Computational issues when testing on the base of functional test statistic: The R package GET

based on joint work with Mari Myllymäki and many others

Tomáš Mrkvička

Faculty of Economics
University of South Bohemia

Examples of functional test statistic

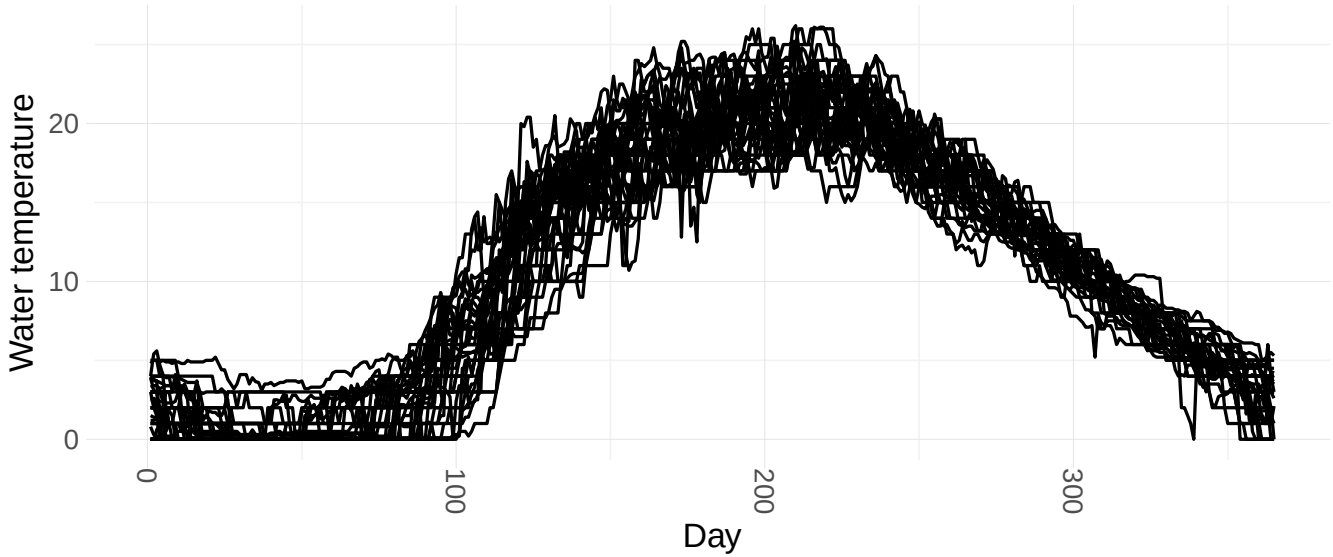


Figure – Annual water temperature curves sampled at the water level of Rimov reservoir in Czech Republic every day from 1979 to 2014.

The functional test statistic is the regression coefficient β computed at every day of the year. $Temperature(Day, Year) \sim \beta_0(Day) + \beta_1(Day) * Year$
This is function-on-scalar regression, but we put no assumption on the model, regarding distribution and homogeneity of the distribution. Here winter days are distributed differently than summer days.
Question : In which days of a year do we observe warming ?

