CENTRE FOR **STOCHASTIC GEOMETRY**
AND ADVANCED **BIOIMAGING**

Ute Hahn

A note on simultaneous Monte Carlo tests

No. 06, May 2015

A note on simultaneous Monte Carlo tests

Ute Hahn

Centre for Stochastic Geometry and Advanced Bioimaging,
Aarhus University, ute@math.au.dk

May 27, 2015

Abstract

In this short note, Monte Carlo tests of goodness of fit for data of the form $X(t), t \in I$ are considered, that reject the null hypothesis if $X(t)$ leaves an acceptance region bounded by an upper and lower curve for some t in I. A construction of the acceptance region is proposed that complies to a given target level of rejection, and yields exact p-values. The construction is based on pointwise quantiles, estimated from simulated realizations of $X(t)$ under the null hypothesis.

Keywords: Cdf transform, functional data, multiple testing, multivariate data, simulation based tests.

1 Introduction

A simultaneous envelope test is a significance test of hypotheses about multivariate data or processes $X(t), t \in I$. It rejects the hypothesis if the observation leaves an acceptance region $A(t)$ at some point $t \in I$. The acceptance region is bounded by a lower and an upper critical curve $X_{\text{low}}(t)$ and $X_{\text{upp}}(t)$, which allows for an immediate graphical interpretation. Such graphical tests are popular in practical applications, in particular in the context of goodness of fit tests for spatial pattern, where the spatial information is summarized by some function which is estimated from the observed pattern. The distribution of the estimated functions $X(t)$ is typically unknown, and therefore one has to resort to Monte Carlo methods and simulate from the hypothesized distribution. It has been unclear for a long time how to construct the acceptance region such that a given rejection rate α is met, therefore pointwise $(1-\alpha)$-quantile envelopes have been plotted and occasionally been misinterpreted as tests to the level α, see Loosmore and Ford (2006); Baddeley et al. (2014).

Only recently, a series of papers emerged that successfully tackle this problem (Grabarnik et al., 2011; Myllymäki et al., 2013a,b, 2015). The solution given in Myllymäki et al. (2013b) allows to construct envelopes with a rejection rate that comes arbitrarily close to the target level α, depending on the number of simulations. The *global rank envelope* proposed in that paper is accompanied with an interval

for the corresponding Monte Carlo p-value. In the examples from goodness of fit tests for point processes discussed there, about 2500 simulated curves are needed to keep the interval reasonably short. In Myllymäki et al. (2015), this solution was supplemented with a non graphical *rank count test* to the level α, which always returns p-values within the p-interval of the envelope test.

In the present short note, I develop an alternative, closely related construction for envelopes. It is as exact as the rank count test, if the distribution of the $X(t)$ is at least partly continuous in the tails. Since it combines exactness with graphical representation, it can replace the global rank envelope when a large number of simulations is not feasible. This note is meant as an addendum to Myllymäki et al. (2015), where other constructions of envelopes known from the literature are thoroughly discussed and compared with the global rank envelope.

2 Monte Carlo goodness of fit tests seen as permutation tests

A goodness of fit test is a test for the hypothesis that the observation x_1 is a realization of a random variable distributed as X_0, that is, a hypothesis of the form $H_0 : X_1 \sim X_0$. Its Monte Carlo version compares the observation x_1 with realizations of i.i.d. variables $X_2, \dots, X_n$, simulated from the model. The hypothesis can then be replaced by the equivalent hypothesis H_0' : "$X_1, X_2, \dots, X_n$ are i.i.d.". The Monte Carlo tests dealt with in the following are essentially tests for the hypothesis H_0^* that the n random variables $X_1, \dots, X_n$ are *exchangeable*, that is, the distribution of the sample $\mathbf{X} := (X_1, \dots, X_n)$ is permutation invariant. Since it is known that $X_2, \dots, X_n$ are i.i.d., the exchangeability hypothesis H_0^* is here equivalent to H_0' and thus to H_0.

