Robust nonparametric hypothesis tests for differences in the covariance structure of functional data

Kelly RAMSAY¹ and Shoja’eddin CHENOURI²

¹Department of Mathematics and Statistics, York University, 4700 Keele Street, Toronto, Ontario, Canada
²Department of Statistics and Actuarial Science, University of Waterloo, 200 University Ave W., Waterloo, Ontario, Canada

Key words and phrases: Covariance operator; functional data; hypothesis testing; nonparametric; robust.

MSC 2020: Primary 62G10; Secondary 62G35.

Abstract: We develop a group of robust, nonparametric hypothesis tests that detect differences between the covariance operators of several populations of functional data. These tests, called functional Kruskal–Wallis tests for covariance, or FKWC tests, are based on functional data depth ranks. FKWC tests work well even when the data are heavy-tailed, which is shown both in simulation and theory. FKWC tests offer several other benefits: they have a simple asymptotic distribution under the null hypothesis, they are computationally cheap, and they possess transformation-invariance properties. We show that under general alternative hypotheses, these tests are consistent under mild, nonparametric assumptions. As a result, we introduce a new functional depth function called $L^2$-root depth that works well for the purposes of detecting differences in magnitude between covariance kernels. We present an analysis of the FKWC test based on $L^2$-root depth under local alternatives. Through simulations, when the true covariance kernels have an infinite number of positive eigenvalues, we show that these tests have higher power than their competitors while maintaining their nominal size. We also provide a method for computing sample size and performing multiple comparisons.

Résumé: Les auteurs de cet article développent un ensemble de tests d’hypothèses robustes et non paramétriques pour détecter les différences entre les opérateurs de covariance de plusieurs populations de données fonctionnelles. Ces tests, appelés tests de Kruskal–Wallis fonctionnels pour la covariance ou tests FKWC, sont basés sur les classements de profondeur de données fonctionnelles. Les tests FKWC sont efficaces même avec des données à queues lourdes, comme le démontrent à la fois les simulations et la théorie. Ils offrent également plusieurs autres avantages : ils ont une distribution asymptotique simple sous l’hypothèse nulle, ils sont peu coûteux en termes de calcul et ils possèdent des propriétés d’invariance par transformation. Les auteurs montrent que dans le cadre d’hypothèses alternatives générales, ces tests sont convergents sous de faibles hypothèses non paramétriques. Suite à quoi, ils introduisent une nouvelle fonction de profondeur fonctionnelle appelée profondeur racine-$L^2$ qui convient parfaitement à l’identification des différences de grandeur entre les noyaux de covariance. Ensuite, ils présentent une analyse du test FKWC basée sur la profondeur de racine-$L^2$ sous des contre-hypothèses locales. Enfin, grâce à des simulations, ils montrent que lorsque les noyaux de covariance réels ont un nombre infini de valeurs propres positives, ces tests ont une puissance supérieure à celle de leurs concurrents tout en maintenant leur seuil nominal. Ils fournissent également une méthode pour le calcul de la taille d’échantillon et la réalisation de comparaisons multiples.
1. INTRODUCTION

Data such that each observation is a smooth curve, called functional data, are increasingly common in a variety of fields. For example, medical images (Aston, Pigoli & Tavakoli, 2017; López-Pintado & Wrobel, 2017), intraday financial asset returns (Cerovecki et al., 2019), and environmental “omics” data (Piña et al., 2018) can all be interpreted as functional data. As such, many functional analogues of univariate and multivariate statistical tools are needed. One such tool is the notion of common variance in the functional context, namely a common covariance operator or covariance kernel. In this article, we introduce new nonparametric functional $k$-sample tests for covariance structure equality. We call them functional Kruskal–Wallis tests for covariance structure or, for short, FKWC tests.

The KW nonparametric ANOVA test has been widely used to test for distributional differences between multiple independent univariate populations (Kruskal, 1952; Gastwirth, 1965). The test statistic is the traditional ANOVA test statistic applied to the combined sample ranks. Therefore, the test statistic will be large when a distributional difference implies a difference in the means of the combined sample ranks. Another benefit of this test is that it is very robust against extreme observations since the ranks are always bounded by the total sample size. Therefore, to use the KW nonparametric ANOVA test in the context of covariance kernel testing with multiple samples, we must reduce the data to ranks such that the distributions of these ranks differ when there are differences in the covariance kernel between samples. To do this, we use ranks based on functional data depth. To contextualize our procedure, we review existing related works.

Panaretos, Kraus & Maddocks (2010) were the first to discuss comparing the covariance structures of two populations of functional data. They present a two-sample test based on the Hilbert–Schmidt norm for integral operators and restrict their attention to that of Gaussian processes. Fremdt et al. (2013), Jarušková (2013), and Zhang & Shao (2015) extended these results to the non-Gaussian setting, the change-point setting, and the setting with dependent data, respectively. Gaines, Kaphle & Ruymgaart (2011) later proposed a test for the equality of two covariance operators based on univariate likelihood ratios and Roy’s union–intersection principle (Roy, 1953).

Up until this point, existing tests were based on the Hilbert–Schmidt metric. Pigoli et al. (2014) presented a discussion of distances between covariance operators, including criticisms of using finite-dimensional distances on functional data. They argued that the Hilbert–Schmidt metric ignores the geometry of the space of covariance kernels and therefore is not an appropriate distance. As a result, they introduced a two-sample permutation procedure, which Cabassi et al. (2017) later extended to the multisample case. In the same vein of resampling, Paparoditis & Sapatinas (2016) proposed a $k$-sample bootstrap test that can detect differences in mean and covariance structure simultaneously.

Guo & Zhang (2016) further studied the multisample test first proposed by Zhang (2013). When the data come from a Gaussian process, under the null hypothesis, the test statistic corresponding to this test is a $\chi^2$-type mixture. This distribution must be approximated in practice. Guo & Zhang (2016) also provided a random permutation method to be used in the case of small samples or non-Gaussian data. Guo, Zhou & Zhang (2018) developed a $k$-sample test inspired by functional ANOVA. One feature of this test is that it does not require any form of dimension reduction. Further, their method is scale-invariant in the sense that rescaling the data at any time $t$ does not affect the test statistic. Similar to the results of Guo & Zhang (2016), the distribution of the test statistic under the null hypothesis must be estimated. The estimation of the critical values relies on parameters that are estimated from the data, which may pose problems if the data are contaminated.

Boente, Rodriguez & Sued (2018) studied a new type of bootstrapping method to calibrate critical values for the problem of testing for covariance kernel differences. They focused on norms between covariance operators. The resulting distribution under the null hypothesis is
based on the eigenvalues of fourth moment operators, which must be estimated. They suggested bootstrapping the eigenvalues of fourth moment operators. This can be problematic if the data are heavy-tailed or contaminated.

Kashlak, Aston & Nickl (2019) provided an analysis of covariance operators based on a concentration inequality that includes a $k$-sample test and a classifier. They used concentration results to develop confidence sets based on $p$-Schatten norms. They then used “tuned” confidence sets to define rejection regions for $k$-sample tests. This test tends to underestimate the confidence level when the data are heavy-tailed.

Some other related works include the following. López-Pintado & Wrobel (2017) used a version of band depth defined on images to test for a difference in dispersion between two sets of images. Their measure of dispersion ignores differences in shape or “wiggyness” between the two samples. Sharipov & Wendler (2020) extended the bootstrapping procedures of Paparoditis & Sapatinas (2016) to change-point problems and dependent data. Flores, Lillo & Romo (2018) presented a test for the homogeneity of two distributions based on depth measures. They explicitly stated that their work was not focused on means or covariance operators. They provided four test statistics based on the deepest functions or absolute values of differences in the depth distributions.

Our FKWC tests have several advantages over these other methods. First, FKWC tests are very robust in the sense that they perform well in the presence of extreme observations and heavy-tailed distributions, a feature that has not often been discussed in other works. FKWC tests are based on rank statistics generated via functional data depth measures. Functional data depth measures are, among other things, used for outlier detection and trimmed means; data depth measures are designed to produce robust inference procedures. A test statistic based on the ranks of data depth measures would then inherit the robustness properties of both the depth measure and rank statistics. We demonstrate this robustness via simulation in Section 5.

Aside from being robust, many functional data depth measures are invariant under certain transformations of the data (Gijbels & Nagy, 2017). If the functional observations are all scaled by an arbitrary measurable function, we would like the test statistic to remain unchanged. Guo, Zhou & Zhang (2018) point out that many existing tests for equal covariance structures are not invariant under this type of transformation, e.g., those of Panaretos, Kraus & Maddocks (2010) and Guo & Zhang (2016). In contrast, many data depth measures remain unchanged if the data are scaled by an arbitrary function. Such invariance properties are then inherited by the FKWC test statistic, provided that derivatives are not included in the calculation of depth; see Section 3. If derivatives are included, then the FKWC tests satisfy a weaker form of transformation invariance.

Furthermore, data depth measures leverage existing consistency results (Nagy & Ferraty, 2019) and permit an asymptotic analysis of the FKWC tests under both the null and alternative hypotheses. We show that under the null hypothesis, the test statistic is a $\chi^2$ random variable. This is a particularly nice feature as it circumvents the need to estimate the distribution of the test statistic under the null hypothesis using the data, which many other tests require, e.g., Pigoli et al. (2014), Paparoditis & Sapatinas (2016), Cabassi et al. (2017), Boente, Rodriguez & Sued (2018), Guo, Zhou & Zhang (2018), and Kashlak, Aston & Nickl (2019). Using data to estimate the distribution of the test statistic under the null hypothesis can be complicated when the data are contaminated. This can also be computationally expensive if resampling methods are used.

Not only is there no need to use data to estimate the distribution of the FKWC test statistic under the null hypothesis but there is also no need to estimate the sample covariance operators in the computation of the FKWC test statistic. This fact implies that one does not need to reduce the dimension of the data via truncated basis expansions of the covariance kernels, as is needed by the methods of Fremdt et al. (2013), Pigoli et al. (2014), Paparoditis & Sapatinas (2016), and Boente, Rodriguez & Sued (2018). An additional byproduct of avoiding estimation of the covariance operators is that we do not require finite fourth moments or any assumptions related
to fourth moments for our theoretical analysis. Such assumptions are required for many other tests. For example, Panaretos, Kraus & Maddocks (2010), Gaines, Kaphle & Ruymgaart (2011), Fremdt et al. (2013), Paparoditis & Sapatinas (2016), Boente, Rodriguez & Sued (2018) all require some type of fourth-moment assumption on the data-generating process.

In terms of the alternative hypothesis, we show that under some mild conditions, the FKWC tests are consistent under a wide class of alternatives. We also provide a method for estimating power and sample size under general alternatives. Some recent works have explored various local alternatives for this testing problem (Gaines, Kaphle & Ruymgaart, 2011; Guo & Zhang, 2016; Boente, Rodriguez & Sued, 2018; Guo, Zhou & Zhang, 2018). We also provide a class of local alternatives under which a particular FKWC test is consistent. This FKWC test is based on a new depth measure, $L^2$-root depth, for which we prove several elementary properties. This depth measure has a particular interpretation in this testing problem that provides a basis for this depth measure’s development.

The rest of this article is organized as follows. Section 2 covers the methodology of the hypothesis tests, including the data model and the intuition behind the test statistic. Section 3 gives a brief overview of functional data depths and their benefits when used in a hypothesis-testing context. Section 4 presents asymptotic results on the behaviour of the test statistic under both the null and alternative hypotheses. Section 5 presents a simulation study in which we compare the FKWC tests against some competing tests, including those of Boente, Rodriguez & Sued (2018) and Guo, Zhou & Zhang (2018). Section 6 shows an application of the FKWC tests to two data sets: intraday financial asset return curves and log periodograms of five groups of recorded syllables.

2. MODEL AND METHODOLOGY

Suppose that we have observed $J$ independent random samples and that for each sample $j \in \{1, \ldots, J\}$ we have the functional observations $X_{j1}, \ldots, X_{jN_j}$. The combined sample size is then $N = N_1 + \cdots + N_J$, where we assume that $N_j/N \to \theta_j$ as $N \to \infty$. We also assume that $E[X_{ji}] = 0$, where $0$ is the zero function. Owing to the translation invariance of depth functions, this assumption is equivalent to that of each observation having the same mean. By “functional observations”, we mean that each observation is a mean-square continuous stochastic process that lies in $\mathcal{L}^2([0, 1], \mathcal{B}, \mu)$. Here, $\mu$ is the Lebesgue measure and $\mathcal{B}$ is the collection of Borel sets of $[0, 1]$. See the Appendix for more details. In the future, we may write $\mathcal{L}^2$ for short. Our methods extend to higher dimensional domains and ranges, i.e., to $X_{ji} : [0, 1]^d \to \mathbb{R}^p$ for which functional data depths are defined, but we restrict our study to the setting with a univariate domain and range.

The goal is to construct a test statistic for the following hypothesis test:

$$
H_0 : \mathcal{K}_1 = \cdots = \mathcal{K}_J \quad \text{vs.} \quad H_1 : \mathcal{K}_j \neq \mathcal{K}_k \text{ for some } j \neq k.
$$

Here, $\mathcal{K}_j$ refers to the covariance kernel of group $j$. The assumptions on the random variables $X_{ji}$ imply that this is equivalent to the hypothesis test

$$
H_0 : \mathcal{H}_1 = \cdots = \mathcal{H}_J \quad \text{vs.} \quad H_1 : \mathcal{H}_j \neq \mathcal{H}_k \text{ for some } k \neq j,
$$

where $\mathcal{H}_j$ is the covariance operator of group $j$. 

