Statistics and Computing manuscript No.
(will be inserted by the editor)

A multivariate uniformity test for the case of unknown support

José R. Berrendero · Antonio Cuevas · Beatriz Pateiro-López

Received: date / Accepted: date

Abstract A test for the hypothesis of uniformity on a support $S \subset \mathbb{R}^d$ is proposed. It is based on the use of multivariate spacings as those studied in Janson (1987). As a novel aspect, this test can be adapted to the case that the support S is unknown, provided that it fulfils the shape condition of λ-convexity. The consistency properties of this test are analyzed and its performance is checked through a small simulation study. The numerical problems involved in the practical calculation of the maximal spacing (which is required to obtain the test statistic) are also discussed in some detail.

Keywords Uniformity · set estimation · multidimensional spacings

1 Introduction

In the univariate goodness-of-fit theory the hypothesis of uniformity is maybe the most important one, only second to normality; see, e.g., Marhuenda *et al.* (2005) for a recent comparison of univariate uniformity tests. The interest in the univariate uniform distribution is mostly associated with the need for checking the different procedures of random number generation which are in the basis of most simulation procedures and therefore in the core of many important procedures in science and technology. Besides this obvious motivation, the uniform distribution reflects a notion of non-informativeness and lack of structure which is interesting to check in many cases. The multivariate counterpart of the theory of uniformity tests is much less developed. It has been mostly considered (often under the name of "randomness") in connection with the theory of point processes, especially for the bi-variate case $d = 2$. In that setup the uniformity hypothesis amounts to assume that the underlying process generating the points is of Poisson type with a constant intensity function. See e.g., Ripley (1979), Moller and Waagepetersen (2004) for the study of randomness in the point processes framework.

However, we are concerned here with the more classical approach in which we want to test whether a $\mathbb{R}^d$-valued random variable X has a uniform distribution on a compact support $S \subset \mathbb{R}^d$ and the available information is given by a random iid sample of $X_1, \ldots, X_n$ drawn from X. Only a few references on this topic can be found in the literature even in the case of a simple null hypothesis, that is, when the support S is completely specified and we want to test

$$H_0 : X \text{ is uniform on } S. \tag{1}$$

Liang *et al.* (2001), propose a class of tests based on discrepancy measures between expectations of the type $E(f(X))$ and their Monte Carlo approximations. The asymptotic behavior of these discrepancies under the uniformity hypothesis is used to derive several uniformity tests for the case $S = [0, 1]^d$ which work for any dimension d. Berrendero *et al.* (2006) propose a method based on the distance from the sample points to the boundary of S. It is computationally efficient and works for a wide class of possible supports S but, still, only the case (1) of known support is analyzed in some detail by these authors. Tenreiro (2007) uses the Bickel-

José R. Berrendero
Universidad Autónoma de Madrid
E-mail: joser.berrendero@uam.es

Antonio Cuevas
Universidad Autónoma de Madrid
E-mail: antonio.cuevas@uam.es

Beatriz Pateiro-López
Universidad de Santiago de Compostela
E-mail: beatriz.pateiro@usc.es

Rosenblatt approach, based in nonparametric density estimation, to derive an uniformity test for the case $S = [0,1]^d$.

Smith and Jain (1984) define a test statistic for testing the uniformity of multidimensional data over some compact convex set. The test is obtained as an extension of the Friedman-Rafsky test (see Friedman and Rafsky, 1979), which determines if two sets of sample points belong to the same distribution. It is based on the Minimum Spanning Tree of the pooled samples.

The case of unknown support

In this paper we will deal with the case of unknown support. This means that we are considering a composite null hypothesis such as

$$H_0 : \text{The distribution of } X \text{ belongs to the class } \mathcal{U}_{\mathcal{C}}, \tag{2}$$

where $\mathcal{U}_{\mathcal{C}}$ is the class of uniform distributions whose support belongs to a family $\mathcal{C}$ of compact connected supports on $\mathbb{R}^d$.

This is the analog, for the uniformity hypothesis, of the classical problem of goodness of fit to a parametric family (which includes for example the important problem of testing whether a random variable has a normal distribution, without specifying the values of the location and scale parameters).