Under the null hypothesis H_0^*, all permutations $\pi(\mathbf{X}) = (X_{\pi(1)}, \dots, X_{\pi(n)})$ are equally likely, given the unordered sample $\breve{\mathbf{X}} := \{X_1, \dots, X_n\}$. This fact is exploited in the test, which is based on a statistic $T(\mathbf{X}) = T(X_1, X_2, \dots, X_n)$. The original value $T(\mathbf{X})$ is compared with the values $T(\pi(\mathbf{X}))$ obtained on the permuted samples. Without loss of generality, assume that a small value of T indicates an extreme observation and casts doubt on the null hypothesis. A general permutation test to the level α rejects the hypothesis if $T(\mathbf{X}) \le T_{(\alpha)}$, where $T_{(\alpha)}$ is a lower α-quantile of the values $T(\pi(\mathbf{X}))$ for all permutations $\pi \in S$, $S = \mathrm{Perm}(\{1, \dots, n\})$, fulfilling

$$\tfrac{1}{\operatorname{card} S} \sum_{\pi \in S} \mathbf{1}\big(T(\pi(\mathbf{X})) < T_{(\alpha)}\big) < \alpha \le \tfrac{1}{\operatorname{card} S} \sum_{\pi \in S} \mathbf{1}\big(T(\pi(\mathbf{X})) \le T_{(\alpha)}\big).$$

It can be found as

$$T_{(\alpha)} = \max\big\{t \in \mathcal{T} : \sum_{\pi \in S} \mathbf{1}(T(\pi(\mathbf{X})) \le t) \ \le \alpha \operatorname{card} S\big\}, \quad \mathcal{T} = \big\{T(\pi(\mathbf{X})) : \pi \in S\big\}.$$

In Monte Carlo goodness of fit tests, the variables $X_2, \dots, X_n$ are i.i.d., therefore only statistics T are used that do not depend on the sequence of the variables $X_2, \dots, X_n$. It is thus sufficient to consider a set $\tilde{S} = \{\pi_1, \dots, \pi_n\}$ of permutations

that only affect the first element, viz.

$$\pi_i(\mathbf{X}) = (X_i, X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n).$$

A Monte Carlo goodness of fit test to the level α has the form

$$\varphi_\alpha(\mathbf{X}) = \mathbf{1}(T(\mathbf{X}) \le T_{(\alpha)}), \tag{2.1}$$

where

$$T_{(\alpha)} = \max\big\{t \in \{T_1, \ldots, T_n\} : \sum_{i=1}^{n} \mathbf{1}(T_i \le t) \le \alpha n\big\}, \quad T_i = T(\pi_i(\mathbf{X})). \tag{2.2}$$

Let $T_{[1]} \le T_{[2]} \le \cdots \le T_{[n]}$ denote the ordered set of values $T_1, \ldots, T_n$. Then $T_{(\alpha)} = T_{[\lfloor \alpha n \rfloor]}$ if and only if $T_{[\lfloor \alpha n \rfloor]} < T_{[\lfloor \alpha n \rfloor + 1]}$. Otherwise, i.e. if $T_{[\lfloor \alpha n \rfloor]} = T_{[\lfloor \alpha n \rfloor + 1]}$, it is $T_{(\alpha)} < T_{[\lfloor \alpha n \rfloor]}$. Consequently, the test (2.1) has rejection rate, given $\breve{\mathbf{X}}$,

$$\mathrm{E}[\varphi_\alpha(\mathbf{X}) \mid \breve{\mathbf{X}}] \quad \begin{cases} = \lfloor \alpha n \rfloor / n, & T_{[\lfloor \alpha n \rfloor]} < T_{[\lfloor \alpha n \rfloor + 1]}, \\ < \lfloor \alpha n \rfloor / n, & T_{[\lfloor \alpha n \rfloor]} = T_{[\lfloor \alpha n \rfloor + 1]}. \end{cases} \tag{2.3}$$

To the significance test φ_α corresponds a *Monte Carlo p-value* given by

$$p(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(T_i \le T_1). \tag{2.4}$$

It follows from the definition of the critical value $T_{(\alpha)}$ in (2.2), that

$$T_1 \le T_{(\alpha)} \iff p(\mathbf{X}) \le \alpha. \tag{2.5}$$

Hence, the test

$$\varphi_\alpha(\mathbf{X}) = \mathbf{1}(p(\mathbf{X}) \le \alpha) \tag{2.6}$$

is exact if and only if $\alpha n \in \mathbb{N}$ and upper ties with the critical value do not occur, i.e.

$$\mathrm{E}\,\varphi_\alpha(\mathbf{X}) = \alpha \iff \alpha n \in \mathbb{N} \quad \text{and} \quad \mathrm{Prob}(T_{[\lfloor \alpha n \rfloor]} = T_{[\lfloor \alpha n \rfloor + 1]}) = 0. \tag{2.7}$$

Otherwise, the test φ is conservative. In order to avoid these ties, the test statistic should be preferably chosen such that $T(\pi_i(\mathbf{X})) \ne T(\pi_j(\mathbf{X}))$ for $i \ne j$ whenever $\breve{\mathbf{X}}$ only contains unique values.