The Canadian Journal of Statistics / La revue canadienne de statistique
DOI: 10.1002/cjs.11767
Typically, this hypothesis test involves estimating $\mathcal{K}_j$. In order for estimates of $\mathcal{K}_j$ to converge weakly, it is typically required that $\mathbb{E} \left[ \left\| X_{ji} \right\|^4 \right] < \infty$; in some cases it is not desirable to make this assumption. The estimation of $\mathcal{K}_j$ is also high dimensional and can be computationally intensive if repeated a number of times, such as in a bootstrap procedure. We take a different approach; we do not aim to estimate $\mathcal{K}_j$. Instead, the idea is to reduce each observation to a one-dimensional rank via a data-driven ranking function. The ranking function is designed such that differences in the samples’ mean ranks are implied by differences in the underlying covariance kernels. We can then use the classical KW test statistic (Kruskal, 1952) and perform a rank test. Specifically, the proposed test statistic is

$$\hat{W}_N = \frac{12}{N(N+1)} \sum_{j=1}^J N_j \left( \bar{R}_j - \frac{N+1}{2} \right)^2. \quad (1)$$

Here, $\bar{R}_j$ is the mean rank of the observations in group $j$, where the ranking mechanism will be explained in the next section. This test statistic also gives, for each sample $j$, a measure of how much its covariance kernel differs from the average sample kernel via $\left( \bar{R}_j - \frac{N+1}{2} \right)^2$.

Alternatively, we can perform FKWC multiple comparisons; see Section 6.2. We can further modify this test statistic using the percentile modification of Gastwirth (1965), who presented a more powerful version of the univariate Wilcoxon rank-sum test. The percentile modification was later extended to multivariate depth-based rank tests (Chenouri, Small & Farrar, 2011). We further extend these methods to the functional setting. The percentile modification is predicated on the fact that it is actually the extreme rank values that allow us to detect differences between samples. In the context of detecting differences in covariance structure, the idea is to remove the middle portion of the data. Therefore, we only use the outlying data or, equivalently, the low depth-based ranks. To this end, let $r \in (0, 1)$ and let $N' = \lfloor rN \rfloor$, where $\lfloor \cdot \rfloor$ refers to the integer-value operator. Let $\delta_j(s) = 1$ if the observation which has rank equal to $s$ is in group $j$ and let $\delta_j(s) = 0$ otherwise. Define the percentile-modified test statistic as

$$\hat{M}_{N,r} = \sum_{j=1}^N \left( 1 - \frac{N_j}{N} \right) K_j \quad \text{with} \quad K_j = \frac{1}{\sigma_j^2} \left( \frac{N'}{N} \right) \left( \sum_{s=1}^{N'} (N' - s + 1) \delta_j(s) - \varphi_j \right)^2, \quad (2)$$

where

$$\varphi_j = \frac{N_j N' (N' + 1)}{2N} \quad \text{and} \quad \sigma_j^2 = \frac{N_j (N - N_j) N' (N' + 1) \left\{ 2N (2N' + 1) - 3N' (N' + 1) \right\}}{12N^2(N-1)}.$$

For guidance on choosing $r$, one can run simulations or follow the recommendations in Gibbons (1973).

3. FUNCTIONAL DEPTH MEASURES

The methods we use for ranking data are based on a functional depth measure $D(\cdot; F)$. Functional data depth measures, extended from multivariate data depth measures (Zuo & Serfling, 2000), are versatile nonparametric tools used in the analysis of functional data. A functional data depth measure $D(\cdot; F) : \mathcal{F} \to \mathbb{R}$ assigns to each value in $x \in \mathcal{F}$ a real number that describes how
“central” $x$ is with respect to some measure $F$ over the Hilbert space $L^2$. Here, “central” is used very loosely; central can be in terms of location but also can be in terms of shape. Often we have that $F = F_N$, the empirical measure of the data, and that the depth values $D(x; F_N)$ describe centrality with respect to the sample. The measure $F_N$ is such that there is $1/N$ mass at each observation in the sample.

Sample ranks based on data depth measures can be calculated simply by ranking the (univariate) depth values of the observations, where depth is calculated with respect to $F_N$. In this article, we rank the observations with respect to $F_{s,N}$, which places equal weight on each element of the combined sample. For some $j \in \{1, \ldots, J\}$ and some $i \in \{1, \ldots, N_j\}$, define the sample-depth-based rank of $X_{ji}$ to be

$$\hat{R}_{ji} = \#\{(\ell, m) : D(X_{\ell m}; F_{s,N}) \leq D(X_{ji}; F_{s,N}), \ell \in \{1, \ldots, J\}, m \in \{1, \ldots, N_\ell\}\}.$$

This method of ranking is analogous to univariate centre-outward ranking in the sense that observations have a high rank when they are deep inside the data cloud.

The motivation for using depth measures as the ranking function follows from the fact that depth has already been shown to have good power for detecting scale changes in surface data (López-Pintado & Wrobel, 2017). Additionally, using depth ranks for such a purpose has been shown to work well in univariate and multivariate settings (Ansari & Bradley, 1960; Siegel & Tukey, 1960; Chenouri, Small & Farrar, 2011; Chenouri & Small, 2012). In fact, it appears that ranks based on depth measures capture second-order differences better in the functional setting than in the multivariate setting. For example, a sign change in a covariance parameter could not be detected by data depth ranks in the multivariate setting (Chenouri, Small & Farrar, 2011). This is not the case in the functional setting; some of our proposed FKWC tests can detect a difference of the type $\mathcal{H}_1 = \mathcal{U} \mathcal{H}_2 \mathcal{U}^*$ for some unitary operator $\mathcal{U}$, where $\mathcal{U}^*$ denotes the adjoint of $\mathcal{U}$. As discussed by Serfling & Wijesuriya (2017), functional depth measures are designed with the goal of describing functional data; this goal differs from that of describing multivariate data. Specifically, functional data depth measures are not required to be affine-invariant and are designed to account for both shape and scale features of the data. Differences between covariance kernels are often exhibited by changes in the shape or scale of the data, which are precisely the features captured by functional data depth functions. For example, Figure 1 shows two samples of 10 realizations of Gaussian processes and their derivatives. Each sample has the same mean but a different covariance kernel. Visually, the distinguishing factor between these two samples is the scale and shape of the curves and their derivatives. Notice that the difference is more pronounced in the derivatives.

There have been many definitions put forth for functional depth measures, especially recently. We restrict ourselves to depth measures that meet criteria suitable for the hypothesis testing problem. Namely we would like the depth measures to admit definitions on multivariate functional data. Multivariate functional data refers to observations whose range is $\mathbb{R}^p$ for some integer $p > 1$. For example, a functional observation and its derivative, $(X, X^{(1)})$, could be referred to as a multivariate functional observation with $p = 2$. Choosing depth measures that admit a multivariate functional definition allows for the use of derivatives, warping functions, or other functions in the depth computation. In this article, we use the derivatives in conjunction with the originally observed curves to compute depth values and show that this leads to better performance.

On top of having multivariate functional definitions, we would like the depth measure to be scale- and translation-invariant with respect to $F_N$. By scale-invariant, we mean that if the data are collectively scaled by some nonzero function $g$, the result of the test is unchanged. Clearly translation invariance is necessary. Many functional depth measures are invariant under the aforementioned transformations of the data. However, suppose that we have a set of univariate
FIGURE 1: Two samples of (a) Gaussian processes and (b) their derivatives. Each process has the same mean but a different covariance kernel. The samples have an exponential covariance kernel with $\alpha = 0.3$ in the first sample and $\alpha = 1$ in the second sample (see Section 5). Notice that the difference in covariance structure is exhibited by differences in the shape of the original curves and in the shape and scale of the derivatives.

functional observations but we take the data to be the observed function along with the observed first derivative $(X, X^{(1)})$. Scaling $X$ by $g$ results in a different transformation of the derivative and produces different observations, namely $(gX, gX^{(1)} + g^{(1)}X)$. If $g$ is a constant function, then a scaling of the data corresponds to a scaling of the derivative and all is well. Therefore, ranking methods that are based on an analysis of the function and its derivative are not generally invariant under scaling by a non-constant function. They are, however, invariant under scaling by a constant function. In other words, the FKWC tests based on the original curves are invariant under scaling by a non-constant function, whereas the FKWC tests based on function–derivative pairs are not invariant under scaling by a non-constant function. However, in order for the derivative to be substantially affected by arbitrary scaling, the derivative of the function $g$ must be large relative to the observed function, i.e., that the function $g$ is quite steep in some places. It is hard to see a reason to scale the data by steep functions in practice.

It is also desirable for depth measures to be uniformly consistent with a rate of $O(N^{-1/2})$ under some general conditions. Then for a large sample size $N$, the test will detect a difference in covariance structure with high probability. So far, this has only been shown by Nagy & Ferraty (2019) for the depth measure introduced by Fraiman and Muniz (2001). We extend their result to the random projection depth in Section 4. Lastly, we use functional data depth measures that have implementations in R or software that can be integrated into R. This last restriction allows us to readily provide simulation results for the tests as well as implementations that can be used in practice.

The first depth measure we introduce is a version of the integrated depth of Fraiman and Muniz (2001), which, in this case, is equivalent to a version of the multivariate functional halfspace depth (MFHD) (Slaets, 2011; Hubert et al., 2012; Claeskens et al., 2014). Let $F_t$ be the marginal distribution of $X(t) \in \mathbb{R}^p$ if $X \sim F$. We define MFHD as

$$\text{MFHD}(x; F) = \int_{[0,1]} \inf_{y \in \mathbb{R}^p, \|y\|=1} \Pr \{ X(t)^T y \leq x(t)^T y \} \, dt.$$ 

MFHD can be interpreted as the average of the point-wise halfspace depth values (Tukey, 1974).
We next consider modified band depth (MBD) (López-Pintado & Romo, 2009). For some \( x \in \mathbb{S} \) and a set of functional observations \( Y_1, \ldots, Y_k \), define
\[
B(x; Y_1, \ldots, Y_k) = \left\{ t : t \in [0, 1], \min_{r \in \{1, \ldots, k\}} Y_r(t) \leq x(t) \leq \max_{r \in \{1, \ldots, k\}} Y_r(t) \right\}
\]
as the set such that \( x \) is in the \( k \)-band delimited by \( Y_1, \ldots, Y_k \). If \( Y_1, \ldots, Y_k \) are independent and come from the distribution \( F \), we can define the MBD with the parameter \( K \) as
\[
\text{MBD}_K(x; F) = \sum_{k=2}^{K} E_F \left[ \mu \left( B(x; Y_1, \ldots, Y_k) \right) \right],
\]
where \( \mu \) is the standard Lebesgue measure. In this article, we choose \( K = 2 \) as this is the most computationally favourable choice of \( K \) and the MBD ordering is very stable under different choices of \( K \) (López-Pintado & Romo, 2009; Sun, Genton & Nychka, 2012). There are two multivariate functional extensions of this depth measure in Ieva & Paganoni (2013) and Lopez-Pintado et al. (2014). We choose the extension of Ieva & Paganoni (2013) because of its existing implementation in \( R \). The multivariate extension of MBD, as defined in Ieva & Paganoni (2013),
\[
\text{MMBD}_f(x; F) = \sum_{k=1}^{p} w_k \text{MBD}_f(x; F^k),
\]
where \( \sum_{k=1}^{p} w_k = 1 \), and \( F^k \) is the marginal distribution of the \( k \)th univariate functional argument in the vector of observations. We use \( \text{MBD}'_K(x; F) \) to refer to \( \text{MMBD}_K \) applied to the pair \((x, x^{(1)})\).

The last of the existing depth measures we introduce is random projection depth (RPD) (Cuevas, Febrero & Fraiman, 2007). The idea behind this depth measure is to choose \( M \) random directions \( u_m(t) \in S \), where \( S \) is the unit sphere in \( \mathbb{S}^2([0, 1], \mathbb{R}, \mu) \). Then for each direction \( u \), a separate depth value based on the projections onto \( u \) is computed, i.e., \( D(\langle x, u \rangle; F_u) \). Here, \( F_u \) is the CDF of the random variable \( \langle X, u \rangle \), where \( X \sim F \). These depths are then averaged to give a final depth value
\[
\text{RPD}_M(x; F) = \frac{1}{M} \sum_{m=1}^{M} D(\langle x, u_m \rangle; F_{u_m}) .
\]
In this article, we typically take \( D(\langle x, u \rangle; F_u) = F_u(x)(1 - F_u(x)) \) and use \( M = 20 \) projections. Cuevas, Febrero & Fraiman (2007) introduce a second version of RPD in which both the projection of the function and the projection of the function’s first derivative are calculated. This approach provides pairs of observations. They then apply a multivariate depth measure to the couples
\[
\text{RPD}'_M(x; F) = \frac{1}{M} \sum_{m=1}^{M} D\left( \left( \langle x, u_m \rangle, \langle x^{(1)}, u_m \rangle \right); F_{u_m, x^{(1)}} \right),
\]
where \( F_{u(x^{(1)})} \) is the bivariate distribution of \((\langle X, u \rangle, \langle X^{(1)}, u \rangle)\). For \( D(\cdot) \), we use the likelihood depth (Müller, 2005), as this is the default in the \texttt{fda} \texttt{.usc} \texttt{R} package. In this article, \( u_1, \ldots, u_M \) are Gaussian processes with exponential variogram \( \gamma(s, t) = \exp(-5|s - t|) \). These processes are then standardized to have unit norm.

Using the ranks of the norms of the observations is another natural approach to this problem. As such, we compare using the depth-based ranks against ranking the norms of the observations.
We can show that ranks derived from squared norms are a kind of depth-based rank. If we define depth as
\[
\text{LTR}(x; F) = \left(1 + E_F[\|x - X\|^2]^{1/2}\right)^{-1},
\]
then the depth-based ranks from this depth function are equivalent to ranking the squared norms, provided that the observations have mean equal to zero. We call this depth the \(L^2\)-root depth, or LTR depth for short. In the Appendix, we demonstrate that LTR depth actually measures depth. This definition of depth is similar to that of the \(L^2\) depths of Zuo & Serfling (2000), except that the square root is outside of the expectation in order to relate the depth values to the covariance operator. This LTR depth does not enjoy the same interpretation as the inverse average distance to a point drawn from \(F\) because, technically, the squared norm function is not a distance. However, the definition is of the same spirit as the \(L^p\) depths of Zuo & Serfling (2000). Framing norm-based ranks as depth-based ranks allows us to easily extend norm-based ranks to include derivative information. LTR depth can be extended to account for \(p\) derivatives as
\[
\text{LTR}_p(x; F) = \left(1 + \frac{1}{p} \sum_{k=0}^{p} \text{MAD}_k \ E_F[\|x^{(k)} - X^{(k)}\|^2]^{1/2}\right)^{-1},
\]
where \(\text{MAD}_k\) is the median absolute deviation with respect to the law of the norm of the \(k\)th derivative of \(X\), where \(X \sim F\).

