

# A ONE-WAY ANOVA TEST FOR FUNCTIONAL DATA WITH GRAPHICAL INTERPRETATION

TOMÁŠ MRKVIČKA, MARI MYLLYMÄKI, MILAN JÍLEK AND UTE HAHN

A new functional ANOVA test, with a graphical interpretation of the result, is presented. The test is an extension of the global envelope test introduced by Myllymäki et al. (2017, Global envelope tests for spatial processes, *J. R. Statist. Soc. B* 79, 381–404, doi: 10.1111/rssb.12172). The graphical interpretation is realized by a global envelope which is drawn jointly for all samples of functions. If a mean function computed from the empirical data is out of the given envelope, the null hypothesis is rejected with the predetermined significance level  $\alpha$ . The advantages of the proposed one-way functional ANOVA are that it identifies the domains of the functions which are responsible for the potential rejection. We introduce two versions of this test: the first gives a graphical interpretation of the test results in the original space of the functions and the second immediately offers a post-hoc test by identifying the significant pair-wise differences between groups. The proposed tests rely on discretization of the functions, therefore the tests are also applicable in the multidimensional ANOVA problem. In the empirical part of the article, we demonstrate the use of the method by analyzing fiscal decentralization in European countries.

*Keywords:* global envelope test, groups comparison, permutation test, Europe, fiscal decentralization, nonparametrical methods

*Classification:* 62H15, 62G10

## 1. INTRODUCTION

Functional data appear in a number of scientific fields, where the process of interest is monitored continuously. Those include e. g. monitoring of the share price, the temperature in a given location or monitoring of a body characteristic. A classical statistical problem is to decide, if there exist differences between the groups of functions (e. g. the control and treatment group). This problem is usually solved by determining the differences among the mean group functions and then we deal with a one-way functional ANOVA problem.

The functional ANOVA problem, both one-way and more complex designs, was previously studied by many authors. For example, [10] introduced an asymptotic version of the ANOVA  $F$ -test, and [37] considered asymptotic or bootstrapped versions of a  $L_2$ -norm based test,  $F$ -type statistic based test and globalizing pointwise  $F$ -test. Fur-

ther, [15] introduced a method based on a basis function representation, [28] described a bootstrap procedure based on pointwise  $F$ -tests, [1] used wavelet smoothing techniques, and [14] used a dimension reduction approach. Furthermore, [9] applied the  $F$ -test on several random univariate projections and bound the tests together through the false discovery rate (FDR). There is also the possibility to transform the functions into single numbers and use the classical ANOVA, but such a procedure can be blind against some alternatives.

Furthermore, nonparametric permutation procedures have been used to address this problem. [17] used a one-dimensional integral deviation statistic to summarize the deviance between groups. Its distribution was obtained by permuting the functions. [23] based the test either on certain pointwise statistics, such as the  $F$ -statistic, and found the distribution of its maxima by permutation, or alternatively used the size of the area which is given by exceeding some given threshold. Since these statistics need to satisfy the homogeneity across the functional domain, [25] recommended to concentrate on the  $p$ -values which are implicitly homogeneous across the domain and find the distribution of its minima by permutation. This  $p$ -min and also  $F$ -max methods are able to identify the regions of rejections by identifying a threshold of the statistics of the interest.

To identify the regions of rejections also other methods were developed in the literature. [8] developed a method similar to the  $p$ -min procedure, but specifically designed for functional data. They applied the Westfall–Young randomization method to correct for multiple testing. A global  $p$ -value can not be obtained for this method. [36] partitioned the domain of interest and applied multiple testing correction on the individual parts of the domain. A disadvantage is that partition has to be prespecified. [26] developed a functional multi-way ANOVA framework for determining the regions of rejections under the interval wise control of the error rate. This approach controls false rejection on any interval, whereas typically the family wise error rate, i. e. false rejection of any pointwise hypothesis, is controlled. Also the regions with differences between groups can not be determined by this method. [7] defined a graphical representation of confidence bands for functional data inference, which can be used for two functional sample testing.

However, none of the available methods is able to give a graphical interpretation of group specific differences with respect to the null hypothesis or pairwise group differences, together with providing the global  $p$ -value in the family wise error rate sense. This interpretation can help the user to understand what are the reasons of potential rejections, when or where the potential differences appear. Our new proposed method which has such a graphical interpretation is based on the global rank envelope test and the extreme rank length measure introduced in [21]. Here we extend this procedure for the functional ANOVA setting by using permutations to obtain the simulations from the null model which are required for applying the global rank envelope test. We define the test statistics suitable for testing the functional ANOVA null hypothesis. Further, we introduce here new graphical interpretation of the test based on the extreme rank length directly which allows to merge the graphical interpretation with single extreme rank length  $p$ -value rather than with  $p$ -interval as it was merged in the original definition of global rank envelope test. In this work, we concentrate on the one-way functional ANOVA problem only, because in this case and under homoscedasticity the proposed Monte Carlo test is exact, i. e. its type I error is exactly the prescribed significance level

$\alpha$ , under the assumption of all observed functions being from the same distribution. We also define an extension of our method for the heteroscedastic case.

In Section 2 we introduce two versions of our completely nonparametric method for the one-way functional ANOVA problem. The graphical interpretation of the second version of the test also gives an immediate post-hoc test for finding which of the groups differ from each other. Interestingly, this post-hoc comparison is done simultaneously with the ANOVA test, thus at the exact significance level  $\alpha$ . Therefore, no second comparison is needed to find which groups differ.

In Section 2.2 we also describe the global rank envelope test applied to the pointwise  $F$ -statistics. This test does not have its graphical representation in the space of the functions, but we introduce it as another possibility of applying the global envelope test in the functional ANOVA setting. Further, in Section 2.3 we show, how these methods can be used to testing the homoscedasticity.

In Section 3, we present results of a simulation study that was performed to compare powers of our graphical procedures with the powers of the procedures which were already available. In order to be able to compare the performances, we chose to the comparison only such procedures which were available through the software R [27] and which provided the global  $p$ -value in the sense of the family wise error rate. The simulation study was performed only in order to show that our method, which has a unique graphical interpretation, has comparable power with respect to other global methods. All our proposed methods can be found in the R package GET [20, 21].

In Section 4, we apply our methods to the fiscal decentralization issue in European countries. The empirical analysis aims to capture differences in developments of government expenditure decentralization among different groups of European countries. The assumption is, based on the existing literature, that countries with a longer European integration history and therefore presumably with deeper economic and political integration are supposed to decentralize their government expenditure more extensively. We use the government expenditure centralization ratios of 29 European Union and EFTA countries in period from 1995 to 2016 sorted into three groups according to the presumed level of European economic and political integration.

Section 5 is for further discussion.

## 2. GRAPHICAL FUNCTIONAL ANOVA

Let us assume that we have  $J$  groups which contain  $n_1, \dots, n_J$  functions observed on the finite interval  $R = [a, b]$  and denote the functions by  $T_{ij}, j = 1, \dots, J, i = 1, \dots, n_j$ . Assume that  $\{T_{ij} \in L^\infty, i = 1, \dots, n_i\}$  is an i.i.d. sample from a stochastic process  $SP(\mu_j, \gamma_j)$  with a mean function  $\mu_j$  and a covariance function  $\gamma_j(s, t), s, t \in R$  for  $j = 1, \dots, J$ . We want to test the hypothesis

$$H_0 : \mu_1(r) = \dots = \mu_J(r), r \in R.$$

We do not need to specify the stochastic process  $SP$  in order to define our method and thus our method can be taken as completely nonparametric comparison of groups of functions.

The hypothesis  $H_0$  is equivalent to the hypothesis

$$H'_0 : \mu_{j'}(r) - \mu_j(r) = 0, r \in R, j' = 1, \dots, J - 1, j = j', \dots, J.$$

This hypothesis corresponds to the post-hoc test done usually after the ANOVA test is significant.

In the following we introduce the test statistics both for the hypothesis  $H_0$  and  $H'_0$ , first for the case of equal covariance functions (i.e. for the case of  $\gamma_1(s, t) = \dots = \gamma_J(s, t), s, t \in R$ ) and then for the case of unequal variance functions (i.e. for the case of  $\gamma_1(s, t)/\gamma_1(s, s) = \dots = \gamma_J(s, t)/\gamma_J(s, s), s, t \in R$ ) (Sections 2.1 and 2.2). Then we describe how the permutations are performed under the null hypothesis (Section 2.4) and show how the rank envelope test can be used for these test statistics (Section 2.5).

The implementation of our method relies on the discretization of functions. We assume that all functions are discretized in the same way obtaining values at points  $(r_1, \dots, r_K)$ . If this is not the case, we have to apply smoothing techniques (e.g. those described in [37]) and then make such a necessary discretization. In the simulation study (see Section 3) we study our method with respect to increasing denseness of the discretization. Remark here, that the discretization can be arbitrary and equidistances are not required.

### 2.1. Test vectors

The hypothesis  $H_0$  can be tested by the test vector consisting of the mean of functions in the first group followed by the mean of test functions in the second group, etc. Shortly, the test vector is

$$\mathbf{T} = (\bar{T}_1(\mathbf{r}), \bar{T}_2(\mathbf{r}), \dots, \bar{T}_J(\mathbf{r})), \tag{1}$$

where  $\bar{T}_j(\mathbf{r}) = (\bar{T}_j(r_1), \dots, \bar{T}_j(r_K))$ . Thus, the length of the test vector becomes  $J \times K$ , where  $K$  stands for the number of  $r$  values to which the functions are discretized.

The hypothesis  $H'_0$  can be tested by the test vector consisting of the differences of the group means of functions, i.e. the test vector is

$$\mathbf{T}' = (\bar{T}_1(\mathbf{r}) - \bar{T}_2(\mathbf{r}), \bar{T}_1(\mathbf{r}) - \bar{T}_3(\mathbf{r}), \dots, \bar{T}_{J-1}(\mathbf{r}) - \bar{T}_J(\mathbf{r})). \tag{2}$$

Here the length of the test vector becomes  $J(J - 1)/2 \times K$ .

#### 2.1.1. Correction for an unequal variances

To deal with different variances of functions in different groups, we consider the rescaled functions  $S_{ij}$  instead of the original functions  $T_{ij}$ ,

$$S_{ij}(r) = \frac{T_{ij}(r) - \bar{T}_j(r)}{\sqrt{\text{Var}(T_j(r))}} \cdot \sqrt{\text{Var}(T(r))} + \bar{T}_j(r), \tag{3}$$

where the group sample mean  $\bar{T}_j(r)$  and overall sample variance  $\text{Var}(T(r))$  are involved to keep the mean and variability of the functions at the original scale. The group sample variance  $\text{Var}(T_j(r))$  corrects the unequal variances.

