} - y_{ik}|, \quad \delta_{ij}^{(-)} = \max_{k: y_{jk} < y_{ik}} |y_{jk} - y_{ik}| \quad (4.2)$$

for the city block distance

$$\delta_{ij}^{(+)} = \sum_{k: y_{jk} > y_{ik}} |y_{jk} - y_{ik}|, \quad \delta_{ij}^{(-)} = \sum_{k: y_{jk} < y_{ik}} |y_{jk} - y_{ik}| \quad (4.3)$$

For convenience, we define $\delta_{ij}^{(+)} = 0$ if $y_{jk} \leq y_{ik}$ for all k and likewise $\delta_{ij}^{(-)} = 0$ if $y_{jk} \geq y_{ik}$ for all k . The minimum absolute distance is not suitable for this one-sided situation as it very often results in distances of zero.

Note that this kind of modification can destroy the meaning of a distance measure and may not be applicable for a given situation. This should be checked carefully on a case-by-case basis. Particularly, on compositional data (when variables represent relative frequencies), negative changes in some variables induce positive effects in other variables. One-sided test are not meaningful in this situation.

For this one-sided testing problem, it is important that δ_{between} in (2.10) is defined with the first sample element (index i) from sample 1 and the second (index j) from sample 2. For the two-sided tests of Section 2, this order of the samples does not matter.

The seemingly straightforward way would be the use of the “positive” distance measures $\delta_{ij}^{(+)}$ instead of δ_{ij} in the definitions of the test statistic of Section 2 together with the permutation test of Section 3. However, this results in one-sided tests for the very specific alternative that there is at least one variable with larger values in sample 2 compared to sample 1 regardless of possibly inverse effects in one or even the majority of the other variables. That might be of interest in special situations. Here, we are interested in alternatives with positive effects in at least one variable and essentially no inverse effects in the other variables. Therefore, the use of $\delta_{ij}^{(+)}$ is not further investigated here.

Instead, we propose to reverse the procedure in two details:

1. We use the “negative” distances $\delta_{ij}^{(-)}$ in the definition of the test statistic yielding the values d_0 for the original sample and d_i ($i = 1, \dots, N$) for the permuted samples. Analogously, the alternative test statistics d^* and d^{**} could be used.
2. We calculate the p -value of the permutation test at the lower tail of the null distribution:

$$P^{(-)} = \frac{\#\{i \in \{1, \dots, N\} : d_i \leq d_0\} + 1}{N + 1}. \quad (4.4)$$

In other words, we show that there are less negative effects in the variables than one would expect under the null hypothesis of equality of both distributions. The method rejects H_0 if few variables indicate ‘bad’ results on sample 2 and these inferior results are numerically not much inferior.

This procedure guarantees an exact control of the type I error under the null hypothesis (3.1) of equal multivariate distributions in both samples. Considering the shift hypothesis

$$H_1 : F_2(y) = F_1(y - \theta) \quad \forall y$$

for some shift vector θ as class of alternatives in (3.2), rejections of the null hypothesis also occur outside the positive orthant of θ with a probability above the nominal alpha level. That happens particularly near the borders of the positive orthant where large positive effects in some variables override small negative effects in others. Our proposal shares this problem with other one-sided test versions such as the ALR test of Tang et al. (1993), one-sided applications of O’Brien’s

(1984) test versions or the proposal of Glimm and Läuter (2010). Even in a multiple test procedure with local tests for each component of θ , it would be difficult to achieve a strict type I error control over the whole parameter space outside the positive orthant without dramatic power losses in situations where group differences in the desired direction occur only in some of the variables. One would have to accept a very poor power on the borders of the orthant or to define tolerance limits for acceptable deteriorations in single variables.

In contrast to the two-sided tests in Section 3, there is no direct relation between the two averages δ_{between} and δ_{within} over the permutations of the combined samples. The term δ_{between} includes only those pairs of sample elements where the first element belongs to sample 1 and the second one to sample 2. The “opposite” pairs have different distances with the asymmetric distance measures, and this omitted part is not constant over the permutations. As a consequence, the three statistics d , d^* , and d^{**} give different test decisions (p -values). In situations with a large shift between the two groups in some of the variables, the distances of sample elements from different groups are essentially determined by these (possible few) variables with large shift. In contrast, for pairs of sample elements of the same group, the number of contributing variables (with negative shift) may vary from zero to p , which brings about a large variability of δ_{within} .

Therefore, statistic d is omitted in the simulation study below. The statistics d^* and d^{**} do not suffer from this problem because δ_{within} is not included in d^* , and it occurs in the nominator as well as the denominator of d^{**} levelling out the disturbances.

Figure 2a shows the results of simulation experiments with these test versions with the same parameter settings as in Section 3, only the two location parameters μ_1 and μ_2 controlling the two blocks of variables vary here in larger steps of 0.3 between -3 and $+3$. Again, we consider the situation of two blocks of variables of an equal size of 10 and the situation with block 1 being a single variable and block 2 comprising the remaining 19 variables. The city block distance gives very similar results as the squared Euclidean distance and is therefore omitted here. Both remaining distance measures are investigated, each in combination with the two test statistics d^* and d^{**} . In this one-sided situation, two comparators are considered in Figure 2b. The first one is the one-sided maxT test (Union-intersection test: “Is there at least one multiplicity adjusted univariate significance?”), the second one is the combination of the one-sided t tests by the intersection-union principle (“Are all unadjusted univariate tests significant?”).

As expected, in all distance-based test versions the contour lines for the power level of 0.05 (nominal test level) cross the point $\mu_1 = \mu_2 = 0$. The parameter region with a power above 0.05 also comprises points outside the positive orthant as discussed above. However, the versions based on the statistic d^{**} are more focused on this orthant than those based on d^* . Furthermore, the versions based on the maximum absolute distance are concentrated a little more in the positive orthant than versions based on the squared Euclidean distance. The second panel shows, however, that even in the most focused version with maximum absolute distance and statistic d^{**} , negative deviations in a single variable can be disguised by large positive deviations in the other variables.

The choice of the test version involves a trade-off. On the one hand, a large power in the positive orthant is desirable. That includes the boundaries (except the origin) where positive effects in a part of the variables are meaningful even if other variables do not show effects. On the other hand, (substantial) negative effects should be ruled out.

For comparison, Figure 2b shows the contour lines in the same situation for two extreme strategies. The one-sided maxT test considers positive effects in at least one variable, no matter what other variables do. That is the situation discussed above for the use of the “positive” one-sided distance measures, which is usually not the objective in one-sided multivariate tests. The stringent exclusion of negative effects—as the other extreme—is ensured with the intersection union principle where the multivariate decision is significant only if a positive effect is proven for each single variable. For practical purposes, this requirement is too strict, leading to very low power in general and particularly for low or absent effects in some variables.

5 | TESTS FOR EQUIVALENCE

The classical way to test for equivalence in a multivariate context would be to show equivalence in each variable. According to the intersection-union principle, this requires no α -adjustment (Hothorn & Oberdoerfer, 2006). However, the power to declare equivalence is decreasing with the number of variables, hence this requirement is unreasonably strict in high-dimensional data. Therefore, multivariate distance-based tests are applied for this type of test problem here. The two groups are considered equivalent if the multivariate distance measure (defined as the expectation of one of the test statistics) can be shown to be smaller than a pre-defined tolerance threshold θ , that is, if the null hypothesis $H : E(D) > \theta$ can