4. THEORETICAL RESULTS

This section is devoted to characterizing the behaviour of the FKWC tests under the null and alternative hypotheses when the sample size is large. Proofs of the theorems presented in this section can be found in the Appendix. We also remind the reader that, throughout this article, including in all theorem statements, we assume that the observations have mean zero and that \(N_j/N \to \vartheta_j\) as \(N \to \infty\). We denote the probability measure over \(\mathcal{L}^2\), which describes the random behaviour of \(X_{ji}\), by \(F_j\). Let \(F^* = \sum_{j=1}^{J} \vartheta_j F_j\), be a mixture of probability measures over \(\mathcal{L}^2\). Alternatively, \(F^*\) can be interpreted in the stochastic process sense such that the finite-dimensional distributions of an element from the combined sample \(X: (X(t_1), \ldots, X(t_k))\) for \(k \in \mathbb{N}\), are mixtures of the \(J\) finite-dimensional distributions of each group, with weights \(\vartheta_1, \ldots, \vartheta_J\). These finite-dimensional distributions can be identified by \(F^*\) and \(F_j\). For a sequence of random variables \(Z_N\) and a distribution \(F\) over \(\mathbb{R}^d\), take \(Z_N \leadsto F\) to mean that \(Z_N\) converges in distribution to \(F\). The following theorem characterizes the behaviour of the FKWC tests under the null hypothesis.

**Theorem 1.** Suppose that for any \((j, i) \neq (k, \ell)\), \(\Pr(\widehat{R}_{ji} = \widehat{R}_{k\ell}) = 0\). Let the statistics \(\widehat{W}_N\) and \(\widehat{M}_N\) be as defined in Equations (1) and (2), respectively. Under the null hypothesis, as \(N \to \infty\), for any \(0 < r < 1\), \(\widehat{W}_N \leadsto \chi^2_{J-1}\) and \(\widehat{M}_{N,r} \leadsto \chi^2_{J-1}\).

Tied ranks can be randomly broken to meet the requirements of Theorem 1. The asymptotic behaviour of the test statistic under the null hypothesis is remarkably simple for such a complex testing problem. Therefore, the asymptotic critical values can easily be obtained independently of the data, which improves accuracy and computation time. The fact that the asymptotic critical values are independent of the data is important for robustness, as there is no need to assess the robustness of procedures used to approximate the asymptotic null distribution.

Under the alternative hypothesis, we must impose additional assumptions in order for the test to be consistent.
**Assumption 1.** For all $j$, $\Pr(D (X_j; F_s) \leq v)$ as a function of $v$, is a Lipschitz function.

**Assumption 2.** It holds that $\mathbb{E} \left[ \sup_{x \in \mathcal{B}} |D (x; F_{N, *} ) - D (x; F_*)| \right] = O(N^{-1/2})$.

Assumption 1 is generally satisfied when the finite-dimensional distributions corresponding to $F_*$ are continuous. Assumption 2 has been shown to be satisfied for MFHD; see Nagy & Ferraty (2019). We can extend the results of Nagy & Ferraty (2019) to RPD. Suppose that the unit vectors are drawn from some probability measure $\nu$ on the unit sphere $S$ and that $M_N = O(N)$. Let

$$\text{RPD}_\infty(x; F) = \int_S F_u(\langle x, u \rangle)(1 - F_u(\langle x, u \rangle))d\nu(u).$$

(3)

Then

$$\mathbb{E} \left[ \sup_{x \in \mathcal{B}} |\text{RPD}_{M_N} (x; F_N ) - \text{RPD}_\infty(x; F)| \right] = O(N^{-1/2}).$$

(4)

Note that the same analysis applies when $F_u(\langle x, u \rangle)(1 - F_u(\langle x, u \rangle))$ is replaced with $1/2 - |1 - F_u(\langle x, u \rangle)|$ in Equation (3). Assumption 2 need not be satisfied for LTR depth because the sample ranks are already based on LTR $(\cdot; F)$. Now, if there exists $k \in \{1, \ldots, J\}$ such that

$$\sum_{j=1}^J \theta_j \Pr(D (X_k; F_s) > D (X_j; F_*)) \neq \frac{1}{2},$$

(5)

then under Assumptions 1 and 2

$$\lim_{N \to \infty} \Pr(\hat{\mathcal{W}}_N > \delta) = 1$$

(6)

for any $\delta > 0$. This result shows that the set of alternative hypotheses that induce consistency are characterized completely by (5). First, for the FKWC tests to detect a change in covariance, we must have that a difference in the covariance operator between the groups implies that there is a difference in the location of the depth values between the groups. Note that we only need at least one pair of groups to differ in depth location: specifically, that the median of $D (X_{\ell 1}; F_s) - D (X_{j 1}; F_*)$ is nonzero for some $\ell \neq j$. The condition in (5) has an additional caveat which says that the differences between the groups do not perfectly “cancel” each other out. This caveat applies to all tests based on the KW statistic and is a minor technicality. This is not an issue if $J = 2$.

As previously mentioned, we must demonstrate that changes in the covariance kernels produce, on average, a location difference in the depth values. This can be argued qualitatively, seeing as changes in the covariance operator elicit changes in the shape or magnitude of the data (Liu & Singh, 2006). Given that depth measures rate the observation on how close it is in shape and magnitude to the combined sample, then it is intuitive that observations with a different covariance kernel will have different depth values.

Beginning with the two-sample case, we analyze the relationship between the covariance operator and the depth values for RPD. Let the sample rank of $X_{ji}$ based on the true distribution $F_*$ be defined as

$$R_{ji} = \# \left\{ (\ell', m) : D (X_{\ell'm}; F_s) \leq D (X_{ji}; F_*) , \ell' \in \{1, \ldots, J\}, m \in \{1, \ldots, N_j\} \right\} .$$

(6)
We define $\mathcal{W}_N$ as the test statistic based on these ranks, which are unknown except in the special case of ranks based on LTR depth. The next theorem gives the desired result for RPD.

**Theorem 2 (RPD).** If $X \sim F_\star$, let $F_{u,\star}$ be the distribution of $(X, u)$. Suppose that $u_1, \ldots, u_{MN}$ are drawn independently from a probability measure $\nu$ on the unit sphere $S$ in $\mathbb{R}^2$ and that $M_N = O(N)$. Assume that, for any $u$, $F_{u,\star}$ is three times differentiable and that the first three derivatives of $F_{u,\star}$ are bounded functions in $u$. Suppose that Assumption 1 holds for $\text{RPD}_\infty$, $J = 2$, $E \left[ ||X_{ji}||^3 \right] < \infty$ for all $X_{ji}$, and

$$E \left[ \text{RPD}_\infty(X_{11}; F_\star) - \text{RPD}_\infty(X_{21}; F_\star) \right] \neq 0$$

implies

$$\text{Med} \left( \text{RPD}_\infty(X_{11}; F_\star) - \text{RPD}_\infty(X_{21}; F_\star) \right) \neq 0.$$

Then the FKWC test based on $\hat{\mathcal{W}}_N$ using $\text{RPD}_{MN}$ with $D(z; F) = F(z)(1 - F(z))$ is consistent in the sense of Equation (6) under alternatives of the form

$$H_1 : \int_S h(F_{u,\star})(\mathcal{X}_1 u, u) d\nu(u) + R_1 \neq \int_S h(F_{u,\star})(\mathcal{X}_2 u, u) d\nu(u) + R_2,$$

where $h$ is a function of a univariate distribution function such that

$$h(F) = \frac{1}{2} f'(0) - (F(0)f'(0) - f^2(0)),$$

and the remainder term $R_j$ can be written as

$$R_j = \frac{1}{6} \int_S E \left[ f^{(3)}(X_{j1}, u) f^{(2)}(t)(1 - 2F_{u,\star}(t)) - 6f_{u,\star}(t)f^{(1)}_{u,\star}(t) \right] (X_{j1}, u - t)^3 dt d\nu(u).$$

We expect $R_j$ to be small because $E [X_{ji}] = 0$. In terms of $h(F_{u,\star})$, as $F_{u,\star}$ approaches symmetry, $h(F_{u,\star})$ approaches $f_{u,\star}^3(0)$, the squared height of the projected density at 0. Symmetry of the projected distributions implies a form of symmetry of $F_\star$. This is a natural definition of symmetry in the sense that it is analogous to halfspace symmetry in the multivariate setting (Zuo & Serfling, 2000). This discussion implies that the FKWC test paired with RPD will work better for $F_\star$ that are symmetric. Theorem 2 highlights an important difference between the functional data setting and the multivariate setting. Suppose that $\mathcal{X}_1$ is unitarily equivalent to $\mathcal{X}_2$, i.e., $\mathcal{X}_1 = \mathcal{U} \mathcal{X}_2 \mathcal{U}^T$, for some unitary operator $\mathcal{U}$. Assume that the samples are multivariate, i.e., $X_i \in \mathbb{R}^b$. The covariance operators can then be represented by the covariance matrices $\Sigma_1$ and $\Sigma_2$. Unitary equivalence in the multivariate setting corresponds to the case where $\Sigma_1 = \mathcal{U} \Sigma_2 \mathcal{U}^T$ for a rotation matrix $\mathcal{U}$. Let $\tilde{u} = u^T \mathcal{U}$, where it is clear that $||\tilde{u}|| = 1$. It is then easy to see that

$$\int_{S^{b-1}} h(F_{u,\star}) u^T \Sigma_1 u d\nu(u) = \int_{S^{b-1}} h(F_{u,\star}) \tilde{u}^T \Sigma_2 \tilde{u} d\nu(u) \approx \int_{S^{b-1}} h(F_{u,\star}) \tilde{u}^T \Sigma_2 \tilde{u} d\nu(u),$$

which is why we are unable to detect changes characterized by rotations of the data via data depth in the multivariate setting. This equivalence does not exist in the functional setting because the measure $\nu$ is not uniform on $S$. Therefore, it is not necessarily true that

$$\int_S h(F_{u,\star})(\mathcal{X}_1 u, u) d\nu(u) \approx \int_S h(F_{u,\star})(\mathcal{X}_2 u, u) d\nu(u).$$

DOI: 10.1002/cjs.11767
We verify this fact in Section 5: in finite-dimensional simulation scenarios 1–3, the covariance operators differ between the samples but are unitarily equivalent. We see that the FKWC test based on RPD can detect the difference.

The following is a consistency theorem for the FKWC test based on LTR depth ranks.

**Theorem 3 (LTR).** Suppose that Assumption 1 holds, \( J = 2 \), and that

\[
E \left[ \|X_{11}\|^2 - \|X_{21}\|^2 \right] \neq 0 \quad \text{implies} \quad \text{Med} \left( \|X_{21}\|^2 - \|X_{11}\|^2 \right) \neq 0. \tag{7}
\]

Then the test based on \( \mathcal{W}_N \) and using ranks based on LTR depth is consistent in the sense of Equation (6) under alternatives of the form \( H_1 : \|\mathcal{K}_1\|_{TR} \neq \|\mathcal{K}_2\|_{TR} \), where \( \|\cdot\|_{TR} \) refers to the trace norm.

The above theorem features \( \mathcal{W}_N \) and not \( \hat{\mathcal{W}}_N \). Recall that for an FKWC test, \( \mathcal{W}_N \) is the test statistic based on the ranks of the population depth values at the sample points. In the special case of the FKWC test based on LTR depth where the derivatives are not incorporated, \( \mathcal{W}_N \) is known. As such, we study \( \mathcal{W}_N \) in the case of LTR depth. Also note that \( \mathcal{K}_j \) are trace class since the observed processes are mean-square-continuous, which implies that the kernel is continuous. Theorem 3 shows that the FKWC test based on the ranks of the squared norms is consistent if the trace norm of the covariance operators differs. The alternative hypothesis \( \|\mathcal{K}_1\|_{TR} \neq \|\mathcal{K}_2\|_{TR} \) is equivalent to \( \sum_{k=1}^\infty \lambda_k \neq \sum_{k=1}^\infty \lambda_k' \), where \( \{\lambda_k\}_{k=1}^\infty \) and \( \{\lambda_k'\}_{k=1}^\infty \) are decreasing sequences of singular values resulting from the singular value decompositions of \( \mathcal{K}_1 \) and \( \mathcal{K}_2 \), respectively. Clearly, if the covariance operators are equal, then \( \|\mathcal{K}_1\|_{TR} = \|\mathcal{K}_2\|_{TR} \) and (7) is not satisfied.

It is useful to mention a few cases where (7) is surely satisfied. If \( \|X_{ji}\|^2 \) has a symmetric distribution for all \( j \), then (7) is satisfied. If \( E \left[ \|X_{21}\|^4 \right] < \infty \) and

\[
\|\mathcal{K}_1\|_{TR} - \|\mathcal{K}_2\|_{TR} \geq \text{var} \left( \|X_{11}\|^2 \right)^{1/2} + \text{var} \left( \|X_{21}\|^2 \right)^{1/2},
\]

then (7) is satisfied (Page & Murty, 1982). If the distribution of the squared norms is unimodal, Basu & DasGupta (1997) gives a sharper bound:

\[
\|\mathcal{K}_1\|_{TR} - \|\mathcal{K}_2\|_{TR} \geq \left( \frac{3}{5} \left( \text{var} \left( \|X_{11}\|^2 \right) + \text{var} \left( \|X_{21}\|^2 \right) \right) \right)^{1/2}.
\]

If \( X_{ji} \) are Gaussian processes \( GP(0, \mathcal{K}_j) \), then \( \|X_{11}\|^2 = \sum_{k=1}^\infty \lambda_k V_k \) and \( \|X_{21}\|^2 = \sum_{k=1}^\infty \lambda_k' V_k' \), where \( V_k, V_k' \) are independent and \( V_k \sim \chi^2_1 \). It follows that the random variables \( \|X_{11}\|^2 \) and \( \|X_{21}\|^2 \) are stochastically ordered, which implies that (7) is satisfied.