For small samples, the sample variance estimators can have big variance, which may influence the test procedure undesirably. In order to deal with this problem the variances can be smoothed by applying moving average (MA) to the estimated variance with

a chosen window size  $b$ . Thus, the rescaled functions take the form

$$S_{ij}(r) = \frac{T_{ij}(r) - \bar{T}_j(r)}{\sqrt{\text{MA}_b(\text{Var}(T_j(r)))}} \cdot \sqrt{\text{MA}_b(\text{Var}(T(r)))} + \bar{T}_j(r). \quad (4)$$

After transformation, the test vectors are composed in the same way as in the case of the equal covariance functions but with rescaled functions:

$$\mathbf{T}_s = (\bar{S}_1(\mathbf{r}), \bar{S}_2(\mathbf{r}), \dots, \bar{S}_J(\mathbf{r})). \quad (5)$$

where  $\bar{S}_j(\mathbf{r}) = (\bar{S}_j(r_1), \dots, \bar{S}_j(r_K))$ , and

$$\mathbf{T}'_s = (\bar{S}_1(\mathbf{r}) - \bar{S}_2(\mathbf{r}), \bar{S}_1(\mathbf{r}) - \bar{S}_3(\mathbf{r}), \dots, \bar{S}_{J-1}(\mathbf{r}) - \bar{S}_J(\mathbf{r})). \quad (6)$$

## 2.2. Rank envelope $F$ -type test

When a graphical interpretation for group specific differences is not of interest but the area of rejection is, one can utilize the  $F$ -type test for each  $r \in R$  separately and form the test vector from the  $r$ -wise  $F$ -statistics,

$$\mathbf{T}_F = (F(r_1), F(r_2), \dots, F(r_K)),$$

where  $F(r_k)$  stands for the  $F$ -statistic computed from the univariate ANOVA model at point  $r_k$ . In this case the correction for unequal variances can be done by choosing the variance corrected  $F$ -statistic.

## 2.3. Homoscedasticity tests

Following the Levene's test of homoscedasticity, it is possible to test the equality of variances in functional ANOVA design by setting the test vector  $\mathbf{T}_V = \mathbf{T}$ ,  $\mathbf{T}_V = \mathbf{T}'$  or  $\mathbf{T}_V = \mathbf{T}_F$ , where instead of the original functions  $T_{ij}$ , the functions

$$Z_{ij}(r) = |T_{ij}(r) - \bar{T}_j(r)| \quad (7)$$

are used. Similarly, in order to test the equality of lag  $s$  covariance, it is possible to set the test vector  $\mathbf{T}_C = \mathbf{T}$ ,  $\mathbf{T}_C = \mathbf{T}'$  or  $\mathbf{T}_C = \mathbf{T}_F$ , where instead of the original functions  $T_{ij}$ , the functions

$$W_{ij}(r) = \sqrt{|(T_{ij}(r) - \bar{T}_j(r))(T_{ij}(r+s) - \bar{T}_j(r+s))|} \cdot \text{sign}[(T_{ij}(r) - \bar{T}_j(r))(T_{ij}(r+s) - \bar{T}_j(r+s))], \quad (8)$$

for  $r \in [a, b-s]$ , are used.

## 2.4. Permutations and exchangeability of the test vectors

The most important aspect of the permutation tests is the manner in which data are shuffled under the null hypothesis. In all our one-way ANOVA tests, we perform the simple permutation of raw functions among the groups. That is, if  $G$  is a vector of

group indices of length  $N = \sum_{j=1}^J n_j$ , then the permutation  $\mathbf{P}$  is  $N \times N$  matrix that has all elements being either 0 or 1, each row and column having exactly one 1. Pre-multiplying the group indices  $G$  by the matrix  $\mathbf{P}$  permutes the group indices. Note that the possible correction for unequal variances is performed prior to the permutation and the permutations are consequently done for the rescaled functions (3) or (4).

We say that a test vector  $\mathbf{T}$  is exchangeable if the observed and simulated (permuted) test vectors  $\mathbf{T}^1, \dots, \mathbf{T}^s$  are exchangeable, i.e. the joint distribution of  $\mathbf{T}^1, \dots, \mathbf{T}^s$  is not affected by permutation.

**Proposition 2.1.** Under the assumption that all the functions  $T_{ij}, j = 1, \dots, J, i = 1, \dots, n_j$  follow the same stochastic process, the test vectors  $\mathbf{T}, \mathbf{T}', \mathbf{T}_F$  and also  $\mathbf{T}_V, \mathbf{T}_C$  are exchangeable for permutations  $\mathbf{P}$ . Under the assumption of normality of stochastic processes  $SP(\mu_j, \gamma_j)$  the test vectors  $\mathbf{T}_s, \mathbf{T}'_s$  and  $\mathbf{T}_F$  with variance corrected  $F$ -statistics are asymptotically exchangeable for permutations  $\mathbf{P}$  for the case of unequal variances and the null hypothesis of equal means. The asymptotics is taken over  $\min_{j=1}^J n_j$ .

*Proof.* Since the permutations are performed on the whole functions (i.e. the block permutation scheme is used) and we assume that the functions form an i.i.d. sample from a stochastic process, the joint distribution of  $\mathbf{T} (\mathbf{T}', \mathbf{T}_F, \mathbf{T}_V, \mathbf{T}_C)$  is equal to the joint distribution of  $\mathbf{T} (\mathbf{T}', \mathbf{T}_F, \mathbf{T}_V, \mathbf{T}_C)$  for permuted groups  $\mathbf{P}G$ .

In the case of unequal variances the functions are first scaled by the sample group variance for computation of  $\mathbf{T}_s$  and  $\mathbf{T}'_s$ . The sample group variance  $\text{Var}(\bar{T}_j(r))$  converges a.s. to the true group variance. This holds similarly for the group sample mean and overall sample variance. Thus the stochastic process  $S_{ij}$  converges in distribution to  $SP(\mu, \gamma_j(s, t)\gamma(s, s)/\gamma_j(s, s))$ , where  $\gamma(s, s)$  is the overall variance. Under the null hypotheses of equal means and unequal variances and assumption of normality these stochastic processes are equal and thus the test vectors are asymptotically exchangeable. A similar proof can be made for  $\mathbf{T}_F$  in the case of unequal variances.  $\square$

### 2.5. Global rank envelope test

The idea of the global rank envelope was introduced in [21] for testing in spatial statistics. Further [19] extended the notion of this global envelope for general multivariate test vectors. This extension applies, e.g., to the case where the multivariate vector consists of values of two or more functions at once. We first recall the measures and associated  $p$ -values introduced in [21]. Second, we define the global extreme rank length envelope as a refinement of the global rank envelope.

Assume the general multivariate vector of the form

$$\mathbf{V} = (V_1, \dots, V_d).$$

Let  $\mathbf{V}_1, \dots, \mathbf{V}_s$  be the multivariate vectors generated by permutations under the null hypothesis. Let  $\mathbf{V}_1$  denote the vector obtained by identical permutation.

First we define the extreme rank  $R_i$  of the vector  $\mathbf{V}_i = (V_{i1}, \dots, V_{id})$  as the minimum of its pointwise ranks, namely

$$R_i = \min_{k=1, \dots, d} R_{ik}, \tag{9}$$

where the pointwise rank  $R_{ik}$  is the rank of the element  $V_{ik}$  among the corresponding elements  $V_{1k}, V_{2k}, \dots, V_{sk}$  of the  $s$  vectors such that the lowest ranks correspond to the most extreme values of the statistics. How the pointwise ranks are determined, depends on whether a one-sided or a two-sided envelope test is to be performed: Let  $r_{1k}, r_{2k}, \dots, r_{sk}$  be the raw ranks of  $V_{1k}, V_{2k}, \dots, V_{sk}$ , such that the smallest  $V_{ik}$  has rank 1. In the case of ties, the raw ranks are averaged. The pointwise ranks are then calculated as

$$R_{ik} = \begin{cases} r_{ik}, & \text{for one-sided test, small } V \text{ is considered extreme} \\ s + 1 - r_{ik}, & \text{for one-sided test, large } V \text{ is considered extreme} \\ \min(r_{ik}, s + 1 - r_{ik}), & \text{for two-sided test.} \end{cases} \tag{10}$$

The extreme ranks can contain many ties, e. g. in a one-sided test with  $d$ -variate vectors, up to  $d$  out of the  $s$  vectors can take the rank 1. Therefore we need to break these ties in an efficient way. Ordering of the vectors by the extreme rank length [21] refines the extreme rank ordering in order to minimize the possibility of ties.

Consider the vectors of pointwise ordered ranks  $\mathbf{R}_i = (R_{i[1]}, R_{i[2]}, \dots, R_{i[d]})$ , where  $\{R_{i[1]}, \dots, R_{i[d]}\} = \{R_{i1}, \dots, R_{id}\}$  and  $R_{i[k]} \leq R_{i[k']}$  whenever  $k \leq k'$ . The extreme rank given in (9) corresponds to  $R_i = R_{i[1]}$ . The extreme rank length measure of the vectors  $\mathbf{R}_i$  is equal to

$$R_i^{\text{erl}} = \frac{1}{s} \sum_{i'=1}^s \mathbf{1}(\mathbf{R}_{i'} \prec \mathbf{R}_i), \tag{11}$$

where

$$\mathbf{R}_{i'} \prec \mathbf{R}_i \iff \exists n \leq d : R_{i'[k]} = R_{i[k]} \forall k < n, R_{i'[n]} < R_{i[n]}.$$

We remark here that [22] independently defined the two-sided extreme rank length measure as a functional depth and called it extremal depth.

### 2.5.1. $p$ -values

We distinguish three different  $p$ -values attached to the rank envelope test. All the  $p$ -values are based on Monte Carlo testing principles. The conservative and liberal  $p$ -values are given as

$$p_+ = \sum_{i=1}^s \mathbf{1}(R_i \leq R_1) / s, \quad p_- = \sum_{i=1}^s \mathbf{1}(R_i < R_1) / s. \tag{12}$$

The  $p$ -value based on the extreme rank length ordering is given as

$$p^{\text{erl}} = 1 - \sum_{i=1}^s \mathbf{1}(\mathbf{R}_1 \prec \mathbf{R}_i) / s. \tag{13}$$

According to [21, Proposition 6.1], it holds that  $p_- < p^{\text{erl}} \leq p_+$ .