Examples of functional test statistic

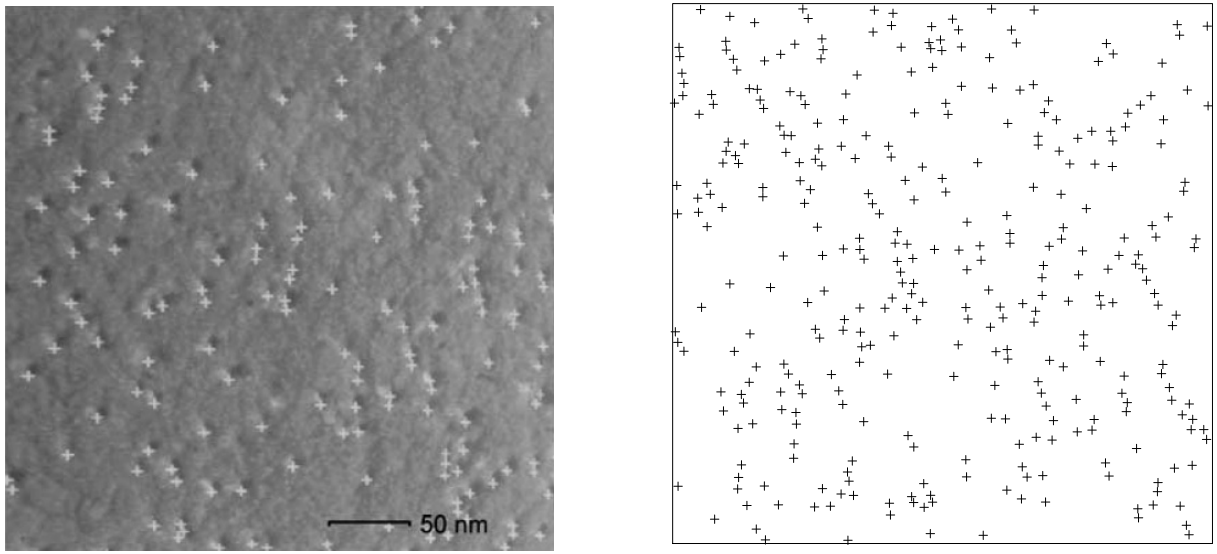


Figure – Electron micrographs of intramembraneous particles.

The functional test statistic is the transformation of Ripley's K function - The expected number of points up to a distance r divided by intensity of points.
Question : Which particle distances deviates from complete spatial randomness ?

Monte Carlo test / Permutation test

- We observe test statistic $T_0(r)$.
- Assume we can generate replicates $T_1(r), \ldots, T_s(r)$ of the test statistic under the null hypothesis. Either by reproducing data (Monte Carlo) or permuting data.
- Sort the function from most extreme to least extreme using a measure M. $T_i(r) \prec T_j(r)$, i.e. T_i is more extreme than T_j, iff $M_i < M_j$.
- Global p-value is $p = \sum_{i=0}^{s} \mathbf{1}(M_i \leq M_0)/(s+1)$.

Global envelope test / central region

• If we are able to order the functions $T_0(r), \ldots, T_s(r)$, we are able to produce the central region.

1. Denote $m_{(\alpha)} \in \mathbb{R}$ the largest of the M_i such that the number of $i : M_i < m_{(\alpha)}$ is less or equal αs ;
2. Denote $[\mathbf{T}^{(\alpha)}_{\text{low}}(r) = \min\{ T_i(r) : M_i \geq m_{(\alpha)}\}, \mathbf{T}^{(\alpha)}_{\text{upp}}(r) = \max\{ T_i(r) : M_i \geq m_{(\alpha)}\}]$ the $100(1-\alpha)\%$ central region.

• We are interested ONLY in ordering satisfying the intrinsic graphical interpretation (IGI).

• The 95% central region can be used to produce the global envelope test constructed from $T_0(r)$ test statistic of the data and $T_1(r), \ldots, T_s(r)$ resamples under null hypothesis.

Intrinsic graphical interpretation (IGI)

Intrinsic graphical interpretation with respect to the ordering $\prec$ induced by measure M satisfies :

1. $[\ T_i(r) < \mathbf{T}^{(\alpha)}_{\mathrm{low}}(r)$ or $T_i(r) > \mathbf{T}^{(\alpha)}_{\mathrm{upp}}(r)$ for some r iff $M_i < m_{(\alpha)}]$ for every $i = 1, \ldots, s$;
2. $[\ \mathbf{T}^{(\alpha)}_{\mathrm{low}}(r) \leq T_i \leq \mathbf{T}^{(\alpha)}_{\mathrm{upp}}(r)$ for all r iff $M_i \geq m_{(\alpha)}]$ for every $i = 1, \ldots, s$.