To give more insight into the behaviour of the test statistic, we provide some analysis under local alternatives. Suppose that Assumption 2 and (5) hold. Following Fan, Zhang & Zhang (2011), \( \mathcal{W}_N \) is approximately distributed as a non-central \( \chi^2 \) random variable, denoted \( \chi^2_{J-1}(\tau_N) \), with a non-centrality parameter

\[
\tau_N = \frac{12}{N(N + 1)} \sum_{j=1}^J N_j \left\{ \frac{N}{\sum_{k \neq j} \theta_k} \left( \frac{\text{Pr}(D(X_{ji}; F_1) \leq D(X_{k1}; F_0) - \frac{1}{2})}{\sum_{k \neq j} \theta_k} \right)^2 \right\}.
\]
For LTR depth, we are able to compute \( \Pr(D(X_{k1}; F_s) \leq D(X_{j1}; F_s)) \), which is the same quantity as \( \Pr(\|X_{k1}\| \leq \|X_{j1}\|) \). Therefore, one can compute the power and, consequently, sample sizes for any assumed \( F_s \) using

\[
\Pr\left( \sum_{m=1}^p \left[ \|X_{k1}^{(m)}\| - \|X_{j1}^{(m)}\| \right] \leq 0 \).
\]

This could of course be done by Monte Carlo simulation for complicated models.

**Theorem 4.** Let \( \delta_j \in \mathbb{R} \) for \( j \in \{1, \ldots, J\} \) and let \( G \) be a probability measure on the Borel sets of the real numbers with an associated continuously differentiable density \( g \). Suppose that, for all \( i, j, k \) and \( \ell \)

\[
\|X_{ji}\|^2 \overset{d}{=} \left[ \frac{\sqrt{N} + \delta_k}{\sqrt{N} + \delta_j} \right] \|X_{k\ell}\|^2 \sim G.
\]

Let \( \bar{\delta} = \sum_{j=1}^J \theta_j \delta_j \) and \( \tau = 12\left( \int_{\mathbb{R}} z^2 g(z)^2 d\mu \right)^2 \sum_{j=1}^J \theta_j \left( \delta_j - \bar{\delta} \right)^2 \). Under LTR depth ranks, as \( N \to \infty \), \( \mathcal{W}_N \to \chi^2_{J-1}(\tau) \).

Note that Equation (9) holds when \( \mathcal{X}_j = \mathcal{X}_0 \left[ 1 + N^{-1/2} \delta_j \right] \) and the \( \|X_{ji}\|^2 \)'s form a scale family. For example, if \( X_{ji} \sim \mathcal{GP}(0, \mathcal{X}_0 \left[ 1 + N^{-1/2} \delta_j \right]) \), then Theorem 4 is applicable.

Owing to the fact that Euclidean spaces are Hilbert spaces, the previous results provide some consequences for similar methods using depth-based ranks in the multivariate setting. For example, Chenouri, Mozaﬁfari & Rice (2020b) and Ramsay & Chenouri (2020) provide methodologies for detecting single and multiple changepoints in the covariance matrix of multivariate data based on depth ranks. Theorem 3 provides justification for assuming the hypothesis of Theorem 2 of Chenouri, Mozaﬁfari & Rice (2020b) as well as for Assumption 4 of Ramsay & Chenouri (2020) when the ranks are based on LTR depth. Note that the definition of spatial depth in Chenouri, Mozaﬁfari & Rice (2020b) provides ranks equivalent to those based on LTR depth. Theorem 3 then implies that the methods of Chenouri, Mozaﬁfari & Rice (2020b) and Ramsay & Chenouri (2020) can detect changes in the sum of the eigenvalues of the covariance matrix. Similarly, Theorem 2 provides justification for assuming the hypothesis of Theorem 2 of Chenouri, Mozaﬁfari & Rice (2020a) and Assumption 4 of Ramsay & Chenouri (2020) under the multivariate depth studied by Cuevas & Fraiman (2009). Liu & Singh (2006) provide a \( k \)-sample test for the covariance matrix of multivariate data. Theorems 2–4 give analogous results for this multivariate \( k \)-sample test.

5. SIMULATION RESULTS

5.1. Models and Settings

In this section we evaluate the finite-sample performance of the FKWC tests on both infinite-dimensional and finite-dimensional models. We compare the performance of the tests using the different depth functions discussed in Section 3 as well as the effect of the percentile modification discussed in Section 2. We further compare the FKWC tests against seven other tests: the test of and Boente, Rodriguez & Sued (2018) (Boen), the \( L^2 \)-norm tests of Guo & Zhang (2016) (Tmax, L2nv, L2br, and L2rp), and the ANOVA-inspired tests of Guo, Zhou & Zhang (2018) (GPFnv, GPFrp, and Fmax). For the test of Boente, Rodriguez & Sued (2018), we used 10 principal components and 5000 bootstrap samples. For the tests of Guo & Zhang (2016) and Guo, Zhou & Zhang (2018), we used 1000 permutations.
We simulated data from both infinite-dimensional and finite-dimensional models. For the infinite-dimensional models, we simulated observations from $J = 2$ and $J = 3$ samples. The results from the simulations with three groups were the same as with two groups and are omitted. We tested sample sizes of $N = 100$, $N = 200$, and $N = 500$ for the two-sample case, where the first sample size was $N_1 = \lfloor qN \rfloor$ for $q \in \{0.2, 0.3, 0.4, 0.5\}$. For the three-sample case, we used $N = 150$ and $N = 300$, with $N_1 = N_3$ and $N_2 = \lfloor qN \rfloor$ for $q$ as above. In each infinite-dimensional case, the data were sampled from a Gaussian process, a Student’s $t$ process with three degrees of freedom, or a skewed Gaussian process. For the infinite-dimensional case we used the squared exponential covariance kernel

$$K(s, t; \alpha, \beta) = \beta \exp\left(-\frac{(s - t)^2}{2\alpha^2}\right).$$

The sample differences were controlled via the $\alpha$ (shape difference) and $\beta$ (scale difference) parameters.

For the finite-dimensional models, the data were simulated from a Gaussian process where we directly specified $K$ nonzero eigenvalues of the covariance operator. See the Appendix for details on the eigenvalue specifications. A Fourier basis was used for the eigenfunctions. Here we only examined the two-sample version of the test. We used the same sample sizes described above. When we examined the effect of outliers, we used the same outlier scenarios described below. We ran nine scenarios, which resulted from combining short linear, long linear, and long exponential eigenvalue decay with either a unitary operator difference, a scale difference, or no difference.

We also tested the effects of different kinds of outliers on the different tests at different levels of contamination. The contamination level was measured as a percentage of the total sample size $N$. We examined contamination levels of 1%, 2.5%, and 5%. We present the results at a 2.5% contamination level of the total sample size $N_1 + N_2$. Smaller contamination levels showed similar but less pronounced results. In addition, higher percentages of contamination open up the debate of outliers versus an observed mixture distribution. The three kinds of outliers were linear drift, oscillating outliers, and spike-type outliers; see Figure 2. When examining the performance of the tests under contamination, we only considered the two-sample versions of the tests. Additionally, in our simulation we considered two cases. In the first case, both samples contained outliers. In the second case, only one of the samples contained outliers. Lastly, we tested the effect of missing portions of the curves and the effect of the number of directions on RPD. The results of these simulations can be found in the Appendix.

We ran each simulation scenario 200 times. A grid size of 100 was used to simulate the functions. The codes can be seen on Github (Ramsay, 2019). Note that the FKWC tests require the observations to be on the same grid in order to compute the sample depth values. If the

![Figure 2: Five uncontaminated observations compared to (left) a drift outlier, (middle) a wavy outlier, and (right) a scale outlier.](image-url)
functions are not observed on the same grid, then it is necessary to interpolate the curves in some way so that they can be brought to the same grid. For example, one can smooth the curves and then rediscretize them if necessary. To test the effect of missing portions of the curve, we interpolated the data with splines using the \texttt{zoo} package in R.

5.2. Results

Unanimously, the methods that incorporated the derivatives performed better than those that did not. This includes when the data contained outliers or had unequal group sizes. Also, with the exception of RPD, the methods not using derivative information do not work for shape differences. In addition, the percentile modification had little effect on the power and size in all simulation runs. Therefore, we proceed by only presenting the results from the FKWC methods based on $\hat{\mathcal{W}}_N$ that incorporated the derivatives.

The depth functions performed more or less similarly in all respects. That being said, the LTR depth and MBD functions had the most power for detecting scale differences; see Figure 3. The latter also shows that, for detecting shape differences, RPD with likelihood depth was the best, especially under heavy tails. LTR depth and MBD also performed well under shape differences. Under the very-low-dimensional models (with only three nonzero eigenvalues), RPD with likelihood depth did not work as well as the other depth functions. The best performing depth functions in this scenario were RPD with simplicial depth and the MFHD; see Table 2. When the group sizes differed or when outliers were present, the depth functions were all

![Power curves](image)

**Figure 3:** Empirical power curves for the different hypothesis tests as (left) the $\beta$ parameter and (right) the $\alpha$ parameter of the second sample move away from the null hypothesis. Here, $N_j = 100$. The black curves correspond to the different FKWC tests, and the blue curves correspond to the competing tests.

DOI: 10.1002/cjs.11767 The Canadian Journal of Statistics / La revue canadienne de statistique
TABLE 1: Empirical power of the different tests for $J = 2$ and $N = 500$ when the group sample sizes were unequal.

| $N_1/N$ : | Gaussian | Student’s $t$ | Skewed Gaussian |
|-----------|----------|--------------|----------------|
|           | 0.2  | 0.3  | 0.4  | 0.2  | 0.3  | 0.4  | 0.2  | 0.3  | 0.4  |
| FKWC      | MFHD   | 1.00 | 1.00 | 1.00 | 0.87 | 0.95 | 0.98 | 1.00 | 1.00 | 1.00 |
| RPD       | 1.00  | 1.00 | 1.00 | 0.88 | 0.94 | 0.98 | 1.00 | 1.00 | 1.00 |
| MBD       | 1.00  | 1.00 | 1.00 | 0.88 | 0.96 | 0.98 | 1.00 | 1.00 | 1.00 |
| LTR       | 1.00  | 1.00 | 1.00 | 0.88 | 0.96 | 0.98 | 1.00 | 1.00 | 1.00 |
| RPL       | 1.00  | 1.00 | 1.00 | 0.87 | 0.94 | 0.96 | 1.00 | 1.00 | 1.00 |
| Competing | Boen   | 1.00 | 1.00 | 1.00 | 0.17 | 0.15 | 0.04 | 0.99 | 1.00 | 1.00 |
|           | L2br   | 0.48 | 0.93 | 0.99 | 1.00 | 1.00 | 1.00 | 0.49 | 0.94 | 0.99 |
|           | L2rp   | 0.35 | 0.92 | 0.99 | 0.00 | 0.00 | 0.03 | 0.36 | 0.93 | 0.99 |
|           | Tmax   | 0.68 | 0.97 | 0.99 | 0.01 | 0.01 | 0.06 | 0.63 | 0.97 | 1.00 |
|           | GPFrp  | 0.03 | 0.56 | 0.91 | 0.01 | 0.00 | 0.04 | 0.01 | 0.54 | 0.87 |
|           | Fmax   | 0.12 | 0.47 | 0.83 | 0.00 | 0.01 | 0.10 | 0.11 | 0.51 | 0.85 |

Note: The groups differed in scale with $\beta_1 = 0.5$ and $\beta_2 = 0.71$. Notice that when the sample sizes differ greatly, the competing tests do not perform as well as the FKWC tests. RPL is the FKWC test paired with likelihood depth and not simplicial depth.

FIGURE 4: Empirical power of the tests for detecting Gaussian scale differences for $J = 2$ with $N_1 = N_2 = 100$ under 5% contamination in each of the samples. The legend follows that of Figure 3. The black curves correspond to the FKWC tests, and the blue curves correspond to the competing tests.

We now compare the FKWC tests to the existing tests in the literature. The naive versions of the L2 and GPF tests are omitted since their performances were similar to those of the “biased-reduced” versions. In addition, we only present the results from the FKWC tests based on the derivatives without the percentile modification. For the infinite-dimensional models, Figure 3 shows the empirical power of the FKWC tests compared to the competing tests under the Gaussian and Student’s-$t$ processes. Corresponding results for the skewed Gaussian process results were similar to those for the Gaussian process, and the results at other sample sizes were similar to those shown in Figure 3. Under the infinite-dimensional models, the FKWC tests have a higher empirical power than the competing tests, especially when the distribution is heavy-tailed. In addition, the heavy tails corrupt the L2pr test.
Table 2: Empirical proportion of rejected tests under the finite-dimensional simulation scenarios.

| Test  | Simulation scenario | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  |
|-------|---------------------|----|----|----|----|----|----|----|----|----|
| MFHD  |                     | 0.90 | 1.00 | 1.00 | 0.99 | 1.00 | 1.00 | 0.04 | 0.07 | 0.07 |
| RPD   |                     | 0.92 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.04 | 0.09 | 0.07 |
| MBD   |                     | 0.80 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.06 | 0.06 | 0.06 |
| LTR   |                     | 0.79 | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 0.06 | 0.07 | 0.04 |
| RPD†  |                     | 0.26 | 0.90 | 0.86 | 1.00 | 1.00 | 1.00 | 0.06 | 0.06 | 0.06 |
| Boen  |                     | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 0.09 | 0.04 | 0.06 |
| L2br  |                     | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 0.08 | 0.03 | 0.05 |
| L2rp  |                     | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 0.08 | 0.03 | 0.06 |
| Tmax  |                     | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 0.07 | 0.04 | 0.05 |
| GPFlp |                     | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 0.10 | 0.01 | 0.07 |
| Fmax  |                     | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.07 | 0.04 | 0.05 |

Note: Here, \( N_1 = N_2 = 100 \). The first row indicates the scenario number, whose descriptions are in the Appendix. In scenarios 1–6, the alternative hypothesis was true and in scenarios 7–9 the null hypothesis was true. Thus, the first six columns give the power of the test under different alternatives and the last three columns give the size of the test. RPL is the FKWC test paired with likelihood depth and not simplicial depth. Largest value is bolded in the first 6 meaning highest power, values 0.05 or under are bolded in the last 3, meaning the empirical size matches the theoretical size or is lower.

Table 1 shows the empirical power of the tests for a fixed scale difference between the samples under imbalanced group sizes. Many of the competing tests do not work well under very imbalanced group sizes, whereas the FKWC tests are unaffected. The tables for imbalanced group sizes with a shape change or under the finite-dimensional models are given in the Appendix.