Note here, that there still can appear some ties between  $\mathbf{R}_1$  and  $\mathbf{R}_i, i = 2, \dots, s$ . However, since these ties are unlikely to happen, we define  $p^{\text{erl}}$  as the conservative  $p$ -value. Alternatively these ties could be broken by randomization.

2.5.2. The new graphical envelope

Myllymäki et al. [21] defined the graphical global envelope with respect to the ordering of the extreme ranks  $R_i, i = 1, \dots, s$ . This ordering can have a lot of ties and consequently the graphical envelope based on this ordering requires a lot of permutations in order to be precise. We eliminate this problem in this paper by defining the graphical envelope with respect to the extreme rank length ordering (11).

Assuming that all the  $\mathbf{V}_i$  follow the same joint distribution, we construct rank envelopes with level  $(1 - \alpha)$  as sets  $\{\mathbf{V}_{\text{low}}^{(\alpha)}, \mathbf{V}_{\text{upp}}^{(\alpha)}\}$  such that, the probability that  $\mathbf{V} = (V_1, \dots, V_d)$  falls outside this envelope in any of the  $d$  points is less or equal to  $\alpha$ ,

$$\Pr(\exists k \in \{1, \dots, d\} : V_k \notin [V_{\text{low } k}^{(\alpha)}, V_{\text{upp } k}^{(\alpha)}]) \leq \alpha.$$

Let  $R_{(\alpha)}^{\text{erl}}$  be the largest value in  $\{R_1^{\text{erl}}, \dots, R_s^{\text{erl}}\}$  for which

$$\sum_{i=1}^s \mathbf{1}(R_i^{\text{erl}} < R_{(\alpha)}^{\text{erl}}) \leq \alpha s, \tag{14}$$

and  $I_\alpha = \{i \in 1, \dots, s : R_i^{\text{erl}} \geq R_{(\alpha)}^{\text{erl}}\}$  be the index set of vectors whose extreme rank length measure is larger than or equal to the critical value  $R_{(\alpha)}^{\text{erl}}$ . Then define

$$V_{\text{low } k}^{(\alpha)} = \min_{i \in I_\alpha} V_{ik}, \quad V_{\text{upp } k}^{(\alpha)} = \max_{i \in I_\alpha} V_{ik}$$

for the two-sided test, following the idea of [22]. For one-sided tests, let  $V_{\text{low } k}^{(\alpha)} = -\infty$  or  $V_{\text{upp } k}^{(\alpha)} = \infty$ , respectively, for all  $k = 1, \dots, d$ . This envelope has the graphical interpretation described in the next subsection.

The following theorem states that inference based on the  $p^{\text{erl}}$  and the global envelope specified by  $V_{\text{low } k}^{(\alpha)}$  and  $V_{\text{upp } k}^{(\alpha)}$  are equivalent. Therefore, we can refer to this envelope as the  $100 \cdot (1 - \alpha)\%$  global extreme rank length envelope.

**Theorem 2.2.** Let  $p^{\text{erl}}$  be as given in (13), and  $V_{\text{low } k}^{(\alpha)}, V_{\text{upp } k}^{(\alpha)}$  define the  $100 \cdot (1 - \alpha)\%$  global extreme rank length envelope. Then, assuming that there are no pointwise ties with probability 1, it holds that:

1.  $V_{1k} < V_{\text{low } k}^{(\alpha)}$  or  $V_{1k} > V_{\text{upp } k}^{(\alpha)}$  for some  $k = 1, \dots, d$  iff  $p^{\text{erl}} \leq \alpha$ , in which case the null hypothesis is rejected;
2.  $V_{\text{low } k}^{(\alpha)} \leq V_{1k} \leq V_{\text{upp } k}^{(\alpha)}$  for all  $k = 1, \dots, d$  iff  $p^{\text{erl}} > \alpha$ , and thus the null hypothesis is not rejected;

**Proof.** According to the definition of  $p^{\text{erl}}$  is  $p^{\text{erl}} \leq \alpha$  iff number of  $\mathbf{R}_i$  smaller or equal to  $\mathbf{R}_1$  is smaller or equal to  $\alpha s$ . That is equivalent, according to the definition of  $R_{(\alpha)}^{\text{erl}}$ , to the  $R_1^{\text{erl}} < R_{(\alpha)}^{\text{erl}}$ . This holds iff  $1 \notin I_\alpha$ , which is equivalent to  $V_{1k} < V_{\text{low } k}^{(\alpha)}$  or  $V_{1k} > V_{\text{upp } k}^{(\alpha)}$  for some  $k = 1, \dots, d$  according to the definition of the extreme rank length envelope. The second part of the proof can be proven equivalently.  $\square$



In the following theorem, we will prove that the global extreme rank length envelope is contained in the global rank envelope. The  $l$ th rank envelope was defined in [21] by

$$V_{\text{low } k}^{(l)} = \min_{i=1, \dots, s}^l V_{ik} \quad \text{and} \quad V_{\text{upp } k}^{(l)} = \max_{i=1, \dots, s}^l V_{ik} \quad \text{for } k = 1, \dots, d, \quad (15)$$

where  $\min^l$  and  $\max^l$  denote the  $l$ th smallest and largest values, respectively, and  $l = 1, 2, \dots, \lfloor (s+1)/2 \rfloor$ .

**Lemma 2.1.** Let

$$W_{\text{low } k}^{(l)} = \min_{i \in I_l} V_{ik}, \quad W_{\text{upp } k}^{(l)} = \max_{i \in I_l} V_{ik}$$

for  $I_l = \{i \in 1, \dots, s : R_i \geq l\}$ . Then  $W_{\text{low } k}^{(l)} \geq V_{\text{low } k}^{(l)}$  and  $W_{\text{upp } k}^{(l)} \leq V_{\text{upp } k}^{(l)}$ .

*Proof.* Since all vectors in  $I_l$  have the extreme rank greater or equal to  $l$  and thus  $R_{ik} \geq l$  or  $R_{ik} \leq s - l + 1$  (for two-sided test), the envelope defined by  $W$  is contained in the envelope defined by  $V$ .  $\square$

**Theorem 2.3.** The  $100 \cdot (1 - \alpha)\%$  global extreme rank length envelope is contained in the  $100 \cdot (1 - \alpha)\%$  global rank envelope.

*Proof.* This is a direct consequence of Lemma 2.1 and the fact that  $I_\alpha$  contains a smaller number of functions than  $I_l$  for  $l = l_\alpha = \max\{l : \sum_{i=1}^s \mathbf{1}(R_i < l) \leq \alpha s\}$ , which is the critical rank for the  $100 \cdot (1 - \alpha)\%$  global rank envelope.  $\square$

Remark here that the  $(l+1)$ th rank envelope given by  $V_{\text{low } k}^{(l+1)}$  and  $V_{\text{upp } k}^{(l+1)}$  is not necessarily contained in the  $l$ th extreme rank length envelope given by  $W_{\text{low } k}^{(l)}$  and  $W_{\text{upp } k}^{(l)}$ .

## 2.6. One-way graphical functional ANOVA test

The proposed tests are performed in three steps. First the test vector is chosen. Second  $s$  permutations are applied to the raw functions (or on the rescaled functions in the case of unequal variance) and the chosen test vector is computed for each permutation. Third the global rank envelope test is applied to the set of  $s$  test vectors. The following theorem specifies the graphical interpretation of our proposed tests and claims the exactness of the graphical method.

**Theorem 2.4.** Consider one-way graphical functional ANOVA test with  $\mathbf{T}$ ,  $\mathbf{T}'$  or  $\mathbf{T}_F$  chosen as the test vector. Assume that all the functions  $T_{ij}, j = 1, \dots, J, i = 1, \dots, n_j$  follow the stochastic process  $SP(\mu_j, \gamma)$ . Let  $p^{\text{erl}}$  be as given in (13), and  $T_{\text{low } k}^\alpha$  and  $T_{\text{upp } k}^\alpha$  define the  $100 \cdot (1 - \alpha)\%$  global extreme rank length envelope. Then, assuming that there are no pointwise ties with probability 1 in the stochastic process  $SP(\mu, \gamma)$ , it holds that:

1.  $T_{1k} < T_{\text{low } k}^\alpha$  or  $T_{1k} > T_{\text{upp } k}^\alpha$  for some  $k$  iff  $p^{\text{erl}} \leq \alpha$ , in which case the null hypothesis is rejected;
2.  $T_{\text{low } k}^\alpha \leq T_{1k} \leq T_{\text{upp } k}^\alpha$  for all  $k$  iff  $p^{\text{erl}} > \alpha$ , and thus the null hypothesis is not rejected.

Theorem 2.4 is direct consequence of Proposition 2.1 and Theorem 2.2. The same theorem holds for the proposed homoscedasticity tests. For the case of unequal variances the above interpretation is due to the Proposition 2.1 achieved only asymptotically with the additional assumption of normality.

Since we apply here the global extreme rank length envelope and not the global rank envelope as it was the case in our previous works, a lower number of permutations can be used. Anyway we recommend to use some thousands of permutations at minimum for repeatability. In case of many groups the number of permutations has to increased in order to not loose the power of the test, as it is demonstrated in the simulation study.

It is important to mention here that the graphical interpretation automatically identifies which groups are responsible for the potential rejection and also it identifies which parts of the functions are responsible for the rejection. This is very important for the interpretation of the result of the test.

Note that for the test vector  $\mathbf{T}_F$  the one-sided rank test has to be used, whereas for the other test vectors the two-sided rank test is used.

## 2.7. Comparison to other permutation methods

The nonparametric permutation methods often used in the brain image statistics are similar to our proposed methods, therefore we would like to stress the differences. The single threshold test [23] of a certain statistic whose maximum is permuted is limited to the statistics that are homogeneous across the functional domain, in order to be sensitive in the whole functional domain and not only in the part of the domain where the functions are the most varying. The  $p$ -min permutation procedure used e.g. in [25] solves this problem. This method can be viewed as our rank envelope  $F$ -type test. However, the  $p$ -min permutation procedure uses the conservative  $p$ -value of our rank envelope  $F$ -type test, i.e. the upper bound of the  $p$ -interval,  $p_+$  in (13). On the other hand, our rank envelope  $F$ -type test is equipped also with the extreme rank length  $p$ -value which solves the problem of ties in the  $p$ -min distribution and therefore it significantly reduces the conservativeness of the test.