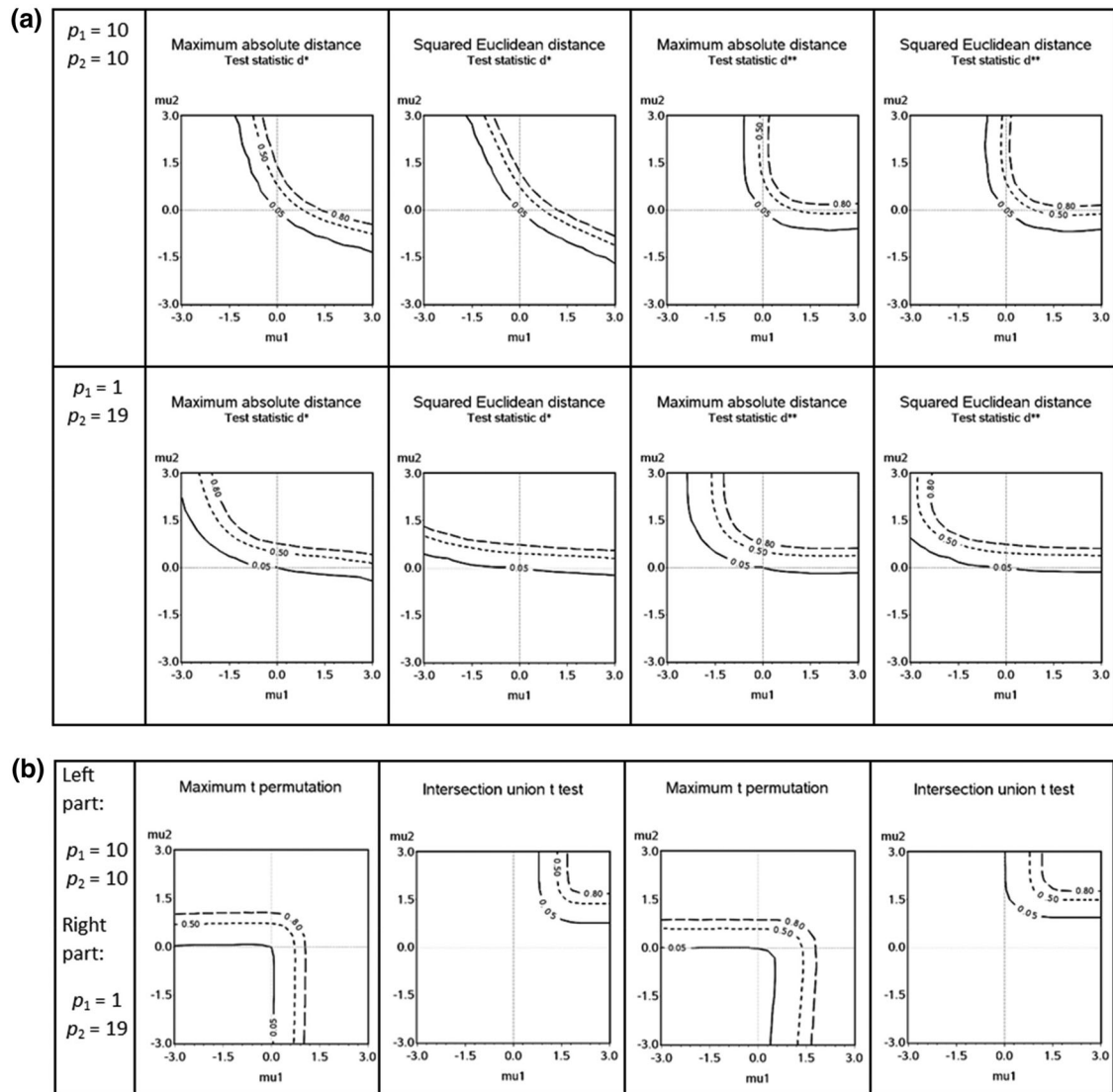


FIGURE 2 (a) Contour plots for the power of distance based tests on one-sided difference between both groups (levels 0.05, 0.5, and 0.8). The parameters of the simulation runs are described in the text. (b) Contour plots for the power of one-sided maxT test and one-sided univariate t tests combined with the intersection-union principle, both as comparators to the results of the distance based tests in Figure 2a (levels 0.05, 0.5, and 0.8)

be rejected at a given significance level. Here and later on in this section, D denotes one of the test statistics (2.12)–(2.14) (or a transformation of them, see below).

Chervoneva et al. (2007) investigated this situation for multivariate normal variables and gave an asymptotic solution that in some sense corresponds to our approach when using the squared Euclidean distance as the distance measure between two sample vectors.

Unfortunately, we did not find a way to use permutation tests here. The null hypothesis is not a point hypothesis. The same multivariate distance can arise from various univariate shifts in the p variables and there are shifts that remain within the space of the null hypothesis. Due to this, it is not obvious how shift methods (or permutations of residuals that would allow the adaptation of permutation tests) could be generalized to this situation. A somewhat different approach based on ranks (Pauly et al., 2016) could be considered, but is not further explored here.

Hence, we propose to apply a leave-one-out approach instead. The starting point is one of the test statistics (2.12)–(2.14) or a transformation of them (particularly, we will use the square root of (2.13) later on). The statistic is computed once with the whole sample of all $n = n_1 + n_2$ sample elements and then n times repeatedly under successive exclusion of one sample element in each repetition. Here we use the notation D for the test statistic (one of d , d^* , d^{**} , or $\sqrt{d^*}$) in the

whole sample and D_{-j} ($j = 1, \dots, n$) for the leave-one-out versions for this statistic (see restrictions below). With this notation, Efron and Stein (1981) have proposed a variance estimator

$$\widehat{\text{VAR}} D. = \hat{\sigma}^2 = \frac{n-1}{n} \sum_{j=1}^n (D_{-j} - D.)^2 \quad \text{with} \quad D. = \frac{1}{n} \sum_{j=1}^n D_{-j} \quad (5.1)$$

for the average of the leave-one-out estimates. They have shown in Theorem 2 of their paper that this yields a conservative estimate for any statistic having finite second moment, where $D = D(y_1, \dots, y_n)$ is not necessarily symmetrically defined in its arguments and the y_j are independent but not necessarily identically distributed. Thus, the two-sample situation is covered as long as the finite second moment exists.

Here we assume that the variance of D is approximately equal to the variance of $D.$ and that the test statistic approximately follows a normal distribution. Asymptotically, if the number of variables remains finite while the sample size goes to infinity in such a way that $0 < \frac{n_1}{n_2} < \infty$, these assumptions are covered under mild regularity conditions by standard central limit theorems (Sen, 1985). In our simulation studies, this also seemed acceptable for the test statistic d^{**} from (2.14) for the finite sample sizes we considered. The statistic d^* from (2.13) has a skewed distribution but can sufficiently be normalized using the square root $\sqrt{d^*}$. We did not find a suitable transformation for statistic d from (2.12), so that it cannot be used here. Under this normality assumption, we calculate an approximate one-sided $(1 - \alpha)$ -confidence interval for the expectation of D by

$$\left(-\infty, D + t_{1-\alpha, n-2} \sqrt{\hat{\sigma}^2}\right). \quad (5.2)$$

In order to complete the equivalence test, a threshold θ for the expectation of the test statistic has to be defined that constitutes the maximal distance between both populations to be still considered as equivalent. If the upper limit of the confidence interval is below the threshold, that is, $D + t_{1-\alpha, n-2} \sqrt{\hat{\sigma}^2} < \theta$, then equivalence is concluded. The actual type I error control of this procedure is checked for selected normal data situations subsequently in this section and in more general settings in Section 7.

Since it is already a challenge to define such a threshold in the univariate situation, it is even more difficult in this setup with complex multivariate distance measures and test statistics derived from them. In Antweiler et al. (2017), additional features of the experimental design have been used to derive a threshold. A selection of those data is reconsidered here as Example 2 in Section 8. But such additional features are not always available in an experiment. Test statistic d^{**} in (2.13) can still be calculated if such features are missing. The statistic d^{**} describes the increase of between-group distances over within-group distances, relative to the within-group distances. For such a relative measure it might be possible to find a general agreement similar to the 80% / 125% limits in bioequivalence studies.

Again, the simulation settings of the previous section are used to demonstrate the performance of the proposed tests. We present the results for the test statistics d^{**} (as relative measure) and $\sqrt{d^*}$ (as absolute measure). Both yielded at least partly approximate normal distributions over the simulation runs for fixed parameter settings (histogram plots over the simulation replications, not shown here). As a consequence, the type I error was kept in good approximation in these situations (see below).

Figure 3a shows the results of simulation series with both statistics. As in the previous sections, the power over the range of both shift parameters μ_1 and μ_2 is presented in contour plots (contour lines for 5%, 50%, and 80%). The threshold for the relative statistic d^{**} was chosen as 0.25 (inspired by the usual univariate thresholds in bioequivalence studies). For the statistic $\sqrt{d^*}$, the threshold (2.0 with maximal absolute distance, 10.0 with squared Euclidean distance) was fixed pragmatically at values that led to a sufficient variation of power values in the considered range of the expectation parameters μ_1 and μ_2 . Of, course, the largest power values are around the origin.

Additionally, for each μ_1 - μ_2 -combination, the expectation of the test statistic was estimated as the arithmetic mean of the calculated values of the test statistic D over the 4000 independent simulation runs for a parameter setting. The contour line for this expectation at the level of the tolerance threshold is added to the figures. This allows to distinguish between parameter settings under the formal null hypothesis (expectation of test statistic is larger than the threshold) and those under the alternative hypothesis (expectation of test statistic is small enough to accept equivalence). Comparing the contour lines of the power and that of the expected test statistic enables a check of type I error control.

