New methods for multiple testing in permutation inference for the general linear model

Tomáš Mrkvička1 | Mari Myllymäki2 | Mikko Kuronen2 | Naveen Naidu Narisetty3

1Department of Applied Mathematics and Informatics, University of South Bohemia, České Budějovice, Czech Republic
2Natural Resources Institute Finland (Luke), Helsinki, Finland
3Department of Statistics, University of Illinois, Urbana-Champaign, Champaign, Illinois

Correspondence
Tomáš Mrkvička, Department of Applied Mathematics and Informatics, University of South Bohemia, Studentská 13, 37005 České Budějovice, Czech Republic.
Email: mrkvicka.toma@gmail.com

Funding information
Academy of Finland, Grant/Award Numbers: 295100, 306875, 327211; Grantová Agentura České Republiky, Grant/Award Number: 19-04412S

Abstract
Permutation methods are commonly used to test the significance of regressors of interest in general linear models (GLMs) for functional (image) data sets, in particular for neuroimaging applications as they rely on mild assumptions. Permutation inference for GLMs typically consists of three parts: choosing a relevant test statistic, computing pointwise permutation tests, and applying a multiple testing correction. We propose new multiple testing methods as an alternative to the commonly used maximum value of test statistics across the image. The new methods improve power and robustness against inhomogeneity of the test statistic across its domain. The methods rely on sorting the permuted functional test statistics based on pointwise rank measures; still, they can be implemented even for large data. The performance of the methods is demonstrated through a designed simulation experiment and an example of brain imaging data. We developed the R package GET, which can be used for the computation of the proposed procedures.

KEYWORDS
function-on-scalar regression, general linear model, global envelope test, graphical method, multiple testing correction

1 INTRODUCTION

General linear models (GLMs) are among the most commonly used statistical tools for analyzing functional or image data. One of the approaches to perform statistical inference for the GLMs is to use parametric methods which rely on stringent assumptions, such as the normality of the random errors in the GLM, homogeneity, and independence of the random errors across the function or image. Alternatively, one can use nonparametric methods with weaker assumptions about the data.1 Such methods are popular in neuroscience,2 but they are also used for other biomedical functional data.3

Permutation tests are a class of nonparametric methods, which have a long history going back to Fisher.4 Fisher demonstrated that the null hypothesis could be tested by observing, how often the test statistic computed from permuted observations would be more extreme than the same statistic computed without permutation. Even though the data we are interested in are functions or images, the same principle still holds; a detailed introduction to the permutation inference for the GLMs for neuroimage data can be found in Winkler et al.1 Another advantage of the nonparametric methods is that they do not suffer from the false positive rates reported by Eklund et al for parametric approaches.5

Analysis of functional or image data based on permutation tests consists of several crucial steps. First, a suitable test statistic that is computed for each single location (spatial point, voxel, vertex, or face) of a function or image has to be
chosen such that it is informative about the studied null and alternative hypotheses and homogeneous across the function or image. Second, the appropriate permutation scheme has to be used to generate the permutations from the studied null hypothesis. Finally, an appropriate multiple testing correction has to be applied to decide which spatial locations are significant.

Commonly used test statistics include the t- and F-statistics, or the G statistic, which generalizes the classical statistics into various cases with heteroscedasticity. The maximum of the chosen test statistic is then usually treated as a test statistic in a permutation test. However, a serious limitation of the F or G statistics is that they are not pivotal across different locations of the function or image in terms of the entire distribution, but only in terms of the first and second moments when the errors are not normally distributed. Therefore, if the error distribution is non-normal and inhomogeneous across the function or image, the distribution of these test statistics varies, which causes the desired quantiles of the test statistics to vary as well. If this heterogeneity is ignored, for example, by taking the maximum F-statistic across the spatial points as the test statistic, some or all spatial points where the null model is broken may be overlooked. In a quantitative manner, this can bring a substantial loss of power, as we demonstrate in our simulation study. In a qualitative manner, Figure 1 demonstrates this effect on a simple toy example similar to the case (e) of the simulation study. The upper two panels show two groups of functions that were drawn from different distributions on [0, 0.5] and [0.5, 1], with an abrupt change at 0.5 for simplicity: on [0, 0.5] the functions were drawn from the lognormal random field with \( \mu = 1 \), \( \sigma = 1 \), and scale of correlation \( \rho = 0.3 \), while on [0.5, 1], the lognormal random field had parameters \( \mu = 1 \), \( \sigma = 1 \), and \( \rho = 0.3 \). Further, the mean of functions in the second group (on the right in Figure 1) was adjusted to have a small bump around the location \( r = 0.10 \), making it to deviate from the control group 1 (on the left in Figure 1). The bottom panel of Figure 1 shows the F-statistic for the simple functional two-sample test on [0, 1] (solid line). It shows the F-max threshold (solid straight line) obtained from 5000 permutations of the functions in the two groups, and it shows the output of our proposed method based on the same 5000 permutations: the shaded area gives the threshold of the new proposed method and it demonstrates the variability of the quantiles of the F-statistic in different parts of the domain. The F-statistic is clearly below the F-max threshold since the F-max threshold is driven by the right part, which achieves higher quantiles, but our proposed method identifies the difference around 0.1 (red color used for the part where the F-statistic exceeds the threshold) since it is adjusted to the variability of the quantiles of the F-statistic across the functional domain.

Due to the non-pivotal nature of the F and G statistics, a more suitable choice is the permutation p-value, which is automatically a pivotal statistic, that is, its distribution does not change across the function or image. Therefore, it would be convenient to perform the multiple testing correction by taking the minimum p-value across the function or image as the test statistic. However, the permutation p-value has another serious disadvantage: due to its discreteness, the (minimum) p-values obtained from permutations contain ties, and the resulting test tends to be conservative, leading to loss of power as well. This disadvantage is enormous, especially in the case of high-resolution images where the number of permutations cannot be large due to computational limitations. This is further demonstrated in the introductory example in Section 2.5. In the rest of this article, we concentrate on such high-resolution images, even though all methods discussed in this article are applicable also for one-dimensional functions or low-resolution images.

Besides the p-min or F-max approaches, a local area above a given threshold around a spatial location can be used as a test statistic. Another problem arises for such test statistics. Namely, if the spatial autocorrelation of the data is inhomogeneous across the image, then the sizes of the areas above the thresholds in different parts of the image are not comparable. This problem was recently treated for the special case of cluster size permutation tests. In this article, we discuss solutions for the case of inhomogeneous spatial autocorrelation and also for the case of inhomogeneous distribution of the test statistic.

The classical multiple testing procedures, computing the maximum F (F-max) or the minimum p-value (p-min) statistic across all spatial points, control the family-wise error rate (FWER). This article introduces three new multiple testing corrections for the permutation inference for the GLMs, controlling the FWER in a similar manner.

The new methods are alternative solutions to the ties problem of p-values. The first correction is based on the extreme rank length (ERL) measure that was originally proposed by Myllymäki et al and Narisetty and Nair for spatial and functional data analysis. This correction was also used for comparing two groups of functions in biomedical applications. The other two corrections rely on the continuous rank measure (Cont) introduced in the technical report of Hahn and the area rank measure (Area) presented in this work first time. We carefully describe the homogeneity assumptions of the F-max, p-min, and new methods and further their sensitivity to different types of extremeness of the test statistic. By a simulation study, we illustrate the power of the new and existing tests under different scenarios, both when the homogeneity assumptions of the multiple correction methods are met and when not. A further asset of the methods, namely a graphical interpretation of the test results in a similar manner as in global envelope testing, is illustrated by a real
data example of autism brain imaging data.\textsuperscript{14} We conclude that the choice of the measure for multiple correction should be made on the basis of reasonable assumptions of different types of homogeneity and the expected type of extremeness of the test statistic.

The proposed tests are implemented in the R library GET.\textsuperscript{9,15}

2 | MULTIPLE TESTING CORRECTION FOR PERMUTATION METHODS

Let us consider the GLM

\[ Y(r) = X(r)\beta(r) + Z(r)\gamma(r) + \epsilon(r), \]  

(1)
where the argument \( r \in I \subset \mathbb{R}^d \) determines a spatial point of the image or, generally, a location where the \( d \)-dimensional functions, \( d \geq 1 \), are observed. The number of locations \( r \in I \) will be denoted by \( N \). For every location \( r \), we consider a one-dimensional GLM with \( X(r) \) being a \( n \times k \) matrix of regressors of interest, \( Z(r) \) being a \( n \times l \) matrix of nuisance regressors, \( Y(r) \) being a \( n \times 1 \) vector of observed data, and \( e(r) \) being a \( n \times 1 \) vector of random errors with mean zero and finite variance \( \sigma^2(r) \) for every \( r \in I \). Further assumptions about the error structure will be given along with the definitions of the multiple correction methods below. Further, \( \beta(r) \) and \( \gamma(r) \) are the regression coefficient vectors of dimensions \( k \times 1 \) and \( l \times 1 \), respectively. Often factors are given for the whole image, as in our examples as well. In this case, the regressors do not depend on the index \( r \), and so the matrices \( X(r) \) and \( Z(r) \) are identical for every \( r \in I \). However, this simplification is not necessary. The null hypothesis to be tested is

\[
H_0 : C(r) \beta(r) = 0, \quad \forall r \in I,
\]

where \( C(r) \) is a \( t \times k \) matrix of \( t \) contrasts of interest.