Further, our graphical functional ANOVA test gives the graphical interpretation in the original space of functions and for each group of functions, whereas the  $p$ -min test gives it only in the transformed space of  $p$ -values and for all groups simultaneously. Therefore the  $p$ -min test is able only to identify the regions of rejection. Our graphical functional ANOVA test is also equipped with the global extreme rank length envelope which informs the user about the variability of the curves in the study. Finally, the graphical functional ANOVA test is defined here also for combining several post-hoc tests together in one test and therefore it indicates which two groups are different and where they are different.

## 3. SIMULATION STUDY

Our simulation study has four parts. First we compared our methods with some existing methods on a design taken from the study of [10] in order to check if our methods are comparable in power and significance level to the existing methods. We chose methods which were available in the software R, especially in the packages `fda.usc` [13] and

fdANOVA [16], and which are fundamentally different of each other. Second we checked the robustness of the studied methods with respect to heteroscedasticity. Third we changed the design from comparing three groups into comparing ten groups in order to check how much of the power is lost by having a long test vector  $\mathbf{T}$  or  $\mathbf{T}'$  with respect to other procedures. Fourth we studied the dependence of the powers on the level of discretization of the functions. The tests included in our study are listed in Table 1.

Abbr.	Introduced/described in	Test description
AsF	[10]	a bootstrapped version of the asymptotic $F$ -test
RPM	[9]	a random univariate projection method
Fb	[37]	a bootstrapped version of the $F$ -type statistic test
GPF	[37]	a globalizing pointwise $F$ -test
FP	[15]	a method based on a basis function representation
IPT	[17]	a one-dimensional integral permutation test
F-max	[23]	the $F$ -max permutation procedure
p-min	[25]	the $p$ -min permutation procedure
GFAM	here	the graphical functional ANOVA based on the test vector (1) (group means)
GFAC	here	the graphical functional ANOVA based on the test vector (2) (group mean contrasts)
REF	here	the rank envelope $F$ -type test

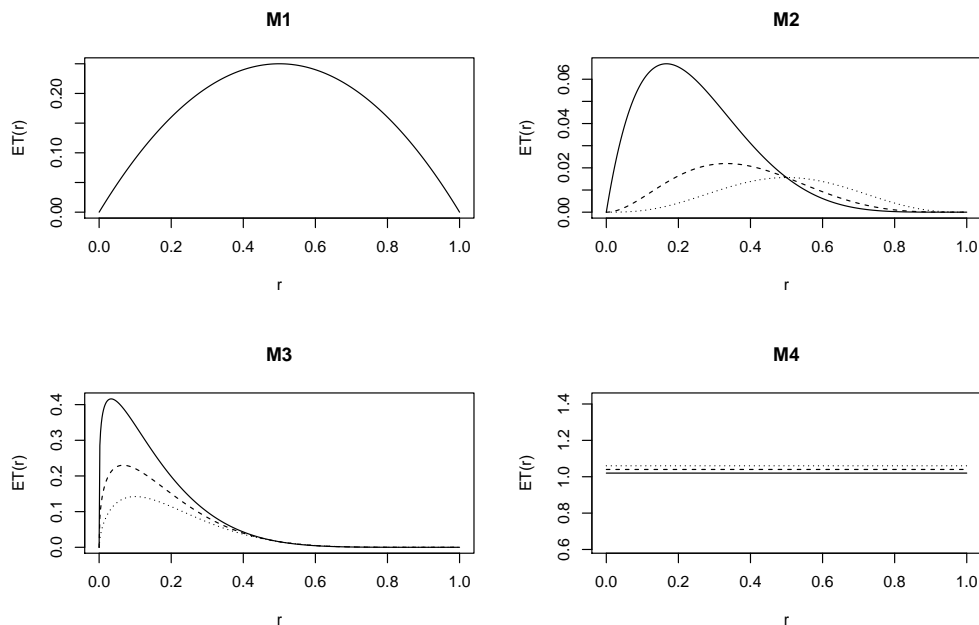
**Tab. 1.** List of tests included in the simulation study with their abbreviations (Abbr.) and short description.

First we compared all the tests of Table 1 using an artificial example of  $J = 3$  groups and  $n = 10$  functions in each group observed in the interval  $[0, 1]$  through 100 evenly spread discrete points. Four different models with two different autocorrelation error structures were considered. The models were

- M1:  $T_{ij}(r) = r(1 - r) + e_{ij}(r)$ ,  $i = 1, 2, 3, j = 1, \dots, 10$ ,
- M2:  $T_{ij}(r) = r^i(1 - r)^{6-i} + e_{ij}(r)$ ,  $i = 1, 2, 3, j = 1, \dots, 10$ ,
- M3:  $T_{ij}(r) = r^{i/5}(1 - r)^{6-i/5} + e_{ij}(r)$ ,  $i = 1, 2, 3, j = 1, \dots, 10$ ,
- M4:  $T_{ij}(r) = 1 + i/50 + e_{ij}(r)$ ,  $i = 1, 2, 3, j = 1, \dots, 10$ .

The mean function for each model and group is shown in Figure 1.

The first autocorrelation structure of errors was modelled by the Gaussian random process with exponential correlation structure with scale parameter equal to 0.1 and standard deviation  $\sigma$ . In the second structure, the errors  $e_{ij}(r)$  were modelled by the Brownian process with dispersion parameter  $\sigma$ . For each combination of the four models and two autocorrelation structures, we considered six different contaminations for the deterministic part of the model given by six different standard deviations  $\sigma_1 = 0.05, \sigma_2 = 0.1, \sigma_3 = 0.15, \sigma_4 = 0.2, \sigma_5 = 0.4, \sigma_6 = 0.8$ . Every standard deviation was twice the



**Fig. 1.** The mean functions of the models M1, M2, M3 and M4. The first group ( $i = 1$ ) corresponds to the solid line, second ( $i = 2$ ) dashed and third ( $i = 3$ ) dotted. In M1, the three lines coincide.

previous one, except  $\sigma_3$  which was added in the middle of  $\sigma_2$  and  $\sigma_4$  in order to increase the sensitivity of the simulation study.

The model M1 corresponds to the situation where  $H_0$  is true. Thus, in this case, the predetermined significance level was estimated. It was set to  $\alpha = 0.05$  in all cases. The other models represent different situations where  $H_0$  is false. The mean functions in the models M2 and M3 have different shape, whereas the mean functions in the model M4 are constant.

Although the permutation tests are exact, they contain variability in sampled permutations and further in the resulted  $p$ -value. [18] recommended to use at least 1000 Monte Carlo replicates in order to control this variability. Therefore, all the tested procedures were run using 2000 Monte Carlo replications or permutations in order to keep the running time manageable and fulfil the above requirement. The RPM was run with 30 random projections, as it was found by [9] to be high enough in a general case, and the false discovery rate  $p$ -value computed out of these projections was used as a final output of this procedure. The extreme rank length  $p$ -value was used as the output of all our new tests (GFAM, GFAC, REF). We performed 1000 simulations for each combination of model, autocorrelation structure and standard deviation, and we computed the proportion of rejections to obtain the estimates of significance levels and powers. All the results are summarized in Table 3 for the Gaussian process cases and in Table 2 for the Brownian error cases.

/ <b>M1</b>	$\sigma_1 = 0.05$	$\sigma_2 = 0.1$	$\sigma_3 = 0.15$	$\sigma_4 = 0.2$	$\sigma_5 = 0.4$	$\sigma_6 = 0.8$
AsF	0.066	0.063	0.060	0.060	0.060	0.069
RPM	0.055	0.039	0.042	0.042	0.038	0.045
Fb	0.033	0.032	0.038	0.032	0.030	0.039
GPF	0.068	0.071	0.068	0.071	0.071	0.077
FP	0.047	0.056	0.050	0.048	0.048	0.054
IPT	0.050	0.044	0.053	0.056	0.049	0.062
F-max	0.058	0.041	0.059	0.050	0.047	0.058
p-min	0.046	0.042	0.050	0.047	0.045	0.057
GFAM	0.047	0.047	0.056	0.050	0.052	0.054
GFAC	0.053	0.050	0.066	0.044	0.049	0.053
REF	0.053	0.037	0.054	0.048	0.048	0.057
<b>M2</b>	$\sigma_1 = 0.05$	$\sigma_2 = 0.1$	$\sigma_3 = 0.15$	$\sigma_4 = 0.2$	$\sigma_5 = 0.4$	$\sigma_6 = 0.8$
AsF	0.660	0.134	0.091	0.080	0.077	0.062
RPM	0.993	0.623	0.255	0.134	0.065	0.049
Fb	0.393	0.065	0.048	0.037	0.042	0.026
GPF	1.000	0.663	0.279	0.171	0.102	0.071
FP	0.645	0.106	0.066	0.064	0.065	0.049
IPT	1.000	0.600	0.227	0.129	0.074	0.049
F-max	1.000	0.958	0.599	0.337	0.110	0.059
p-min	1.000	0.954	0.584	0.330	0.105	0.050
GFAM	1.000	0.949	0.548	0.321	0.112	0.053
GFAC	1.000	0.930	0.540	0.308	0.116	0.048
REF	1.000	0.955	0.598	0.338	0.108	0.053
<b>M3</b>	$\sigma_1 = 0.05$	$\sigma_2 = 0.1$	$\sigma_3 = 0.15$	$\sigma_4 = 0.2$	$\sigma_5 = 0.4$	$\sigma_6 = 0.8$
AsF	1.000	0.506	0.194	0.122	0.073	0.055
RPM	1.000	0.997	0.894	0.652	0.159	0.064
Fb	0.997	0.230	0.098	0.047	0.040	0.031
GPF	1.000	1.000	1.000	0.996	0.329	0.099
FP	1.000	0.455	0.158	0.092	0.057	0.042
IPT	1.000	1.000	1.000	0.994	0.259	0.073
F-max	1.000	1.000	1.000	1.000	1.000	0.608
p-min	1.000	1.000	1.000	1.000	1.000	0.592
GFAM	1.000	1.000	1.000	1.000	0.996	0.436
GFAC	1.000	1.000	1.000	1.000	0.994	0.476
REF	1.000	1.000	1.000	1.000	1.000	0.598
<b>M4</b>	$\sigma_1 = 0.05$	$\sigma_2 = 0.1$	$\sigma_3 = 0.15$	$\sigma_4 = 0.2$	$\sigma_5 = 0.4$	$\sigma_6 = 0.8$
AsF	0.746	0.226	0.132	0.096	0.105	0.063
RPM	0.920	0.288	0.110	0.072	0.060	0.040
Fb	0.590	0.144	0.077	0.060	0.053	0.038
GPF	1.000	0.667	0.295	0.190	0.115	0.074
FP	0.691	0.193	0.116	0.077	0.079	0.049
IPT	1.000	0.596	0.240	0.143	0.091	0.052
F-max	1.000	1.000	0.981	0.808	0.188	0.075
p-min	1.000	1.000	0.978	0.791	0.180	0.068
GFAM	1.000	0.999	0.903	0.628	0.155	0.070
GFAC	1.000	0.997	0.893	0.648	0.146	0.069
REF	1.000	1.000	0.981	0.798	0.187	0.074

**Tab. 2.** The proportions of rejections (at level 0.05) over 1000 runs in the case of Brownian errors for models M1, M2, M3, M4. See text for the model specifications and Table 1 for descriptions of different test abbreviations.