Table 2 shows the empirical power of the tests for the finite-dimensional models. The competing tests all had power equal to 1 or very close to 1. The FKWC tests had power equal to 1 in many of the simulation scenarios except in two cases. The first case is when RPD was paired with likelihood depth. The second case is under simulation scenario 1, which is the lowest dimensional model (only three nonzero eigenvalues) where the difference was generated by applying a unitary operator to the first sample’s covariance operator. The competing tests may have had higher power in scenario 1 because the lower complexity of the data-generating process allowed for easier approximation of the distribution of the test statistic under the null hypothesis. When there were only three nonzero eigenvalues, the competing tests did, however, have a slightly inflated power; see scenario 7. We conclude that the competing tests work better for very-low-dimensional, light-tailed models. Table 3 shows the size of the tests under the infinite-dimensional models; the sizes of both the FKWC tests and the competing tests are relatively close to 0.05, with the exception of the Boen test and the L2br test.

As mentioned previously, we also simulated models where portions of the data were contaminated with outliers. Figure 2 shows the three different kinds of outliers used to contaminate the data. Figure 4 shows the empirical power of the tests for detecting Gaussian scale differences under 5% contamination in each sample with \( N_1 = N_2 = 100 \). For reference, this means that five curves in total were contaminated. We expect contamination in both samples to reduce the power of the tests. The drift and spike outliers have especially negative effects on the power of the tests. When there are drift or spike outliers in both samples, the powers of the
Table 3: Empirical sizes for $J = 2$ for different tests under the infinite-dimensional models.

| Test  | Gaussian | Student’s-t | Skewed Gaussian |
|-------|----------|-------------|-----------------|
|       | 50       | 100         | 250             | 50   | 100   | 250 | 50   | 100   | 250 | 50   | 100   | 250 |
| FKWC  | MFHD     | 0.06        | 0.03            | 0.06 | 0.03  | 0.06 | 0.07 | 0.08  | 0.07  | 0.06 |
|       | RPD      | 0.07        | 0.04            | 0.06 | 0.04  | 0.04 | 0.06 | 0.05  | 0.06  | 0.06 |
|       | MBD      | **0.05**    | **0.02**        | **0.04** | **0.03** | 0.04 | 0.08 | 0.06  | 0.06  | 0.06 |
|       | LTR      | **0.05**    | **0.04**        | 0.06 | 0.04  | 0.04 | 0.08 | 0.06  | 0.06  | **0.04** |
|       | RPL      | 0.06        | **0.03**        | 0.07 | 0.04  | 0.03 | 0.07 | 0.07  | **0.04** | 0.06 |
| Competing | Boen    | 0.21        | 0.24            | 0.34 | **0.00** | **0.00** | **0.00** | 0.16  | 0.21  | 0.32 |
|       | L2br     | 0.04        | 0.06            | 0.06 | 1.00  | 1.00 | 1.00 | **0.04** | **0.04** | 0.08 |
|       | L2rp     | 0.06        | 0.06            | 0.06 | 0.05  | **0.04** | **0.04** | 0.04  | 0.04  | 0.07 |
|       | Tmax     | **0.04**    | 0.06            | 0.07 | 0.04  | 0.04 | 0.04 | 0.04  | **0.03** | 0.06 |
|       | GPFrp    | 0.06        | 0.06            | 0.06 | 0.06  | 0.04 | 0.04 | 0.05  | 0.03  | 0.06 |
|       | Fmax     | **0.03**    | **0.04**        | **0.05** | **0.04** | **0.04** | **0.03** | 0.06  | 0.02  | 0.06 |

Note: The first row indicates the underlying process, and the second row indicates the sample size of each group. RPL is the FKWC test paired with likelihood depth and not simplicial depth. Values 0.05 or under are bolded, meaning the empirical size matches the theoretical size or is lower.

Table 4: Empirical sizes of the tests when one sample was contaminated by each type of outlier.

| MFHD | RPD | MBD | LTR | RPL | L2br | L2rp | Tmax | GPFrp | Fmax |
|------|-----|-----|-----|-----|------|------|------|-------|------|
|      | 0.04 | 0.04 | 0.04 | 0.05 | 0.03 | 0.81 | 0.47 | 0.25  | 0.36  |
|      | 0.04 | 0.06 | 0.04 | 0.04 | 0.03 | 0.08 | 0.10 | 0.06  | 0.10  |
|      | 0.04 | 0.06 | 0.04 | 0.05 | 0.02 | 0.29 | 0.08 | 0.06  | 0.05  |

Note: Specifically, 5% of the first sample was contaminated. The sizes of the competing tests are inflated in almost all scenarios. RPL is the FKWC test paired with likelihood depth and not simplicial depth.

Suppose now that the contamination is only in one sample. We expect contamination in a single sample to increase the size of the tests. Table 4 shows the size of the tests when one sample was contaminated by the different kinds of outliers. The sizes of the competing tests are inflated in almost all scenarios, even under the wavy-type outliers. The drift outlier seems to have especially negative effects on the competing tests, with the exception of the Fmax test. We have omitted the Boen test in Table 4, as it did not perform well on the uncontaminated data. In conclusion, the FKWC tests perform very well compared to the existing tests. This is especially true in the presence of imbalanced groups, heavy tails, and drift or spike type outliers.

6. APPLICATIONS TO REAL DATA

In this section we present an application of our methodology to two functional datasets. One contains intraday stock prices and the other contains data related to digitized speech.
6.1. F-GARCH Residual Analysis of Intraday Stock Price Curves

We analyze the daily asset price curves of $J = 3$ stocks (TWTR, FB, and SNAP) starting on June 24, 2019 and ending on March 20, 2020, which gives $N_1 = 207$ and $N_2 = N_3 = 208$. The price of each stock was measured over the course of the trading day at 1-min intervals for a total of 390 min per day. In order to account for edge effects from curve-smoothing, we trimmed 10% of the minutes from the beginning of the day and 5% of the minutes from the end of the day. This resulted in 332 min of stock prices. In actuality, we analyzed the log returns

$$X_{ji}(t) = \ln(Y_{ji[331t]+1}) - \ln(Y_{ji[331t]}),$$

where $Y_{jik}$ is the $j$th asset price on the $i$th day at minute $k$. Figure 5a shows the intraday log return curves of Facebook (FB) stock. A model was fit using a B-spline basis with 50 basis functions; see smooth.basis in the fda R package.

The magnitudes of the curves vary widely. Figure 5b displays the squared norms of the daily curves as a function of the day $i$. We can see that the magnitude of each observation is related to the day on which it was observed. For example, around March 2020 the norms are higher, likely due to the volatility that resulted from the COVID-19 pandemic.

Figure 5: (a) Daily log differenced intraday return curves for FB stock, from June 24, 2019 to March 20, 2020. (b) Daily squared norms of the intraday returns. These norms vary with the time period; the curves exhibit heteroskedastic features. For example, the most recent month of returns are much more variable. (c) Residuals for the FB log returns after fitting a functional GARCH(1,1) model. (d) Squared norms of the residuals over time.
To handle heteroscedasticity and any serial correlation present in the data, we employed a functional GARCH(1,1) model (Aue et al., 2017; Cerovecki et al., 2019) and applied the FKWC tests to the residuals. The idea is to decompose the data into conditional volatility $\eta^2_{ij}(t)$ and independent error $\epsilon_{ij}(t)$, which can be approximated by the residuals $\hat{\epsilon}_{ij}(t)$. Unlike in the univariate GARCH model, the second-order behaviour of $\epsilon_{ij}(t)$ can differ between different assets; $E[\epsilon^2_{ij}(t)] = 1$ for all $t$ is assumed for an identifiable model (Cerovecki et al., 2019) but nothing is assumed about $E[\epsilon_{ij}(t)\epsilon_{ij}(s)]$ for $s \neq t$. Thus, it is also of interest to investigate the properties of $E[\epsilon_{ij}(t)\epsilon_{ij}(s)]$. For example, if the errors come from the same distribution, then the residuals can be pooled and bootstrapped to provide standard errors.

Given that this type of data is typically heavy-tailed, a robust test is suitable. In order to check the condition that $\eta^2_{ij}$ completely encapsulates the serial dependence in the data, we use the tests described by Rice, Wirjanto & Zhao (2019). Specifically, we fit the functional GARCH(1,1) model to each series of intraday returns using quasi-maximum likelihood (Cerovecki et al., 2019). We assumed that the volatility curves could be represented as linear combinations of $M$ Bernstein basis functions. The value of $M$ was chosen on the basis of the Box–Jenkins-type test for the functional GARCH model (Rice, Wirjanto & Zhao, 2019) and graphical assessments of the fit of the raw mean of the squares, see Cerovecki et al. (2019). We also ensured that the number of basis functions was the same for each stock. This resulted in choosing $M = 4$; our results in terms of testing the residuals for a difference in covariance structure were insensitive to the number of basis functions. Figure 5c shows the residuals of the resulting GARCH(1,1) model fitted to FB log stock returns, and Figure 5d shows the norms of those residuals as a function of the day $i$. Notice that both the residuals and their norms are fairly uniform over time, especially when compared with the raw data. Figure 5d also shows that there are some outliers in the data.

Figure 6 shows contour plots of the estimated covariance kernels of the residuals of each functional time series. Given that the outliers affect the estimation of the covariance kernels, the observations (and their corresponding derivatives) that had the lowest 5% RPD values were trimmed. To clarify, the outliers were only trimmed for the estimation of the covariance kernels given in Figure 6; the outliers were included in the rest of the analysis.

Notice that the estimated covariance kernel of the residuals of FB differs from the other two assets visually. We conducted the FKWC test based on RPD which incorporates the derivative information at a 5% level of significance. The means of the ranks were 244.4203, 424.2837, and 266.9712 for the TWTR, FB, and SNAP residuals, respectively; $\hat{W}_{N} = 120.37$ implies a $P$-value of essentially zero, and we reject the hypothesis that these three series have the same covariance kernels. The means of the ranks are similar between the TWTR and SNAP stocks, but that for the FB stock differs substantially. This is apparent in Figure 6, where the diagonal of the covariance kernels is

![Figure 6: Covariance kernels $K_{ij}(s, t)$ of the residuals resulting from the GARCH(1,1) model fit to the log returns of (a) TWTR, (b) FB, and (c) SNAP.](image-url)
kernels of the TWTR and SNAP stocks are much brighter than that of the FB stock. This is especially true later in the day, where the diagonal of the covariance kernel of the FB stock is the dimmest. This suggests that the TWTR and SNAP stocks are more variable than the FB stock, especially later in the day.

6.2. Comparing Speech Variability with Phoneme Periodograms

In this section we analyze the Phoneme dataset, where the observations are log periodograms of digitized speech. The data can be retrieved from the fda R package (Hastie, Buja & Tibshirani, 1995). The data are split into five groups representing the syllables “aa”, “ao”, “dcl”, “iy”, and “sh”. The goal is to characterize the differences between the syllables’ distributions in order to aid the understanding of speech as well as to help improve the performance of speech recognition models. The data have obvious location differences; for example, Figure 7a shows the periodograms for two different syllables on the same plot. We centred the data by the deepest curve, as measured by RPD, within each group. We used a robust measure of centre to account for outliers; for example, there is a large outlier in the “dcl” syllable group. From Figure 7b, we might suspect that there are differences in the covariance kernels between the curves. To this end, we can run the FKWC tests on the five groups of syllables. We run two versions of the test and compare the results. We run the FKWC test paired with RPD with the derivative information and the FKWC test paired with LTR depth. These tests will differ if there are differences of the form \( K_j(s, t) \neq K_k(s, t) \) for \( s \neq t \) and \( j \neq k \). Both tests result in \( P \)-values smaller than \( 2 \times 10^{-16} \).

We can further examine differences between groups by performing multiple comparisons. There are two obvious routes for multiple comparisons. One method is to directly use the precalculated joint sample ranks, analogous to the univariate method of Dunn (1964). Here, for large \( N \), one is essentially assessing the behaviour of the random variables \( D(X_{j1}; F^*) - D(X_{k1}; F^*) \) through combined sample ranks. The other method is to extend the methods of Steel (1960) and compare the mean ranks of \( D(X_{ji}; F_{jk}) \) and \( D(X_{ki}; F_{jk}) \), where

\[
F_{jk} = \frac{\theta_j}{\theta_j + \theta_k} F_j + \frac{\theta_k}{\theta_j + \theta_k} F_k.
\]

Given that Theorems 2–3 imply that differences in covariance structures will be exhibited in the pairwise ranks of the depth values when the depth values are taken with respect to the empirical estimate of \( F_{jk} \), it seems natural to use the methods of Steel (1960). A second argument in support of the methods of Steel (1960) is as follows. Suppose that one group \( j' \) has a very different covariance structure when compared to the remaining groups. If the depth values are computed

![Figure 7](image-url)
with respect to the combined sample, then this group $j'$ may “wash away” any differences between the remaining groups. In other words, for some $j, k \in \{1, \ldots, j' - 1, j' + 1, \ldots, J\}$, it is possible that there is a difference between the random variables $D(X_{j'i}; F_\ast)$ and $D(X_{k'i}; F_\ast)$, but that this difference is small relative to the combined sample and therefore may not be detected. The multiple comparisons procedure is as follows: For each pair of groups, $(j, k)$, compute the combined two-sample depth values

$$\left\{D(X_{j1}; F_{jk,N}), \ldots, D(X_{jn}; F_{jk,N}), D(X_{k1}; F_{jk,N}), \ldots, D(X_{kn}; F_{jk,N})\right\},$$

where $F_{jk,N}$ is the empirical distribution of $X_{j1}, \ldots, X_{jn}, X_{k1}, \ldots, X_{kn}$. Next, perform the Wilcoxon rank-sum test on the depth values for each pair. Lastly, correct the final $P$-values using the Šidák correction (Šidák, 1967) (or any other multiple-testing correction). Again, Theorems 1–3 justify this procedure. Table 5 shows the Šidák-corrected $P$-values of pairwise “functional Steel tests” performed on the centred curves. The $P$-values are corrected for the tests done across both hypothesis tests, i.e., across 22 tests. The results show that the syllables “dcl” and “sh” differ from the remaining syllables and from each other in terms of the variability of the magnitude of their log frequencies. The results of the FKWC test paired with LTR depth give that “dcl” is similar to the other syllables. This implies that the trace norm of the covariance operator of “dcl” is similar to that of the other syllables, with the exception of “ao”. However, the results of the FKWC test paired with RPD' test give that the syllable “dcl” differs from all other syllables. We can interpret this as the log periodograms of “dcl” being more or less “wiggly” when compared to the other syllables, or the frequencies having high variability are different for the syllable “dcl” than the other syllables. This suggests that the centred squared differences of the log periodograms may be useful in distinguishing “dcl” and “sh” from both each other and the other syllables.