3 Envelope tests

Consider now data of the form $X(t)$, where $t \in I$ is a continuous argument in the case of functional data or an index in the case of multivariate data. A test

$$\varphi_{\mathrm{env}+}(X) = \mathbf{1}\big(X(t) \notin [X^{(\alpha)}_{\mathrm{low}}(t), X^{(\alpha)}_{\mathrm{upp}}(t)] \text{ for some } t \in I\big) \tag{3.1}$$

shall be called an *inclusive envelope test*, since the bounding curves are included in the acceptance region, whereas

$$\varphi_{\mathrm{env}-}(X) = \mathbf{1}\big(X(t) \notin (X^{(\alpha)}_{\mathrm{low}}(t), X^{(\alpha)}_{\mathrm{upp}}(t)) \text{ for some } t \in I\big) \tag{3.2}$$

will be denoted an *exclusive envelope test.*

For a *Monte Carlo envelope test of goodness of fit* of the hypothesis $X_1 \sim X_0$, the bounding curves $X_{\text{low}}^{(\alpha)}(t)$ and $X_{\text{upp}}^{(\alpha)}(t)$ are calculated from $\mathbf{X} = (X_1, \ldots, X_n)$, where $X_2(t), \ldots, X_n(t) \sim X_0(t)$ are i.i.d. simulated curves. To the Monte Carlo goodness of fit test (2.1) using a test statistic T corresponds an inclusive envelope, if

$$T(\mathbf{X}) \leq T_{(\alpha)} \iff \exists t \in I : X_1(t) < X_{\text{low}}^{(\alpha)}(t) \quad \text{or} \quad X_1(t) > X_{\text{upp}}^{(\alpha)}(t), \tag{3.3}$$

and an exclusive envelope correspondingly if

$$T(\mathbf{X}) \leq T_{(\alpha)} \iff \exists t \in I : X_1(t) \leq X_{\text{low}}^{(\alpha)}(t) \quad \text{or} \quad X_1(t) \geq X_{\text{upp}}^{(\alpha)}(t). \tag{3.4}$$

Remark 3.1. Although the an envelope test fulfilling (3.3) or (3.4) rejects if the observation leaves the acceptance region on either side, it is essentially a one sided test through the test statistic T, and it is not controlled how many of the rejections are due to the observation exceeding the upper envelope or going below the lower envelope.

Myllymäki et al.'s (2013b) global rank test

The global rank test compares observed and simulated curves by pointwise upward and downward ranks. Let $R_i^{\uparrow}(t)$ denote the rank of $X_i(t)$ among $X_1(t), \ldots, X_n(t)$, starting with rank 1 for the smallest value, and $R_i^{\downarrow}(t)$ the corresponding downward rank. In the case of ties, mid or maximal ranks are used. From these ranks, the pointwise extreme rank

$$R_i^{\diamond}(t) = \min(R_i^{\uparrow}(t), R_i^{\downarrow}(t))$$

is calculated that measures how "extreme" $X_i(t)$ is within the sample $X_1(t), \ldots, X_n(t)$. The overall test statistic is then

$$R_i = \min_{t \in I} R_i^{\diamond}(t).$$

The pointwise ranks $R_i^{\diamond}(t)$ cannot take other values than $1, 2, \ldots, \lceil n/2 \rceil$. Consequently, the distribution of R_i is also concentrated on these values, and ties occur with probability one. To account for the ties, an interval $(p_-, p_+]$ is reported instead of a single p-value, with

$$p_- = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(R_i < R_1) \quad \text{and} \quad p_+ = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(R_i \leq R_1).$$

The corresponding tests

$$\varphi_{\text{lib}}(\mathbf{X}) = \mathbf{1}(p_- \leq \alpha) \quad \text{and} \quad \varphi_{\text{cons}}(\mathbf{X}) = \mathbf{1}(p_+ \leq \alpha)$$

are liberal and conservative, respectively. Note that p_+ is equal to the conservative Monte Carlo p-value given in (2.4), and thus φ_{cons} is equal to the test φ_α given in (2.6). Myllymäki et al. (2013b) describe the construction of an envelope that matches this p-interval. They determine a critical rank $k_\alpha \in \mathbb{N}$ that fulfils

$$\sum_{i=1}^{n} \mathbf{1}(R_i < k_\alpha) \leq \alpha n < \sum_{i=1}^{n} \mathbf{1}(R_i < k_\alpha + 1)$$

and let

$$X^{(\alpha)}_{\text{low}}(t) = X(t)_{[\uparrow k_\alpha]} \quad \text{and} \quad X^{(\alpha)}_{\text{upp}}(t) = X(t)_{[\downarrow k_\alpha]},$$

where $X(t)_{[\uparrow k]}$ is the k-th smallest and $X(t)_{[\downarrow k]}$ is the k-th largest value among $X_1(t), \ldots, X_n(t)$. The authors show that

- the inclusive envelope test is equivalent to the conservative test based on p_+, i.e., $\varphi_{\text{env}+}(X_1) = \varphi_{\text{cons}}(\mathbf{X})$ for all $\mathbf{X}$,
- X_1 stays inside the exclusive envelope if and only if $p_- > \alpha$,
- $X_1(t)$ coincides with the bounding curves but does not exit the envelope if and only if $p_- \leq \alpha < p_+$.

