Доклады Международной конференции "Математическая биология и биоинформатика". Под ред. В.Д. Лахно. Том 7. Пущино: ИМПБ РАН, 2018. Статья № e91. doi: 10.17537/icmbb18.111

The generic envelope test and its modifications

Grabarnik P.[1], Myllymäki M.[2]

[1]*Institute of Physicochemical and Biological Problems in Soil Science of the Russian Academy of Sciences, Pushchino, Moscow Region, Russian Federation*
[2]*Natural Resources Institute Finland (Luke), Helsinki, Finland*

pavel.grabarnik@gmail.com, mari.myllymaki@luke.fi

The generic envelope test for spatial point patterns, practical variants of which were introduced in Myllymäki et al., 2017, is presented. We distinct two forms of the envelope test: quantile and functional and discuss their equivalence.

Key words: spatial point patterns, functional statistic, multiply testing problem.

Базовая форма критерия огибающих и его модификации

Грабарник П.[1], Миллимяки М.[2]

[1]*Институт физико-химических и биологических проблем почвоведения, РАН, Пущино, Московская область, Российская Федерация*
[2]*Институт природных ресурсов Финляндии (Луке), Хельсинки, Финляндия*

В работе представлена базовая форма критерия огибающих для анализа пространственных точечных структур, практический вариант которой был введен в работе Myllymäki et al., 2017. Мы рассматриваем две модификации этого теста: квантильную и функциональную и обсуждаем их эквивалентность.

1. Introduction

The simulation envelope method is one of main techniques in the spatial statistics toolbox being a popular solution for goodness-of-fit and mark independence testing problems. Its great advantage which has not seen in other methods is that it provides a graphical display of testing results.

Popular summary statistics to describe the spatial arrangement are functions $T(r)$ which depend on inter-point distances r. The distribution of the statistics $T(r)$ is generally unknown. Therefore we are needed to resort to the Monte Carlo simulation method. The next step of testing is to define critical bounds for a deviation of observed $T(r)$ from the expected value for each chosen r and to display the result graphically. When making conclusions from the testing results, it has often been missed in applications that rigorous statistical inferences have to be done for all r's simultaneously. This multiple testing problem recently attracted a close attention of several authors [1–4].

For an exposition of the simulation envelope method, it is important to understand its relation to a classical problem in stochastic processes theory – the computation of the probability that a random process remains between two given boundaries. This difficult question was a target of efforts of many researchers over long time. So far only in a few cases some analytical progress was made. In case of the Brownian motion a reader can be referred to e.g., [5–7] and in case of Gaussian processes to [8].

To see the connection between the envelope test and the cross-boundaries problems we point out that the functional statistic $T(r)$ on continuous domain $[0, r_{\max}]$ can be considered as a real valued process $(T_r)_{r\in[0,r_{\max}]}$, and therefore the generic envelope may be defined as a strip with non-random boundaries $e_{\text{upp}}(r)$ and $e_{\text{low}}(r)$ for which holds

$$\Pr(e_{\text{low}}(r) \le T(r) \le e_{\text{upp}}(r),\ r\in[0,r_{\max}]) = 1-\alpha,$$

where $0 < \alpha < 1$ is a some prescribed level.

In case when some transformation of the process $T(r)$ leads to a process (T_r^*) which has a constant variance, a band with constant width can be taken as the strip $[e_{\text{upp}}(r), e_{\text{low}}(r)]_{r\in[0,r_{\max}]}$.

In a general case a choice of the boundaries $e(r)$ has to be dependent on an alternative hypothesis. When alternative hypotheses are vague, the most reasonable choice of the boundary curves would be quantiles of the marginal distributions $F_{T(r)}$ of statistics $T(r)$, i.e.

$$e_{\text{upp}}(r) = F_{T(r)}^{-1}(1-\beta),$$

$$e_{\text{low}}(r) = F_{T(r)}^{-1}(\beta),$$

where β is the pointwise level of the marginal distributions. This type of the envelope boundaries may be called the Ripley's envelope to distinct it from other forms of the envelope method. Thus, the problem of a construction of the envelope test is reduced to an estimation of the distributions $F_{T(r)}$ and finding such a value β that the overall level (global type I error probability) would be equal to α.