The GLM setup includes a wide range of applications in neuroimaging.\(^1\) Under various setups, a nonparametric pointwise permutation test can be applied to obtain a set of \( p \)-values, \( \{ p(r) : r \in I \} \). We refer to the detailed description of pointwise permutation tests in Winkler et al.\(^1\) Shortly, a test statistic \( T(r) \) is first chosen from a vast number of possibilities, reflecting the tested null model \( H_0 \) and heteroscedasticity. A common choice is the \( F \)-statistic,\(^1\) but in principle any statistic where extreme values reflect evidence against the null hypothesis could be used. Then permutations are used to obtain the distribution of \( T(r) \) under the null hypothesis \( H_0 \). The choice of the permutation scheme is important for the performance of the method. We chose the permutation of the residuals under the null model,\(^17\) which is approximate, but according to Anderson and Ter Braak\(^18\) it is the most precise permutation method in the case of nuisance effects.\(^1\) If there are no nuisance regressors, except the constant, permutation of raw data can be used instead; it is however equivalent to performing the Freedman-Lane algorithm since the fit of the null model is constant in this case. In this case, the test is exact. The last step is to apply a multiple testing correction.

The following sections describe various multiple testing corrections in these permutation procedures. We assume that the same \( J \) permutations have been applied for every location \( r \in I \). We denote the test statistic calculated for the original data by \( T_0(r) \) and the corresponding test statistics calculated for each permutation \( j, j = 1, \ldots, J \), by \( T_j(r) \). Further, \( p_0(r), r \in I \), denotes the set of pointwise \( p \)-values for the test statistic \( T_0(r) \) of the original data, and by \( p_j(r) \) the corresponding \( p \)-values for each permutation \( j, j = 1, \ldots, J \).

### 2.1 The \( p \)-min approach

In the \( p \)-min approach, the minimum of the pointwise \( p \)-values,

\[
p_j^\text{min} = \min_{i \in I} p_i(r),
\]

are calculated for all \( j = 0, 1, \ldots, J \), and the Monte Carlo \( p \)-value is obtained as

\[
P_j^\text{min} = \frac{1}{J + 1} \sum_{k=0}^{J} \mathbb{I}(p_k^\text{min} \leq P_j^\text{min}).
\]

Here \( \mathbb{I} \) denotes the indicator function.

The pointwise \( p_j(r) \) -values are pivotal; therefore, the \( p \)-min approach needs no further assumptions about the random error \( e(r) \). (Remark here, that the assumption of subset pivotality, assumed in the Westfall and Young step down procedure,\(^19\) is not required here, since we consider the \( p \)-min approach as a one-step test.) However, since the \( p \)-values of the permutation test can achieve only values \( \frac{1}{J+1}, \frac{2}{J+1}, \ldots, 1 \), many ties can appear among \( p_0^\text{min}, p_1^\text{min}, \ldots, p_J^\text{min} \), especially for large \( N \), leading to a conservative test. Remark here that we use this definition of permutation \( p \)-values, that is, \( \frac{1}{J+1} \sum_{j=0}^{J} \mathbb{I}(T_j(r) \leq T_0(r)) \), in accordance with the recommendation of Phipson and Smyth,\(^20\) which is shown to be slightly conservative in cases of permutations with replacement. On the contrary, the \( p \)-value defined as \( \frac{1}{J} \sum_{j=1}^{J} \mathbb{I}(T_j(r) \leq T_0(r)) \) is liberal.\(^20\)
2.2 | F-max approach

Usually the statistic $T(r)$ is one of the statistics used in the parametric GLMs, that is, $t$-statistic, $F$-statistic, or variance weighted $F$-statistic. In these cases, an immediate solution to the problem of ties is to replace $p_j^{\text{min}}$ statistic in the Monte Carlo test directly by the $F_j^{\text{max}} = \max_{r \in I} T_j(r)$ statistic, but with the price of losing the pivotality and, consequently also the power. See Section 3 for details.

2.3 | Refinements of the $p$-min approach

The $p$-min approach is equivalent to the global rank envelope test defined by Myllymäki et al.\textsuperscript{9} for spatial processes. That is, the $p_j^{\text{min}}$ is after a simple normalization equivalent to the extreme rank measure $R_j$:

\[(J + 1)p_j^{\text{min}} = R_j = \min_{r \in I}(J + 2 - R_j(r)),\]

where $R_j(r)$ is the pointwise rank of the test statistic $T_j(r)$ among $\{T_0(r), \ldots, T_J(r)\}$ such that the smallest value obtains the rank 1. The measures and the corresponding global envelopes are defined here as one-sided since the extremeness of the test statistic $T(r)$ used in the GLM is usually realized only for high values. This one-sided alternative also leads to $J + 2$ in the right hand side of the formula (5).

In the sequel, we propose the following three methods for breaking the ties between the $p_j^{\text{min}}$ values:

- The ERL approach takes into account the number of pointwise $p$-values equaling the minimal $p$-value $p_j^{\text{min}}$,
- the continuous rank (Cont) approach measures the maximal size of extremeness of $T_j(r)$ that is associated with $p_j^{\text{min}}$, and
- the area rank (Area) accumulates the extremeness of $T_j(r)$ across the spatial points $r$ where $p_j^{\text{min}}$ is obtained.

2.3.1 | Extreme rank length

The ERL measure refines the $p$-min approach in the sense that not only the minimal $p$-value $p_j^{\text{min}}$, but also the size of the domain where $p_j(r)$ are equal to $p_j^{\text{min}}$ is taken into account. Typically, based on our experience, the ties are already broken by considering the counts of the smallest pointwise $p$-values among $r \in I$. However, to define the ERL measure formally,\textsuperscript{5,10,13} the complete vectors of pointwise ordered $p$-values $p_j = (p_{j[1]}, p_{j[2]}, \ldots, p_{j[n]})$, where the pointwise $p$-values $p_j(r)$ observed at $r_1, \ldots, r_n$ are arranged from smallest to largest such that $p_{j[k]} \leq p_{j[k']} \text{ whenever } k \leq k'$, are considered and ordered by reverse lexical ordering: The ERL measure is equal to

\[e_j = \frac{1}{J + 1} \sum_{j=0}^{J} (p_{j'} < p_j),\]

where

\[p_{j'} \prec p_j \Leftrightarrow \exists l \leq n : p_{j[k]} = p_{j[k']} \forall k < l, \quad p_{j[l]} < p_{j[l]}\]

and $J + 1$ scales the values to the interval from 0 to 1. Thus, the ERL measure uses in fact all the pointwise $p$-values to order the test statistics $T_j(r)$ from most extreme to the least extreme one, while the $p$-min procedure utilizes only $p_j^{\text{min}} = p_{j[1]}$.

Since the probability of having a tie in the ERL measure is rather small, the application of Monte Carlo test on the ERL solves the ties problem.\textsuperscript{9} Thus the final $p$-value is $p_{\text{ERL}} = 1 \sum_{j=0}^{J} (e_j \leq e_0)$.

Because often even the counts of the most extreme $p$-values break the ties, the measure can be efficiently implemented utilizing the counts of a certain number (eg, six) of smallest pointwise $p$-values (see Section 4).
2.3.2 | Continuous rank

Another refinement of $p^{\text{min}}$ is the continuous rank measure

$$c_j = \frac{1}{J+1} \min_c C_j(r) \quad \text{with} \quad C_j(r) = J + 1 - c_j(r), \quad (7)$$

where $c_j(r)$ is the continuous pointwise rank that is a refinement of the ordinary rank $R_j(r)$ based on the relative magnitude of the test statistic $T_j(r)$ with respect to the other $T_j(r), j' = 0, \ldots, J$. The continuous rank $c_j(r)$ is defined such that the smallest value of $T_0(r), \ldots, T_J(r)$ obtains the smallest continuous rank: Let $T_{[0]}(r) \leq T_{[1]}(r) \leq \cdots \leq T_{[J]}(r)$ denote the ordered set of values $T_j(r), i = 0, 1, \ldots, J$. Then the continuous rank of $T_{[j]}(r)$ is

$$c_{[j]}(r) = j + \frac{T_{[j]}(r) - T_{[j-1]}(r)}{T_{[J]}(r) - T_{[j-1]}(r)}, \quad \text{for} \quad j = 1, 2, \ldots, J - 1,$$

and

$$c_{[0]}(r) = \exp \left( \frac{-T_{[1]}(r) - T_{[0]}(r)}{T_{[1]}(r) - T_{[0]}(r)} \right),$$
$$c_{[J]}(r) = J - \exp \left( \frac{-T_{[J]}(r) - T_{[J-1]}(r)}{T_{[J-1]}(r) - T_{[0]}(r)} \right). \quad (8)$$

If the probability to have ties among $T_j(r), j = 0, \ldots, J$, is zero, then the probability of ties among $c_j(r)$ is zero as well. If ties appear among $T_j(r), j = 0, \ldots, J$, that is, $T_{[j-1]}(r) < T_{[j]}(r) = \cdots = T_{[J]}(r) < T_{[j+1]}(r)$, then the continuous rank is defined as $c_{[j]}(r) = \frac{1 + j}{2} \frac{m_j}{J}$ for $l = j, \ldots, J'$. The continuous ranks were first suggested by Hahn,11 and are defined here with a slight modification in the scaling of $c_{[0]}(r)$ and $c_{[J]}(r)$.

Finally, the univariate Monte Carlo test is performed based on $c_j$. Thus the final $p$-value is $p^{\text{cont}} = \frac{1}{J+1} \sum_{j=0}^{J} \mathbb{I}(c_j \leq c_0)$.