<b>M1</b>	$\sigma_1 = 0.05$	$\sigma_2 = 0.1$	$\sigma_3 = 0.15$	$\sigma_4 = 0.2$	$\sigma_5 = 0.4$	$\sigma_6 = 0.8$
AsF	0.031	0.029	0.023	0.031	0.026	0.035
RPM	0.057	0.061	0.055	0.057	0.077	0.056
Fb	0.011	0.005	0.009	0.009	0.007	0.007
GPF	0.061	0.063	0.048	0.058	0.067	0.062
FP	0.059	0.055	0.040	0.056	0.060	0.050
IPT	0.062	0.061	0.042	0.052	0.060	0.047
F-max	0.058	0.063	0.046	0.051	0.052	0.048
p-min	0.034	0.041	0.025	0.028	0.035	0.029
GFAM	0.062	0.058	0.041	0.046	0.060	0.054
GFAC	0.047	0.061	0.038	0.043	0.058	0.051
REF	0.054	0.062	0.046	0.046	0.055	0.048
<b>M2</b>	$\sigma_1 = 0.05$	$\sigma_2 = 0.1$	$\sigma_3 = 0.15$	$\sigma_4 = 0.2$	$\sigma_5 = 0.4$	$\sigma_6 = 0.8$
AsF	0.615	0.133	0.058	0.048	0.026	0.024
RPM	0.515	0.143	0.088	0.071	0.069	0.062
Fb	0.437	0.055	0.022	0.017	0.007	0.005
GPF	0.729	0.209	0.113	0.086	0.055	0.056
FP	0.717	0.197	0.104	0.084	0.050	0.054
IPT	0.699	0.192	0.104	0.074	0.051	0.054
F-max	0.576	0.128	0.085	0.067	0.051	0.050
p-min	0.486	0.096	0.053	0.043	0.030	0.028
GFAM	0.613	0.166	0.094	0.065	0.063	0.055
GFAC	0.600	0.144	0.084	0.062	0.055	0.057
REF	0.586	0.134	0.084	0.063	0.054	0.057
<b>M3</b>	$\sigma_1 = 0.05$	$\sigma_2 = 0.1$	$\sigma_3 = 0.15$	$\sigma_4 = 0.2$	$\sigma_5 = 0.4$	$\sigma_6 = 0.8$
AsF	1.000	0.479	0.190	0.094	0.035	0.029
RPM	0.991	0.442	0.201	0.118	0.069	0.059
Fb	0.994	0.319	0.086	0.033	0.009	0.006
GPF	1.000	0.610	0.284	0.165	0.073	0.054
FP	1.000	0.615	0.275	0.152	0.065	0.057
IPT	1.000	0.586	0.260	0.150	0.069	0.050
F-max	1.000	0.628	0.230	0.141	0.068	0.050
p-min	1.000	0.527	0.168	0.096	0.041	0.036
GFAM	1.000	0.637	0.260	0.146	0.071	0.057
GFAC	1.000	0.659	0.267	0.160	0.063	0.059
REF	1.000	0.634	0.259	0.138	0.064	0.049
<b>M4</b>	$\sigma_1 = 0.05$	$\sigma_2 = 0.1$	$\sigma_3 = 0.15$	$\sigma_4 = 0.2$	$\sigma_5 = 0.4$	$\sigma_6 = 0.8$
AsF	0.813	0.202	0.086	0.053	0.035	0.025
RPM	0.686	0.195	0.096	0.066	0.061	0.052
Fb	0.694	0.113	0.041	0.018	0.010	0.006
GPF	0.888	0.290	0.134	0.106	0.077	0.054
FP	0.876	0.280	0.131	0.096	0.065	0.052
IPT	0.861	0.265	0.125	0.093	0.069	0.048
F-max	0.548	0.156	0.098	0.069	0.068	0.043
p-min	0.468	0.116	0.061	0.060	0.042	0.029
GFAM	0.617	0.185	0.095	0.081	0.067	0.045
GFAC	0.623	0.181	0.097	0.072	0.067	0.056
REF	0.574	0.162	0.098	0.066	0.067	0.042

**Tab. 3.** The proportions of rejections (at level 0.05) over 1000 runs in the case of Gaussian process errors for models M1, M2, M3, M4. See text for the model specifications and Table 1 for descriptions of different test abbreviations.

The empirical significance level should be in the interval (0.037, 0.064) with the probability 0.95 (given by the 2.5% and 97.5% quantiles of the binomial distribution with parameters 1000 and 0.05). This was satisfied, except for the AsF and Fb procedures in the case of the Gaussian process errors. This exception was caused by the high number of discretized points. For 20 discretized points these methods did not show this feature.

Our three tests had very good power in the Brownian error case and slightly lower power than the best methods (GPF, FP, IPT) in the Gaussian process error case. This different behaviour was obviously caused by the fact that the Gaussian process error is better captured by integral based methods, whereas Brownian error case with increasing variability is better captured by maximum based methods. Surprisingly the REF test did not have greater powers than the RECM and RECMD tests which are fully nonparametric (the errors were normally distributed in the considered models).

In the second part of our simulation study, we studied the robustness of all studied methods to heteroscedasticity. We computed the empirical significance levels of all tests in the case of Gaussian process error structure and a) with standard deviations  $\sigma \cdot 0.5^{(i-1)}$ , where  $i$  is the group indicator, and b) with the same standard deviations  $\sigma \cdot 0.5^{(i-1)}$  and further with scale parameter of the Gaussian error process equal to  $i/10$ . The setting a) corresponds to the case of unequal variances, where our tests are asymptotically exact. In the setting b), the variances as well as covariance structures are unequal. First, we used the tests without any corrections (Table 4 rows 1-9). Second, we explored the corrected versions of our three tests (Table 4 rows 12-14) as well as of IPT and RPM for comparison. For IPT, a variance transformation similar to the transformation (4) employed for GFAM and GFAC tests was used. The RPM method relies instead on the variance correction of the  $F$ -statistic, similarly as REF method. The other methods of the table were not corrected for heteroscedasticity due to the fact that their implementation in R did not support it, but it is shown in [10] that the AsF is robust to heteroscedasticity.

The second part of the study shows that the methods based on the maximum ( $F$ -max,  $p$ -min, GFAM, GFAC, REF) are much more sensitive to the heteroscedasticity than the methods based on the integral principle (Fb, GPF, IPT). The FP method, which is based on the basis representation, was least affected by heteroscedasticity. The AsF and Fb methods were clearly conservative as in the homoscedastic case. Considering the methods with correction for unequal variances, i. e. GFAMU, GFACU, REFU, IPT, RPM (Table 4 rows 10-14), the GFACU test was the least liberal method in the small sample case of ten functions per group. The GFACU test was even less liberal than FP. The inequality of covariance structures did not affect the liberality of the methods. Thus, we conclude that our variance correction by transformation (4) of functions can be used even for small sample sizes with expecting small liberality.

In the third part of our simulation study we took the model M3 and extended it for ten groups, considering

$$M: T_{ij}(r) = r^{i/5}(1-r)^{6-i/5} + e_{ij}(r), \quad i = 2, \dots, 11, j = 1, \dots, 10.$$

We used again the two correlation structures in the model and six levels of contamination as in the first part of the study. Table 5 summarizes the results both for the Gaussian process error (upper part) and the Brownian error case (lower part). The relations

Unequal variances				Unequal variances and covariances			
	$\sigma_1 = 0.05$	$\sigma_2 = 0.1$	$\sigma_3 = 0.15$		$\sigma_1 = 0.05$	$\sigma_2 = 0.1$	$\sigma_3 = 0.15$
AsF	0.031	0.035	0.017	AsF	0.032	0.029	0.032
Fb	0.007	0.008	0.006	Fb	0.011	0.006	0.011
GPF	0.103	0.100	0.084	GPF	0.090	0.081	0.076
FP	0.087	0.095	0.072	FP	0.077	0.073	0.066
F-max	0.282	0.267	0.272	F-max	0.285	0.279	0.272
p-min	0.181	0.184	0.180	p-min	0.208	0.202	0.195
GFAM	0.173	0.163	0.178	GFAM	0.154	0.158	0.147
GFAC	0.204	0.194	0.184	GFAC	0.198	0.186	0.160
REF	0.241	0.226	0.234	REF	0.244	0.247	0.228
RPM	0.107	0.103	0.117	RPM	0.096	0.096	0.101
IPT	0.109	0.116	0.103	IPT	0.104	0.102	0.093
GFAMU	0.111	0.107	0.081	GFAMU	0.067	0.071	0.079
GFACU	0.074	0.063	0.054	GFACU	0.042	0.048	0.030
REFU	0.156	0.168	0.163	REFU	0.144	0.143	0.124

**Tab. 4.** The proportions of rejections (at level 0.05) over 1000 runs for model M1 for the Gaussian error process and heteroscedastic case.

The results shown in last three rows correspond to our three tests with correction for unequal variances. Also the RPM and IPT methods (below the horizontal line) were corrected, while the other tests (above the horizontal line) were used without corrections.

between powers of different methods were the same as in the case of the three groups. Also there was no observable decrease in the power for the GFAM and GFAC methods with respect to the REF and  $p$ -min methods in the Gaussian process error case. On the other hand, there was such decrease in the Brownian error case. This loss of power in the GFAM and GFAC tests can be prevented by increasing number of permutations: We performed the experiment also with 10000 permutations and obtained very similar powers as with 2000 permutations for all the other methods except GFAM and GFAC: the power of GFAM increased from 0.360 to 0.884 and the power of GFAC from 0.360 to 0.688 in the case of  $\sigma_6$  and Brownian errors. Thus the power of the graphical tests was comparable to the power of the REF and  $p$ -min tests with 10000 permutations.