7. CONCLUSION

We have introduced a new type of multisample hypothesis test called the FKWC test, which detects differences in the covariance operator between samples. These tests are based on data depth, ranks, and the derivatives of the curves. We have shown that FKWC tests have several benefits, including being nonparametric, robust, and computationally fast. We have shown various theoretical properties of these tests, including the distribution of the test statistic under the null hypothesis as well as consistency under certain alternative hypotheses. Through simulations, we showed that these tests are robust against heavy tails, outliers, different group sizes, and skewed distributions. In addition, we have also shown that these tests work very well

### Table 5: Šidák-corrected $P$-values of pairwise functional Steel tests performed on the centred curves.

| Syllable | RPD' | LTR |
|-----------|------|-----|
|           | aa   | ao  | dcl | iy  | sh  | aa   | ao  | dcl | iy  | sh  |
| aa        | 1.00 | 0.85| **0.00**| 0.97| **0.00**| 1.00 | 0.13| 0.40| 1.00| **0.00**|
| ao        | 0.85 | 1.00| **0.00**| 0.27| **0.00**| 0.13 | 1.00| **0.00**| 0.29| 0.11|
| dcl       | **0.00**| **0.00**| 1.00| **0.00**| **0.00**| 0.40 | **0.00**| 1.00| 0.29| 0.11|
| iy        | 0.97 | 0.27| **0.00**| 1.00| **0.00**| 1.00 | 0.18| 0.29| 1.00| **0.00**|
| sh        | **0.00**| **0.00**| **0.00**| **0.00**| 1.00| **0.00**| **0.00**| 0.11| **0.00**| 1.00|

**Note:** Values below the corrected significance level are bolded.
when compared to existing tests. In line with Hubert, Rousseeuw & Segaert (2015), this study has shown that the derivatives of the observed curves provide valuable information about the sample. In addition, we have shown a mathematical link between the covariance kernel and certain data depth functions, which can be further exploited. Future directions include incorporating testing for relevant differences, the theoretical behaviour of depth computed on derivatives, and FKWC tests for dependent data.

ACKNOWLEDGEMENTS
The authors would like to thank the editor and reviewers for their insightful comments which improved this article. They would also like to acknowledge Boente, Rodriguez & Sued (2018) and Guo, Zhou & Zhang (2018) for kindly providing the code for their methods. This work was supported by the NSERC, Canada.

REFERENCES
Ansari, A. R. & Bradley, R. A. (1960). Rank-sum tests for dispersions. *The Annals of Mathematical Statistics*, 31(4), 1174–1189.
Aston, J. A. D., Pigoli, D., & Tavakoli, S. (2017). Tests for separability in nonparametric covariance operators of random surfaces. *The Annals of Statistics*, 45(4), 1431–1461.
Aue, A., Horváth, L., & Pellatt, F. D. (2017). Functional generalized autoregressive conditional heteroskedasticity. *Journal of Time Series Analysis*, 38(1), 3–21.
Basu, S. & DasGupta, A. (1997). The mean, median, and mode of unimodal distributions: A characterization. *Theory of Probability & Its Applications*, 41(2), 210–223.
Boente, G., Rodriguez, D., & Sued, M. (2018). Testing equality between several populations covariance operators. *Annals of the Institute of Statistical Mathematics*, 70(4), 919–950.
Cabassi, A., Pigoli, D., Secchi, P., & Carter, P. A. (2017). Permutation tests for the equality of covariance operators of functional data with applications to evolutionary biology. *Electronic Journal of Statistics*, 11(2), 3815–3840.
Cerovecki, C., Francq, C., Hörmann, S., & Zakoïan, J.-M. (2019). Functional GARCH models: The quasi-likelihood approach and its applications. *Journal of Econometrics*, 209(2), 353–375.
Chenouri, S., Mozaffari, A. & Rice, G. (2020a). Multiple change point detection based on standard and wild rank-cusum binary segmentation (Unpublished technical report, University of Waterloo).
Chenouri, S., Mozaffari, A., & Rice, G. (2020b). Robust multivariate change point analysis based on data depth. *Canadian Journal of Statistics*, 48(3), 417–446.
Chenouri, S. & Small, C. G. (2012). A nonparametric multivariate multisample test based on data depth. *Electronic Journal of Statistics*, 6, 760–782.
Chenouri, S., Small, C. G., & Farrar, T. J. (2011). Data depth-based nonparametric scale tests. *Canadian Journal of Statistics*, 39(2), 356–369.
Claeskens, G., Hubert, M., Slaets, L., & Vakili, K. (2014). Multivariate functional halfspace depth. *Journal of the American Statistical Association*, 109, 411–423.
Cuevas, A., Febrero, M., & Fraiman, R. (2007). Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics*, 22, 481–496.
Cuevas, A. & Fraiman, R. (2009). On depth measures and dual statistics. A methodology for dealing with general data. *Journal of Multivariate Analysis*, 100(4), 753–766.
Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics*, 6(3), 241–252.
Fan, C., Zhang, D., & Zhang, C.-H. (2011). On sample size of the Kruskal-Wallis test with application to a mouse peritoneal cavity study. *Biometrics*, 67(1), 213–224.
Flores, R., Lillo, R., & Romo, J. (2018). Homogeneity test for functional data. *Journal of Applied Statistics*, 45(5), 868–883.
Fraiman, R. & Muniz, G. (2001). Trimmed means for functional data. *Test*, 10(2), 419–440.
Fremdt, S., Steinebach, J. G., Horváth, L., & Kokoszka, P. (2013). Testing the equality of covariance operators in functional samples. *Scandinavian Journal of Statistics*, 40(1), 138–152.
Gaines, G., Kaple, K., & Ruyymgaart, F. (2011). Application of a delta-method for random operators to testing equality of two covariance operators. *Mathematical Methods of Statistics*, 20(3), 232.
Gastwirth, J. L. (1965). Percentile modifications of two sample rank tests. *Journal of the American Statistical Association*, 60(312), 1127–1141.

Gibbons, J. D. (1973). Comparisons of asymptotic and exact power for percentile modified rank tests. *Sankhyā The Indian Journal of Statistics, Series B (1960-2002)*, 35(1), 15–24.

Gijbels, I. & Nagy, S. (2017). On a general definition of depth for functional data. *Statistical Science*, 32(4), 630–639.

Guo, J. & Zhang, J.-T. (2016). A further study of an $L^2$-norm based test for the equality of several covariance functions. arXiv e-prints, arXiv:1609.04231.

Guo, J., Zhou, B., & Zhang, J.-T. (2018). New tests for equality of several covariance functions for functional data. *Journal of the American Statistical Association*, 114, 1–13.

Hastie, T., Buja, A., & Tibshirani, R. (1995). Penalized discriminant analysis. *The Annals of Statistics*, 23(1), 73–102.

Hsing, T. & Eubank, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*, Wiley Series in Probability and Statistics, Wiley, Chichester, UK.

Hubert, M., Claeskens, G., Ketelaere, B. D. & Vakili, K. (2012). A new depth-based approach for detecting outlying curves. In *Proceedings of COMPSTAT 2012*, 329–340.

Hubert, M., Rousseeuw, P. J., & Segaert, P. (2015). Multivariate functional outlier detection. *Statistical Methods and Applications.*, 24(2), 177–202.

Ieva, F. & Paganoni, A. M. (2013). Depth measures for multivariate functional data. *Communications in Statistics-Theory and Methods*, 42(7), 1265–1276.

Jarušková, D. (2013). Testing for a change in covariance operator. *Journal of Statistical Planning and Inference*, 143(9), 1500–1511.

Kashlak, A. B., Aston, J. A. D., & Nickl, R. (2019). Inference on covariance operators via concentration inequalities: $k$-sample tests, classification, and clustering via Rademacher complexities. *Sankhyā The Indian Journal of Statistics, Series A*, 81(1), 214–243.

Kruskal, W. H. (1952). A nonparametric test for the several sample problem. *The Annals of Mathematical Statistics*, 23(4), 525–540.

Liu, R. & Singh, K. (2006). *Rank Tests for Multivariate Scale Difference Based on Data Depth*, DIMACS: Series in Discrete Mathematics and Theoretical Computer Science, Vol. 72, American Mathematical Society, Providence, Rhode Island, 17–35.

López-Pintado, S. & Romo, J. (2009). On the concept of depth for functional data. *Journal of the American Statistical Association*, 104(486), 718–734.

López-Pintado, S., Sun, Y., Lin, K. J., & Genton, M. G. (2014). Simplicial band depth for multivariate functional data. *Advances in Data Analysis and Classification*, 8, 321–338.

López-Pintado, S. & Wrobel, J. (2017). Robust non-parametric tests for imaging data based on data depth. *Stat.*, 6(1), 405–419.

Müller, C. H. (2005). Depth estimators and tests based on the likelihood principle with application to regression. *Journal of Multivariate Analysis*, 95(1), 153–181.

Nagy, S. & Ferraty, F. (2019). Data depth for measurable noisy random functions. *Journal of Multivariate Analysis*, 170, 95–114.

Page, W. & Murty, V. N. (1982). Nearness relations among measures of central tendency and dispersion: Part I. *The Two-Year College Mathematics Journal*, 13(5), 315–327.

Panaretos, V. M., Kraus, D., & Maddocks, J. H. (2010). Second-order comparison of Gaussian random functions and the geometry of DNA minicircles. *Journal of the American Statistical Association*, 105, 670–682.

Paparoditis, E. & Sapatinas, T. (2016). Bootstrap-based testing of equality of mean functions or equality of covariance operators for functional data. *Biometrika*, 103(3), 727–733.

Pigoli, D., Aston, J. A. D., Dryden, I. L., & Secchi, P. (2014). Distances and inference for covariance operators. *Biometrika*, 101(2), 409–422.

Piña, B., Raldúa, D., Barata, C., Portugal, J., Navarro-Martín, L., Martínez, R., Fuertes, I., & Casado, M. (2018). *Functional Data Analysis: Omics for Environmental Risk Assessment*, Comprehensive Analytical Chemistry, Vol. 82, Elsevier, London, England, 583–611.

Ramsay, K. & Chenouri, S. (2020). Robust multiple change-point detection for multivariate variability using data depth. arXiv e-prints, arXiv:2011.09558.

Ramsay, K. A. (2019). FKWC: GitHub repository, https://github.com/12ramsake/FKWC
APPENDIX

Technical Considerations for Functional Observations

We define “functional” observations by the following assumptions. First, we assume that each $X_{ji}$ is a mean-square-continuous stochastic process, meaning that for each $t \in [0, 1]$, $X_{ji}(t)$ is measurable with respect to some probability space $(\Omega, \mathcal{A}, P)$ and that $\lim_{t \to s} E[|X(t) - X(s)|^2] = 0$. Second, for each $\omega \in \Omega$, $X_{ji}(\cdot, \omega)$ is a continuous function. We use $\mathcal{F}$ to denote the space of such processes. These assumptions imply that $X_{ji}(t, \omega)$ is jointly measurable with respect to the product $\sigma$-field $\mathcal{B} \times \mathcal{A}$, where $\mathcal{B}$ denotes the Borel sets of $[0, 1]$. This joint measurability implies that each $X_{ji}$ can be interpreted as a random element in $L_2([0, 1], \mathcal{B}, \mu)$, where $\mu$ is the Lebesgue measure on $[0, 1]$. For more details, see Chapter 7 of Hsing & Eubank (2015). Some variants of the proposed test involve derivatives, and for those tests we will additionally require that the derivative of $X_{ji}$, which we denote by $X_{ji}'(t)$, exists on the interval $(0, 1)$ and satisfies the same continuity assumptions imposed on $X_{ji}$.

LTR Depth

The following theorem lists some properties of LTR depth.

**Theorem 5.** Let $X \sim F$, where $F$ is a measure over $L_2([0, 1], \mathcal{B}, \mu)$; $F_N$ be the empirical measure corresponding to a random sample of size $N$ from $F$; $a, b \in \mathbb{R}$; and $c, c' \in \mathbb{R}^+$. LTR depth satisfies the following properties:

1. Sample ranks based on LTR depth are invariant when the sample is transformed by a linear function $g$, such that $g(x) = ax + b$;

DOI: 10.1002/cjs.11767
2. If \( X \overset{d}{=} -X \), then \( \sup_x LTR(x; F) = LTR(0; F) \);
3. If \( X \overset{d}{=} -X \), then \( LTR(c; x; F) \) is decreasing in \( c \);
4. \( \lim_{c \to \infty} LTR(c; x; F) = 0 \);
5. If \( E \|X\|^2 < \infty \), then \( \sup_x |LTR(x; F_N) - LTR(x; F)| = o(1) \) almost surely.

The proof of this theorem is given below. We remark that \( LTR(x; F) \) is not invariant under the linear transformation by \( g \) in property 1 of Theorem 5. It is easy to see that

\[
LTR(x; F) = \frac{1}{1 + c^T} \Rightarrow LTR(ax; aF) = \frac{1}{1 + \|a\|^T c^T},
\]

where \( aX \sim aF \) if \( X \sim F \). This fact implies that hypothesis tests based on LTR depth values themselves are not invariant under linear transformations. However, Theorem 5 shows that hypothesis tests based on ranks of these depth values are invariant under linear transformations. Further, a median based on this depth would be equivariant under the linear transformation \( g \) in property 1.

In this setting, ranking the observations based on LTR depth is equivalent to ranking the observations based on their norms. Under the assumption of a zero mean, we have that

\[
E_{F_n} \left[ \|X_{ji} - X\|^2 \right] = E_{F_n} \left[ \|X_{ji}\|^2 \right] + E_{F_n} \left[ \|X\|^2 \right] = \|X_{ji}\|^2 + \sum_{j=1}^{J} \theta_j \mathcal{X}_j + o(1),
\]

from which it is easily seen that the ranks are equivalent to those based on \( E_{F_n} \left[ \|X_{ji}\|^2 \right] \). We emphasize that this relationship relies on the assumption of a zero mean as well as on the data being centred. In this context, ranks generated from this depth function do not need to be estimated; we can compute ranks based on \( D (\cdot; F_n) \) directly.