2.3.3 | Area rank

The last refinement is the area rank measure $a_j$ defined as:

$$a_j = \frac{1}{J+1} \left( R_j - \frac{1}{N} \sum_r (R_j - C_j(r)) \mathbb{1}(C_j(r) < R_j) \right), \quad (9)$$

where $N$ is equal to the number of $r \in I$. Thus, the area measure refines the extreme rank $R_j$ by reducing from it the scaled sum of the differences between $R_j$ and $C_j(r)$ across the spatial points $r$ where the minimal $p$-values (or ranks) are obtained. Thus, the larger the amount of those critical $r$ and the smaller the pointwise continuous ranks $C_j(r)$ at those critical $r$ are, the smaller is the value of the area measure, that is, the more extreme the corresponding test statistic $T_j(r)$. The univariate Monte Carlo test is performed based on $a_j$ with $p^{\text{area}} = \frac{1}{J+1} \sum_{j=0}^{J} \mathbb{I}(a_j \leq a_0)$.

2.4 | Illustration of the new measures

Figure 2 shows a small set of one-dimensional functions (test statistics) $T(r)$ for describing the behavior of the new measures. Only large values of $T(r)$ are considered significant. In this set of functions, the variance of the distribution of $T(r)$ increases along $r$ (from left to right). According to the $F$-max method, the most extreme test statistic is the function 1 that reaches the overall highest value across the locations $r$. On the other hand, the $p$-min method gives the minimal $p$-value to six functions (1, 3, 4, 5, 7, and 9). The refinement measures decide the extremeness of these functions by different rules.

The most extreme function according to the ERL measure is the function 5 (brown) because it obtains the largest value, that is, the minimal $p$-value, over the largest domain. On the other hand, according to the continuous and area rank measures, the most extreme function is the function 9 (blue): The deviation of the functions 3 (red) and 9 (blue)
Figure 2  A set of one-dimensional functions (test statistics) for describing the different measures. The variance of the distribution of the test statistics increases from left to right. The most extreme function according to ERL measure is the function 5; according to continuous and Area rank measures, it is the function 9. The function 3 is not as extreme according to these measures because the overall variation near its peak is higher than for the function 9.

Table 1  Sources of information gathered by different measures

|       | \( p_{j}^{\min} \) | Domain with \( p_{j}^{\min} \) | Value |
|-------|---------------------|------------------------|------|
| F-max | No                  | No                     | Yes  |
| p-min | Yes                 | No                     | No   |
| ERL   | Yes                 | Yes                    | No   |
| Cont  | Yes                 | No                     | Yes  |
| Area  | Yes                 | Yes                    | Yes  |

Note: While F-max depends on the maximum value of the test statistic \( T(r) \) across all \( r \in I \), the other measures utilize the values of \( T(r) \) at the locations \( r \) with \( p_{j}^{\min} \).

from the second most extreme functions is of the same size at those locations \( r \) where these functions reach the minimal \( p \)-value. However, the scaling of the continuous rank by the overall variation (see Equation 8) leads to smaller pointwise continuous ranks for the function 9, therefore also to a smaller overall continuous rank. Because the two functions reach the minimal \( p \)-values approximately at domains of the same size, also the area measure deduces the function 9 as the most extreme function.

This example illustrated two issues: (1) The fact that the ERL, continuous and area measures are all refinements of the \( p \)-min method allows them to regard as most extreme also functions which reach their minimal \( p \)-value, \( p_{j}^{\min} \), at locations where the variation of the functions is small. (2) The refinement measures break the ties by different rules as summarized in Table 1.

2.4.1  | Graphical Interpretation: 100(1 – \( \alpha \))% global envelope

Consider any of the three measures we defined and denote them using a common notation \( M_{j}, j = 0, \ldots, J \). Let \( I_{\alpha} = \{ j \in 0, \ldots, J : M_{j} \geq M_{(\alpha)} \} \) be the index set of the test statistics, where the threshold \( M_{(\alpha)} \) is chosen to be the largest value in...
Figure 3 Upper envelope (blue) of the global area rank test and significant voxels (red) on slice 45 when testing the effect of the group variable in abide brain data subset using 100,000 permutations.

\{M_0, \ldots, M_J\} for which

\[ \sum_{j=0}^{J} \mathbb{I}(M_j < M_{(\alpha)}) \leq \alpha (J + 1), \tag{10} \]

that is, the set \(I_\alpha\) contains the indices of the 100(1 – \(\alpha\))% of the least extreme test statistics \(T_j(r)\). The global envelope is defined to be

\[ T_{\text{low}}^{(\alpha)}(r) = -\infty, \quad T_{\text{upp}}^{(\alpha)}(r) = \max_{j \in I_\alpha} T_j(r). \]

As a consequence, the probability under null hypothesis that \(T_0(r)\) falls outside this global envelope in any of the \(r\) points is less or equal to \(\alpha\),

\[ \Pr \left( \exists \ r \in I : T_0(r) \notin \left[ T_{\text{low}}^{(\alpha)}(r), T_{\text{upp}}^{(\alpha)}(r) \right] \right) \leq \alpha. \]

The following theorem states that inference based on the \(p\)-value \(p^M\) (where \(p^M\) stands for \(p^\text{el}, p^\text{cont}, \) and \(p^\text{area}\)) and on the global envelope specified by \(T_{\text{low}}^{(\alpha)}(r)\) and \(T_{\text{upp}}^{(\alpha)}(r)\) with respect to the appropriate measure are equivalent. Therefore, we can refer to these envelopes as the 100(1 – \(\alpha\))% global ERL envelope, continuous rank envelope, and area rank envelope. Figure 3 shows an example of the ERL upper envelope together with red points indicating the area where the test function crossed this envelope. Thus the graphical interpretation identifies the voxels which cause the rejection.

**Theorem 1.** Assume that there are no pointwise ties with probability 1 among \(T_j(r), j = 0, \ldots, J\). Then

1. \(T_0(r) > T_{\text{upp}}^{(\alpha)}(r)\) for some \(r \in I\) iff \(p^M \leq \alpha\), in which case the null hypothesis is rejected;
2. \(T_0(r) \leq T_{\text{upp}}^{(\alpha)}(r)\) for all \(r \in I\) iff \(p^M > \alpha\), and thus the null hypothesis is not rejected.

**Proof.** Since 1. holds iff 2. holds, it is enough to show 1. According to the definition of \(p^M\), \(p^M \leq \alpha\) iff number of \(M_j\) smaller or equal to \(M_0\) is smaller or equal to \(\alpha (J + 1)\). That is equivalent, according to the definition of \(M_{(\alpha)}\) to \(M_0 < M_{(\alpha)}\). In order
to finish the proof, we need to prove that \( M_0 < M(a) \) iff \( T_0(r) > T^{(a)}_{\text{upp}}(r) \) for some \( r \in I \). Since \( T_0(r) > T^{(a)}_{\text{upp}}(r) \) implies that \( M_0 < M(a) \), we only need to prove the other direction. Suppose now that \( M_0 < M(a) \). This implies that the extreme ranks (5) satisfy \( R_0 < R(a) \), where \( R(a) \) is the threshold for the extreme ranks defined similar to \( M(a) \). If \( R_0 < R(a) \), then \( T_0(r) > T^{(a)}_{\text{upp}}(r) \) for any \( r \) that satisfies \( R_0(r) < R(a) \). On the other hand, if \( R_0 = R(a) \), then no pointwise ties implies \( T_0(r) > T^{(a)}_{\text{upp}}(r) \) for any \( r \) that satisfies \( R_0(r) = R(a) \). Namely, let \( r \) be such that \( R_0(r) = R(a) \) and \( j \in I_a \). Then \( M_j \geq M(a) \) by definition of \( I_a \). By the definition of measure \( M \), \( R_j \geq R(a) \) for all considered measures. The definition of \( R_j \) implies that \( R_j(r) \geq R_j \geq R(a) = R_0(r) \). No pointwise ties implies that \( R_j(r) \neq R_0(r) = R_0 \). Thus \( R_j(r) > R(a) \) which is equivalent to \( T_j(r) > T_0(r) \). Since this holds for any \( j \), we have shown that \( T_0(r) > T^{(a)}_{\text{upp}}(r) \) for any \( r \) that satisfies \( R_0(r) = R(a) \).

2.4.2 Adjusted p-values for multiple testing

In this article, we are mostly interested in the global null hypothesis (2). The proposed tests are exact, that is, the type I error is precisely \( \alpha \), according to Lemma 1 in Myllymäki et al.,\(^9\) if the permutations are exchangeable, for example, for the functional one-way ANOVA. When the Freedman and Lane procedure is involved in the generation of permutations, the procedures are only approximately exact.

Let’s now look at the problem of FWER control from the point of view of multiple hypothesis testing. The statement in the previous paragraph corresponds to the weak control of FWER. If the exchangeability is satisfied, the continuous rank satisfies the strong control similarly to the \( p \)-min approach since it is defined as the minimum of continuous pointwise ranks. Since the ERL and Area procedures accumulate extremeness of the test statistic across \( r \in I \), they do not satisfy the strong control of FWER. However, because they converge to the \( p \)-min procedure when the number of permutations is increased, the strong control is satisfied asymptotically for them.