In the last part of the simulation study we studied the powers of all test with respect to increasing discretization of the functions. For this purpose we simulated the model M3 with Gaussian process error with standard deviation equal to 0.15. Table 6 shows the estimated powers for five different levels of discretizations expressed by the number of observations of functions. The powers did not decrease with increasing level of discretization. The only exception was the AsF method, whose power decreased, which is correspondence with our finding in the first part of the study that this method was conservative for resolution with 100 points.



<b>iid</b>	$\sigma_1 = 0.05$	$\sigma_2 = 0.1$	$\sigma_3 = 0.15$	$\sigma_4 = 0.2$	$\sigma_5 = 0.4$	$\sigma_6 = 0.8$
AsF	1.000	0.950	0.493	0.262	0.057	0.045
RPM	1.000	0.982	0.573	0.275	0.106	0.055
Fb	1.000	0.899	0.271	0.065	0.007	0.005
GPF	1.000	0.996	0.743	0.421	0.085	0.092
FP	1.000	0.996	0.725	0.414	0.077	0.071
IPT	1.000	0.990	0.715	0.394	0.102	0.107
F-max	1.000	0.997	0.775	0.408	0.112	0.065
p-min	1.000	0.997	0.704	0.344	0.095	0.049
GFAM	1.000	1.000	0.786	0.449	0.120	0.077
GFAC	1.000	1.000	0.842	0.521	0.119	0.078
REF	1.000	1.000	0.760	0.425	0.106	0.078
<b>Brown</b>	$\sigma_1 = 0.05$	$\sigma_2 = 0.1$	$\sigma_3 = 0.15$	$\sigma_4 = 0.2$	$\sigma_5 = 0.4$	$\sigma_6 = 0.8$
AsF	1.000	0.996	0.656	0.364	0.092	0.064
RPM	1.000	1.000	1.000	0.996	0.544	0.104
Fb	1.000	0.944	0.208	0.148	0.036	0.012
GPF	1.000	1.000	1.000	1.000	0.724	0.148
FP	1.000	1.000	0.564	0.268	0.048	0.052
IPT	1.000	1.000	1.000	1.000	0.656	0.124
F-max	1.000	1.000	1.000	1.000	1.000	0.912
p-min	1.000	1.000	1.000	1.000	1.000	0.892
GFAM	1.000	1.000	1.000	1.000	0.976	0.360
GFAC	1.000	1.000	1.000	1.000	1.000	0.360
REF	1.000	1.000	1.000	1.000	1.000	0.904

**Tab. 5.** The proportions of rejections (at level 0.05) over 1000 runs for model M. The Gaussian error process cases are shown in the upper part and the Brownian error cases are shown in the lower part of the table. See text for the model specification and Table 1 for descriptions of different test abbreviations.

#### 4. FISCAL DECENTRALIZATION EXAMPLE

The topic of fiscal federalism was brought about into the normative theory of public finance in the middle of twentieth century. The main issue was to solve the extent to which fiscal competences and responsibilities should be decentralized from central to sub-central levels of government. The gradual development of the theory of fiscal decentralization led to distinguishing between the first and the second generation theories of fiscal decentralization, as explained in details by [24] and [35].

Generally, there are two types of empirical studies on fiscal decentralization. Within the first type, the concern is in the consequences of fiscal decentralization in terms of economic growth and the growth of public sector. The second type of studies deal with the determinants of fiscal decentralization including a growing body of literature dealing with the issues of globalization, economic and political integration and its consequences on decentralization or secession. Our empirical example deals with the effects of European integration on fiscal decentralization dynamics in individual European countries.

$K$	25	50	100	200	400
AsF	0.157	0.16	0.164	0.160	0.165
RPM	0.183	0.192	0.178	0.19	0.182
Fb	0.07	0.075	0.077	0.075	0.078
GPF	0.254	0.264	0.273	0.276	0.274
FP	0.234	0.248	0.254	0.255	0.250
IPT	0.235	0.248	0.252	0.253	0.253
F-max	0.228	0.252	0.255	0.246	0.236
p-min	0.210	0.196	0.190	0.194	0.033
GFAM	0.234	0.254	0.268	0.267	0.256
GFAC	0.229	0.254	0.268	0.254	0.264
REF	0.233	0.245	0.249	0.238	0.231

**Tab. 6.** The proportions of rejections (at level 0.05) over 1000 runs for model M3, Gaussian process error with standard deviation equal to 0.15 and various levels of discretizations ( $K$  = number of discretized values of functions).

#### 4.1. Decentralization under European economic and political integration

The decentralization has been a characteristic feature of social development in many democratic countries since the last decades of 20th century. The usual presumption is that the federated countries are more decentralized than unitary ones. The process of decentralization, however, is not derived only from the switch to the federal structure, but it is usually more gradual and it is influenced by a number of factors. Even though the institutional (constitutional) changes might be crucial, there are usually many gradual and subtle changes in de facto decentralization. Moreover, as [4] noted, constitutional changes are discrete events which in certain contexts may be difficult culturally and politically to achieve. The gradual changes are more likely to be reflected in a continuous measure, such as the ratio of state and local governments in total general government expenditures or revenue.

Recently, the issue of centralization versus decentralization of government has attracted attention in Europe. On the one hand, there are efforts to further integrate or even federalize the European Union (EU), on the other hand there is resistance to further integration, or even process of (br)exit from the EU (the case of United Kingdom). Also some secession tendencies at the sub-national level of individual countries, i. e. in Spain, Italy, Belgium or the UK, are quite strong [30]. The prominent argument behind these tendencies is an insufficient (fiscal) decentralization. The EU and its member countries are experiencing two parallel tendencies of decentralization and centralization.

[2] claims that there is a continuum between international treaties based integration and the federal constitutions and thus between international organizations and federations. European Community, even before it became the European Union, has moved a long way along the continuum towards the federation. This is true mostly with regards to the capacity to legislate or regulate. The EU budget is still rather small, amounting to approximately one percent of gross national income (GNI). According to [31], the EU

is entity *sui generis*, not a federation, although this might be contested. The EU already has institutions of a federation including a Second Chamber in the form of the European Council and powerful exclusive policy competencies in competition and commerce.

A deep insight into the causes of decentralization or even secession trends is provided by [3]. They formulated the trade-off between the benefits of large jurisdictions and the cost of heterogeneity of large and diverse populations, which determine the size of countries. By reducing the political and economic transaction cost, economic integration extends the size of market and lowers the benefits of large jurisdictions, thus enhancing incentives to secession. They conclude that the democratization (as is in the case of the EU) lead to secessions. Decentralization can however be associated with lower cost and may be preferred over secession [5]. Because the EU enlargements, whilst keeping the European diversity, have gradually created large single internal market, the EU economy as a whole became much less open compared to individual countries. Therefore, we can hypothesize tendency to the European integration conditioned decentralization of government budgets in individual EU countries. Similar conclusions of positive effects of economic, political and social integration on fiscal decentralization were found by [33] and [11] for OECD countries. National governments in the integrating EU therefore run the risk of getting squeezed between the supranational and the sub-national levels of government. Lower government expenditure centralization ratio may be expected in this case.

[29] formulated the positive effect of economic openness on the size of the public sector as well as on the fiscal centralization due to higher government expenditures for centralized redistribution and macroeconomic stabilization. If the economic integration strengthens incentives to fiscally decentralize, under some circumstances, an increased integration may cause fiscal centralization. Such circumstances may come with macroeconomic imbalances in economic crisis.

#### 4.2. Decentralization characteristic, data and hypothesis formulation

To analyze the fiscal decentralization, a suitable characteristic is needed. There are variety of approaches of expressing the fiscal decentralization phenomenon (for complex overview see [32, 34]). This paper follows in principle the approach of [6] and [4], using the ratio of centralization. The advantage of this simple approach is twofold: it avoids the problems with various, complicated and not easily comparable structures of decentralized levels of governments, and it gives the largest dataset.

We use the government expenditure centralization ratio (GEC) in percent. It is the ratio of central government expenditure to the total general government expenditure. Because it includes all kinds of government expenditure (consumption, investment and transfers), it is the most general measure of expenditure decentralization. Data were collected from the [12] database. European countries were selected in order to achieve the maximum size of dataset. Only those countries were included, where the data were available from 1995 to 2016 without interruption. Finally, 29 countries were classified into three groups in the following way:

Group 1: Countries joining EC between 1958 and 1986 (Belgium, Denmark, France, Germany (until 1990 former territory of the FRG), Greece, Ireland, Italy,

Luxembourg, Netherlands, Portugal, Spain, United Kingdom. These countries have long history of European integration, representing the core of integration process.

Group 2: Countries joining the EU in 1995 (Austria, Sweden, Finland) and 2004 (Malta, Cyprus), except CEEC (separate group), plus highly economically integrated non-EU countries, EFTA members (Norway, Switzerland). Countries in this group have been, or in some case even still are standing apart from the integration mainstream. Their level of economic integration is however very high.

Group 3: Central and Eastern European Countries (CEEC), having similar features in political and economic history. The process of economic and political integration have been initiated by political changes in 1990s. CEEC joined the EU in 2004 and 2007 (Bulgaria, Czech Republic, Estonia, Hungary, Latvia, Lithuania, Poland, Romania, Slovakia, Slovenia, data for Croatia joining in 2013 are incomplete, therefore not included).

Finally, based on the discussion presented in Section 4.1, we formulate the null hypothesis that the trend of the fiscal centralization is same in all three groups of countries.

### 4.3. Data analysis

The data were first centred with respect to country average in order to remove the differences in absolute values of GEC between countries and in order to keep the shape of GEC functions. Thus we study the functions

$$CGEC_{ij}(r) = GEC_{ij}(r) - \frac{1}{22} \sum_{r=1995}^{2016} GEC_{ij}(r), j = \{1, 2, 3\}, i = 1, \dots, n_j,$$

where  $n_1 = 12, n_2 = 7, n_3 = 10$  and  $r = 1995, 1996, \dots, 2016$ . The curves are shown in Figure 2.

First we checked the assumption of equality of covariance structure which is required by our tests. This we propose to do by testing the equality of lag 1 covariances. Figure reffig:CovarianceTest shows the result of the REF test applied to the transformed functions (8) ( $p = 0.392$ ).