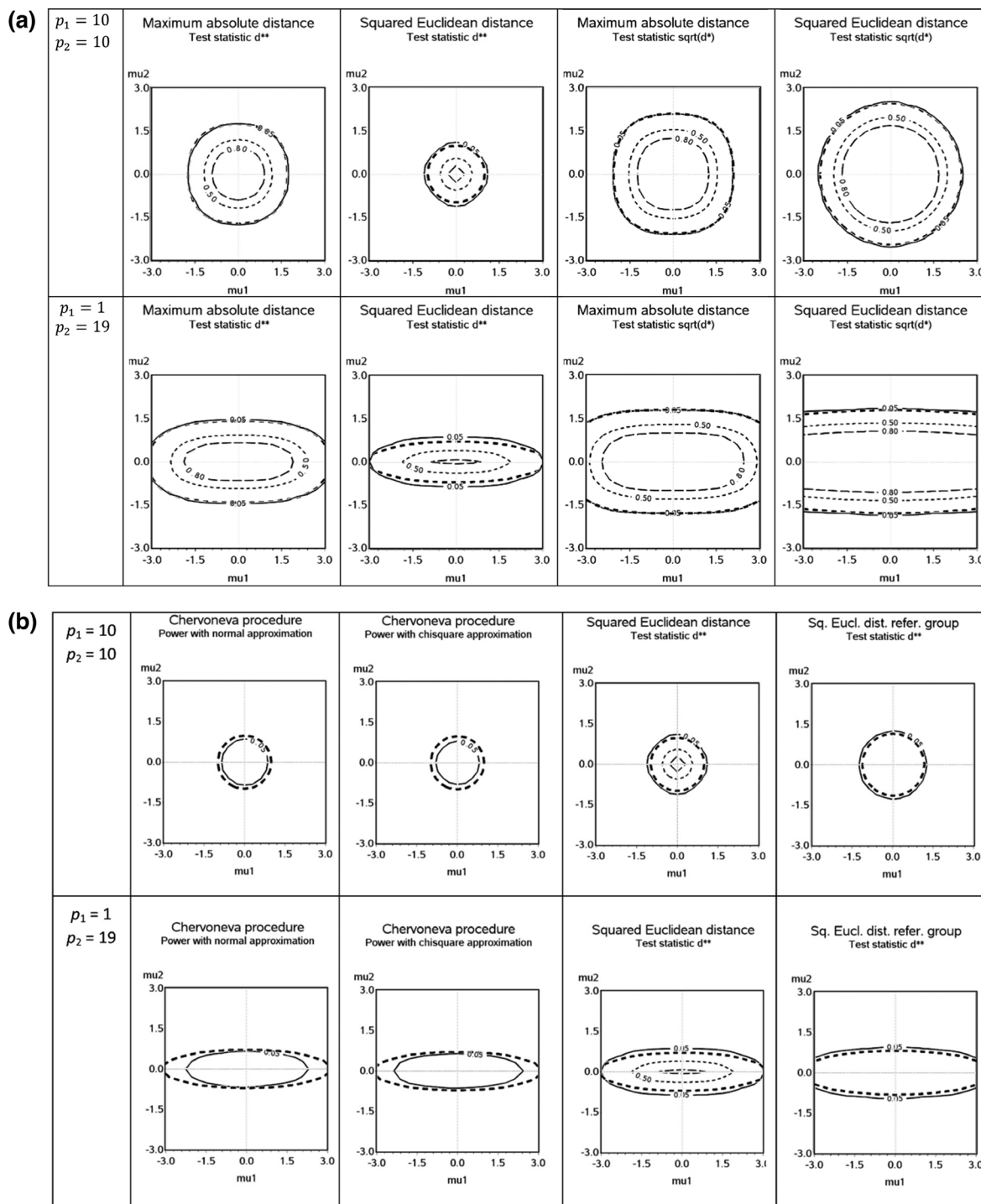


FIGURE 3 (a) Contour plots for the power of distance-based tests on (two-sided) equivalence between both groups (levels 0.05, 0.5, and 0.8). The parameters of the simulation runs are described in the text. The figure includes two versions of distances (Maximum absolute distance and squared Euclidean distance), both in combination with one of the test statistics $\sqrt{d^*}$ or d^{**} and for two independent blocks of variables of size 10/10 or 1/19, respectively. (b) Contour plots for the power of the test proposals by Chervoneva et al. (2007) using both the normal and the chi-square approximation compared to the above test best on the squared Euclidean distance and the relative statistic d^{**} and a modified version where the within-group distances are computed only from the reference group. Additionally, the limits of that part of the parameter space under the null hypothesis are presented as bold dashed lines

For the tests based on the maximum absolute distance, the contour lines for the power of 5% and for the expectation of the test statistic at the threshold value agree very well indicating a good formal type I error control. The statistic d^{**} based on squared Euclidean distance showed a very slight anticonservative behavior as the region within the expectation threshold is a bit smaller than the region with a power above 5%. In the simulation runs, additionally the coverage rates of the region (5.2) for the expected value of the test statistics have been obtained at each grid point in the μ_1 - μ_2 -plane. These graphs are not presented here because they agree very well with the considerations so far: the coverage rates were very close to 95% in most situations but were only between 90% and 95% in some regions of the considered μ_1 - μ_2 -plane for the statistic d^{**} with the squared Euclidean distance. When increasing the sample sizes to 20 per group, these deviations disappeared.

The regions of power above the considered reference values for the tests with the relative statistic d^{**} are larger with the maximum absolute distance than with the squared Euclidean distance. That indicates a larger power for the maximum absolute distance. However, one has to keep in mind that the same threshold for the expectation of the test statistic has a different meaning in terms of the expectation of the raw data even though both test versions use the relative presentation d^{**} . That can be seen from the contour lines for the expectation of the test statistic at the level of the equivalence threshold (bold dashed lines). An allowed deviation of 25% is more stringent in a statistic based on a quadratic measure than with a linear one. Power comparisons between the versions based on d^{**} and $\sqrt{d^*}$ and also between the $\sqrt{d^*}$ -version with squared Euclidean distance and with maximum absolute distance are not possible for the same reason that the equivalence thresholds are not comparable.

Although the upper panel with equal number of variables in the two blocks with expectation μ_1 and μ_2 , respectively, show symmetric graphs with respect to the two axes, the contours in the lower panel have larger extent in the μ_1 -direction. A large difference in a single variable is not ignored, but has less influence than moderate differences in many variables.

Figure 3b compares the test proposal by Chervoneva et al. (2007) with the above test based on the squared Euclidean distance and the relative statistic d^{**} . Chervoneva et al. propose two asymptotic approximations, a normal and a chi-square approximation. Our threshold of 0.25 corresponds to a threshold of 0.5 in their notation. Despite the common basis of multivariate normal data, the use of squared Euclidean distances and a ratio statistic, there are two conceptual differences. Although Chervoneva et al. consider the ratio of the expectations of numerator and denominator, we use the expectation of the ratio. More importantly, Chervoneva et al. use only the reference group for the calculation of within-group distances. Therefore, Figure 3b includes a modification of our proposal, where the calculation of within-group differences is restricted to the reference group (sample 1) as well. The contour lines for the power of the procedures are again presented together with the contour lines for the limits of the parameter space under the null hypothesis.

Both approximations for the test of Chervoneva et al. show a low power, so that only the contour line for a power of 0.05 is presented. The comparison with the limit of the parameter space under the null hypothesis shows an anticonservative behavior in both approximations (cf. Section 7 for some additional results). Our original test proposal for the squared Euclidean distance reaches a power of 0.8 near the origin. In the modified version with within-group distances calculated only from the reference group, the power is similar to that of the Chervoneva procedure.

6 | TESTS FOR NON-INFERIORITY

For a test of noninferiority, we use the leave-one-out approaches from Section 5 in combination with the asymmetric distance measures (4.1)–(4.3). In contrast to Section 4, we no longer aim to show that group 2 is better than group 1, but rather that it is not relevantly worse. Following the approach in Section 4, again an inversion of the one-sided (asymmetric) measures has to be used in order to rule out a strong deterioration in one or several variables. The averages δ_{within} and δ_{between} and the test statistic are computed according to (2.10), (2.11) and one of (2.12)–(2.14) from the distance measures (4.1), (4.2), or (4.3). This is repeated in a leave-one-out cycle in order to determine the one-sided confidence interval for the expectation as in (5.1)/(5.2) and finally to compare the interval with the tolerance threshold. Remember that the inverted one-sided distance measures assess the deterioration of group 2 with respect to group 1. Therefore, the *upper* limit of the confidence interval for the expectation of the statistic D is used in the test for noninferiority as well.