Global envelope test for temperature data

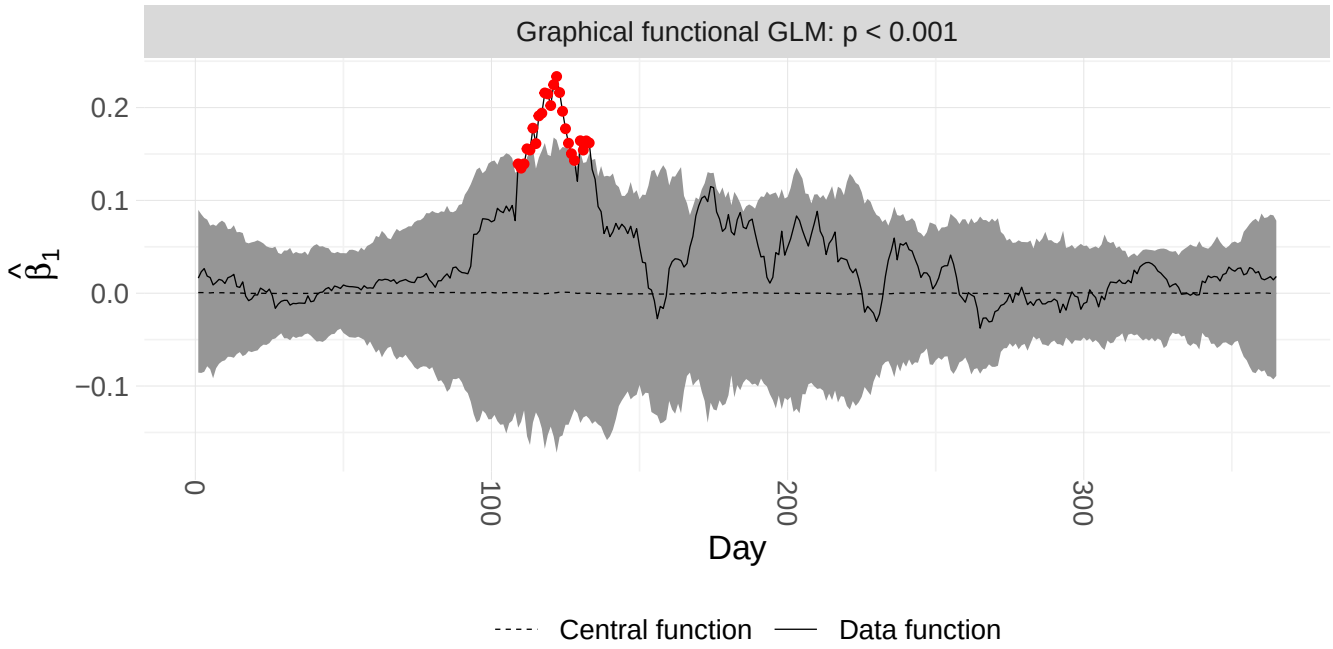


Figure – The output of the global ERL envelope test ($p < 0.001$) for testing the effect of the year on the temperatures. The grey area represents the 95% global ERL envelope and the red dots show the days where the data function exceeds the envelope.