One can additionally be interested in the adjusted \( p \)-value \( p_0^A(r) \) at the location \( r \in I \). The adjustment for the multiple testing is made here again through the ordering of whole functions. Let \( m_j \) denote the order of the \( j \)th function with respect to the measure \( M \) among functions \( T_0 \), ..., \( T_J \) such that the most extreme function obtains the value 1. Inspired by Xu and Reiss,\(^2\) who defined adjusted \( p \)-values for the \( p \)-min procedure, we define adjusted \( p \)-values for any of the proposed measures as

\[
p_0^A(r) = \frac{\min\{k : T_0(r) \leq \max\{T_j(r) : j = 0, \ldots, J \text{ and } m_j \geq k\}\}}{J + 1}.
\]

Clearly, \( p_0^A(r) \leq \alpha \) iff \( T_0(r) \notin [T^{(a)}_{\text{low}}(r), T^{(a)}_{\text{upp}}(r)] \), thus all locations determined by any of our envelopes as significant have adjusted \( p \)-value smaller than \( \alpha \).

2.5 Illustratory example

To illustrate the proposed methods, we studied the autism brain imaging data collected by resting state functional magnetic resonance imaging (R-fMRI).\(^1\) The raw fMRI data contain measurements from 539 individuals with the autism spectrum disorder (ASD) and 573 typical controls (TC). When the number of individuals in the analysis is large, the test statistics can be well approximated by the limiting \( F \) distribution. Therefore, in order to show the differences between the methods, we have chosen a small number of individuals, namely the patients measured in Oregon Health and Science University with 13 subjects with ASDs, 15 TCs, and homogeneity in sex. In fact, when the whole dataset of 1112 patients was used in our analysis, the \( F \)-max procedure worked equally to our proposed methods. The imaging measurement for local brain activity at the resting state was the fractional amplitude of low frequency fluctuations.\(^2\) The whole brain consisting of 175493 voxels was analyzed. The difference between the groups was studied while age was taken as a nuisance factor. The \( F \)-max, \( p \)-min, and new tests were performed using the same 50 000 and 100 000 permutations. The experiment was run using the R package GET\(^1\) and took 10 hours using two cores and 10 GB memory of an ordinary laptop. Most of this time was allocated to calculating the \( F \)-statistics for each permutation and each voxel, which were needed for all tests.

Table 2 shows the \( p \)-values of all the investigated methods. While the differences in the \( p \)-values are small between the \( F \)-max and new methods, only the ERL and Area tests rejected the null hypothesis of no group effect at the strict significance level 0.05. An essential difference between the \( F \)-max and new methods is in their critical bounds (see also
TABLE 2  The $p$-values for the test of the effect of the group variable in the example brain data subset computed by investigated methods

| Number of permutations | $F$-max | $p$-min | ERL | Cont | Area |
|------------------------|---------|---------|-----|------|------|
| 50 000                 | 0.058   | 0.592   | 0.032 | 0.105 | 0.040 |
| 100 000                | 0.058   | 0.420   | 0.041 | 0.053 | 0.040 |

Note: The ERL, Cont, and Area refer to the new methods to be introduced in this article.

FIGURE 4  Proportion of rank 1 ties in the $p$-min procedure with respect to the number of permutations, when testing the effect of the group variable in the example brain data

the toy example in Figure 1): Figure 3 shows the 95% global upper envelope of the global Area test for one slice of the brain; Appendix B presents the results for the whole brain. This upper envelope is adjusted to the variability of the 95% global quantile of the $F$-statistics, ranging from 20 to 70 for the given slice. If the empirical $F$-statistic crosses the 95% global quantile, the null hypothesis is rejected. On the other hand, the critical bound of the $F$-max test is constant (48.4). Therefore, the $F$-max procedure can overlook significant voxels where the variability of the $F$-statistics is smaller.

The $p$-min procedure, which is also adjusted to the variability of the statistics across the image, was highly conservative for 50 000 or 100 000 permutations (Table 2). Figure 4 further shows the proportion of permutations that achieved the most extreme $F$-statistic at least for one voxel for different numbers of permutations. This value corresponds to the minimal $p$-value that the $p$-min procedure can achieve. This example with a total of 10 000 000 permutations was run on a computing server with hundreds of cores. Thus if one needs the $p$-min procedure to give an answer for the test of the whole brain, 6 000 000 permutations have to be used in minimum to achieve the minimal $p$-value 0.02 (which is still not a very precise procedure), and this is typically too much. This article gives the three solutions for breaking the ties of the $p$-min procedure, which can be used with a reasonable number of permutations, that is, in the time comparable to the time required by the $F$-max procedure. The proposed methods also provide the correct significance level as it is shown by a simulation study in Section 3 and also by an experiment similar to one proposed in Eklund et al.\(^5\) which we report in Appendix A.

3 | SIMULATION EXPERIMENT

To compare the power and robustness of the proposed methods and the existing multiple comparison methods under different scenarios, we generated synthetic imaging data mimicking real data from neuroimaging studies. We considered
a categorical factor \( g \) taking the values 1 or 2 according to the group to which the image belongs, and a continuous factor \( z \) that was generated from the uniform distribution on \((0, 1)\). We simulated images \( Y(r) \) in the square window \([-1, 1]^2\) on a grid of \( 51 \times 51 \) pixel resolution from the following GLM models:

\[
\begin{align*}
M_0: & \quad Y(r) = \epsilon(r), \\
M_1: & \quad Y(r) = \exp(-10 \cdot ||r||) \cdot g + \epsilon(r), \\
M_1': & \quad Y(r) = \exp(-200 \cdot ||r||) \cdot g + \epsilon(r), \\
\end{align*}
\]

and

\[
M_2: \quad Y(r) = \exp(-10 \cdot ||r||) \cdot (g + z) + \epsilon(r).
\]

The image resolution is obviously small in comparison to the brain data (see Section 2.5) but was regarded as sufficient for the designed experiment to show differences between the measures. Here \( ||r|| \) denotes the Euclidean distance of the pixel to the origin, and \( \epsilon(r) \) is a zero-mean correlated error. The model \( M_0 \) has no factors and generates purely noisy images, the models \( M_1 \) and \( M_1' \) generate two groups of images depending on the categorical factor \( g \), and the model \( M_2 \) generates images that depend on both the categorical and continuous factors. The models \( M_1 \) and \( M_1' \) correspond to a simple comparison of two groups of images, where the “bump” at the center of the image is two times higher for the second group than for the first group. The area where the two groups differ is about a hundred times smaller in the model \( M_1' \) than in \( M_1 \). Because the departures from the null model in \( M_1 \) occur for many pixels \( r \) and the ERL and Area methods accumulate information across the pixels, the ERL and Area methods are expected to have an advantage over the other methods in this model. On the other hand, the departures from the null model are expected only for a few pixels \( r \) in \( M_1' \) and, thus, such an advantage does not occur in this model.

The model \( M_0 \) is a null model where the images in the two groups are from the same model (no factors), and it was used for estimating the significance levels of the different tests. The model \( M_2 \) is similar to \( M_1 \), but the groups are disturbed by the continuous factor \( z \). The models \( M_1, M_1', \) and \( M_2 \) were used for power estimation. The permutation of raw data was used in models \( M_1 \) and \( M_1' \). The Freedman and Lane permutation scheme was used in \( M_2 \). Because the model \( M_2 \) revealed consistent results with model \( M_1 \), the results are not presented.

For both models \( M_1 \) and \( M_1' \), seven different correlated error structures \( \epsilon(r) \) were considered:

(a) Gaussian error \( \epsilon(r) = G_{0.15}(r) \), where \( G_{\rho}(r) \) is a Gaussian random field with the exponential correlation structure with scale parameter \( \rho \) and standard deviation \( \sigma \) which will take several values,

(b) skewed error \( \epsilon(r) = \exp(G_{0.15}(r)) \),

(c) shape inhomogeneous distribution with increasing bimodality \( \epsilon(r) = \frac{1}{4} \text{sign}(G_{0.15}(r))|G_{0.15}(r)|^{1/(2||r||+1)}, \)

(d) \( \epsilon(r) = \frac{1}{2}G_{0.15}(r)^{1/5} + \frac{1}{2}||r|| \leq 0.5G_{0.15}(r) \), that is, inhomogeneous distribution over \( I \) with the bimodal and normal errors in the periphery and in the middle of the image, respectively,

(e) \( \epsilon(r) = \frac{1}{2}||r|| \leq 0.5 \exp(3G_{0.15}(r)) + \frac{1}{2}||r|| > 0.5(G_{0.15}(r) + 1), \) that is, inhomogeneous distribution over \( I \) with the skewed and normal errors in the middle and periphery of the image, respectively,

(f) \( \epsilon(r) = \frac{1}{2}||r|| \leq 0.5G_{0.05}(r) + \frac{1}{2}||r|| > 0.5G_{0.3}(r) \), that is, Gaussian distribution with inhomogeneity in the correlation structure (scale parameters 0.05 and 0.3 in the middle and periphery of the image),

(g) \( \epsilon(r) = \frac{1}{2}||r|| \leq 0.5 \frac{1}{2}G_{0.05}(r)^{1/5} + \frac{1}{2}||r|| > 0.5G_{0.3}(r)^{1/5}, \) that is, bimodal distribution with inhomogeneity in the correlation structure.