As in Section 5, the “relative” test statistics d^{**} and the transformed “absolute” statistic $\sqrt{d^*}$ showed approximate normal distributions in the respective histogram plots when considered in simulation experiments with the same data

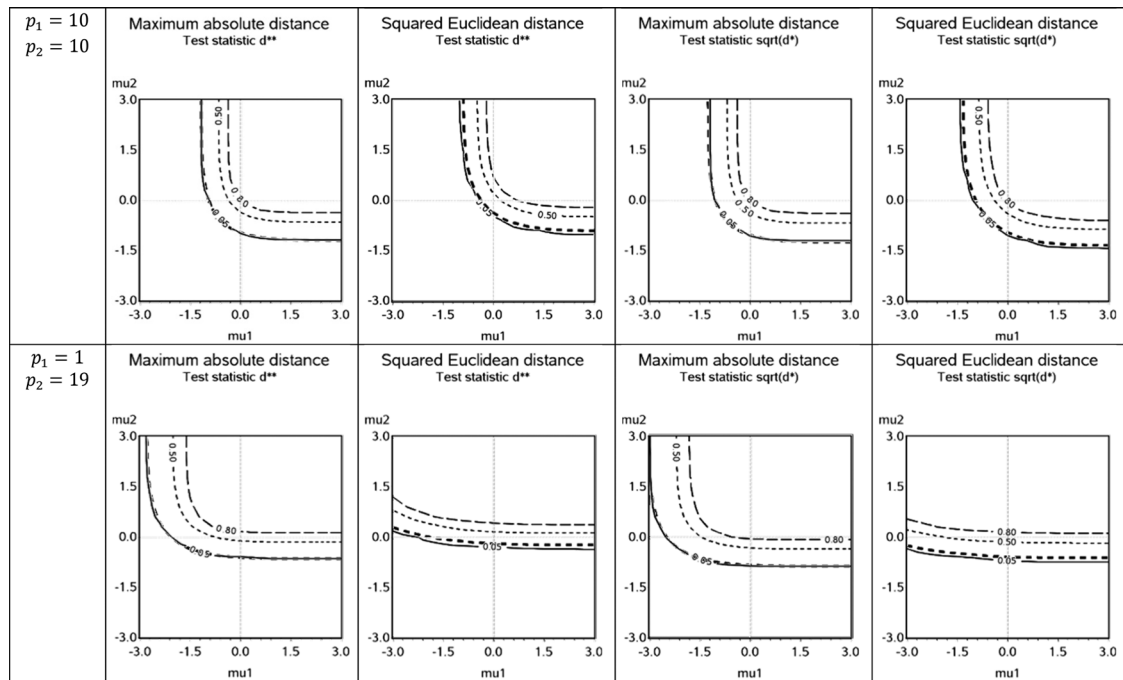


FIGURE 4 Contour plots for the power of distance based tests for noninferiority of group 2 with respect to group 1 (levels 0.05, 0.5, and 0.8). The parameters of the simulation runs are described in the text. The upper panels show the results for two versions of distances (Maximum absolute distance and squared Euclidean distance), both in combination with the test statistic d^{**} and $\sqrt{d^*}$ for two independent blocks of 10 variables each. The lower panel presents the analogous results for the case that the first block of variables comprises only of one variable and the second one of 19 variables. In addition to the contour lines for a power of 5%, 50%, and 80% (all thin lines), the contour lines for the expectation of the multivariate measure ($\sqrt{d^*}$ or d^{**}) at the threshold values of the equivalence test are presented as bold dashed lines, derived as average over the 4000 independent simulation runs. The threshold values used are 0.25 for d^{**} (with both distance measures), 6 for $\sqrt{d^*}$ with squared Euclidean distance, and 1.75 for $\sqrt{d^*}$ with maximum absolute distance

settings as in Sections 4 and 5. For the simulation series with the relative statistic d^{**} , the same tolerance threshold of 0.25 is used as for the equivalence tests. For the transformed absolute statistic $\sqrt{d^*}$, smaller values for the threshold are used than in the last section because the one-sided distance measures give smaller values than the two-sided ones. Therefore, the threshold 1.75 is used in combination with the maximum absolute distance, and the value 6 for the squared Euclidean distance.

Figure 4 shows the results of the simulation series. Again, the contour plots for the power over varying values of the two shift parameters μ_1 and μ_2 are combined with the contour line for the expectation of the test statistics at the level of the tolerance threshold.

The contour line for the expectation of the test statistic at the used threshold level coincides rather well with the contour line of the power 0.05 in all plots (a bit better for the maximum absolute distance). The nominal type I error is reasonably well kept in these analyses.

In the upper panel, the four plots are similar. Of course, this relies on the choice of the tolerance threshold values. Particularly, the thresholds for the statistic $\sqrt{d^*}$ were chosen pragmatically to give similar plots. In contrast, the choice of the threshold 0.25 for statistic d^{**} is less arbitrary. As in the simulation runs for the equivalence tests, one can observe that the same threshold for this “relative” statistic yields slightly different contour lines for the expectation of the statistics and for the power. These lines are shifted into the negative direction on both axes for the maximum absolute distance, so that for only slightly negative shifts noninferiority can be proven with this measure but not with the squared Euclidean distance. The differences are much more obvious in the lower panel with only one variable with expectation μ_1 . In the analyses with the squared Euclidean distance, changes in this one variable have only little effect on the test result. In contrast, the maximum absolute distance reflects such changes much more (as expected from the definition).

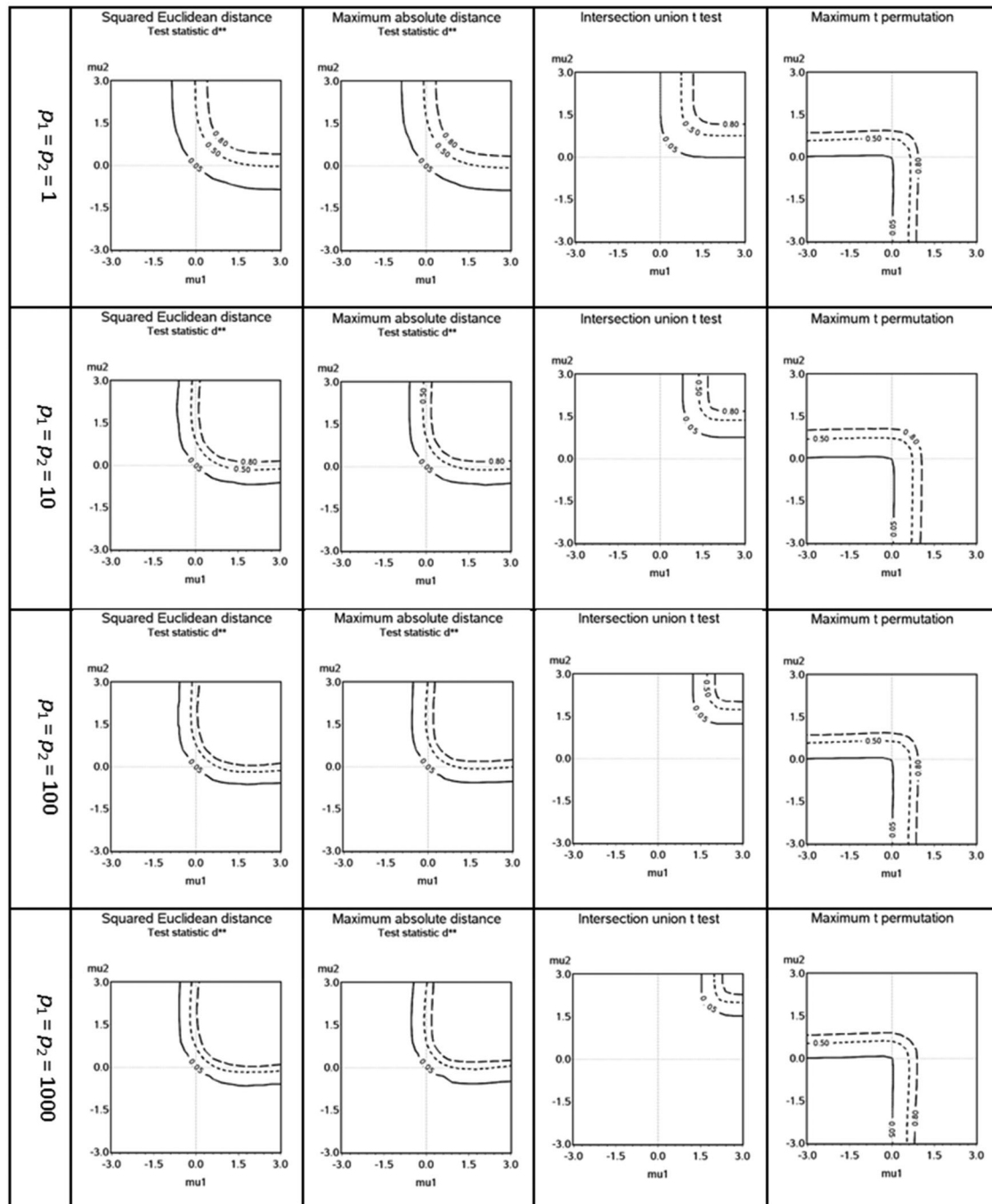


FIGURE 5 Contour plots for the power of one-sided tests for superiority (levels 0.05, 0.5, and 0.8) with varying number of variables in the two uncorrelated blocks of variables. Other parameters are used as before