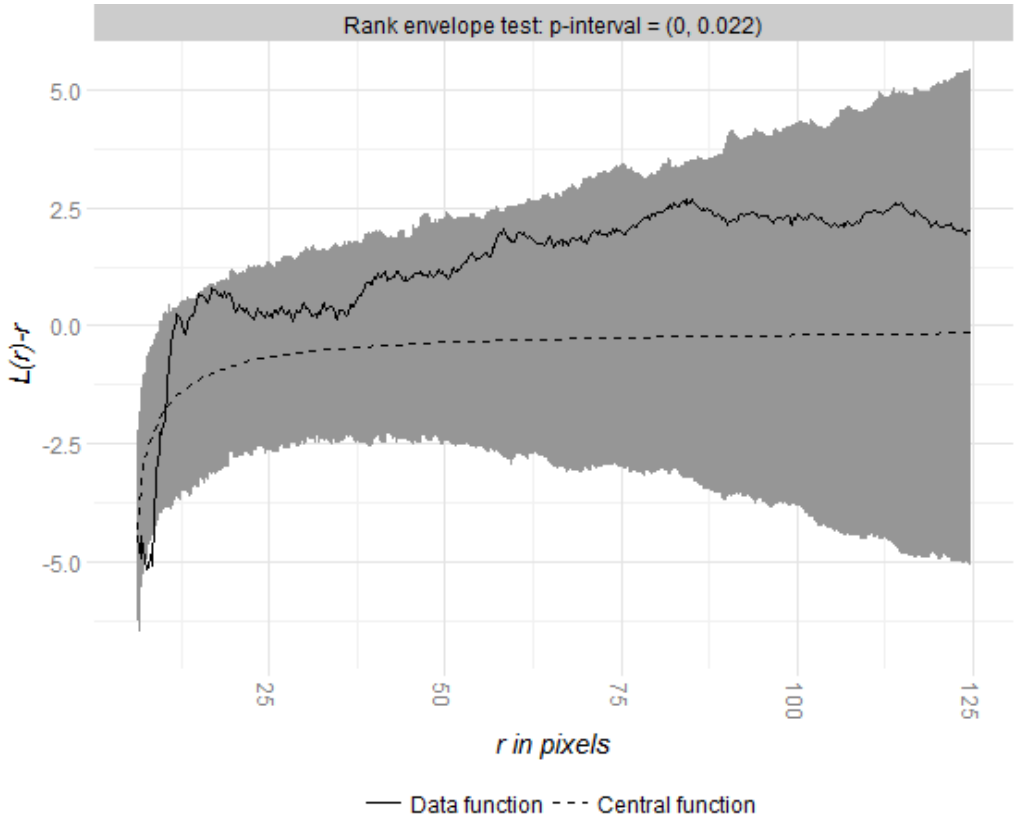
Global envelope test

• Implementation in R package GET. (Myllymäki M., Mrkvička T. : GET : Global envelopes in R. arXiv:1911.06583 [stat.ME])
• Provides p-value.
• Provides envelopes which specifies r distances leading to the rejection.
• Helps to detect reasons why the data contradict the null hypothesis.
• Does not suffer from the uneven variance and asymmetry of $T(r)$.
• It is a completely nonparametric procedure - It gives exactly same weight to every r.
• It does not rely on assumptions about distribution of the test statistic, homogeneity of the test statistic across the image, so it can be used for any test statistic.

Another perspective

• The functions $T_i(r)$ have to be discretized, then $T_i(r)$ is a multivariate vector of dimension d.
• Then we have d point-wise tests and
• Global envelope test is multiple test comparison procedure with family wise error rate (FWER).
• The global envelope test uses the whole correlation structure between the test, since whole functions are resampled. Thus it is very powerful multiple testing procedure.
• It can be treated as the p-min multiple testing procedure, with resolved ties problem of p-mins. The solution is based on Extreme rank length measure (ERL) or Continuous rank measure (Cont) or Area rank measure (Area).
• In FDA the F-max test is also often used $F^{\max} = \max_{r \in I} T(r)$.
• The statistic $T(r)$ should be pivotal statistic, i.e. the distribution of the statistic must not change for different r values, in order to have same contribution of every r in F-max.
• But even basic statistics like t or F are not pivotal, but only second order pivotal statistic.

Global envelope test for goodness-of-fit of intramembraneous particles



- The null model is hard-core point pattern model
- It indicates the deviation for short distances and suggests that a soft core model should be used.

GET :package

GET : Global envelopes in R. : Myllymäki M., Mrkvička T. can be downloaded from CRAN (https://cran.r-project.org/package=GET).
Content :

1. General GET - graphical multiple testing with control of FWER
2. Functional GLM / Functional ANOVA (which groups differ)
3. Goodness-of-fit testing in spatial statistics
4. Functional box plot and central region
5. Functional clustering
6. Local independence test
7. Testing difference between 2 or more CDF
8. Local spatial correlation test
9. Still to be developed :)

3D - function - brain image (fMRI) - 50.000 voxels - 100.000 permutations

Requirements : 10 hours - 2 cores - 10GB

Genome-wide Association Studies - 1.000.000 genes - 2.000.000 permutations

Requirement : 20 hours - 2000 cores - 10 TB

The global envelope considers the correlation between genes, whereas usual multiple-testing correction methods do not.

s - # permutations
d - # genes
t - time for computation of single test statistic

- time problem - $O(sdt + sd\log(s))$ - massive paralleling
- space problem - $O(sd)$ - clever algorithms