The next step is to decide if the correction for unequal variances should be used. Figure 4 shows the results of the GFAM test applied to the transformed functions (7). Since the global  $p$ -value is 0.163, we have no evidence that the group variances differ and we therefore prefer to use no correction. Figure 4 also shows the mean absolute deviation across the three groups with the global 95% envelope reflecting the overall variation of the mean absolute deviation among all groups over the years.

Finally, we performed the GFAM and GFAC tests for equality of means. The Figure 5 shows the mean centred GEC functions across the groups and their time developments together with the global 95% envelope reflecting the overall variation of the mean function among all groups. Figure 6 shows the group differences of the mean centred GEC functions and their time developments together with the global 95% envelope reflecting the variation of these differences. Both tests shows the deviation from the null

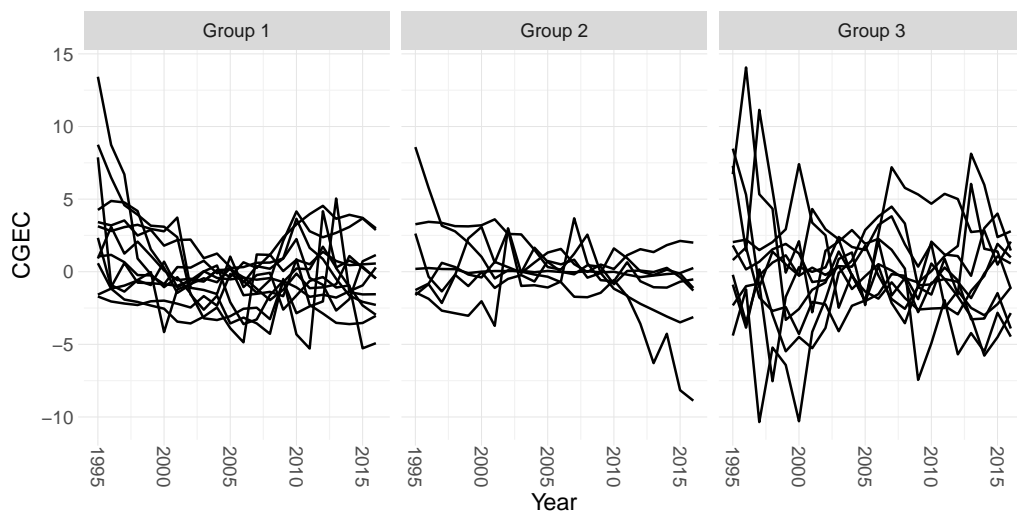


Fig. 2. The centred GEC index for the years from 1995 to 2016 in the three groups.

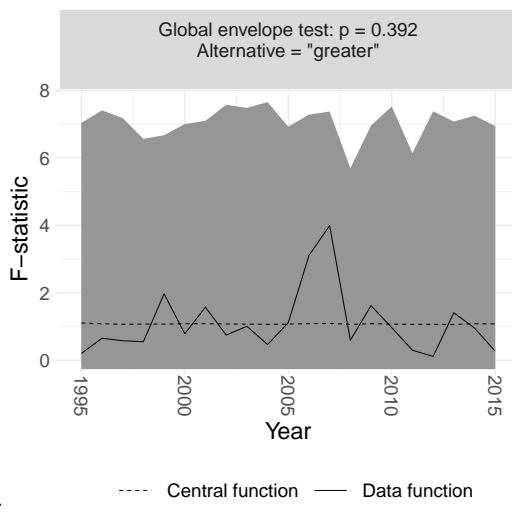
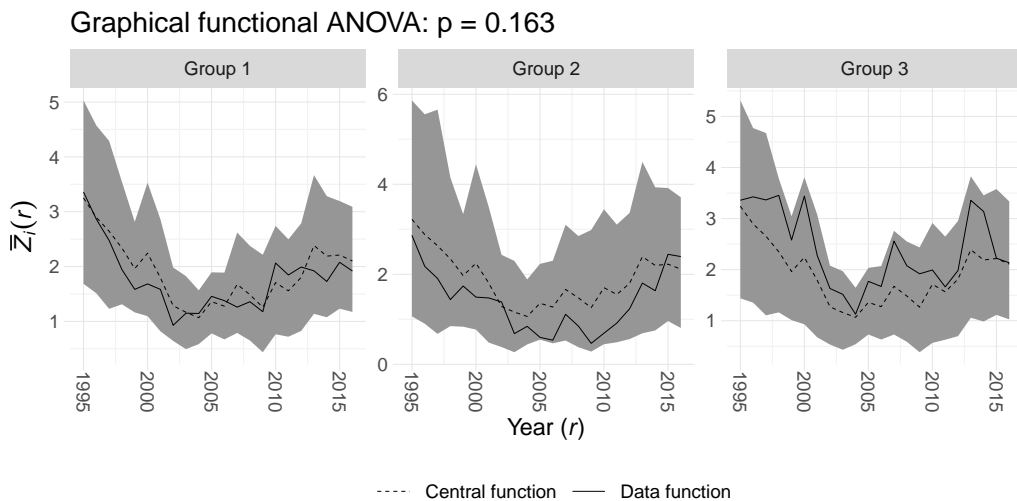


Fig. 3. The test for equality of lag 1 covariances of the centred GEC indices in the three groups using the rank envelope  $F$ -type test (REF).

hypothesis in the year 2006 and the difference of the groups is supported by  $p$ -values equal to 0.045 and 0.021 respectively. According to the GFAM test, the first group is significantly different from the other two. According to the GFAC test in the post-hoc fashion, only groups 1 and 3 significantly differ. All the shown test were performed with 9999 permutations.



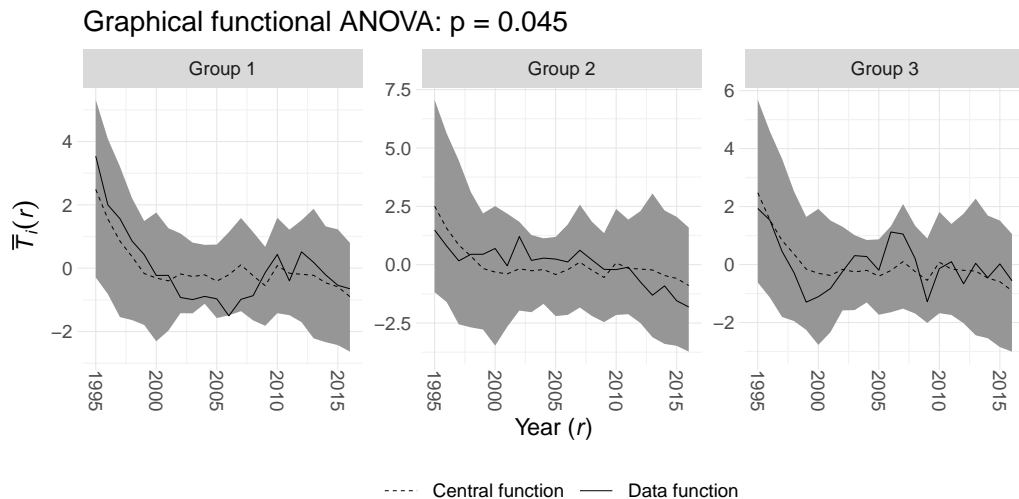
**Fig. 4.** The GFAM test for equality of variances of the centred GEC indices in the three groups.

For comparison, we also performed the other tests of the simulation study (see Table 1). First, our REF test was borderline significant ( $p = 0.055$ ). The data statistic had a peak very close to envelope for the year 2006, but the test was not able to reject the null hypothesis at the strict significance level 0.05 (figure omitted). The  $p$ -values of the other tests were for AsF equal to 0.467, for RPM method equal to 0.375, Fb 0.531, GPF 0.304, FP 0.559, IPT 0.265 and  $F$ -max 0.046. Since the deviation is present mainly only in one year, the integral based methods are not capable to detect this small difference, whereas the maximum type method  $F$ -max and all our methods are able to detect this difference.

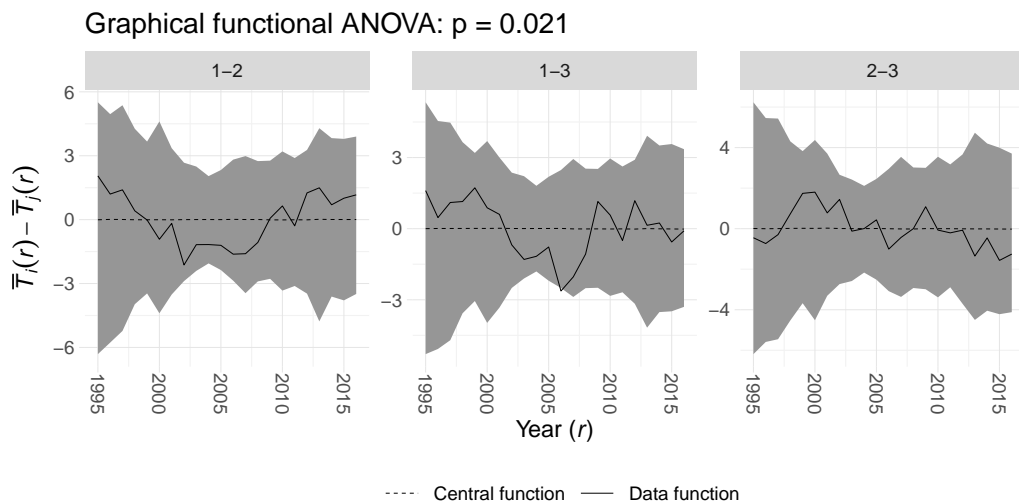
#### 4.4. Interpretation of the tests results

The permutation based tests  $F$ -max, GFAM and GFAC found differences between the groups. The GFAM test (Figure 5) shows greater downward departure from the central function of government expenditure centralization in the period between 2001 and 2008 in the group 1 than in other groups. This tendency is supported by significant difference observed in the year 2006. This period roughly corresponds to the period between Euro introduction and start of financial crises. The tendency of the countries in the first group to centralize after the year 2008 could be caused by the need of central governments to increase the control over the general government budgets at the onset of the crisis. Also the well known phenomenon of rather pro-cyclical fiscal behaviour of sub-national governments supports this argument.

The GFAC test (Figure 6) shows that the difference in trends of mean group government expenditure centralization is realized between groups 1 and 3. This result is consistent with the expectation, because the difference in the depth of integration between these groups is the largest.