In the presented paper we focus on the approach of the envelope test construction based on Monte Carlo simulations. This approach allows to work with any model of spatial processes, whereas the price we have to pay for the universality is computational load which may be very high.

We present two ways to construct envelope tests. The first is based on a rule that the null hypothesis is rejected if the observed functional statistic $T(r)$ crosses the Ripley's envelope boundaries. The second envelope construction approach is based on ordered simulated functions $T_i(r)$, $i = 1, \dots, s + 1$ as suggested by [4, 12]. This approach is in a close connection with methods used in functional data analysis, where observed functions are ordered by one of depth measures [9]. Unlike the objective of the functional data analysis where the main goal to find the most central function, we use the ordering to select the most extremal functions.

2. The envelope test formulation

A point pattern is a finite set $\mathbf{x} = \{x_1, \dots, x_n\}$ of locations of n objects in a region D, which is a compact subset of $\mathbf{R}^d$.

We want to test the hypothesis H_0 that the observed point pattern $\mathbf{x}_W$ is a realization of a point process X in a window $W \subseteq D$. To test H_0, the complex information contained in the pattern is represented by a test function $T(\mathbf{x}_W)$, that is typically an estimated summary function that captures the spatial arrangement of points. Most often in spatial point patterns analysis such functions as the Ripley's K-functions, the closely related L-function or van Lieshout and Baddeley's J-function (see e.g [10]) are used.

These functions enjoy the property of the second-order homogeneity of the point process X that allows to consider the summary functions $T(r)$ as a function of inter-point distances r only. The null and alternative model may suggest which summary function is suitable as a test function.

To construct a test we have to agree which functions $T_{\mathbf{x}}(r)$ would be typical and which untypical or "extremal" if the observed point pattern was sampled from the true model X. It is often natural to reject the null hypothesis if the summary function $T_{\mathbf{x}}(r)$ exceeds some limits within an interval of distances r, i.e. $I = [r_{\min}, r_{\max}]$. This rejection rule leads to the so-called *envelope* test.

Let E be a region in a plane R^2

$$E_I(U, L) = \{(r, y) \in I \times R : L(r) \le y \le U(r)\}$$

where U is the upper boundary and L is the lower boundary, and I is a fixed interval.

Let $G(T)$ be a graph of a function $T(r)$, and its restriction on the interval I denote $G_I(T)$, i.e. $G_I(T) = \{(r, T(r)), r \in I\}$.

D e f i n i t i o n. A region $E_I(U, L)$ is *the envelope of level* α, $\text{Env}_I(L_\alpha, U_\alpha)$ if

$$\mathbf{P}(G_I(T) \subset E_I(U, L)) = 1-\alpha$$

D e f i n i t i o n. A tes t φ is called the *envelope* test at level α if

$$\varphi_{\text{env}}(T) = \mathbf{1}(G_I(T) \not\subset \text{Env}_I(L_\alpha, U_\alpha))$$

This definition does not determine the envelope test at the certain significance level uniquely because the envelope $\text{Env}_I(L_\alpha, U_\alpha)$ depends on a choice of the upper and lower boundaries. We could specify the test uniquely if to impose some constraints on the form of the envelope. For example, such constraints could be determined by alternative hypotheses to maximise the Type II error probability. However, when an alternative model is not specified some *ad hoc* methods can be appropriate. One possibility is to choose boundaries such that the area of the envelope would be minimal. Very popular envelope is based on Kolmogorov–Smirnov's boundaries of a constant width

$$\max_{r \in I} |T(r) - T_0(r)|,$$

where $T_0(r)$ is the expected value under the null hypothesis.