The homogeneous error distributions (a) and (b) represent cases where all methods should perform well in identifying alternative hypotheses. The skewed error (b) represents a situation where the permutation inference is necessary because the assumptions needed for parametric methods are not met. Further, the errors (c), (d), and (e) are inhomogeneous across the image and illustrate cases where the variability of the test statistic \( T(r) \) in the periphery of the image mask the signal from the center of the image. The error (c) is rather normal in the center of the image, while the increasing bimodality appears closer to the periphery. The bimodal error corresponds to the sharp changes in the images, whereas Gaussian error corresponds to the smooth changes in the images. The errors (d) and (e) are normal at different places in order to have bigger quantiles of the test statistic at the periphery of the image. On the other hand, the errors (f) and (g) are used to investigate the effect of inhomogeneity in the correlation structure on the results.
TABLE 3  Empirical significance levels of all studied methods based on 1000 replicates for model M0 with error (a) to (g) with standard deviation $\sigma_1 = 0.1$

|       | (a) $\sigma_1$ | (b) $\sigma_1$ | (c) $\sigma_1$ | (d) $\sigma_1$ | (e) $\sigma_1$ | (f) $\sigma_1$ | (g) $\sigma_1$ |
|-------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| $F$-max | 0.054          | 0.056          | 0.040          | 0.046          | 0.053          | 0.058          | 0.043          |
| $p$-min | 0.000          | 0.000          | 0.000          | 0.000          | 0.000          | 0.000          | 0.000          |
| ERL    | 0.060          | 0.052          | 0.045          | 0.043          | 0.047          | 0.057          | 0.045          |
| Cont   | 0.058          | 0.056          | 0.031          | 0.043          | 0.045          | 0.049          | 0.053          |
| Area   | 0.058          | 0.047          | 0.040          | 0.040          | 0.043          | 0.061          | 0.045          |

Note: The chosen standard deviation $\sigma$ for each case is the one with maximal contrast between methods.

TABLE 4  Empirical powers of all studied methods based on 1000 replicates for model M1 with error (a) to (g)

|       | (a) $\sigma_5$ | (b) $\sigma_5$ | (c) $\sigma_4$ | (d) $\sigma_4$ | (e) $\sigma_7$ | (f) $\sigma_5$ | (g) $\sigma_6$ |
|-------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| $F$-max | 0.395          | 0.464          | 0.267          | 0.334          | 0.294          | 0.632          | 0.421          |
| $p$-min | 0.000          | 0.000          | 0.000          | 0.000          | 0.000          | 0.000          | 0.000          |
| ERL    | 0.606          | 0.594          | 0.948          | 0.943          | 0.751          | 0.776          | 0.984          |
| Cont   | 0.378          | 0.399          | 0.228          | 0.363          | 0.335          | 0.522          | 0.358          |
| Area   | 0.544          | 0.568          | 0.879          | 0.909          | 0.577          | 0.749          | 0.979          |

All the images in the simulation study had the resolution $51 \times 51$ pixels. The number of permutations used throughout the whole simulation study was set to 2000. The estimated significance levels and powers were recorded for various cases for all studied multiple comparison methods, that is, for $F$-max, $p$-min, and the proposed ERL, continuous, and area rank methods. To capture the behavior of the methods for various levels of significance, seven different standard deviations $\sigma_1 = 0.1$, $\sigma_2 = 0.2$, $\sigma_3 = 0.3$, $\sigma_4 = 0.5$, $\sigma_5 = 0.75$, $\sigma_6 = 1$, and $\sigma_7 = 1.25$ were used in all studied cases. Each model was simulated with 10 images per group and each experiment was repeated 1000 times in order to obtain estimated significance levels and powers.

Tables 3 to 5 show the results for models M0, M1, and M1’ in a shorted way. The detailed results are presented in Appendix C.

The estimated significance levels revealed the same structure for all errors (a) to (g) (Table 3): The $p$-min was enormously conservative. The $F$-max and all our proposed methods achieved the preset significance level $\alpha = 0.05$ in all cases. (The 95% confidence interval for 1000 simulations with success probability 0.05 is (0.037, 0.063).) The only exception was the error (c), where Cont was slightly conservative.

The conservative $p$-min method had no power (Tables 4 and 5). The ERL and Area methods had uniformly higher power than $F$-max for all errors in the case of the model M1. This is caused by the fact that ERL and Area summarize the extremeness of the test statistic from all spatial points. This advantage led to higher power even in the cases (f) and (g) with inhomogeneous autocorrelation.

The situation is more complicated for model M1’, where the area of extremeness is rather small: the ERL and Area cannot benefit from their feature of collecting information from all spatial points. For the homogeneous cases with normal error (a) and skewed error (b), the $F$-max is slightly more powerful than our proposed methods. When the inhomogeneity of errors is set in such a way that the values of the test statistic quantiles are larger at the periphery where no differences are present—cases (c), (d), and (e)—the ERL and Area methods are significantly more powerful than $F$-max (Table 5). Finally, when the range of correlation is larger at the periphery where no differences are present, the methods cumulating information from the neighborhood should be affected by this fact. Really, in case (f), ERL appears to be less powerful than $F$-max, but the Area method which uses different sources of information (see Table 1) is affected by this fact only a little. This decrease of power is not further seen in case (g), where the normal distribution is replaced by the bimodal distribution. At last, we observe that Cont and $F$-max methods were rather equivalent in our study with 2601 spatial points and 2000 permutations.
TABLE 5  Empirical powers of all studied methods based on 1000 replicates for model M1\' with error (a) to (g)

|     | (a) $\sigma_1$ | (b) $\sigma_2$ | (c) $\sigma_7$ | (d) $\sigma_3$ | (e) $\sigma_4$ | (f) $\sigma_3$ | (g) $\sigma_1$ |
|-----|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| $F$-max | 0.917 | 0.562 | 0.377 | 0.549 | 0.190 | 0.963 | 0.412 |
| $p$-min | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| ERL | 0.732 | 0.365 | 0.972 | 0.752 | 0.235 | 0.720 | 0.894 |
| Cont | 0.799 | 0.461 | 0.339 | 0.445 | 0.221 | 0.899 | 0.408 |
| Area | 0.825 | 0.500 | 0.915 | 0.800 | 0.259 | 0.924 | 0.924 |

Note: The chosen standard deviation $\sigma$ for each case is the one with maximal contrast between methods.

4  COMPUTATIONAL DETAILS: HOW TO IMPLEMENT THE METHODS FOR LARGE DATA

The $p$-min, ERL, Cont, and Area measures are based either on ordinary or continuous pointwise ranks of the test statistics $T_j(r), J = 0, 1, \ldots, J$, among each other. A naive algorithm would first calculate all the ranks and save them in the memory in order to calculate the measures thereafter. However, the space complexity of the naive algorithm is $O(Jn)$, and for large data, all the ranks may not fit comfortably to the memory of an ordinary computer. However, for all the proposed measures, it is possible to split the task by the locations (voxels) $r \in I$. In this case, the space complexity is just $O(Js)$, where $s$ is the number of subjects. We note that the space complexity can be even further reduced to $O(J)$ if the permutations are not saved. This can be achieved by resetting the seed of the pseudo-random number generator for each split. This is typically not necessary.

There is some extra computational complexity of the rank based multiple comparisons in comparison to the $F$-max method because the test statistics need to be ranked. Namely, letting $S$ denote the time complexity of the test statistic to be computed, the complexity of computing rank based multiple comparison correction is $O(JnS + Jn \log J)$, while for $F$-max it is $O(JnS)$. Thus, asymptotically the rank based corrections are slower than $F$-max. This is the same whether the naive or location-wise calculations are used.

Algorithm 1 presents the algorithm for location-wise (voxel by voxel) calculation of the rank-based measures. The update rules and calculation of the final measures are described thereafter for the different measures in Sections 4.1 to 4.3. The implementation is obvious for the $p$-min and Cont measures, while for ERL and Area, some more complicated calculations are needed. The computations are implemented in the function partial_order of the R package GET.\footnote{15}{Section 4.4 explains how the global envelope can be calculated given the measures.}

Algorithm 1. An algorithm to compute functional ordering voxel by voxel. Below ranks refers to the pointwise ranks, either ordinary (for $p$-min and ERL) or continuous (for Cont and Area). The update rules as well as final calculations are described in Section 4.1 for $p$-min and Cont, in Section 4.2 for Area and in Section 4.3 for ERL.

**Algorithm 1**

```
Initialize the (augmented) measure $M_j$ for $j = 0, \ldots, J$
Generate $J$ permutations
for each voxel do
    $T_0 \leftarrow$ $(F)$-statistic for data
    for $j \leftarrow 1, \ldots, J$ do
        $T_j \leftarrow$ $(F)$-statistic for permutation $j$
    end for
    $(m_0, m_1, \ldots, m_J) \leftarrow$ ranks$(T_0, T_1, \ldots, T_J)$
    for $j \leftarrow 0, \ldots, J$ do
        $M_j \leftarrow$ update$(M_j, m_j)$
    end for
end for
Compute the final measures from the augmented measures
```
4.1 | Update rule for $p$-min and Cont

For the $p$-min and Cont measures, the update rule is simple:

$$ M_j \leftarrow \min(M_j, m_j). $$

When all locations $r \in I$ have been handled, the final measure, either $p_{j}^{\min}$ or $c_j$, is obtained by dividing the $M_j$ by $J + 1$.