7 | BEHAVIOR IN LARGER DIMENSIONS AND ROBUSTNESS

Although the concept of the distance-based tests is not restricted in the number of variables, we have so far considered rather small dimensions in the simulation settings, mainly for reasons of computing time. Here we will consider some selected test versions for varying number of variables. Again, we use two samples of size 10 each and two uncorrelated blocks of variables of size p_1 and p_2 with within-block correlations of 0.3 as before. The shift between both samples is constant within the blocks, but varies between the blocks (μ_1 and μ_2 , respectively). However, now the number of variables per block varies as $p_1 = p_2 = 1, 10, 100, 1000$. Figure 5 shows the results for selected one-sided tests (4000 replications).

The tests with the squared Euclidean distance and with the maximum absolute distance (both one-sided versions and combined with the relative test statistic d^{**}) show a distinct improvement in power over the positive orthant as $p_1 = p_2 = 1$ increases to $p_1 = p_2 = 10$. For further increases of $p_1 = p_2$ power does not increase a lot anymore (in contrast to corresponding results for two-sided tests that are not presented here). For the univariate t tests combined with the intersection-union principle used as comparator here, the regions with power above the considered thresholds 0.05, 0.5, and 0.8 drift more and more towards the upper right corner of the parameter field. This is the price that has to be paid for the strict one-sided error control in each variable in comparison with the aggregated assessment of variables. Again, the maximum t permutation test rejects also in such constellations where only one block of variables has a positive shift but the other has a negative shift. That behavior would usually not be considered as useful in one-sided multivariate tests.

Figure S1 shows analogous results for some selected versions of equivalence tests. The tests based on the squared Euclidean distance and the maximum absolute distance, both as two-sided versions combined with the relative test statistic d^{**} , show an anticonservative behavior for variable blocks of size 1. This can be seen from the fact that the contour lines for the power at level 0.05 are outside the region limited by the bold dashed lines, that is, within the parameter region under the null hypothesis. With increasing number of variables, this anticonservative behavior completely disappears for the maximum absolute distance and to a large extent for the squared Euclidean distance. In contrast, both versions of Chervoneva's approach (normal and chi-square approximation) become increasingly conservative with larger dimensions.

With respect to robustness, we consider two aspects here: deviations from the normal distribution of the data and imbalanced sample sizes. Additional figures for these situations are given in the Supplemental Material. Here, we present the investigated scenarios and the summary of the results.

The tests for difference or superiority use permutation techniques and do not depend on the distributions with regard to type I error control. However, the shapes of the contour lines for the power might depend on the distribution. In the tests for equivalence and noninferiority, there is the additional question of the performance of the normal approximation in the leave-one-out solution for the confidence interval. In order to investigate these issues, we generated data with slight deviations from the normal distribution in a first step. In particular, slightly skewed data were generated by the Gamma $(2, \frac{1}{2}\sqrt{2})$ distribution, deviations in the excess by the Laplace $(0, \frac{1}{2}\sqrt{2})$ distribution. Both have standard deviation 1 as used before for the normal data. After generating sample matrices with *iid* elements from these distributions, the correlation structure was introduced by multiplying the matrix by the Cholesky root of the same covariance matrix that was used in the simulation of normally distributed data above. This appended multiplication disturbs the marginal distributions, but with the considered relative small correlation values, the disturbance is small. We restrict here to the case $p_1 = p_2 = 10$, the results for $p_1 = 1, p_2 = 19$ were similar.

As shown in Figures S2 and S3, the contour lines for power in the tests for difference and superiority change only very little with these mild modifications of the distribution for nearly all test versions considered in this paper. Only tests with the maximum absolute distance lose a bit of power, whereas those with the minimum absolute distance gain a bit (but are still less powerful than most other test versions). The shapes of the contour lines for power and the type I error control in tests for equivalence and non-inferiority in Figures S4 and S5 are also very similar to the normal case. The most obvious change is the asymmetric shape of the contour lines for power with the Chervoneva procedure in the case of Gamma distributed data. This situation has also been studied with unbalanced sample sizes ($n_1 = 5, n_2 = 15$). As shown in Figure S6, the test versions based on the squared Euclidean distance show an anticonservative behavior (more distinctly with the statistic d^{**} , but in tendency also for $\sqrt{d^{**}}$) with normal and Laplace distributed data. Tests based on the maximum absolute distance approximately keep the type 1 error with these symmetric distributions, but are also slightly liberal with the skewed Gamma distribution.

If the distributions are more skewed, other problems may become relevant. This is demonstrated for log-normal data in Figure 6 for equivalence and noninferiority tests based on the squared Euclidean distance and the maximum absolute difference, considered with sample sizes of 10 for each group as before and for the version with two blocks of variables, each of size 10. Column (a) of Figure 6 shows the contour plots for log-normal data (exponentiated normal data with the correlation structure as before and a subsequent shift with parameters μ_1 and μ_2 for the two blocks of variables). The contour lines for the power have approximately the expected shape, even though these lines are not as close to the limits of the first orthant as with normal data. The tests show a conservative behavior in some cases. As the log-normal distribution has variance greater than 1 here, the border of the parameters under the null hypothesis is outside of the presented section of the parameter space in the second row of the figure. More importantly, however, adding a shift to a log-normal distribution gives a distribution that might be considered very artificial. For example, negative values might