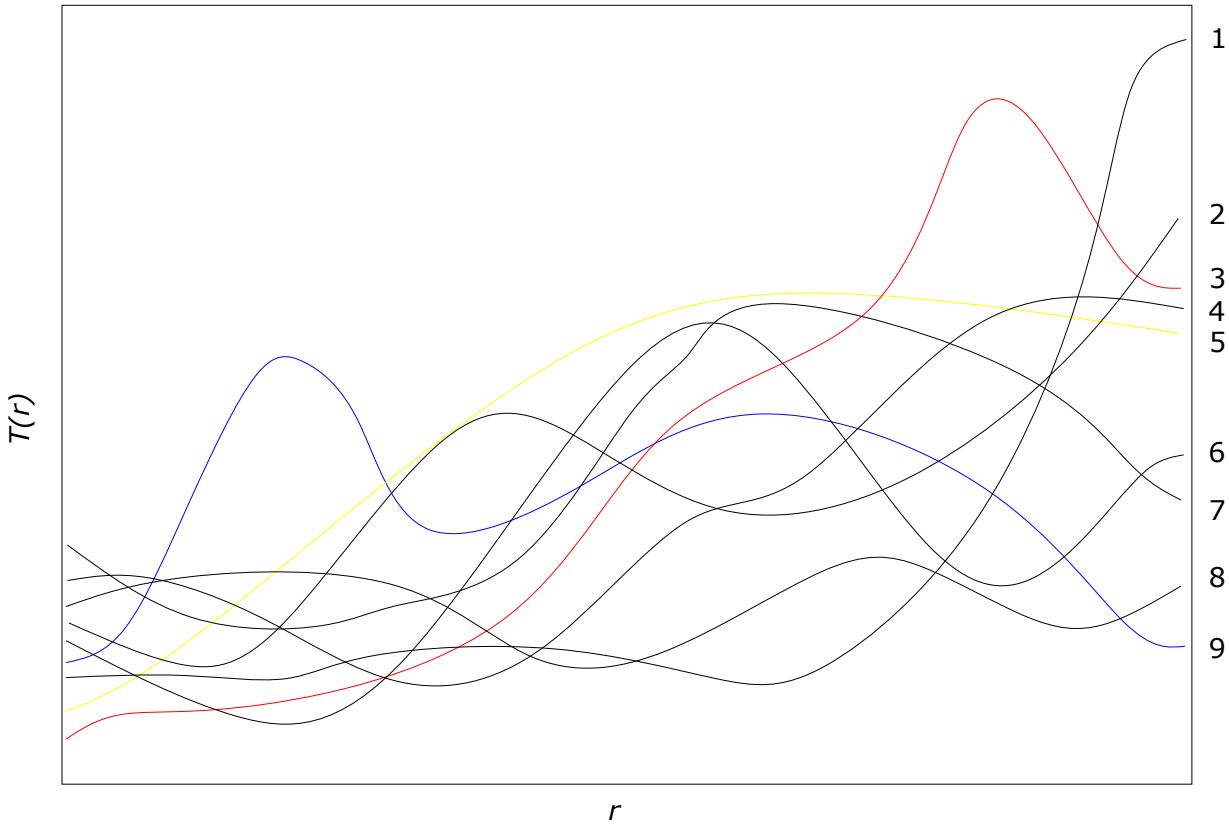
IGI ordering

GET :package - IGI ordering

1. Extreme rank ordering
2. Extreme rank length ordering
3. Continuous rank ordering
4. Area rank ordering

Choice of the type of global envelopes

- Large number of functions/simulations, $s \to$ rank, erl, cont, area lead to the same output, any choice OK
- with a low number of functions, the measure playes a role



Most extreme one : Rank : 1, 3, 4, 5, 7 and 9 ; ERL 5 ; Cont 9 ; Area 9 ; Unscaled 1

Extreme rank ordering - IGI

Let $r_{1j}, r_{2j}, \ldots, r_{sj}$ be the raw ranks of $T_{0j}, T_{2j}, \ldots, T_{sj}$, such that the smallest T_{ij} has rank 1. In the case of ties, the raw ranks are averaged. The two-sided pointwise ranks are then calculated as $R_{ij} = \min(r_{ij}, s + 1 - r_{ij})$.
The extreme rank R_i of the vector $\mathbf{T}_i$ is defined as the minimum of its pointwise ranks, namely

$$R_i = \min_{k=1,\ldots,m} R_{ik}, \tag{1}$$

- Problem - ties
- It is equivalent to the p-min approach in multiple testing literature.