Another quite natural choice is to specify the boundary functions taking into account the marginal distributions $F_{T(r)}$ of the functional statistic $T(\cdot)$.

D e f i n i t i o n. A test φ is called the *quantile envelope* test of a level α if

$$\varphi_{\text{env}_q}(T) = \mathbf{1}(\exists\, r \in I : T(r) \notin (F_{T(r)}^{-1}(\beta), F_{T(r)}^{-1}(1-\beta)),$$

where β is chosen such that the envelope $\text{Env}_I(L, U)$ with $L(r) = F_{T(r)}^{-1}(\beta)$ and $U(r) = F_{T(r)}^{-1}(1-\beta)$ has the overall level α.

This envelope is the theoretical counterpart of a practical way to construct the envelope test by means of Monte Carlo simulations.

3. Simulation envelope tests

Definition of the pointwise simulation envelope.

Monte Carlo simulation allows to replace the theoretical unknown distribution $F_{T(r)}(y)$ by its bootstrapped (simulation) analogue based on simulated functions $T_1(r), \dots, T_{s+1}(r)$ at a distance r,

$$\hat{F}_{T(r)}(y) = \frac{1}{s+1} \sum_{i=1}^{s+1} \mathbf{1}(T_i(r) \le y).$$

Let the lower and upper boundaries of an envelope be defined as

$$L^{p}_{k,s+1}(r) = \hat{F}^{-1}_{T(r)}(k/(s+1))$$

and

$$U^{p}_{k,s+1}(r) = \hat{F}^{-1}_{T(r)}(1-k/(s+1)),$$

where $k/(s+1) = \beta$. We call this envelope as *pointwise* $(1-\beta)$ *simulation envelope,*

$$\mathrm{Env}_p(k, s+1) = \{(L^{p}_{k,s+1}(r), U^{p}_{k,s+1}(r)), r \in I\}.$$

The global level α depends on the pointwise level β and can be estimated. In [11] it was proposed to estimate α by a batch of further simulations.

Definition of the functional simulation envelope.

First, we give a definition of a functional depth measure which allows to order simulated functions $T_i(\cdot)$, $i = 1, \ldots, s+1$. The bootstrapped distribution function $\hat{F}$ is related to ranks $R_1, \ldots, R_{s+1}$ of the variables $T_1(r), \ldots, T_{s+1}(r)$ (ordered in ascending order) by

$$R_j(r) = (s+1)\hat{F}_{T(r)}(T_j(r)).$$

D e f i n i t i o n. Let $\tau = \{T_1, \ldots, T_{s+1}\}$ be a set of functions. The depth measures

$$ER(T_j \mid \tau) = \min_{r \in I} \min(R_j(r), s+1-R_j(r)+1)$$

is called the *extremal rank* of the function $T_j(\cdot)$. This depth measure was first proposed in [4].

Now we introduce another simulation envelope which covers simulated functions more tightly than the pointwise simulation envelope which has been adjusted to the global level α. This envelope was proposed by [12].

Note, that functions ordered by the extremal rank measure have many ties. We need to break these ties, for example, by adding a random variable. More sophisticated way to do it was proposed in [4]. Let ER^* be the extended extremal rank measure which orders all functions without ties. Let $Z(m)$ be a set of functions $T_j(\cdot)$, the extremal rank depth of which $ER^*(T_j|T_1, \ldots, T_{s+1})$ is equal or larger than m.

Let $Z(m) = \{Z_{m,1}(\cdot), \ldots, Z_{m,n_m}(\cdot)\}$, where $Z_{m,i}(\cdot)$ denotes one of the functions $T_j(\cdot)$ having the depth larger or equal to m, and $n_m = |Z(m)|$ is the number of such functions. Define an envelope delimited by these functions as follows

$$\mathrm{Env}_f(\alpha(m, s+1)) = \{(r, y) \in I \times R : \min_{i=1,\ldots n_m} Z_{m,i}(r) \le y \le \max_{i=1,\ldots n_m} Z_{m,i}(r)\}.$$

Therefore, corresponding the lower and upper boundaries of the envelope are

$$L^{f}_{m,s+1}(r) = \min_{i=1,\ldots n_m} Z_{m,i}(r)$$

and

$$U^{f}_{m,s+1}(r) = \max_{i=1,\ldots n_m} Z_{m,i}(r).$$

The test based on the envelope $\mathbf{Env}_f(\boldsymbol{m}, \boldsymbol{s}+\mathbf{1})$ is called *functional simulation envelope* test, the global level α of which can be estimated by simulations.