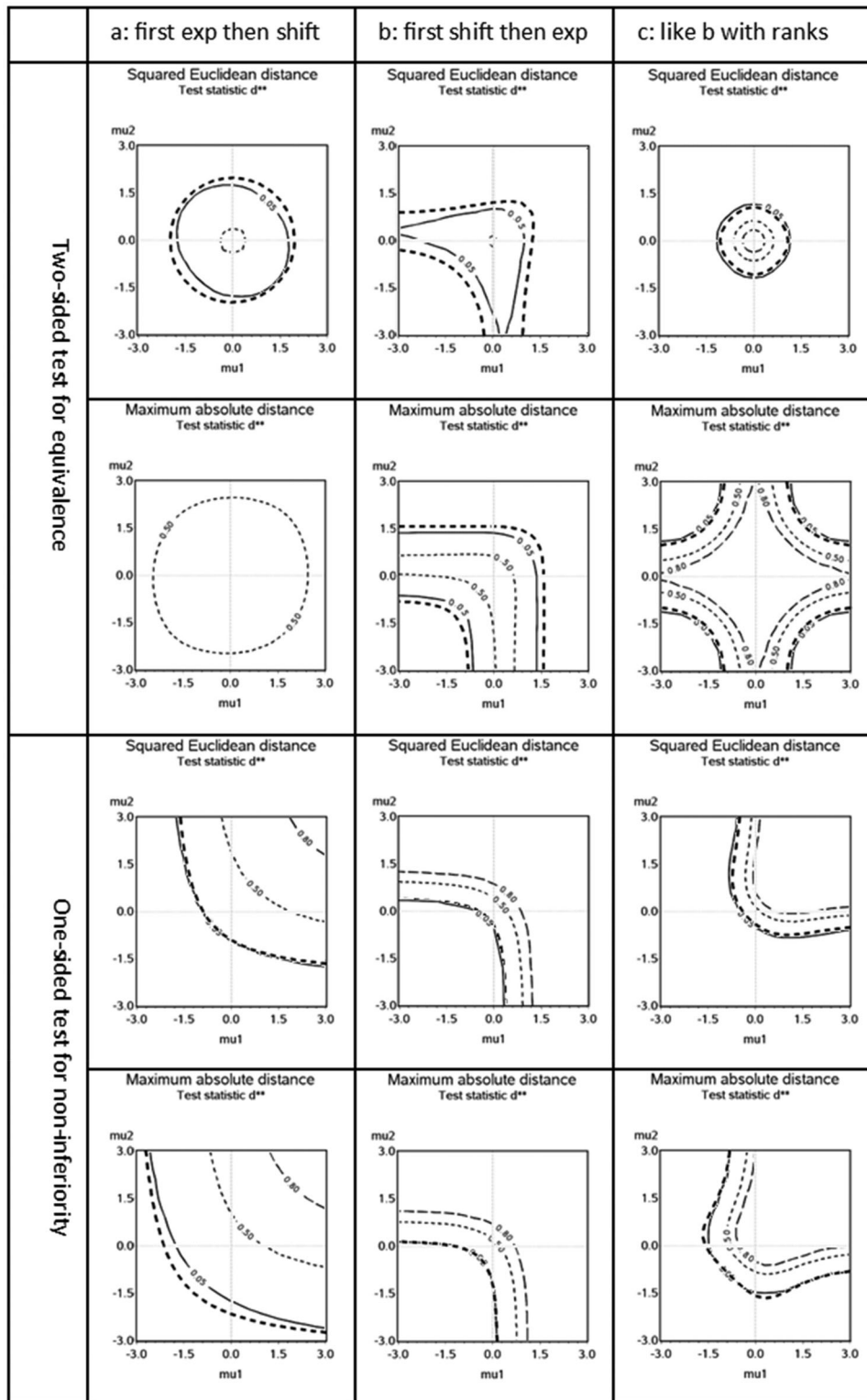


FIGURE 6 Contour plots for the power of tests for equivalence and noninferiority with levels 0.05, 0.5, and 0.8 with log-normal distributed data ($n_1 = n_2 = 10$, $p_1 = p_2 = 10$, other parameters as before). The contour lines for the region under the null hypothesis are added as bold dashed lines. For the description of columns a, b, and c, see the text

occur with a negative shift. Furthermore, in many fields of application, ratios are preferred to differences of observations when the distributions on the original scale are skewed. Therefore, column (b) of Figure 6 considers the situation where the shift by μ_1 and μ_2 is applied to the normal data before the transformation by $\exp(\cdot)$. Then the geometrical distances in the figure correspond to the suspected practical impact, but the shapes of the contour lines differ strongly from the desired shapes. That is particularly obvious in the tests for noninferiority, where large parts of the second and fourth orthant lead to false decisions (with respect to the intended interpretation of the test as checking non-inferiority). Figure S7 shows the results of analogous tests for undirected differences or superiority that support the above discussion. In order to avoid such misbehavior, the measurement scale should be chosen appropriately or it should be corrected before applying the test. In the present case, taking the logarithms would trace back the situation to the case of normal distribution as considered in the previous sections. Also a rank transformation of the skewed distribution works well here, particularly in combination with the squared Euclidean distance, as shown in column (c). Obviously, the combination of all variables in the calculation of the multivariate distances smooths the discrete rank transformation sufficiently well. In contrast, the maximum absolute difference shows deformations of the contour lines in some of the considered situations. There is some analogy to the proposal by Marozzi et al. (2020) who apply ranks to the distances, whereas we apply the ranks to the raw data before starting the tests. Of course, a rank transformation has to be taken into consideration when defining tolerance thresholds for equivalence or noninferiority, and one has to decide if a rank based threshold is appropriate for the given research question.

We conclude this section with a remark on the random error associated with the presented simulation study. All series have been performed with 4000 replications. The same set of random numbers has been used for each parameter combination (μ_1, μ_2) in order to prevent random fluctuations in the contour lines. In Figure S8, we oppose some of the previously shown figures with versions generated with different seeds for the random number generation. The calculated power values for each parameter combination (not shown here) vary in line with what can be expected from the performed number of simulations according to the corresponding binomial rejection probabilities. Regarding the figures in this paper and the Supplemental Material, the influence on the presented contour lines is so small that it can hardly be seen even when the corresponding figures are overlaid.

8 | EXAMPLES

8.1 | Example 1—gene expression data

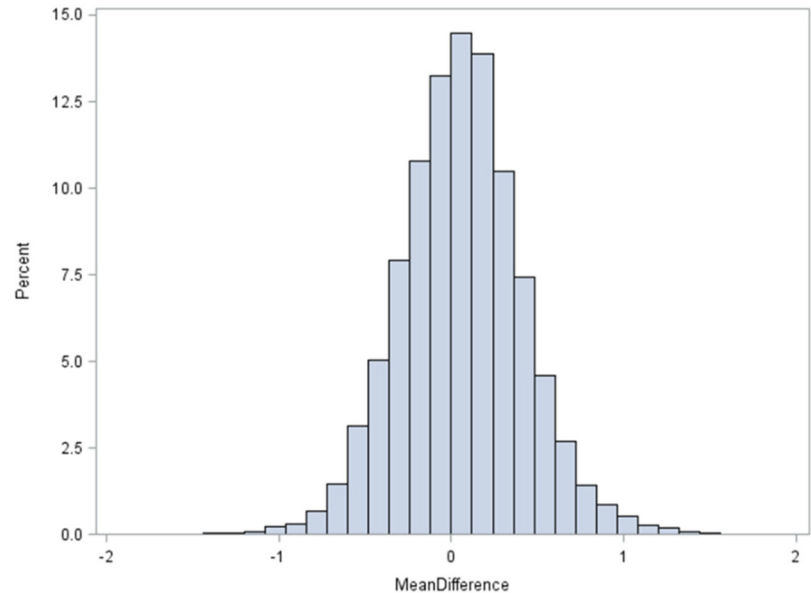
The first example uses gene expression data for thyroid nodules from a study of Prof. Paschke and colleagues at the Medical Department III, University of Leipzig, Germany (Eszlinger et al., 2001). Originally, 15 patients with hot nodules and 15 with cold nodules were included in the study. Here we use only the data of the female patients (12 hot, 9 cold nodules) thus eliminating possible confounding by gender effects and reducing the power of all included tests versions to make them distinguishable. The tissue samples were analyzed by Affymetrix GeneChips considering gene expression levels of 12,625 genes. The data and the SAS code for the analyses are presented in the Supplemental Material.

We applied a log transformation to the expression values in order to reduce skewness and to adapt to a fold change scale that is often applied in gene expression analyses.

The primary research question concerns the presence of differences in gene expression between hot and cold nodules. For our test proposals, we use the city block distance. Technically, we used the relative statistic d^{**} in a permutation test with 99,999 random permutations. However, as mentioned in Section 2, all tests statistics deliver identical p -values in the test for undirected differences. The value of the test statistic is $d^{**} = 0.045$ with the corresponding p -value of 0.00356. Using maxT tests by the SAS procedure MULTTEST as one of the common methods for the analysis of high-dimensional expression data as competing approach, we end up with the smallest permutation adjusted p -value of 0.0046, which is slightly larger than that of our distance-based test.

For biological reasons, one might suspect that the observed differences are mostly due to larger expression values in the hot nodules. Therefore, we also apply the one-sided test version to compare hot versus cold nodules here. Again we use the city block distance (one-sided version) and the test statistic d^{**} . The test statistic is $d^{**} = -0.053$ with the corresponding p -value of 0.2309, which is far from significance. Figure 7 shows a histogram of the differences of the mean log values from hot and cold nodules for all included genes. The slight tendency toward larger expression values in the hot nodules is not strong enough to render the one-sided test significant.

FIGURE 7 Histogram of the difference of the mean ranks of gene expression data of hot and cold nodules (hot minus cold) for the 12,625 genes



8.2 | Example 2—sequencing data

For the test of equivalence, we reconsider data of Antweiler et al. (2017) with a modified test statistic. The data originate from a biological study of the Institute for Epidemiology and Pathogen Diagnostics of the Julius Kühn-Institut in Braunschweig, Germany. The purpose of the study was to investigate if an antifungal biocontrol treatment with *P. jessenii* RU47 against the target pathogen *R. solani* affects the fungal composition in arable soil in the next season. It was hoped that the study can rule out disturbances of the fungal composition by the treatment. Three different soils were investigated in an experimental plot system in Großbeeren near Berlin. For each soil type, eight samples were taken from plots with RU47 treatment in the previous season and eight samples from control plots. The samples were analyzed by high-throughput ITS amplicon sequencing and sequences were allocated to operational taxonomic units (OTUs) by two different approaches. Here, we use only the data classified by the database-dependent strategy (DBDS) and from soil type loess loam (LL) for which eight additional samples from an external field under control condition were available.