Note that there is a sort of "qualitative jump" in the difficulty of the problem (2) from $d = 1$ to $d > 1$. In the first case the class of possible connected supports coincides with the class of bounded closed intervals. Therefore it is "finite-dimensional" or "parametric"; see Baringhaus and Henze (1990). However, in the multivariate case $d > 1$ the class of possible connected supports is really huge so that the hypothesis (2) will be typically non-parametric, unless drastic assumptions are imposed on $\mathcal{U}_{\mathcal{C}}$. In fact, the problem of testing a composite hypothesis of type (2) remains largely unexplored, at least from the theoretical point of view. Berrendero *et al.* (2006) briefly outline a possible extension (based on set estimation ideas) of their distance-to-boundary method, valid for the problem (2) of unknown support. This idea is currently under development. Along the lines of Smith and Jain (1984), Jain *et al.* (2002) propose a method for the uniformity testing problem with unknown support. Their approach does not rely on any explicit assumption on the shape of the support. Still this interesting idea is not fully developed from the theoretical point of view.

The purpose of this paper

We propose here a uniformity test valid for the problem (2) of unknown support. It is based on the use of multivariate spacings as those analyzed in Janson (1987). This author proves a deep result which completely establishes the asymptotic behavior of the largest gap (of a prescribed shape) left by the sample points inside the support S, under extremely wide conditions for S. Relying on Janson's theorem a first uniformity test (for the case of known support) is proposed in Section 2 and their consistency properties are analyzed.

Section 3 is devoted to adapt this spacings-based test to the testing problem with unknown support (2). Our proposal is based on set estimation ideas and will assume that the unknown support belongs to the wide class of λ-convex sets which includes the compact convex supports.

The practical problems involved in the calculation of the multivariate largest spacing are far from trivial. They are analyzed in Section 4.

A small simulation study is given in Section 5.

Section 6 illustrates the application of the new uniformity test to a set of real data.

Section 7 is devoted to the proofs.

2 A test based on multivariate spacings

In the univariate case, the spacings defined by a random sample of points $X_1, \ldots, X_n$ in a support interval $S = [a,b]$ are defined as the gap lengths left by the sample points in the interval. They are calculated in a simple way in terms of differences between consecutive order statistics. The use of univariate spacings in the problem of testing uniformity is known since long time ago; see, e.g., Jammalamadaka and Goria (2004) and references therein.

In the multivariate case $X_1, \ldots, X_n \in S \subset \mathbb{R}^d$, the definition of spacings is not so straightforward. However, there still is a natural way to define the largest (or maximal) spacing Δ_n in such a way that some valuable properties can be derived for it.

Before going on, we need a bit of notation. In what follows let $S \subset \mathbb{R}^d$ be a compact support with $\mu(\partial S) = 0$, where μ denotes the Lebesgue measure on $\mathbb{R}^d$ and ∂S will stand for the boundary of S. Let $X_1, \ldots, X_n$ be a sample drawn on S with distribution P. Denote $\aleph_n = \{X_1, \ldots, X_n\}$. The shape of the considered spacings will be defined by a given set $A \subset \mathbb{R}^d$. For the validity of the theoretical results it is sufficient to assume that A is a bounded convex set with non-empty interior. For practical purposes the usual choices are $A = [0,1]^d$ or $A = B(0,1)$, where $B(x,r)$ denotes the closed ball

of radius r and center x. We will assume throughout this latter choice $A = B(0,1)$ which provides spherical spacings.

Then, the formal definition of maximal spacing is that used by Janson (1987); see also Deheuvels (1983):

$$\begin{aligned}\Delta_n(S;P) &= \sup\{r:\ \exists x \text{ with } x + rA \subset S \setminus \aleph_n\} \\ &= \sup\{r:\ \exists x \text{ with } B(x,r) \subset S \setminus \aleph_n\}. \end{aligned} \quad (3)$$

Let now B be a ball included in S. We denote by $\Delta_n(S,B;P)$ the maximum spacing in B generated by the sample $X_1,\ldots,X_n$ drawn on S. Of course, given the shape element A, the value of the maximal spacing depends only on S and on the sample points in $\aleph_n$ but the notation $\Delta_n(S;P)$ allows us to emphasize that the distribution of this random variable depends on the distribution P of the data points.

The Lebesgue measure (volume) of the balls with radii $\Delta_n(S,B;P)$ and $\Delta_n(S;P)$ will be denoted, respectively, by $V_n(S,B;P)$, $V_n(S;P)$. If P is omitted we will understand that it is uniform. S may be also omitted when no confusion is possible. When the notation $V_m(T)$ is used for a set $T \neq S$ we will understand that it represents the volume of the maximal spacing generated by a uniform sample of size m drawn on T, independently from the original sample $X_1,\ldots,X_n$ used to evaluate $V_n(S;P)$. Note however that the reasonings in the proof of Theorem 1 below do not depend on any independence assumption between V_m and V_n involved statistics as they only rely on the corresponding marginal distributions.