4.2 | Update rule for Area

For the Area and ERL measures, it is necessary to augment the measure with some auxiliary information during the computation. Namely, for the Area measure, the extreme rank $R_j$ and the difference between the extreme rank $R_j$ and the pointwise continuous rank $C_j(r)$ has to be saved and updated for data and each permutation $j$. Let $D_j$ denote the difference. Initially $M_j = (R_j, D_j) = (\infty, 0)$. For the update, there are three possibilities:

$$ M_j = (R_j, D_j) \leftarrow \begin{cases} (R_j, D_j) & \text{if } \text{ceil}(m_i) > R_i, \\ (R_j, D_j + \text{ceil}(m_j) - m_j) & \text{if } \text{ceil}(m_j) = R_j, \\ (\text{ceil}(m_j), \text{ceil}(m_j) - m_j) & \text{if } \text{ceil}(m_j) < R_j. \end{cases} $$

The final measure is

$$ a_j = \frac{1}{J + 1}(R_j - D_j/N). $$

4.3 | Update rules for ERL

For ERL, the augmented measure $M_j$ is a vector of the chosen number of most extreme pointwise ranks of $T_j(r)$ and their counts among $r \in I$. We record the six most extreme ranks. Thus let

$$ M_j = (R_1, R_2, \ldots, R_6, c_1, c_2, \ldots, c_6) = (\infty, \ldots, \infty, 0, \ldots, 0) $$

be the initialized vector. Then the updates are obtained by the following rules:

- If $m_j = R_i$ for some $i = 1, \ldots, 6$, then $c_i \leftarrow c_i + 1$.
- If $R_{i-1} < m_j < R_i$ for some $i = 1, \ldots, 6$, then

$$ R_k \leftarrow R_{k-1}, c_k \leftarrow c_{k-1} \text{ for } k = i + 1, \ldots, 6 $$

and $R_i \leftarrow m_j, c_i \leftarrow 1$.

The final measure for ERL is the ranking of the augmented measures $M_j$ according to a special ordering that is defined as follows. Let $R_i, c_i$ and $R'_i, c'_i$ be the augmented measures for two functions 1 and 2.

1. Start with the most extreme rank $i \leftarrow 1$.
2. if $R_i < R'_i$ then function 1 is more extreme.
3. if $R_i > R'_i$ then function 2 is more extreme.
4. if $R_i = R'_i$ and $c_i < c'_i$ then function 2 is more extreme.
5. if $R_i = R'_i$ and $c_i > c'_i$ then function 1 is more extreme.
6. if $R_i = R'_i$ and $c_i = c'_i$ then set $i \leftarrow i + 1$ and go to step 2.

The final normalized ERL measures (6) are obtained by dividing by $J + 1$. 
4.4 | Computation of the global envelopes

After the final measures have been calculated and the p-value of the test is less than the chosen significance level, it is possible to construct the envelope to detect which parts of the domain have led to the rejection of the null hypothesis. Algorithm 2 describes the computation.

**Algorithm 2.** An algorithm to compute global envelopes after the functional orderings have been calculated

Let $M_j$ be the final measure for $j = 0, \ldots, J$ and $M^{(a)}$ the critical value, that is, the largest of the $M_i$ such that the number of those $j$ for which $M_j < M^{(a)}$ is less or equal to $a(J + 1)$

Use the same $J$ permutations as when computing the functional orderings

for each voxel $v$
do

$T_0 \leftarrow (F^*)$-statistic for data

for $j \leftarrow 1, \ldots, J$
do

$T_j \leftarrow (F^*)$-statistic for permutation $j$

end for

$U_v \leftarrow \max \{ T_j : M_j > M^{(a)} \}$

end for

$U_v$ contains the upper envelope.

5 | DISCUSSION

All the permutation based multiple testing procedures reach the prescribed significance level if the permutation strategy leads to exchangeable test statistics. They even reach the prescribed level approximately in the presence of nuisance factors if the Freedman and Lane permutation strategy is used. This fact was demonstrated here by a simulation study and the Eklund et al\textsuperscript{5} type of experiment presented in Appendix A.

The measures used in the permutation test are different in two ways, their sensitiveness toward different types of extremeness and their robustness against different types of inhomogeneities of the error term $\epsilon(r)$ across $r \in I$. The sensitivites can be understood from the sources of information which are gathered by the different measures (see Table 1): The integral type of extremeness is detected by those measures (ERL, Area) that gather information across the locations $r \in I$ with the minimal p-values, while the measures based on the value of extremes (Cont, Area) are sensitive to the maximum type of extremeness. The categorization is given by the construction of the measures, and it was further supported by the simulation study results for cases M1 and M1'. Table 4 shows that the gain in power by using ERL or Area with respect to other procedures in cases when the extremeness is of integral type can be enormous.

Table 6 summarizes the investigated measures and their problems with the identification of the alternative hypothesis, that is, with ties and various inhomogeneities of the error term. The inhomogeneities can decrease the power of the procedures, as stated in the table. As it was stated already in Section 1, the inhomogeneity and non-normality of error term $\epsilon(r)$ across $r \in I$ together lead to inhomogeneous quantiles of the $F$-statistic. Therefore, the $F$-max procedure can be blind to the departures where the quantiles of $T(r)$ are less variable. This was demonstrated in Section 2.5 and in the simulation study by cases (c), (d), and (e) in Table 5. Since the Cont measure also depends on the values of $T(r)$ and not only ranks, the Cont measure has the same problem as $F$-max with inhomogeneity of $T(r)$. However, the problem of Cont depends on the number of ties to break: the bigger their number is, the less sensitive Cont is expected to be to inhomogeneity.

On the other hand, in the presence of the inhomogeneity of the correlation structure of $\epsilon(r)$ across $r \in I$, the ERL measure can be blind to the departures from the null that occur in areas where the range of correlation of $\epsilon(r)$ is smaller. This was demonstrated by the simulation study cases (f) and (g) in Table 5.

The Area rank measure is a combination of Cont and ERL, and thus it should be sensitive to both inhomogeneities, but as the simulation study shows, the decrease of power in cases (c), (d), (e), (f), and (g) was not as apparent as for the $F$-max, ERL, or Cont procedures. Thus the Area measure appears to be the most robust method with respect to inhomogeneities of the error term $\epsilon(r)$.

Further, the simulation study case (g) showed that the rank based measures are much more robust with respect to extreme non-normality than the $F$-max test as the decrease of power between cases (f) and (g) is apparent only for $F$-max procedure.
### TABLE 6 Problems of different measures

|       | Ties | Inhomogeneity and non-normality of $\epsilon(r)$ | Inhomogeneity of correlation structure of $\epsilon(r)$ |
|-------|------|-------------------------------------------------|-----------------------------------------------------|
| $F$-max | No   | Yes                                             | No                                                   |
| $p$-min | Yes  | No                                              | No                                                   |
| ERL   | No   | No                                              | Yes                                                  |
| Cont  | No   | Yes                                             | No                                                   |
| Area  | No   | Partial                                         | Partial                                              |

## 6 CONCLUSIONS

We presented three new multiple comparison methods for the permutation inference for the GLM. We showed by a designed experiment that the proposed methods have the desired significance level, unlike the $p$-min method that was highly conservative unless an enormous amount of permutations was used. The $p$-min method can well be recommended if a sufficiently large amount of permutations can be afforded. Otherwise, other methods must be used. The choice of the method from $F$-max, ERL, Cont, and Area rank measures should depend on the assumptions about error term $\epsilon(r)$ that can be made and the expected type of extremeness of the test statistic. Both homogeneity of the distribution of $\epsilon(r)$ (HD) and homogeneity of the correlation structure of $\epsilon(r)$ (HCS) across the image play a role (see Table 6). When we assume

- Both HD and HCS, then all methods can be applied. The ERL or Area methods should be preferred when the test statistic is expected to be extreme over a large area of the image $I$; $F$-max, Cont, or Area when extremeness is expected only on a small area. The same recommendations hold if the assumption of homogeneity of $\epsilon(r)$ is replaced by the assumption of the normality of the error $\epsilon(r)$.
- HCS but not HD, then only ERL can be applied without worry of losing power to identify alternative hypotheses. However, in our designed experiment, Area showed good robustness with respect to inhomogeneity of the distribution of $\epsilon(r)$, whereas $F$-max and Cont did not.
- HD but not HCS, then $F$-max and Cont can be applied without worry of losing power to identify alternative hypotheses. Again Area showed good robustness with respect to inhomogeneity of the correlation structure of $\epsilon(r)$, whereas ERL did not.
- Neither HD nor HCS, then none of the methods can be applied without worry of losing power to identify alternative hypotheses, but Area showed good robustness with respect to both types of inhomogeneities.

Thus, our experience suggests that the Area method can be used in a general situation without worry of losing much power. Its dual nature of being based both on the extremeness of the test statistic as well as on summarizing this extremeness across spatial points makes it sensitive to different departures from the null model and robust to different kinds of inhomogeneity. This was demonstrated by a designed experiment as well as an example of a real data study. For these reasons, we believe that the ERL and Area methods can help to find further understanding of the phenomena studied by functions or images through GLMs.

### ACKNOWLEDGEMENTS

Tomáš Mrkvička has been financially supported by the Grant Agency of Czech Republic (Project No. 19-04412S), and Mari Myllymäki and Mikko Kuronen by the Academy of Finland (project numbers 295100, 306875, 327211). The authors wish to acknowledge CSC – IT Center for Science, Finland for computational resources.

### DATA AVAILABILITY STATEMENT

We confirm that the data of autism brain imaging from ABIDE are freely available at http://fcon_1000.projects.nitrc.org/indi/abide/abide_I.html.