Additional Information on the Simulation Study

This section contains some additional simulation results. The finite-dimensional models are as follows. We ran nine scenarios, with \( \lambda_{jk} \) denoting an eigenvalue of the covariance operator of group \( j \). These are

1. reversed short linear decay, with \( \lambda_{1k} = k \) and \( \lambda_{2k} = 3 - k + 1 \) for \( k < 4 \) and \( \lambda_{jk} = 0 \) for \( k \geq 4 \);
2. reversed long linear decay, with \( \lambda_{1k} = k \) and \( \lambda_{2k} = 11 - k + 1 \) for \( k < 12 \) and \( \lambda_{jk} = 0 \) for \( k \geq 12 \);
3. reversed long exponential decay, with \( \lambda_{1k} = 2^k \) and \( \lambda_{2k} = 2^{11-k+1} \) for \( k < 12 \) and \( \lambda_{jk} = 0 \) for \( k \geq 12 \);
4. scaled short linear decay, with \( \lambda_{1k} = k \) and \( \lambda_{2k} = 1.5 \lambda_{1k} \) for \( k < 4 \) and \( \lambda_{jk} = 0 \) for \( k \geq 4 \);
5. scaled long linear decay, with \( \lambda_{1k} = k \) and \( \lambda_{2k} = 1.5 \lambda_{1k} \) for \( k < 12 \) and \( \lambda_{jk} = 0 \) for \( k \geq 12 \);
6. scaled long exponential decay, with \( \lambda_{1k} = 2^k \) and \( \lambda_{2k} = 1.5 \lambda_{1k} \) for \( k < 12 \) and \( \lambda_{jk} = 0 \) for \( k \geq 12 \);
7. null hypothesis short linear decay, with \( \lambda_{1k} = \lambda_{2k} = k \) for \( k < 4 \) and \( \lambda_{jk} = 0 \) for \( k \geq 4 \);
8. null hypothesis long linear decay, with \( \lambda_{1k} = \lambda_{2k} = k \) for \( k < 12 \) and \( \lambda_{jk} = 0 \) for \( k \geq 12 \);
9. null hypothesis long exponential decay, with \( \lambda_{1k} = \lambda_{2k} = 2^k \) for \( k < 12 \) and \( \lambda_{jk} = 0 \) for \( k \geq 12 \).

Tables C1 and C2 present simulation results under imbalanced group sizes.

On the Number of Directions for RPD

We also studied the effect of the number of directions used in the RPD calculation on the FKWC test. Figure C1 shows that the test is stable with respect to the number of directions.
Table C1: Empirical power of the tests for detecting a shape difference with $\alpha_1 = 0.05$ and $\alpha_2 = 0.071$.

| $N_1/N$ | Gaussian | Student’s $t$ | Skewed Gaussian |
|---------|----------|--------------|-----------------|
|         | 0.2      | 0.3          | 0.4             |
| FKWC    | MFHD     | 1.00         | 1.00            | 1.00            |
| RPD     | 1.00     | 0.91         | 0.94            | 0.97            |
| MBD     | 1.00     | 0.88         | 0.94            | 0.97            |
| LTR     | 1.00     | 0.96         | 0.97            | 0.98            |
| RPL     | 1.00     | 1.00         | 1.00            | 1.00            |
| Competing Boen | 0.95 | 0.99         | 0.99            | 0.99            |
| L2br    | 0.81     | 1.00         | 1.00            | 0.78            |
| L2rp    | 0.78     | 0.04         | 0.04            | 0.08            |
| Tmax    | 0.34     | 0.06         | 0.07            | 0.06            |
| GPFrp   | 0.71     | 0.05         | 0.07            | 0.13            |
| Fmax    | 0.60     | 0.07         | 0.18            | 0.32            |

Note: Here, $J = 2$ and $N = 500$ with unequal group sample sizes. When the sample sizes differ greatly, the competing tests do not perform as well.

Curves with Missing Values

Figure C2 shows that the results are essentially the same if the curves have missing values at random time points.

Proofs

Proof of Equation (4). Let

$$Z_N = \sup_{x \in \overline{x}} |\text{RPD}_{M_N}(x; F_N) - \text{RPD}_{\infty}(x; F)|.$$

Observe that

$$E[Z_N] = E \left[ M_N^{-1} \sum_{m=1}^{M_N} \left| D \left( \langle x, u_m \rangle; F_{N,u_m} \right) - \int_S D \left( \langle x, u \rangle; F_{u} \right) d\nu(u) \right| \right]$$

$$\leq E \left[ M_N^{-1} \sum_{m=1}^{M_N} \left| D \left( \langle x, u_m \rangle; F_{N,u_m} \right) - D \left( \langle x, u_m \rangle; F_{u_m} \right) \right| \right] + O(N^{-1/2})$$

$$\leq E \left[ 4M_N^{-1} \sum_{m=1}^{M_N} E \sup_{z \in \overline{z}} \left| F_{N,u_m}(z) - F_{u_m}(z) \right| |u_1, \ldots, u_{M_N} | \right] + O(N^{-1/2})$$

$$= O(N^{-1/2}),$$

where the first inequality follows from the triangle inequality and Hoeffding’s inequality, and the last equality results from the Dvoretzky–Kiefer–Wolfowitz inequality.
Proof of Theorem 5. Let $aF + b$ be the measure associated with $aX + b$. We have that

$$
LTR(ax + b; aF + b) = \left( 1 + E_F [\|ax + b - aX + b\|^2]^{1/2} \right)^{-1}
$$

$$
= \left( 1 + \|a\| E_F [\|x - X\|^2]^{1/2} \right)^{-1}.
$$

The function $(1 + c'x)^{-1}$ is monotonic for any $c' > 0$ and so for any $x, y \in \mathcal{X}$ such that $LTR(x; F) < LTR(y; F)$, it follows that $LTR(ax + b; aF + b) < LTR(ay + b; aF + b)$ and $LTR(ax + b; aF + b) < LTR(ay + b; aF + b)$. This gives the first property. For the second...
Figure C1: Empirical power of the FKWC test paired with RPD to detect scale and shape differences for $J = 2$ with $N_1 = N_2 = 50$ under different numbers of sampled directions. The top row is for Gaussian data, the middle row is for Student’s-$t$ data with three degrees of freedom, and the final row is for skewed Gaussian data.

property, observe that

$$
\text{LTR}(x; F) = \left(1 + E_F[\|x - X\|^2]^{1/2}\right)^{-1}
$$

$$
= \left(1 + 2^{-1/2}E_F[\|x - X\|^2 + \|x + X\|^2]^{1/2}\right)^{-1}
$$

$$
= \left(1 + 2^{-1/2}E_F[2\|x\|^2 + 2\|X\|^2]^{1/2}\right)^{-1}
$$

$$
= \left(1 + 2^{-1/2}2^{1/2}\|x\|^2 + 2^{-1/2}2^{1/2}E[\|X\|^2]^{1/2}\right)^{-1}
$$

$$
= (1 + \|x\| + c')^{-1},
$$

DOI: 10.1002/cjs.11767
which is maximized at $x = 0$. For the third and fourth properties, we have in a similar manner

$$\text{LTR}(cx; F) = \left( 1 + E_F[\|cx - X\|^2]^{1/2} \right)^{-1} = (1 + c \|x\| + c')^{-1},$$

which is decreasing toward 0 as $c$ increases. Lastly, if $X_1, \ldots, X_N$ is a random sample from $F$, then

$$\frac{1}{N} \sum_{i=1}^N \|x - X_i\|^2 = \|x\|^2 + \frac{1}{N} \sum_{i=1}^N \|X_i\|^2 = 2 \left\langle x, \frac{1}{N} \sum_{i=1}^N X_i \right\rangle = \|x\|^2 + \bar{Y}_{x,N}. $$

We have that

$$|\text{LTR}(x; F_N) - \text{LTR}(x; F)| = \left| \left( 1 + (\|x\|^2 + \bar{Y}_{x,N})^{1/2} \right)^{-1} - \left( 1 + (\|x\|^2 + \|\mathcal{X}\|_{\text{TR}})^{1/2} \right)^{-1} \right|$$

$$= \left| \frac{(\|x\|^2 + \|\mathcal{X}\|_{\text{TR}})^{1/2} - (\|x\|^2 + \bar{Y}_{x,N})^{1/2}}{\left( 1 + (\|x\|^2 + \bar{Y}_{x,N})^{1/2} \right) \left( 1 + (\|x\|^2 + \|\mathcal{X}\|_{\text{TR}})^{1/2} \right)} \right|$$

$$\leq \left| \frac{\|\mathcal{X}\|_{\text{TR}} - \bar{Y}_{x,N}^{1/2}}{\left( 1 + (\|x\|^2 + \bar{Y}_{x,N})^{1/2} \right) \left( 1 + (\|x\|^2 + \|\mathcal{X}\|_{\text{TR}})^{1/2} \right)} \right|,$$

where the third line comes from the fact that $\sqrt{x} - \sqrt{y} \leq \sqrt{|x - y|}$. Now, suppose that $\|x\| < c' N^{1/2} (\ln N)^{-1}$. Then

**Figure C2:** Empirical power of the tests for detecting (top row) Gaussian and (bottom row) Student’s-$t$ differences for $J = 2$ and $N_1 = N_2 = 100$ when 20% of the curves were uniformly randomly missing. The missing values were first interpolated with splines. The legend is the same as that in Figure 3.
\[ |\text{LTR}(x; F_N) - \text{LTR}(x; F)| \leq \|X\|_{\text{TR}} - \sqrt{\frac{1}{N}} \sum_{i=1}^{N} \|X_i\|^2 - c' N^{-1/2} (\ln N)^{-1} \sum_{i=1}^{N} \int X_i \, dt \quad \leq \quad o(1) \quad \text{a.s.,} \]

where the last line follows from the strong law of large numbers and the law of the iterated logarithm. Note that \( \int X_i \, dt \) has finite variance for all \( i \in \{1, \ldots, N\} \) and that the second inequality does not depend on \( x \). Suppose now that \( \|x\| \geq c' N^{1/2} (\ln N)^{-1} \). Then it is easy to see that \( |\text{LTR}(x; F_N) - \text{LTR}(x; F)| \to 0 \) by the vanishing at infinity property. \( \blacksquare \)

**Proof of Theorem 1.** Observe that the \( \hat{R}_{ji} \)s are identically distributed under the null hypothesis. This implies that the rank vector is uniformly distributed with a \( \frac{1}{N!} \) probability of each outcome. This is the same setup as Kruskal (1952), and so it follows immediately that \( \mathcal{W}_N \rightsquigarrow \chi^2_{J-1} \). Similarly, it follows directly from Theorem 2 of Chenouri, Small & Farrar (2011) that \( \mathcal{M}_{N,r} \rightsquigarrow \chi^2_{J-1} \). \( \blacksquare \)

**Proof of Equation (6).** Let \( \hat{\sigma}^2_N = N(N+1)/12 \). We can rewrite \( \hat{\mathcal{W}}_N \) as

\[ \hat{\mathcal{W}}_N = \frac{1}{\hat{\sigma}^2_N} \sum_{j=1}^{J} N_j \bar{R}_j^2 - 3(N+1). \]

Now, define

\[ \mathcal{W}_N = \frac{1}{\sigma^2_N} \sum_{j=1}^{J} N_j \bar{R}_j^2 - 3(N+1) \quad \text{with} \quad \bar{R}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} R_{ji}. \]

Under the alternative hypothesis, Assumption 1, and (5), the \( \mathcal{D}(X_{ji}; F^*) \)s are equivalent to the univariate random variables studied by Kruskal (1952). For all \( \delta > 0 \),

\[ \lim_{N \to \infty} \Pr(\mathcal{W}_N > \delta) = 1. \]

We now show that \( |\hat{\mathcal{W}}_N - \mathcal{W}_N| = O_p(1) \), which will complete the proof. To this end, note that the sequences \( R_{j1}, \ldots, R_{jN_j} \) and \( \hat{R}_{j1}, \ldots, \hat{R}_{jN_j} \) are both triangular arrays of exchangeable random variables. This fact allows us to apply the central limit theorem of Weber (1980). Specifically

\[ \sqrt{\frac{N_j}{\text{var}(R_{j1})}} (\bar{R}_j - \mathbb{E}[R_{j1}]) = O_p(1) \quad \text{and} \quad \sqrt{\frac{N_j}{\text{var}(\hat{R}_{j1})}} (\hat{R}_j - \mathbb{E}[\hat{R}_{j1}]) = O_p(1). \quad (D1) \]

We now relate these two quantities with \( \hat{R}_{ji} = R_{ji} + \mathcal{E}_{ji} \), where

\[ \mathcal{E}_{ji} = \sum_{\ell=1}^{J} \sum_{m=1}^{N_j} I(B_{ji,\ell m}) - I(A_{ji,\ell m}), \]

\[ A_{ji,\ell m} = \left\{ \mathcal{D}(X_{ji}, F^*) \leq \mathcal{D}(X_{\ell m}, F^*) \right\} \cap \left\{ \mathcal{D}(X_{ji}, F_{\ell,N}) > \mathcal{D}(X_{\ell m}, F_{\ell,N}) \right\}. \]
and

\[ B_{ji,\ell m} = \{ D (X_{ji}, F_s) > D (X_{\ell m}, F_s) \} \cap \{ D (X_{ji}, F_{s,N}) \leq D (X_{\ell m}, F_{s,N}) \}. \]

Observe that \( \mathbb{E} [\mathbb{I}(B_{ji,\ell m})] = O(N^{-1/2}) \) and \( \mathbb{E} [\mathbb{I}(A_{ji,\ell m})] = O(N^{-1/2}) \) due to Assumptions 1–2; see the paragraph between A4 and A5 of Chenouri, Mozaffari & Rice (2020b) where it is noted that Assumption 1 implies that \( \Pr(|D (X_{ji}; F_s) - D (X_{k1}; F_s)| \leq \nu) \) is Lipschitz in \( \nu \). It then follows that \( \mathbb{E}[\mathcal{E}_{ji}] = \mathbb{E}[\hat{R}_{ji}] - \mathbb{E}[R_{ji}] = O(N^{1/2}) \). With this fact in mind, we now show that

\[
\text{var}(\hat{R}_{ji})/\text{var}(R_{ji}) = O(1) \quad \text{and} \quad \text{var}(R_{ji})/\bar{\sigma}_N^2 = O(1). \tag{D2}
\]