The following neat and general result due to Janson (1987) will be essential in what follows:

Janson's Theorem.- *If the X_i are iid uniform on S, with $\mu(S) = 1$, $\mu(\partial S) = 0$, and the volume element A is a bounded convex set with non-empty interior, then the following weak convergence holds*

$$nV_n - \log n - (d-1)\log\log n - \log\beta \xrightarrow{w} U \quad (4)$$

where V_n is the volume associated with the largest spacing Δ_n defined in (3), β is a known constant (depending on A), convergence in distribution is denoted as $\xrightarrow{w}$, and U is a random variable with distribution $\mathbb{P}(U \le u) = \exp(-\exp(-u))$, for $u \in \mathbb{R}$. Also

$$\liminf \frac{nV_n - \log n}{\log\log n} = d-1, \text{ almost surely (a.s.)}$$

and

$$\limsup \frac{nV_n - \log n}{\log\log n} = d+1, \text{ a.s.}$$

Just two remarks on this result:

(a) From the conclusions of this theorem we directly obtain,

$$\lim \frac{nV_n}{\log n} = 1, \text{ a.s.} \quad (5)$$

which will be very useful in the proofs of Theorems 1 and 2 below.

(b) The value of the constant $\beta = \beta(A)$ is explicitly given by Janson (1987) for the most important particular choices of A. For example it turns out that $\beta = 1$ whenever A is a cube. When A is a ball, the expression of β is more complicated and depends on the gamma function. However, in the case $d = 2$ we have $\beta = 1$ not only for the ball but also for any A with the centrosymmetric property, that is, for all A such that $A - x = -(A - x)$ for some x.

A uniformity test

Janson's theorem suggests a uniformity test on S, which would reject, at a significance level α, the null hypothesis (1) whenever

$$V_n(S;P) > \frac{u_\alpha + \log n + (d-1)\log\log n + \log\beta}{n}, \quad (6)$$

where u_α is the $1-\alpha$ quantile defined by $P(U > u_\alpha) = \alpha$.

In the general case that $\mu(S) = a$ the α-critical region would be

$$V_n(S;P) > \frac{a(u_\alpha + \log n + (d-1)\log\log n + \log\beta)}{n}. \quad (7)$$

The rationale behind this critical region is quite simple: if the distribution P fails to be uniform then there must exist a ball B such that $P(B \cap S)$ would be smaller than the corresponding uniform probability on $B \cap S$. This "low probability region" would be asymptotically detected by an unusually large spacing.

This is made explicit in the following consistency result whose proof is in Section 7.

Theorem 1.- *Assume, without loss of generality, $\mu(S) = 1$. The test (6) based on spacings is consistent against an alternative hypothesis of type*

H_1: P is absolutely continuous with density f such that there exists a ball $B \subset S$ and a constant c with $f(x) < c < 1$ for all $x \in B$.

This means that the probability of having the inequality (6) tends to 1, as $n \to \infty$ if H_1 is true.

3 The case of unknown support

We now concentrate on our main target of obtaining a test for the hypothesis (2) in which the support S is not given in advance.

We will use set estimation ideas; see Cuevas and Fraiman (2009) for a recent overview of this topic. The general idea is quite simple. We will just estimate S by an estimator S_n based on the sample $X_1, \ldots, X_n$. Then the maximal spacing $\Delta_n(S; P)$ will be estimated by means of

$$\widehat{\Delta}_n = \sup\{r : \exists x \text{ with } B(x, r) \subset S_n \setminus \aleph_n\}, \tag{8}$$

and the critical region (7) would be replaced with

$$\widehat{V}_n > \frac{a_n(u_\alpha + \log n + (d-1)\log\log n + \log\beta)}{n}. \tag{9}$$

Here $\widehat{V}_n$ denotes the volume of the ball with radius $\widehat{\Delta}_n$, given in (8) and $a_n = \mu(S_n)$. See Section 5 for a discussion on the effective calculation of a_n.

If $\widehat{\Delta}_n$ is close enough to Δ_n, its asymptotic distribution will also be given by (4). To have this we need to find a "good" estimator S_n of S, for which the estimated boundary ∂S_n is close enough to ∂S. This can be typically achieved by imposing some shape conditions on S and choosing an appropriate estimator based on them.