4 Envelope test based on cdf transforms

The global rank test always yields more or less wide p-intervals, as a result of ties in the test statistic R. These ties are due to the fact that the pointwise ranks that constitute R only allow for a small set of possible values. Ties can be avoided by replacing the pointwise ranks with pointwise estimated cdf transforms as described in the following.

Instead of the upward pointwise rank $R_i^{\uparrow}(t)$, we use an empirical cdf transform $U_i^{\uparrow}(t) = F^{\uparrow}(X_i(t))$, and the downward rank is replaced by the value $U_i^{\downarrow}(t)$ of an empirical survival function $F^{\downarrow}$. The functions $F^{\uparrow}$ and $F^{\downarrow}$ are *bijective* empirical versions of the cdf or survival function of the random variable $X_0(t)$, and map the support of X_0 onto $[0, 1]$. They have the form $F(.; x_1, \ldots, x_n)$, based on realizations $x_1, \ldots, x_n$ of $X_0(t)$, and are required to fulfil

$$\begin{aligned} y \in \{x_1, \ldots, x_n\} \quad \Longrightarrow \quad & F^{\uparrow}(y; x_1, \ldots, x_n) = \frac{1}{n+1} \sum_{i=1}^{n} \mathbf{1}(x_i \leq y) \quad \text{and} \\ & F^{\downarrow}(y; x_1, \ldots, x_n) = \frac{1}{n+1} \sum_{i=1}^{n} \mathbf{1}(x_i \geq y). \end{aligned} \tag{4.1}$$

The corresponding inverse functions are denoted ${F^{\uparrow}}^{-1}$ and ${F^{\downarrow}}^{-1}$. How such functions can be constructed is explained later.

Similar as with the pointwise upward and downward ranks in the rank based envelope test, a pointwise transform $U_i^{\diamond}(t)$ is calculated for each of the curves $X_i(t)$, $i = 1, \ldots, n$, with

$$U_i^{\diamond}(t) = \min(U_i^{\uparrow}(t), U_i^{\downarrow}(t)). \tag{4.2}$$

In order to avoid ties, the empirical functions used to calculate $U_i^{\diamond}$ are based on the *remaining* curves, that is

$$U_i^{\uparrow}(t) = F^{\uparrow}(X_i(t); X_1(t), \ldots, X_{i-1}(t), X_{i+1}(t), \ldots, X_n(t)) \tag{4.3}$$

and

$$U_i^{\downarrow}(t) = F^{\downarrow}(X_i(t); X_1(t), \ldots, X_{i-1}(t), X_{i+1}(t), \ldots, X_n(t)). \tag{4.4}$$

The smallest pointwise value

$$U_i = \inf_{t \in I} U_i^\diamond(t) \tag{4.5}$$

can be seen as a measure of how extreme the curve X_i appears among $X_1, \ldots, X_n$. Since $F^\uparrow$ and $F^\downarrow$ are strictly monotone, ties in the U_i can only occur where there are pointwise ties in the original data. This event has zero probability if the distribution of the $X_i(t)$ is absolutely continuous in all t. This is not always the case in applications — for example, estimates of the K-function used in point process testing take the value 0 for small argument t with positive probability. The property (4.1) of $F^\uparrow$ and $F^\downarrow$ however ensures that in these cases, the corresponding $U_i^\diamond(t)$ are relatively large and thus unlikely to contribute to U_i and in particular to the critical bound $U_{(\alpha)}$ which is determined as explained in Section 2. Since ties between the $U_{[\lfloor \alpha n \rfloor]}$ are virtually impossible, it is $U_{[\lfloor \alpha n \rfloor]} < U_{[\lfloor \alpha n \rfloor + 1]}$ with probability one or close to one, and according to (2.7), the corresponding test φ_α is (close to) exact.