Omitting OTUs with sequences in less than five of the soil samples, 585 OTUs were available for the analysis. The SAS data set `RU47_LL_with_external_samples.sas7bdat` in the Supplemental Material contains the relative frequencies of the OTUs multiplied with the factor 1000.

We first compare the groups using the test statistic $\sqrt{d^*}$. We choose city-block distances applied to the relative frequencies, which is proportional to the relative Bray-Curtis distance that is very popular in ecological studies (De Caceres et al., 2013). The comparison of RU47-treated samples and the internal control group yields $\sqrt{d^*} = 26.341$ with the upper limit of the confidence interval of 27.428. To give an ecological argument for a tolerance threshold, we utilize the external control samples (one of two choices in Antweiler et al., 2017). If we compare the internal and the external control sample in terms of $\sqrt{d^*}$, then the lower limit of a 95%-CI yields a value of 35.400. Using this lower limit as the tolerance threshold, we see that the upper confidence limit for RU-47-treated samples versus the internal control group is much lower, indicating equivalence ($p = 3 \cdot 10^{-10}$).

In order to find more proper comparisons with competing methods that are more restrictive in their choice of distance measure, we also use the relative test statistic d^{**} with the threshold $\theta = 0.25$ as in the simulation studies before. Now the analysis is based on squared Euclidean distances, and the data are log-transformed ($\ln(r \cdot 1000 + 1)$, $r =$ relative frequency) for a better approximation to the normal distribution. The comparison of the RU47-treated samples with the (internal) controls yields the test statistic $d^{**} = -0.0120$ with a p -value of 0.0007 in the test against the threshold 0.25 and the upper limit of the confidence interval of 0.1050. Thus, the test for equivalence is highly significant again.

Using the proposal of Chervoneva et al. (2007) as first competing method, with the control group as reference group and the adapted tolerance threshold of 0.5 according to their notation, does not give a significant results. We embedded their proposals to compute the upper limit of the confidence intervals for fixed test level α in a nonlinear optimization procedure (NLPHQN procedure in SAS/IML) to find the test level where the confidence level is just on the border of the

tolerance region. This way, we obtained the p -values 0.6171 for the normal approximation and 0.6141 for the chi-squared approximation. Both are far from significance presumably mainly due to the strongly conservative behavior for large number of variables (cf. Section 7 and Figure S1) and the fact that this method uses only the reference group for the calculation of within-group distances.

As a second competitor, we consider univariate tests for all variables combined with the intersection-union principle that is, of course, a much stricter criterion. In the special case of only one included variable, equal sample sizes $n_1 = n_2 = n_0$ in both groups and the squared Euclidean distance as measure, a simple algebraic calculation yields $d^{**} = \frac{1}{2} \left(\frac{\bar{y}_1 - \bar{y}_2}{s} \right)^2 - \frac{1}{n_0}$, where \bar{y}_1 and \bar{y}_2 are the two sample means and s is the pooled standard deviation. Thus, a test for d^{**} with the tolerance threshold θ is comparable to a test of $\left(\frac{\bar{y}_1 - \bar{y}_2}{s} \right)^2$ with the threshold $2\left(\theta + \frac{1}{n_0}\right)$. Then, at the border of the tolerance region, the test statistic $F_0 = \left(\frac{\bar{y}_1 - \bar{y}_2}{s} \right)^2 \frac{n_1 n_2}{n_1 + n_2}$ follows a noncentral F -distribution with degrees of freedom $df_1 = 1$, $df_2 = n_1 + n_2 - 2$ and noncentrality parameter $\lambda = 2\left(\theta + \frac{1}{n_0}\right) \frac{n_1 n_2}{n_1 + n_2}$. The p -value follows as $P(F(df_1, df_2, \lambda) \leq F_0)$. From the 585 variables, 191 have p -values below 0.05, 259 below 0.10 and 462 below 0.25. The maximum p -value was 0.933. As that is also the p -value of the intersection-union test, this much stricter approach to equivalence testing does not yield significance here. However, in high-dimensional data with partly limited knowledge on the importance of single variables as in the present situation, the summarizing multivariate approach as proposed here seems appropriate and insistence on variable-wise equivalence as in the intersection-union test would be unreasonable.

The SAS code for the analyses is contained in the Supplemental Material.

9 | DISCUSSION

In this paper, we consider a class of tests that is based on pairwise multivariate distance measures between sample vectors. Such tests are in widespread use as tests for differences, particularly in the field of ecology. They can as well be used in other fields where a group of variables shall be investigated jointly. The choice of the distance measure determines the degree to which single or a few variables influence the outcome of the test and allows a flexible use in different fields of application.

The modifications for asymmetric distance measures presented here facilitate directed tests. In every application, one has to check if the modified measure still reflects the research question. Particularly, they are not suitable for relative frequencies because of the inherent indirect relationships between the variables. Both versions for directed or undirected differences can be implemented as permutation tests.

We also present extensions for multivariate equivalence tests or tests for noninferiority. Instead of checking equivalence for each variable separately, we make the decision on the basis of the multivariate distance measures that is a weakened claim but might be appropriate particularly in high-dimensional data. Since permutation techniques are difficult to implement and interpret for multivariate tests of entire parameter regions under the null hypothesis, we propose a leave-one-out technique that estimates the variance of the test statistic and provides a confidence interval for the expectation of the test statistic that is used to decide about equivalence or non-inferiority.

The permutation tests are distribution-free, but the confidence interval via leave-one-out variance estimate requires an approximate normal distribution of the test statistic in the considered populations. In the simulation experiments, it was difficult to find a suitable transformation for the difference of the means of within-group and between-group distances. A simple root transformation did quite well for the mean of between-group differences (as graphically checked by histogram plots over the simulation runs). The ratio of between-group and within-group distances by the within-group distances also yielded a sufficient approximation. An attempt to overcome the restriction of approximate normality by additional bootstrap-loops within the leave-one-out loop was not successful and has been omitted here.

The two selected test procedures showed robustness with respect to mild deviations from normality of the raw data. In severely skewed data, the robustness can be checked in adapted simulation studies. Furthermore, one has to check then if the chosen distance measure reflects the intentions for the aggregation over the variables. Basically, the distance measures considered here are based on the idea of similarly scaled linear variables. Generally in many situations, a transformation of the original data may be useful. If, for example, a new drug shall be compared to the standard drug for a disease with respect to adverse events in a noninferiority test on the basis of counts in a large list of different adverse effects—some of them occurring more often, others only rarely—then the log transformation $\ln(x + 1)$ might be useful. This would give a better approximation of normality and also a more convenient weighting of small and large values.

The ratio-based test statistic has the advantage that a standardized tolerance threshold could be defined for it. Otherwise, additional experimental extensions or experience from former research is required for the tolerance threshold definition.

In analyses of the composition of microbial species in agricultural experiments, Pearson correlation coefficients r are often used as similarity measures between two sample elements instead of distance measures. These analyses are based on counts of the species. However, for reasons of technological and financial limitations in the sampling and the sample analysis techniques, the total count is not biologically or statistically meaningful in these experiments. Only relative counts (percentages) are meaningful, which is automatically respected in the calculation of correlation coefficients. In the framework presented here, the correlation coefficient r could be handled by using $(1 - r^2)$ as the distance measure. Alternatively, similarity measures such as r could be used in the permutation tests of Section 3 or the leave-one-out versions for tests of equivalence in Section 5. One has just to take in account the reversed orientation (similar sample elements have large correlations but small distances). The derivation of one-sided tests (Sections 4 and 6) based on Pearson correlations, however, seems less straightforward to us.

As already mentioned in Antweiler et al. (2017), it is possible to include several criteria in tests for equivalence or noninferiority. For example, one could claim that two treatment groups are considered equivalent only if the differences expressed by the squared Euclidean distance (targeting a smoothing over several variables) and additionally expressed by the maximum absolute distance (targeting the influence of single variables) are both below corresponding thresholds. It may also be of interest to address additionally the influence of some variable of particular interest. As in these cases, equivalence or noninferiority is only accepted if all partial criteria are fulfilled (intersection-union principle), each of these partial tests can be done at the unadjusted error level α . However, the power analysis has to be adjusted for the additional criteria.

In this paper, only the simple case of two groups has been considered. The tests for difference can be easily extended to multi-factorial designs including covariables. Anderson (2001) gave an overview in the nonparametric setup. Recent developments use permutations of residuals (Pauly et al., 2015). In the classical multivariate normal situation, permutation tests can be replaced by rotation tests yielding exact parametric tests for small samples (Kropf & Adolf, 2009). Both versions might be applicable to one-sided tests for difference, equivalence, or noninferiority based on distances between sample elements. These generalizations require further research.