The latter identity follows easily from the fact that \( N_j/N = \theta_j + o(1) \) and \( \text{var}(R_{ji}) = O(N^2) \) for any \( j \in \{1, \ldots, J\} \) and \( i \in \{1, \ldots, N_j\} \). For the former identity, we can write

\[
\text{var}(\hat{R}_{ji}) = \text{var}(R_{ji} + \mathcal{E}_{ji})
\]

\[
= \text{var}(R_{ji}) + \text{var}(\mathcal{E}_{ji}) + 2 \text{cov}(\mathcal{E}_{ji}, R_{ji}) \\
\leq \text{var}(R_{ji}) + \text{var}(\mathcal{E}_{ji}) + 2 \mathbb{E} [|\mathcal{E}_{ji} - \mathbb{E}[\mathcal{E}_{ji}]|] N \\
= \text{var}(R_{ji}) + \text{var}(\mathcal{E}_{ji}) + O(N^{3/2}) \\
= \text{var}(R_{ji}) + \mathbb{E} \left[ \sum_{\ell=1}^J \sum_{m=1}^{N_j} [\mathbb{I}(B_{ji,\ell m}) - \mathbb{I}(A_{ji,\ell m})] \right]^2 + O(N) + O(N^{3/2}) \\
\leq \text{var}(R_{ji}) + \mathbb{E} \left[ \sum_{\ell=1}^J \sum_{m=1}^{N_j} \mathbb{I}(B_{ji,\ell m}) + \mathbb{I}(A_{ji,\ell m}) \right] + O(N) + O(N^{3/2}) \\
\leq \text{var}(R_{ji}) + O(N^{3/2}),
\]

where the fourth line follows from Equation (A5) of Chenouri, Mozaffari & Rice (2020b). The last line follows from the fact that \( \mathbb{E} [\mathbb{I}(B_{ji,\ell m})] = O(N^{-1/2}) \) and \( \mathbb{E} [\mathbb{I}(A_{ji,\ell m})] = O(N^{-1/2}) \). Now,

\[
\lim_{n \to \infty} \frac{\text{var}(\hat{R}_{ji})}{\text{var}(R_{ji})} = \lim_{n \to \infty} \frac{\text{var}(\hat{R}_{ji})/N^2}{\text{var}(R_{ji})/N^2} = \lim_{n \to \infty} \frac{\text{var}(R_{ji})/N^2 + o(1)}{\text{var}(R_{ji})/N^2} = 1.
\]

It then follows from Slutsky’s lemma, the continuous mapping theorem, and the central limit theorem of Weber (1980) that

\[
\hat{W}_N - W_N = \frac{1}{\hat{\sigma}_N^2} \sum_{j=1}^J N_j \left( \hat{R}_j^2 - \bar{R}_j^2 \right)
\]

\[
= \sum_{j=1}^J \left( \frac{\sqrt{N_j}}{\hat{\sigma}_N} \hat{R}_j \right)^2 - \left( \frac{\sqrt{N_j}}{\hat{\sigma}_N} \bar{R}_j \right)^2
\]
\[
O_p(1) + \frac{1}{\hat{\sigma}_N^2} \sum_{j=1}^{J} N_j \left[ E[\hat{R}_j] - E[R_j] \right]^2 + E[R_j] E[\hat{R}_j] - E[R_j] E[\hat{R}_j] \]
\]
\[
= O_p(1).
\]

The third line results from adding and subtracting \( \sqrt{N_j} E[\hat{R}_j]/\hat{\sigma}_N \) and \( \sqrt{N_j} E[R_j]/\hat{\sigma}_N \) inside the left and right sets of parentheses, respectively, multiplying out, and applying Equations (D1) and (D2).

\( \blacksquare \)

\textbf{Proof of Theorem 3.} In view of Equation (6), it is only necessary to show that, under the alternative hypothesis, (5) holds. First, the function \((1 + a^{1/2})^{-1}\) is monotonic in \(a\), which implies that we can use ranks based on \( E[\|X_{ji} - Z\|^2|X_{ji}] \), where \( Z \sim F_u \). We can then write

\[
\tilde{Y}_{ji} = \left( \|X_{ji} - Z\|^2 \right) X_{ji} = \|X_{ji}\|^2 + E[\|Z\|^2] - E[2 \langle X_{ji}, Z \rangle |X_{ji}] \\
= \|X_{ji}\|^2 + \|\mathcal{K}_Z\|_{TR},
\]

where \( \mathcal{K}_Z \) is the covariance operator corresponding to \( Z \) and the last equality comes from the fact that \( E[Z] = 0 \). We can recentre \( \tilde{Y}_{ji} \) by \( \|\mathcal{K}_Z\|_{TR} \) to use the ranks generated by \( Y_{ji} = \|X_{ji}\|^2 \). Therefore, we have that

\[
E[\|X_{11}\|^2 - \|X_{21}\|^2] = \|\mathcal{K}_1\|_{TR} - \|\mathcal{K}_2\|_{TR} \neq 0.
\]

\( \blacksquare \)

\textbf{Proof of Theorem 2.} First, let \( u \in S \), where \( S = \{u : \|u\| = 1, u \in \mathbb{X}\} \), and let \( Y_{u,j} = \langle X_{j1}, u \rangle \). Observe that \( E[Y_{u,j}] = 0 \). We can define

\[
\sigma_{j,u}^2 = E[Y_{u,j}^2] = E \left[ \int_{[0,1]} \int_{[0,1]} X_{j1}(t)u(t) \cdot X_{j1}(s)u(s) \, ds \, dt \right] = \langle \mathcal{K}_u, u, u \rangle,
\]

where we can take the expectation inside the integrals because of Lebesgue’s dominated convergence theorem. Namely \( \langle X_{j1}, u \rangle^2 \leq \|X_{j1}\|^2 \), which has a finite expectation. One should also recall that in \( \text{RPD}_\infty \), we take

\[
D \left( (x, u); F_u \right) = F_u((x, u))(1 - F_u((x, u))
\]

as the univariate depth. For the remainder of the proof we suppress the \( \infty \) in \( \text{RPD}_\infty \) for brevity. Now, following the same argument as in the first paragraph of the proof of Theorem 3, it is only necessary to verify that

\[
\Pr(\text{RPD}(X_{11}; F_u) > \text{RPD}(X_{21}; F_u)) \neq \frac{1}{2}.
\]

Under the conditions of the theorem, this is equivalent to showing that

\[
E[\text{RPD}(X_{11}; F_u) - \text{RPD}(X_{21}; F_u)] \neq 0.
\]
We can write
\[ E \left[ \text{RPD}(X_{11}; F_u) - \text{RPD}(X_{21}; F_u) \right] = E \left[ \int_S D \left( Y_{u,1} ; F_{u,v} \right) \, d\nu(u) - \int_S D \left( Y_{u,2} ; F_{u,v} \right) \, d\nu(u) \right] \]
or, equivalently
\[ E \left[ \text{RPD}(X_{11}; F_u) - \text{RPD}(X_{21}; F_u) \right] = E \left[ \int_S \left[ F_{u,v}(Y_{u,1})(1 - F_{u,v}(Y_{u,1})) - F_{u,v}(Y_{u,2})(1 - F_{u,v}(Y_{u,2})) \right] \, d\nu(u) \right]. \]

Clearly, since 0 < \( F_{u,v}(Y_{u,1}) < 1 \)
\[ F_{u,v}(Y_{u,1})(1 - F_{u,v}(Y_{u,1})) - F_{u,v}(Y_{u,2})(1 - F_{u,v}(Y_{u,2})) \leq 1/2. \]

By Lebesgue’s dominated convergence theorem
\[ E \left[ \text{RPD}(X_{11}; F_u) - \text{RPD}(X_{21}; F_u) \right] = \int_S E \left[ F_{u,v}(Y_{u,1})(1 - F_{u,v}(Y_{u,1})) \right] - E \left[ F_{u,v}(Y_{u,2})(1 - F_{u,v}(Y_{u,2})) \right] \, d\nu(u). \]

Using the fact that \( F_{u,v} \) is thrice differentiable for all \( u \), we can write
\[ E \left[ F_{u,v}(Y_{u,j}) \right] = F_{u,v}(0) + \frac{1}{2} f_{u,v}^{(1)}(0)\sigma^2_{j,u} + R_{u,j,1} \]
and
\[ E \left[ F_{u,v}^2(Y_{u,j}) \right] = F_{u,v}^2(0) + (F_{u,v}(0)f_{u,v}^{(1)}(0) + f_{u,v}^2(0))\sigma^2_{j,u} + R_{u,j,2}, \]
with
\[ R_{u,j,1} = E \left[ \frac{1}{6} \int_0^{Y_{u,j}} f_{u,v}^{(2)}(t)(Y_{u,j} - t)^3 \, dt \right] \]
and
\[ R_{u,j,2} = E \left[ \frac{1}{3} \int_0^{Y_{u,j}} (3 f_{u,v}(t)f_{u,v}^{(1)}(t) + F_{u,v}(t)f_{u,v}^{(2)}(t))(Y_{u,j} - t)^3 \, dt \right]. \]

Letting \( R_{u,j,3} = R_{u,j,1} - R_{u,j,2} \), it follows that
\[ E \left[ D \left( Y_{u,j} ; F_{u,v} \right) \right] = F_{u,v}(0) + \frac{1}{2} f_{u,v}^{(1)}(0)\sigma^2_{j,u} - F_{u,v}^2(0) - (F_{u,v}(0)f_{u,v}^{(1)}(0) - f_{u,v}^2(0))\sigma^2_{j,u} + R_{u,j,3} \]
\[ = h(F_{u,v})\sigma^2_{j,u} + F_{u,v}(0) - F_{u,v}^2(0) + R_{u,j,3}, \]
where
\[ h(F) = \frac{1}{2} f^{(1)}(0) - (F(0)f^{(1)}(0) - f^2(0)). \]

We can now write
\[ E \left[ D \left( Y_{u,1} ; F_{u,v} \right) - D \left( Y_{u,2} ; F_{u,v} \right) \right] = h(F_{u,v})(\sigma^2_{1,u} - \sigma^2_{2,u}) + R_{u,1,3} - R_{u,2,3}. \]
To conclude, under univariate simplicial depth

$$
E \left[ \text{RPD}(X_{11}; F_*) - \text{RPD}(X_{21}; F_*) \right] = \int_S h(F_{u,*})(\sigma^2_{1,u} - \sigma^2_{2,u}) + R_{u,1.3} - R_{u,2.3} \, dv(u)
$$

$$
= \int_S h(F_{u,*})(\mathcal{K}_1 u - \mathcal{K}_2 u, u) + R_{u,1.3} - R_{u,2.3} \, dv(u)
$$

$$
= \int_S h(F_{u,*})(\mathcal{K}_1 u - \mathcal{K}_2 u, u)dv(u) + R_1 - R_2,
$$

where $R_j < \infty$ by the fact that the integrand is bounded in $u$.

\(\square\)

**Proof of Theorem 4.** In view of the proof of Theorem 3, we can use the fact that LTR-depth-based ranks are equivalent to ranks generated by $Y_{ji} = ||X_{ji}||^2$. Now, the $Y_{ji}$s are univariate observations from a scale family, meaning that $Z_{ji} = \left(1 + \delta_j / \sqrt{N}\right)^{-1} Y_{ji}$ with $Z_{ji} \sim G$. Now, let $\tau = \lim_{N \to \infty} \tau_N$. It follows from Fan, Zhang & Zhang (2011) that $W_N \Rightarrow \chi^2_{J-1}(\tau)$. We have that

$$
\tau_N = \frac{12}{N(N+1)} \sum_{j=1}^J N_j \left[ N \sum_{k \neq j} \theta_k \left( \Pr(Y_{k1} \leq Y_{j1}) - 1/2 \right) \right]^2.
$$

Let $b_{jk} = (\sqrt{N} + \delta_j)/\sqrt{N} + \delta_k$. We have that

$$
\Pr(Y_{k1} \leq Y_{j1}) = \Pr(Z_{k1} \leq Z_{j1} b_{jk}) = \int_{\mathbb{R}} \Pr(Z_{k1} \leq z b_{jk}) g(z)dz.
$$

(D3)

Now, note that $z b_{jk}$ is in a neighbourhood of $z$. A Taylor expansion of $G$ about $z$ at the point $z b_{jk}$ is

$$
\Pr(Z_{k1} \leq z b_{jk}) = G(z) + z \left[1 - b_{jk}\right] g(z) + O(N^{-1}).
$$

Substitution into Equation (D3) yields

$$
\int_{\mathbb{R}} \Pr(Z_{k1} \leq z b_{jk}) g(z)dz = \int_{\mathbb{R}} \left[ G(z) + z \left[1 - b_{jk}\right] g(z) + O(N^{-1}) \right] dG(z)
$$

$$
= \frac{1}{2} + \left[1 - b_{jk}\right] \int_{\mathbb{R}} zg(z)^2dz + O(N^{-1})
$$

$$
= \frac{1}{2} + \left[1 - b_{jk}\right] \Delta_G + O(N^{-1}),
$$

where

$$
\Delta_G = \int_{\mathbb{R}} zg(z)^2dz.
$$

Now, substituting the above identity into Equation (8) gives

$$
\tau_N = \frac{12}{N(N+1)} \sum_{j=1}^J N_j \left[ N \sum_{k \neq j} \theta_k \left( \left[1 - b_{jk}\right] \Delta_G + O(N^{-1}) \right) \right]^2.
$$

DOI: 10.1002/cjs.11767

The Canadian Journal of Statistics / La revue canadienne de statistique
which then immediately implies that

$$\lim_{N \to \infty} \tau_N = 12 \Delta^2 G \sum_{j=1}^{J} \theta_j \left( \delta_j - \bar{\delta} \right)^2,$$

which completes the proof.

\[\square\]

Received 25 January 2022
Accepted 5 October 2022