**Fig. 5.** The GFAM test for equality of means of the centred GEC indices in the three groups using the group means.



**Fig. 6.** The GFAC test for equality of means of the centred GEC indices in the three groups using the group mean contrasts (comparison of groups 1–2, 1–3 and 2–3).

## 5. DISCUSSION

A new one-way graphical functional ANOVA test was introduced in this paper. Under the assumption that the all the data come from the same distribution, the test has exact type I error. It provides a graphical interpretation which is essential for the interpretation of the results. Also tests and corrections for heteroscedasticity were presented. Two different test vectors can be used leading to two different graphical interpretations. The first option compares every group with the rest of the groups. The second option compares differences between every pair of groups similarly like a Tukey post-hoc test in the univariate ANOVA. A positive side effect of our method is that the post-hoc test is provided together with the main ANOVA procedure at the given significance level.

Since the proposed test works in a highly dimensional multivariate settings, no smoothness of the functions is required. On the other hand, the same discretization of functions is required for every function. If this is not the case, a smoothing technique has to be applied followed by further identical discretization of functions.

Our new methods were compared to the other functional ANOVA procedures, available through the software R, with respect to their power. The new graphical tests had comparable power with respect to other procedures in our simulation study. It was shown that the presented correction for unequal variances, which transforms the functions, can be used even with low sample sizes. Its robustness with respect to unequal covariance structure was also shown. In addition, the proposed methods did not loose their power when the functions were more densely discretized.

Importantly, our simulation study shows that there is no procedure which would be uniformly more powerful than our proposed tests. Therefore, we believe that our tests are useful in practice due to their graphical interpretation and post-hoc nature. As shown by our simulation study, our methods can loose some power when the number of groups to compare is large. In such a case where the test vector is very long, the number of permutations has to be increased in order to eliminate this problem.

Our new tests were designed for the one-way functional ANOVA design. A question of our future research is how these procedures can be extended into multi-way design. Since the permutation of the functional residuals leads to a liberal method, the problem has to be solved in a more complex way.

Our methods show graphically the region of rejections in the family wise error rate. The recent method of [26] can also show the regions of rejections but in the sense of interval wise control of the error rate. We plan to compare these two approaches in the future.

## ACKNOWLEDGEMENTS

Mrkvička has been financially supported by the Grant Agency of Czech Republic (Project Nos. 16-03708S and 19-04412S) and Mari Myllymäki by the Academy of Finland (project numbers 295100, 306875). Hahn's research has been supported by the Centre for Stochastic Geometry and Advanced Bioimaging, funded by the Villum Foundation.



## REFERENCES

- 
- [1] F. Abramovich and C. Angelini: Testing in mixed-effects FANOVA models. *J. Statist. Planning Inference* *136* (2006), 4326–4348. DOI:10.1016/j.jspi.2005.06.002
- [2] B. Ackerman: The rise of world constitutionalism. *Virginia Law Rev.* *83* (1997), 771–797. DOI:10.2307/1073748
- [3] A. Alesina and E. Spolaore: On the number and size of nations. *Quarterly J. Econom.* *112* (1997), 1027–1056. DOI:10.1162/003355300555411
- [4] M. Arzaghi and J.V. Henderson: Why countries are fiscally decentralizing. *J. Public Econom.* *89* (2005), 1157–1189. DOI:10.1016/j.jpubeco.2003.10.009
- [5] P. Bolton and G. Roland: The breakup of nations: A political economy analysis. *Quarterly J. Econom.* *112* (1997), 1057–1090. DOI:10.1162/003355300555420
- [6] F. Cerniglia: Decentralization in the public sector: quantitative aspects in federal and unitary countries. *J. Pol. Model.* *25* (2003), 749–776. DOI:10.1016/s0161-8938(03)00069-3
- [7] H. Choi and M. Reimherr: A geometric approach to confidence regions and bands for functional parameters. *J. Royal Statist. Soc.: Series B (Statist. Methodology)* *80* (2018), 239–260. DOI:10.1111/rssb.12239
- [8] D.D. Cox and J.S. Lee: Pointwise testing with functional data using the Westfall–Young randomization method. *Biometrika* *95* (2008), 621–634. DOI:10.1093/biomet/asn021
- [9] J. Cuesta-Albertos and M. Febrero-Bande: A simple multiway ANOVA for functional data. *Test* *19* (2010), 537–557. DOI:10.1007/s11749-010-0185-3
- [10] A. Cuevas, M. Febrero and R. Fraiman: An anova test for functional data. *Computat. Statist. Data Analysis* *47* (2004), 111–122. DOI:10.1016/j.csda.2003.10.021
- [11] B. Ermini and R. Santolini: Does globalization matter on fiscal decentralization? New evidence from the OECD. *Global Econom. Rev.* *43* (2014), 153–183. DOI:10.1080/1226508x.2014.920240
- [12] Eurostat 2018: Government revenue, Expenditure and Main Aggregates (gov\_10a\_main). Retrieved from <https://ec.europa.eu/eurostat/data/database> (26/10/2018).
- [13] M. Febrero-Bande and M. Oviedo de la Fuente: Statistical computing in functional data analysis: The R package *fda.usc*. *J. Statist. Software* *51* (2012), 1–28. DOI:10.18637/jss.v051.i04
- [14] F. Ferraty, P. Vieu, and S. Vignier-Pla: Factor-based comparison of groups of curves. *Computat. Statist. Data Analysis* *51* (2007), 4903–4910. DOI:10.1016/j.csda.2006.10.001
- [15] T. Górecki and L. Smaga: A comparison of tests for the one-way ANOVA problem for functional data. *Comput. Statist.* *30* (2015), 987–1010. DOI:10.1007/s00180-015-0555-0
- [16] T. Górecki and L. Smaga: *fdANOVA: Analysis of Variance for Univariate and Multivariate Functional Data*, R package version 0.1.0, 2017.
- [17] U. Hahn: A studentized permutation test for the comparison of spatial point patterns. *J. Amer. Statist. Assoc.* *107* (2012), 754–764. DOI:10.1080/01621459.2012.688463
- [18] N.B. Loosmore and E.D. Ford: Statistical inference using the G or K point pattern spatial statistics. *Ecology* *87* (2006), 1925–1931. DOI:10.1890/0012-9658(2006)87[1925:siutgo]2.0.co;2
- [19] T. Mrkvička, M. Myllymäki, and U. Hahn: Multiple Monte Carlo testing, with applications in spatial point processes. *Statist. Comput.* *27* (2017), 1239–1255. DOI:10.1007/s11222-016-9683-9

- [20] M. Myllymäki and T. Mrkvička: GET: Global envelopes in R. arXiv:1911.06583 [stat.ME], 2019.
- [21] M. Myllymäki, T. Mrkvička, P. Grabarnik, H. Seijo, and U. Hahn: Global envelope tests for spatial processes. *J. Royal Statist. Soc. B* 79 (2017), 381–404. DOI:10.1111/rssb.12172
- [22] N.-N. Narisetty V. J. and Nair: Extremal depth for functional data and applications. *J. Amer. Statist. Assoc.* 111 (2016), 1705–1714. DOI:10.1080/01621459.2015.1110033
- [23] T. E. Nichols and E. Holmes: Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain M* 15 (2001), 1–25. DOI:10.1002/hbm.1058
- [24] Oates, E. Wallace, E: Toward A second-generation theory of fiscal federalism. *Int. Tax Public Finance* 12 (2005), 349–373. DOI:10.1007/s10797-005-1619-9
- [25] D. Pantazis, T. E. Nichols, S. Baillet, and R. M. Leahya: A comparison of random field theory and permutation methods for the statistical analysis of MEG data. *Neuroi* 25 (2005), 383–394. DOI:10.1016/j.neuroimage.2004.09.040
- [26] A. Pini, S. Vantini, B. M. Colosimo, and M. Grasso: Domain-selective functional analysis of variance for supervised statistical profile monitoring of signal data. *J. Royal Statist. Soc.: Series C (Appl. Statist.)* 67 (2001), 55–81. DOI:10.1111/rssc.12218
- [27] R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing. Vienna 2019.
- [28] J. Ramsay and B. Silverman: Functional Data Analysis. Second edition. Springer Series in Statistics, Springer 2006. DOI:10.1007/978-1-4757-7107-7
- [29] D. Rodrik: Why do more open economies have bigger governments? *J. Polit. Economy* 106 (1998), 997–1032. DOI:10.1086/250038
- [30] J. Sedova, H. Lipovska, and Fischer: Fiscal autonomy in the secessionist regions. In: *Current Trends in Public Sector Research*, Masaryk University, Brno 2017.
- [31] P. B. Spahn: Contract Federalisms. Edward Elgar Publishing Limited, Book Section 7, Cheltenham 2015, pp. 144–160.
- [32] D. Stegarescu: Public sector decentralisation: Measurement concepts and recent international trends. *Fiscal Stud.* 26 (2005), 301–333. DOI:10.1111/j.1475-5890.2005.00014.x
- [33] D. Stegarescu: The effects of economic and political integration on fiscal decentralization: Evidence from OECD countries. *Canadian J. Econom. / Revue Canadienne d’Economie* 42 (2009), 694–718. DOI:10.1111/j.1540-5982.2009.01524.x
- [34] D. H. Vo: New Fiscal Decentralization Indices. The University of Western Australia Discussion Paper 08.14, 93, 2008.
- [35] D. H. Vo: The economics of fiscal decentralization. *J. Econom. Surveys* 24 (2010), 657–679. DOI:10.1111/j.1467-6419.2009.00600.x
- [36] O. Vsevolozhskaya, M. Greenwood, and D. Holodov: Pairwise comparison of treatment levels in functional analysis of variance with application to erythrocyte hemolysis. *Ann. Appl. Statist.* 8 (2014), 905–925. DOI:10.1214/14-aos723
- [37] J.-T. Zhang: Analysis of Variance for the functional data. Chapman and Hall, 2014. DOI:10.1201/b15005

*Tomáš Mrkvička, Faculty of Economics, University of South Bohemia, Studentská 13, 37005 České Budějovice. Czech Republic.*

*e-mail: mrkvicka.toma@gmail.com*

*Mari Myllymäki, Natural Resources Institute Finland (Luke), Latokartanonkaari 9, FI-00790 Helsinki. Finland.*

*Milan Jílek, Faculty of Economics, University of South Bohemia, Studentská 13, 37005 České Budějovice, Czech Republic.*

*Ute Hahn, Centre for Stochastic Geometry and Advanced Bioimaging, Department of Mathematics, University of Aarhus, 8000 Aarhus C. Denmark.*