As discussed above, it is difficult to compare the power of the proposed multivariate tests with that of univariate tests combined in a multiple test procedure because the targets differ. This concerns the strictness with which weak effects in the wrong direction in one-sided tests are tolerated and the definition of tolerance thresholds for tests of equivalence or noninferiority. These caveats have to be kept in mind in applications.


ACKNOWLEDGMENTS

This work was funded by a grant of the German Federal Ministry of Education and Research (BMBF, grant 05M10NMA). We are grateful to the research teams headed by Prof. Dr. Ralf Paschke (formerly III. Medical Department, University of Leipzig, Germany) and Prof. Dr. Kornelia Smalla (Institute for Epidemiology and Pathogen Diagnostics, Julius Kühn-Institut, Federal Research Centre for Cultivated Plants, Braunschweig, Germany) for the permission to use their data as examples in the present paper. We would also like to thank an associate editor and two anonymous reviewers whose thoughtful comments helped to substantially improve an earlier version of the paper.

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

OPEN RESEARCH BADGES

 This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the [Supporting Information](#) section.

This article has earned an open data badge “**Reproducible Research**” for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially due to their computational complexity.

ORCID

Ekkehard Glimm  <https://orcid.org/0000-0003-3624-961X>

REFERENCES

- Anderson, M. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26, 32–46.
- Antweiler, K., Schreiter, S., Keilwagen, J., Baldrian, P., Kropf, S., Smalla, K., Grosch, R., & Heuer, H. (2017). Statistical test for tolerability of effects of an antifungal biocontrol strain on fungal communities in three arable soils. *Microbial Biotechnology*, 10(2), 434–449. <https://doi.org/10.1111/1751-7915.12595>
- Bathke, A. C., Harrar, S. W., & Ahmad, M. R. (2009). Some contributions to the analysis of multivariate data. *Biometrical Journal*, 51, 285–303. <https://doi.org/10.1002/bimj.200800196>
- Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems. I. Effect of inequality of variance in the one-way classification. *Annals of Mathematical Statistics*, 25, 290–302. <https://doi.org/10.1214/aoms/117728786>
- De Caceres, M., Legendre, P., & He, F. (2013). Dissimilarity measurements and the size structure of ecological communities. *Methods in Ecology and Evolution*, 4(12), 1167–1177. <https://doi.org/10.1111/2041-210X.12116>
- Chervoneva, I., Hyslop, T., & Hauck, W. W. (2007). A multivariate test for population bioequivalence. *Statistics in Medicine*, 26, 1208–1223. <https://doi.org/10.1002/sim.2605>
- Dempster, A. P. (1958). A high dimensional two sample significance test. *Annals of Mathematical Statistics*, 29, 995–1010. <https://doi.org/10.1214/aoms/1177706437>
- Dempster, A. P. (1960). A significance test for the separation of two highly multivariate small samples. *Biometrics*, 16, 41–50. <https://doi.org/10.2307/2527954>
- Ding, G.-C., Smalla, K., Heuer, H., & Kropf, S. (2012). A new proposal for a principal component-based test for high-dimensional data applied to the analysis of PhyloChip data. *Biometrical Journal*, 54, 94–107. <https://doi.org/10.1002/bimj.201000164>
- Efron, B., & Stein, C. (1981). The Jackknife estimate of variance. *The Annals of Statistics*, 9, 586–596. <https://doi.org/10.1214/aos/1176345462>
- Eszlinger, M., Krohn, K., & Paschke, R., (2001). cDNA expression array analysis suggests a lower expression of signal transduction proteins and receptors in cold and hot thyroid nodules. *The Journal of Clinical Endocrinology and Metabolism*, 86, 4834–4842. <https://doi.org/10.1210/jcem.86.10.7933>
- Glimm, E., & Läuter, J. (2010). Directional multivariate tests rejecting null and negative effects in all variables. *Biometrical Journal*, 52, 757–770. <https://doi.org/10.1002/bimj.200900254>
- Hothorn, L. A., & Oberdoerfer, R. (2006). Statistical analysis used in the nutritional assessment of novel food using the proof of safety. *Regulatory Toxicology and Pharmacology*, 44, 125–135. <https://doi.org/10.1016/j.yrtph.2005.10.001>
- Jurečková, J., & Kalina, J. (2012). Nonparametric multivariate rank tests and their unbiasedness. *Bernoulli*, 18, 229–251. <https://doi.org/10.3150/10-BEJ326>
- Kropf, S., Heuer, H., Grüning, M., & Smalla, K. (2004). Significance test for comparing complex microbial community fingerprints using pairwise similarity measures. *Journal of Microbiological Methods*, 57, 187–195. <https://doi.org/10.1016/j.mimet.2004.01.002>
- Kropf, S., Lux, A., Eszlinger, M., Heuer, H., & Smalla, K. (2007). Comparison of independent samples of high-dimensional data by pairwise distance measures. *Biometrical Journal*, 49, 230–241. <https://doi.org/10.1002/bimj.200510262>
- Kropf, S., & Adolf, D. (2009). Rotation test with pairwise distance measures of sample vectors in a GLM. *Journal of Statistical Planning and Inference*, 11, 3857–3864. <https://doi.org/10.1016/j.jspi.2009.05.024>
- Langsrud, Ø. (2005). Rotation tests. *Statistics and Computing*, 15, 53–60. <https://doi.org/10.1007/s11222-005-4789-5>
- Läuter, J. (1996). Exact *t* and *F* tests for analyzing studies with multiple endpoints. *Biometrics*, 52, 964–970. <https://doi.org/10.2307/2533057>
- Läuter, J., Glimm, E., & Kropf, S. (1996). New multivariate tests for data with an inherent structure. *Biometrical Journal*, 38, 5–22. Erratum: *Biometrical Journal* 40, 1015. <https://doi.org/10.1002/bimj.4710380102>
- Läuter, J., Glimm, E., & Kropf, S. (1998). Multivariate tests based on left-spherically distributed linear scores. *Annals of Statistics*, 26, 1972–1988. Correction: *Annals of Statistics* 27 (1999), 1441.
- Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27, 209–220.
- Marozzi, M. (2015). Multivariate multidistance tests for high-dimensional low sample size case-control studies. *Statistics in Medicine*, 34, 1511–1526. <https://doi.org/10.1002/sim.6418>
- Marozzi, M., Mukherjee, A., & Kalina, J. (2020). Interpoint distance tests for high-dimensional comparison studies. *Journal of Applied Statistics*, 47, 653–665. <https://doi.org/10.1080/02664763.2019.1649374>
- O'Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics*, 40, 1079–1087. <https://doi.org/10.2307/2531158>
- Pauly, M., Asendorf, T., & Konietschke, F. (2016). Permutation-based inference for the AUC: A unified approach for continuous and discontinuous data. *Biometrical Journal*, 58, 1319–1337. <https://doi.org/10.1002/bimj.201500105>
- Pauly, M., Brunner, E., & Konietschke, F. (2015). Asymptotic permutation tests in general factorial designs. *Journal of the Royal Statistical Society /B*, 77(2), 461–473. <https://doi.org/10.1111/rssb.12073>
- Pesarin, F., & Salmaso, L. (2010). *Permutation tests for complex data*. John Wiley & Sons.
- Sen, P. K. (1985). On permutational central limit theorems. In: S. Kotz, N.L. Johnson, & C.B. Read (Eds.), *Encyclopedia of Statistical Sciences* 6, John Wiley & Sons, New York, 683–687.
- Srivastava, M. S., & von Rosen, D. (2004). MANOVA with singular variance matrix. *Acta et Commentationes Universitatis Tartuensis de Mathematica*, 8, 253–269.
- Srivastava, M. S., & Fujikoshi, Y. (2006). Multivariate analysis of variance with fewer observations than the dimension. *Journal of Multivariate Analysis*, 97, 1927–1940. <https://doi.org/10.1016/j.jmva.2005.08.010>

- Tang, D.-I., Geller, N. L., & Pocock, S. J. (1993). On the design and analysis of randomized clinical trials with multiple endpoints. *Biometrics*, 49, 23–30. <https://doi.org/10.2307/2532599>
- Westfall, P. H., & Young, S. S. (1993). *Resampling-Based Multiple Testing*. John Wiley & Sons.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Kropf, S., Antweiler, K., & Glimm, E. (2022). Use of multivariate distance measures for high-dimensional data in tests for difference, superiority, equivalence and non-inferiority. *Biometrical Journal*, 64, 577–597. <https://doi.org/10.1002/bimj.202000367>