Corresponding envelope

The bounding curves of a $(1-\alpha)$-envelope are obtained as the "empirical pointwise quantiles"

$$X_{\text{low}}(t) = {F^\uparrow}^{-1}(U_{(\alpha)}; X_2(t), \ldots, X_n(t)),$$

and

$$X_{\text{upp}}(t) = {F^\downarrow}^{-1}(U_{(\alpha)}; X_2(t), \ldots, X_n(t)). \tag{4.6}$$

Proposition 4.1. *The exclusive envelope test $\varphi_{\text{env}-}$ using bounding curves given by* (4.6) *is equivalent to the Monte Carlo significance test φ_α using test statistic U.*

Proof. We need to show that (3.4) is fulfilled. By (4.5), $U_1 \le U_{(\alpha)} \iff U_1^\diamond(t) \le U_{(\alpha)}$ for some $t \in I$, that is, as $U_1^\diamond(t) = \min(U_1^\uparrow(t), U_1^\downarrow(t))$, either $U_1^\uparrow(t) \le U_{(\alpha)}$ or $U_1^\downarrow(t) \le U_{(\alpha)}$. Since the functions ${F^\uparrow}^{-1}$ and ${F^\downarrow}^{-1}$ are bijective,

$$U_1^\downarrow(t) \le U_{(\alpha)} \iff {F^\downarrow}^{-1}(U_1^\downarrow(t)) \ge {F^\downarrow}^{-1}(U_{(\alpha)}) \iff X_1(t) \ge X_{\text{upp}}(t)$$

and

$$U_1^\uparrow(t) \le U_{(\alpha)} \iff {F^\uparrow}^{-1}(U_1^\downarrow(t)) \le {F^\uparrow}^{-1}(U_{(\alpha)}) \iff X_1(t) \le X_{\text{low}}(t),$$

that is,

$$U_1 \le U_{(\alpha)} \iff \exists t \in I : X_1(t) \notin (X_{\text{low}}(t), X_{\text{upp}}(t)).$$

Here, we wrote short $F(.) = F(.; X_2(t), \ldots, X_n(t))$. □

Construction of the functions $F^\uparrow$ and $F^\downarrow$

By symmetry, it is enough to consider the empirical cdf, $F^\uparrow$. Assuming a fixed set $x_1, \ldots, x_n$ of data, write short $F^\uparrow(y)$ for $F^\uparrow(y; x_1, \ldots, x_n)$. Without loss of generality,

let the set be ordered, $x_1 \leq x_2 \leq \cdots \leq x_n$. The requirement (4.1) yields the restriction

$$F^{\uparrow}(y) = \frac{1}{n+1} \sum_{i=1}^{n} \mathbf{1}(x_i \geq y), \qquad y \in \{x_1, \ldots, x_n\}. \tag{4.7}$$

Between these values, $F^{\uparrow}$ is obtained by linear interpolation,

$$F^{\uparrow}(y) = \frac{(x_{i+1} - y)F^{\uparrow}(x_i) + (y - x_i)F^{\uparrow}(x_{i+1})}{x_{i+1} - x_i}, \qquad x_i < y < x_{i+1}. \tag{4.8}$$

$F^{\uparrow}(y)$ needs to be explained on the whole assumed support S. The upper tail is not relevant for the construction of the tests (and nor so is lower tail of $F^{\downarrow}$). To find the lower tail, we distinguish between bounded and unbounded support. If S is bounded below, the lower tail is simply found by interpolation,

$$F^{\uparrow}(y) = \frac{(x_1 - y)}{x_1 - x_0} F^{\uparrow}(x_1), \qquad y < x_1, \quad x_0 = \min(S). \tag{4.9}$$

If the support is not bounded below, $F^{\uparrow}$ has to be extended by a strictly monotone increasing function with $F^{\uparrow}(y) \to 0$ for $y \to -\infty$. When the form of the tail distribution is known from the model $X_0(t)$, this information can be used. Otherwise a practical solution is to attach a suitable exponential tail,

$$F^{\uparrow}(y) = F^{\uparrow}(x_1) \exp(-\lambda(x_1 - y)), \tag{4.10}$$

where it is natural to set the parameter λ such that $F^{\uparrow}$ is differentiable in x_1.

Acknowledgements

This research was supported by the Centre for Stochastic Geometry and Advanced Bioimaging, funded by the Villum Foundation.

References

Baddeley, A., Diggle, P. J., Hardegen, A., Lawrence, T., Milne, R. K., and Nair, G. (2014). On tests of spatial pattern based on simulation envelopes. *Ecological Monographs*, 84:477–489.

Grabarnik, P., Myllymäki, M., and Stoyan, D. (2011). Correct testing of mark independence for marked point patterns. *Ecological Modelling*, 222(23):3888–3894.

Loosmore, N. B. and Ford, E. D. (2006). Statistical inference using the G or K point pattern spatial statistics. *Ecology*, 87(8):1925–1931.