### ORCID

Tomáš Mrkvička [https://orcid.org/0000-0003-1613-2780](https://orcid.org/0000-0003-1613-2780)

Mari Myllymäki [https://orcid.org/0000-0002-2713-7088](https://orcid.org/0000-0002-2713-7088)
REFERENCES

1. Winkler A, Ridgway G, Webster M, Smith S, Nichols T. Permutation inference for the general linear model. *NeuroImage*. 2014;92:381-397.

2. Nichols TE, Holmes E. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human Brain M*. 2001;15:1-25.

3. Lopez-Pintado S, Qian K. A depth-based global envelope test for comparing two groups of functions with applications to biomedical data. *Stat Med*. 2021;40(7):1639-1652. doi:10.1002/sim.8861

4. Fisher RA. *The Design of Experiments*. Oxford, UK: Oliver & Boyd; 1935.

5. Eklund A, Nichols TE, Knutsson H. Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. *Proc Natl Acad Sci*. 2016;113(28):7900-7905. doi:10.1073/pnas.1602413113

6. Pantazis D, Nichols TE, Baillet S, Leahy RM. A comparison of random field theory and permutation methods for the statistical analysis of MEG data. *NeuroImage*. 2005;25:383-394.

7. Hayasaka S, Phan K, Liberzon I, Worsley KJ, Nichols TE. Nonstationary cluster-size inference with random field and permutation methods. *NeuroImage*. 2004;22(2):676-687. doi:10.1016/j.neuroimage.2004.01.041

8. Salimi-Khorshidi G, Smith SM, Nichols TE. Adjusting the effect of nonstationarity in cluster-based and TFCE inference. *NeuroImage*. 2011;54(3):2006-2019. doi:10.1016/j.neuroimage.2010.09.088

9. Myllymäki M, Mrkvička T, Grabarnik P, Seijo H, Hahn U. Global envelope tests for spatial processes. *J R Stat Soc B*. 2017;79(2):381-404. doi:10.1111/rssb.12172

10. Narisetty NN, Nair VN. Extremal depth for functional data and applications. *J Am Stat Assoc*. 2016;111(516):1705-1714.

11. Hahn U. A note on simultaneous Monte Carlo tests. Technical report. Centre for Stochastic Geometry and Advanced Bioimaging, Aarhus University; 2015.

12. Mrkvička T, Myllymäki M, Hahn U. Multiple Monte Carlo testing, with applications in spatial point processes. *Stat Comput*. 2017;27(5):1239-1255. doi:10.1007/s11222-016-9683-9

13. Mrkvička T, Myllymäki M, Jílek M, Hahn U. A one-way ANOVA test for functional data with graphical interpretation. *Kybernetika*. 2020;56(3):432-458. doi:10.14736/kyb-2020-3-0432

14. Di Martino A, Yan C, Li Q, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol Psychiatry*. 2014;19:659-667.

15. Myllymäki M, Mrkvička T. GET: global envelopes in R; 2020. arXiv:1911.06583 [stat.ME].

16. Christensen R. *Plane Answers to Complex Questions*. New York, NY: Springer; 2002.

17. Freedman D, Lane D. A nonstochastic interpretation of reported significance levels. *J Bus Econ Stat*. 1983;1(4):292-298.

18. Anderson M, Ter Braak C. Permutation tests for multi-factorial analysis of variance. *J Stat Comput Simul*. 2003;73(2):85-113.

19. Westfall PH, Young SS. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. 1st ed., New York, NY: Wiley; 1993.

20. Phipson B, Smyth GK. Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. *Stat Appl Genet Mol Biol*. 2010;9(1):39. doi:10.2202/1544-6115.1585

21. Xu M, Reiss P. Distribution-free pointwise adjusted P-values for functional hypotheses. In: Aneiros G, Horová I, Hušková MPV, eds. *Functional and High-Dimensional Statistics and Related Fields*. Contributions to Statistics. IWFOS 2020. Cham, Switzerland: Springer; 2020.

22. Zou QH, Zhu CZ, Yang Y, et al. An improved approach to detection of amplitude of low-frequency fluctuation (ALFF) for resting-state fMRI: fractional ALFF. *J Neurosci Methods*. 2008;172:137-141.

How to cite this article: Mrkvička T, Myllymäki M, Kuronen M, Narisetty NN. New methods for multiple testing in permutation inference for the general linear model. *Statistics in Medicine*. 2022;41(2):276-297. doi:10.1002/sim.9236

APPENDIX A. STUDY ON FALSE DISCOVERIES

To study the significant levels of the ERL, Cont, and Area tests by an experiment similar to those proposed by Eklund et al., we used the measurements of the 573 typical controls of the autism brain imaging data of Section 2.5. For this experiment, we chose the left and right Crus Cerebellum 1 regions of the brain containing 4783 voxels. We took 1000 samples of 20 control subjects, and each sample was randomly split into two equally sized groups. For each sample, we tested the effect of the group in the model (1) where age was a nuisance factor, using 2499 permutations by the Freedman-Lane algorithm. This number of permutations was chosen to have the number of permutations about half of the number of voxels as in the illustrative example of Section 2.5. All the empirical significance levels were 0.040.
APPENDIX B. THE GLOBAL ENVELOPE FOR THE BRAIN DATA EXAMPLE

Figure B1 shows the 95% global Area envelope and the significant voxels for the whole brain data for the example discussed in Section 2.5. The number of simulations was 100,000 and $p$-value 0.040 for the test of the effect of the group.

**Figure B1** The upper 95% envelope of the voxelwise $F$-statistic of the studied model where the type (ASD or TC) was the interesting factor (blue image) together with the significant voxels detected by the area rank test (red crosses). The brain is shown by 71 2D slices (with depths varying from 7 to 77)
APPENDIX C. FURTHER RESULTS FOR THE SIMULATION STUDY

Tables C1 to C15 contain all results of the simulation study for M1 and M1′.

**TABLE C1**  Estimated significance levels of all studied methods and seven standard deviations for model M0 with error (a)

|       | \( \sigma_1 \) | \( \sigma_2 \) | \( \sigma_3 \) | \( \sigma_4 \) | \( \sigma_5 \) | \( \sigma_6 \) | \( \sigma_7 \) |
|-------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| \( F \)-max | 0.054          | 0.048          | 0.043          | 0.051          | 0.044          | 0.045          | 0.047          |
| \( p \)-min | 0.000          | 0.000          | 0.000          | 0.000          | 0.000          | 0.000          | 0.000          |
| ERL    | 0.060          | 0.036          | 0.042          | 0.048          | 0.052          | 0.046          | 0.040          |
| Cont   | 0.058          | 0.060          | 0.047          | 0.042          | 0.052          | 0.055          | 0.047          |
| Area   | 0.058          | 0.040          | 0.041          | 0.053          | 0.047          | 0.050          | 0.040          |

**TABLE C2**  Empirical powers of all studied methods and seven standard deviations for model M1 with error (a)

|       | \( \sigma_1 \) | \( \sigma_2 \) | \( \sigma_3 \) | \( \sigma_4 \) | \( \sigma_5 \) | \( \sigma_6 \) | \( \sigma_7 \) |
|-------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| \( F \)-max | 1.000          | 1.000          | 1.000          | 0.863          | 0.395          | 0.184          | 0.106          |
| \( p \)-min | 0.000          | 0.000          | 0.000          | 0.000          | 0.000          | 0.000          | 0.000          |
| ERL    | 0.979          | 0.975          | 0.980          | 0.943          | 0.606          | 0.292          | 0.188          |
| Cont   | 0.979          | 0.975          | 0.980          | 0.824          | 0.378          | 0.170          | 0.105          |
| Area   | 0.979          | 0.975          | 0.980          | 0.929          | 0.544          | 0.257          | 0.150          |

**TABLE C3**  Empirical powers of all studied methods and seven standard deviations for model M1 with error (b)

|       | \( \sigma_1 \) | \( \sigma_2 \) | \( \sigma_3 \) | \( \sigma_4 \) | \( \sigma_5 \) | \( \sigma_6 \) | \( \sigma_7 \) |
|-------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| \( F \)-max | 1.000          | 1.000          | 1.000          | 0.886          | 0.464          | 0.242          | 0.151          |
| \( p \)-min | 0.000          | 0.000          | 0.000          | 0.000          | 0.000          | 0.000          | 0.000          |
| ERL    | 0.973          | 0.980          | 0.978          | 0.942          | 0.594          | 0.325          | 0.233          |
| Cont   | 0.973          | 0.980          | 0.978          | 0.844          | 0.399          | 0.215          | 0.163          |
| Area   | 0.973          | 0.980          | 0.978          | 0.927          | 0.568          | 0.303          | 0.215          |