Extreme rank length ordering - IGI - Solving ties

Consider now the vectors of pointwise ordered ranks $\mathbf{R}_i = (R_{i[1]}, R_{i[2]}, \ldots, R_{i[m]})$, where $\{R_{i[1]}, \ldots, R_{i[m]}\} = \{R_{i1}, \ldots, R_{im}\}$ and $R_{i[k]} \leq R_{i[k']}$ whenever $k \leq k'$. The extreme rank length measure of the vectors $\mathbf{R}_i$ is equal to

$$E_i = \frac{1}{s+1} \sum_{i'=0}^{s} \mathbf{1}(\mathbf{R}_{i'} \prec \mathbf{R}_i) \tag{2}$$

where

$$\mathbf{R}'_i \prec \mathbf{R}_i \quad \Longleftrightarrow \quad \exists\, n \leq m : R_{i'[k]} = R_{i[k]} \;\forall\; k < n,\; R_{i'[n]} < R_{i[n]}.$$

Continuous rank ordering - IGI - Solving ties

The continuous rank measure is

$$C_i = \min_{j=1,\dots m} c_{ij} / \lceil s' / 2 \rceil,$$

where c_{ij} are the pointwise continuous ranks defined as the continuous position of i-th function between one above and one bellow. If it is the most extreme rank it is defined as exponential distance to the second most extreme rank.

Area rank ordering - IGI - Solving ties

The area rank measure combines both approaches

$$A_i = \frac{1}{\lceil s'/2 \rceil m} \sum_j \min(R_i, c_{ij}),$$

where
$R_i = \min_j \{R_{ij}\}$ and R_{ij} are two-sided pointwise ranks defined above.

Sources of information for Rank, ERL, Continuous rank and Area rank

measure	extreme rank	Integration of extremes	Value of extremes
Rank (p-min)	Yes	No	No
ERL	Yes	Yes	No
Cont rank	Yes	No	Yes
Area rank	Yes	Yes	Yes
F - max	No	No	Yes

Table – Sources of information.

Clever memory handling - An algorithm to compute functional ordering voxel by voxel

Initialize the (augmented) measure M_j for $j = 0, \ldots, s$
Generate s permutations
For each voxel
$T_0 \leftarrow$ test statistic for data
  For $j \leftarrow 1, \ldots, s$
    $T_j \leftarrow$ test statistic for permutation j
  End for
  $(m_0, m_1, \ldots, m_s) \leftarrow ranks(T_0, T_1 \ldots, T_s)$
  For $j \leftarrow 1, \ldots, s$
    $M_j \leftarrow update(M_j, m_j)$
  End for
End for
Compute the final measures from the augmented measures

Clever memory handling - 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 :

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

The final measure is

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

Clever memory handling - An algorithm to compute global envelopes after the functional orderings have been calculated.

Let M_j be the final measure for $j = 0, \ldots, s$ and $M^{(\alpha)}$ the critical value, i.e. the largest of the M_i such that the number of those j for which $M_j < M^{(\alpha)}$ is less or equal to $\alpha(s+1)$
Use the same s permutations as when computing the functional orderings
For each voxel v
    $T_0 \leftarrow$ test statistic for data
    For $j \leftarrow 1, \ldots, s$
        $T_j \leftarrow$ test statistic for permutation j
    End for
    $U_v \leftarrow \max\{T_j : M_j > M^{(\alpha)}\}$
End for
U_v contains the upper envelope.

References & Thank you for your attantion

- Dai W., Mrkvička T., Sun Y., Genton M. C. (2020). *Functional Outlier Detection and Taxonomy by Sequential Transformations*, Computational statistics and data analysis , 149, 16960.
- Dai W., Athanasiadis S., Mrkvička T. (2021). *A new functional clustering method with joined dissimilarity sources and graphical interpretation*, in Computational statistics : Intex open.
- Dvořák J., Mrkvička T. (2021). *Graphical tests of independence for general distributions*, Computational statistics
- Mrkvička T., Myllymäki M., Hahn U. (2017). *Multiple Monte Carlo Testing with Applications in Spatial Point Processes*, Statistics and computing 27/5, 1239–1255.
- Mrkvička T., Roskovec T., Rost M. (2019). *A nonparametric graphical tests of significance in functional GLM*, Methodology and computing in applied probability.
- Mrkvička T., Myllymäki M., Jílek M., Hahn U. (2020). *A one-way ANOVA test for functional data with graphical interpretation*, Kybernetika.
- Mrkvička T., Myllymäki M., Kuronen M., Narisetty N. (2021) : New methods for multiple testing in permutation inference for the general linear model, https ://arxiv.org/abs/1906.09004.
- Myllymäki, M., Mrkvicka, T., Seijo, H., Grabarnik, P. and Hahn, U. (2017). *Global envelope tests for spatial processes.* JRSSB.