4. Relation between pointwise and functional envelopes and asymptotic result

It is interesting theoretical problem to show that when k and s tend to ∞ such that $k/(s+1) = \beta$ the pointwise simulation envelope approachs the theoretical quantile envelope $\mathrm{Env}_q(\alpha)$, and to find conditions on a class of functions $T(r)$ under which this holds.

Such result would allow to provide a theoretical basis for an empirical evidence that asymptotically, when $s \to \infty$, both simulation envelope tests and the quantile envelope test are equivalent, i.e.

$$\lim_{s\to\infty} \varphi_{\mathrm{env}_f}(T) = \lim_{s\to\infty} \varphi_{\mathrm{env}_p}(T) = \varphi_{\mathrm{env}_q}(T).$$

This equivalence means that if one of the tests rejects the null hypothesis then both other tests reject the null hypothesis as well.

5. Acknowledgments

The work of P.G. was supported by the Russian Science Foundation (project No. 18-14-00362). The work of M.M. was supported by the Academy of Finland (project No. 295100 and 306875).

6. References

1. Loosmore N.B., Ford E.D. Statistical inference using the G or K point pattern spatial statistics. *Ecology*. 2006. V. 87. № 8. P. 1925–1931.
2. Grabarnik P., Myllymäki M., Stoyan D. Correct testing of mark independence for marked point patterns. *Ecological Modelling*. 2011. V. 222. P. 3888–3894.
3. Baddeley A., Diggle P.J., Hardegen A., Lawrence T., Milne R.K., Nair G. On tests of spatial pattern based on simulation envelopes. *Ecological Monographs*. 2014. V. 84. № 3. P. 477–489.
4. Myllymäki M., Mrkvička T., Grabarnik P., Seijo H., Hahn U. Global envelope tests for spatial processes. *Journal of the Royal Statistical Society: Series B*. 2017. V. 79. P. 381–404.
5. Lerche H.R. Boundary Crossing of Brownian Motion. In: *Lecture Notes in Statistics*. Springer-Verlag, 1986. V. 40.
6. Durbin J. The first-passage density of the brownian motion process to a curved boundaries for the brownian motion and poisson processes and techniques for computing the power of the Kolmogorov–Smirnov test. *Journal of Applied Probability*. 1992. V. 29. P. 291–304.
7. Daniels, H. E. Approximating the first crossing-time density for a curved boundary. *Bernoulli*. 1996. V. 2. № 2. P. 133–143.

8. Leadbetter M.R., Lindgren G., Rootzén H. *Extremes and related properties of random sequencesand processe*s. N.Y., Heidelberg, Berlin: Springer-Verlag, 1983.
9. López-Pintado S., Romo J. On the concept of depth for functional data. *Journal of the American Statistical Association*. 2009. V. 104. № 486. P. 718–734.
10. Illian J., Penttinen A., Stoyan H., Stoyan D. *Statistical Analysis and Modelling of Spatial Point Patterns*. John Wiley & Sons, Ltd, 2008.
11. Ripley B.D. Modelling spatial patterns. *Journal of the Royal Statistical Society, Series B*. 1977. V. 39. № 2. P. 172–212.
12. Mrkvička T., Hahn U., Myllymäki M. A one-way ANOVA test for functional data with graphical interpretation. *Cornell University Library*. 2018. URL: arXiv:1612.03608 [stat. ME] (accessed 26.10.2018).