**TABLE C4**  Empirical powers of all studied methods and seven standard deviations for model M1 with error (c)

|       | \( \sigma_1 \) | \( \sigma_2 \) | \( \sigma_3 \) | \( \sigma_4 \) | \( \sigma_5 \) | \( \sigma_6 \) | \( \sigma_7 \) |
|-------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| \( F \)-max | 1.000          | 1.000          | 0.959          | 0.267          | 0.097          | 0.067          | 0.058          |
| \( p \)-min | 0.000          | 0.000          | 0.000          | 0.000          | 0.000          | 0.000          | 0.000          |
| ERL    | 0.986          | 0.976          | 0.979          | 0.948          | 0.520          | 0.269          | 0.175          |
| Cont   | 0.986          | 0.976          | 0.945          | 0.228          | 0.090          | 0.065          | 0.050          |
| Area   | 0.986          | 0.976          | 0.979          | 0.879          | 0.395          | 0.191          | 0.121          |
| TABLE C5 | Empirical powers of all studied methods and seven standard deviations for model M1 with error (d) |
|-----------------|--------------------------------------------------|
| $\sigma_1$ | $\sigma_2$ | $\sigma_3$ | $\sigma_4$ | $\sigma_5$ | $\sigma_6$ | $\sigma_7$ |
| $F$-max | 1.000 | 1.000 | 0.989 | 0.334 | 0.080 | 0.054 | 0.048 |
| $p$-min | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| ERL | 0.980 | 0.986 | 0.982 | 0.943 | 0.617 | 0.299 | 0.199 |
| Cont | 0.980 | 0.986 | 0.958 | 0.363 | 0.097 | 0.058 | 0.054 |
| Area | 0.980 | 0.986 | 0.982 | 0.909 | 0.492 | 0.229 | 0.157 |

| TABLE C6 | Empirical powers of all studied methods and seven standard deviations for model M1 with error (e) |
|-----------------|--------------------------------------------------|
| $\sigma_1$ | $\sigma_2$ | $\sigma_3$ | $\sigma_4$ | $\sigma_5$ | $\sigma_6$ | $\sigma_7$ |
| $F$-max | 1.000 | 0.997 | 0.979 | 0.857 | 0.609 | 0.410 | 0.294 |
| $p$-min | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| ERL | 0.980 | 0.970 | 0.977 | 0.967 | 0.930 | 0.844 | 0.751 |
| Cont | 0.980 | 0.968 | 0.960 | 0.865 | 0.662 | 0.491 | 0.335 |
| Area | 0.980 | 0.970 | 0.975 | 0.947 | 0.867 | 0.714 | 0.577 |

| TABLE C7 | Empirical powers of all studied methods and seven standard deviations for model M1 with error (f) |
|-----------------|--------------------------------------------------|
| $\sigma_1$ | $\sigma_2$ | $\sigma_3$ | $\sigma_4$ | $\sigma_5$ | $\sigma_6$ | $\sigma_7$ |
| $F$-max | 1.000 | 1.000 | 1.000 | 0.998 | 0.632 | 0.296 | 0.143 |
| $p$-min | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| ERL | 0.981 | 0.980 | 0.985 | 0.985 | 0.776 | 0.325 | 0.126 |
| Cont | 0.981 | 0.980 | 0.985 | 0.971 | 0.522 | 0.249 | 0.129 |
| Area | 0.981 | 0.980 | 0.985 | 0.984 | 0.749 | 0.340 | 0.141 |

| TABLE C8 | Empirical powers of all studied methods and seven standard deviations for model M1 with error (g) |
|-----------------|--------------------------------------------------|
| $\sigma_1$ | $\sigma_2$ | $\sigma_3$ | $\sigma_4$ | $\sigma_5$ | $\sigma_6$ | $\sigma_7$ |
| $F$-max | 0.999 | 0.950 | 0.841 | 0.658 | 0.522 | 0.421 | 0.352 |
| $p$-min | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| ERL | 0.986 | 0.980 | 0.977 | 0.979 | 0.981 | 0.984 | 0.978 |
| Cont | 0.985 | 0.917 | 0.790 | 0.604 | 0.448 | 0.358 | 0.300 |
| Area | 0.986 | 0.980 | 0.977 | 0.979 | 0.981 | 0.979 | 0.970 |

| TABLE C9 | Empirical powers of all studied methods and seven standard deviations for model M1’ with error (a) |
|-----------------|--------------------------------------------------|
| $\sigma_1$ | $\sigma_2$ | $\sigma_3$ | $\sigma_4$ | $\sigma_5$ | $\sigma_6$ | $\sigma_7$ |
| $F$-max | 1.000 | 1.000 | 0.917 | 0.277 | 0.089 | 0.068 | 0.045 |
| $p$-min | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| ERL | 0.975 | 0.976 | 0.732 | 0.201 | 0.071 | 0.064 | 0.052 |
| Cont | 0.975 | 0.975 | 0.799 | 0.230 | 0.077 | 0.058 | 0.048 |
| Area | 0.975 | 0.980 | 0.825 | 0.254 | 0.085 | 0.068 | 0.049 |
**TABLE C10** Empirical powers of all studied methods and seven standard deviations for model M1' with error (b)

|        | $\sigma_1$ | $\sigma_2$ | $\sigma_3$ | $\sigma_4$ | $\sigma_5$ | $\sigma_6$ | $\sigma_7$ |
|--------|------------|------------|------------|------------|------------|------------|------------|
| $F$-max| 1.000      | 1.000      | 0.906      | 0.280      | 0.107      | 0.076      | 0.060      |
| $p$-min| 0.000      | 0.000      | 0.000      | 0.000      | 0.000      | 0.000      | 0.000      |
| ERL    | 0.980      | 0.962      | 0.705      | 0.193      | 0.076      | 0.071      | 0.053      |
| Cont   | 0.980      | 0.962      | 0.811      | 0.232      | 0.099      | 0.068      | 0.070      |
| Area   | 0.980      | 0.968      | 0.830      | 0.243      | 0.100      | 0.076      | 0.061      |

**TABLE C11** Empirical powers of all studied methods and seven standard deviations for model M1' with error (c)

|        | $\sigma_1$ | $\sigma_2$ | $\sigma_3$ | $\sigma_4$ | $\sigma_5$ | $\sigma_6$ | $\sigma_7$ |
|--------|------------|------------|------------|------------|------------|------------|------------|
| $F$-max| 1.000      | 1.000      | 1.000      | 1.000      | 0.965      | 0.377      |           |
| $p$-min| 0.000      | 0.000      | 0.000      | 0.000      | 0.000      | 0.000      | 0.000      |
| ERL    | 0.982      | 0.976      | 0.984      | 0.975      | 0.978      | 0.978      | 0.972      |
| Cont   | 0.982      | 0.976      | 0.984      | 0.975      | 0.978      | 0.686      | 0.339      |
| Area   | 0.982      | 0.976      | 0.984      | 0.975      | 0.978      | 0.978      | 0.915      |

**TABLE C12** Empirical powers of all studied methods and seven standard deviations for model M1' with error (d)

|        | $\sigma_1$ | $\sigma_2$ | $\sigma_3$ | $\sigma_4$ | $\sigma_5$ | $\sigma_6$ | $\sigma_7$ |
|--------|------------|------------|------------|------------|------------|------------|------------|
| $F$-max| 1.000      | 0.993      | 0.549      | 0.060      | 0.066      | 0.062      | 0.048      |
| $p$-min| 0.000      | 0.000      | 0.000      | 0.000      | 0.000      | 0.000      | 0.000      |
| ERL    | 0.979      | 0.973      | 0.752      | 0.188      | 0.097      | 0.074      | 0.056      |
| Cont   | 0.979      | 0.907      | 0.445      | 0.080      | 0.053      | 0.054      | 0.047      |
| Area   | 0.979      | 0.974      | 0.800      | 0.194      | 0.085      | 0.072      | 0.048      |

**TABLE C13** Empirical powers of all studied methods and seven standard deviations for model M1' with error (e)

|        | $\sigma_1$ | $\sigma_2$ | $\sigma_3$ | $\sigma_4$ | $\sigma_5$ | $\sigma_6$ | $\sigma_7$ |
|--------|------------|------------|------------|------------|------------|------------|------------|
| $F$-max| 0.782      | 0.496      | 0.351      | 0.190      | 0.113      | 0.086      | 0.065      |
| $p$-min| 0.000      | 0.000      | 0.000      | 0.000      | 0.000      | 0.000      | 0.000      |
| ERL    | 0.752      | 0.505      | 0.384      | 0.235      | 0.167      | 0.084      | 0.093      |
| Cont   | 0.791      | 0.526      | 0.371      | 0.221      | 0.134      | 0.089      | 0.072      |
| Area   | 0.826      | 0.583      | 0.422      | 0.259      | 0.163      | 0.093      | 0.087      |

**TABLE C14** Empirical powers of all studied methods and seven standard deviations for model M1' with error (f)

|        | $\sigma_1$ | $\sigma_2$ | $\sigma_3$ | $\sigma_4$ | $\sigma_5$ | $\sigma_6$ | $\sigma_7$ |
|--------|------------|------------|------------|------------|------------|------------|------------|
| $F$-max| 1.000      | 1.000      | 0.963      | 0.358      | 0.089      | 0.058      | 0.057      |
| $p$-min| 0.000      | 0.000      | 0.000      | 0.000      | 0.000      | 0.000      | 0.000      |
| ERL    | 0.975      | 0.978      | 0.720      | 0.110      | 0.059      | 0.051      | 0.045      |
| Cont   | 0.975      | 0.979      | 0.899      | 0.275      | 0.083      | 0.062      | 0.047      |
| Area   | 0.975      | 0.979      | 0.924      | 0.258      | 0.077      | 0.056      | 0.048      |
|       | σ₁   | σ₂   | σ₃   | σ₄   | σ₅   | σ₆   | σ₇   |
|-------|------|------|------|------|------|------|------|
| $F$-max | 0.412 | 0.228 | 0.146 | 0.101 | 0.084 | 0.073 | 0.067 |
| $p$-min | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| ERL   | 0.894 | 0.589 | 0.311 | 0.193 | 0.186 | 0.132 | 0.120 |
| Cont  | 0.408 | 0.189 | 0.124 | 0.094 | 0.081 | 0.063 | 0.073 |
| Area  | 0.924 | 0.612 | 0.424 | 0.292 | 0.208 | 0.154 | 0.134 |