In particular, we will assume that S is λ-convex for some $\lambda > 0$. This means that S can be expressed as the intersection of the complement sets of a family of open balls of radius λ. In other words S is λ-convex if it coincides with its λ-*convex hull*, defined by

$$C_\lambda(S) = \bigcap_{\text{int}(B(x,\lambda))\cap S=\emptyset} (\text{int}(B(x,\lambda)))^c.$$

This definition is clearly reminiscent of the definition of a closed convex set S as the intersection of the closed half spaces which contain S. It can be seen that a closed convex set is λ-convex for all $\lambda > 0$. The reciprocal is true when S has non-empty interior. Some references on the statistical interest of the λ-convexity assumption and the estimation of λ-convex sets are Walther (1997, 1999), Rodríguez-Casal (2007) and Pateiro-López (2008). Figure 1 shows a λ-convex set and the λ-convex hull of a sample of size $n = 30$ for $\lambda = 0.25$.

There are three main reasons for considering the shape restriction of λ-convexity here. First, the class of λ-convex sets is very broad and, in any case, much wider than that of closed convex sets. In informal terms, this class includes (for different values of λ) most sets

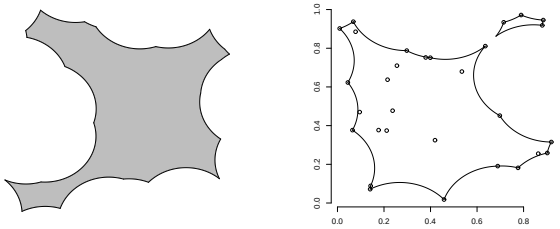

Fig. 1 λ-convex set (left), λ-convex hull of a random sample (right)

one could visualize or draw, provided that they do not have too sharp inlands; see the above mentioned references for technical details. Second, if we assume that a compact support S is λ-convex there is a natural estimator of S from a sample $\aleph_n = \{X_1, \ldots, X_n\}$ which is simply given by the λ-convex hull of the sample (see Rodríguez-Casal 2007 and Pateiro-López 2008),

$$S_n = C_\lambda(\aleph_n).$$

Third, if S is λ-convex we know precise (fast enough) convergence rates for the convergence of S_n, and ∂S_n towards S and ∂S, respectively.

Therefore, we will use the λ-convexity to define the class $\mathcal{C}$ of admissible supports in the hypothesis (2). This class will be made of those supports $S \subset \mathbb{R}^d$ satisfying the following property:

(CS) S is a compact path-connected set with $\text{int}(S) \neq \emptyset$. *Moreover, S and $\overline{S^c}$ are both λ-convex.*

This condition is borrowed from Walther (1997, 1999). It has been used by Rodríguez-Casal (2007) to derive the convergence rates in the approximation of S_n to S that we will need in our Theorem 2 below. The essential point is that the sets fulfilling condition (CS) have a number of properties which make them easier to handle. In particular, condition (CS) implies an intuitive rolling property (a ball of radius λ' rolls freely inside S and $\overline{S^c}$ for all $0 < \lambda' \leq \lambda$) which can be seen as a sort of geometric smoothness statement. See Walther (1997, 1999) for a detailed discussion of these issues.

The following theorem establishes the validity of the uniformity test (9). The proof is given in Section 7.

THEOREM 2.- *Let $X_1, \ldots, X_n$ be a sample of iid observations from a random variable X whose distribution P_X has support $S \subset \mathbb{R}^d$. Let us consider the problem of testing*

H_0 : *P_X has a uniform distribution with support in $\mathcal{C}$,*

where $\mathcal{C}$ is the class of supports satisfying the property (CS). The inequality (9), where $S_n = C_\lambda(\aleph_n)$, is a critical region for H_0 with an asymptotic significance level α, that is

$$\lim_{n\to\infty} \mathbb{P}(\widehat{V}_n > C_{n,\alpha}) = \alpha, \tag{10}$$

under H_0, where $C_{n,\alpha}$ denotes the right-hand side in (9).

Moreover, this test is consistent against the same alternative hypothesis indicated in Theorem 1.

4 Numerical aspects: the effective calculation of spacings

The practical implementation of the proposed tests requires the numerical calculation of the maximal spacings Δ_n (when the support S is known) and $\widehat{\Delta}_n$ (when S is unknown). This is a non-trivial problem closely related with some relevant concepts in stochastic geometry. This section is devoted to discuss this problem both when the support S is assumed to be known and when S is unknown. Recall that, in both cases, calculating the maximal spacing essentially means finding the largest ball contained in a set (either S or an estimation of S), that does not intersect the sample. Let us restrict ourselves to the two dimensional case $S \subset \mathbb{R}^2$. The algorithm we propose consists basically of two stages. First, based on the Voronoi diagram and Delaunay triangulation of the sample, we determine an initial radius, stored as a candidate to be the maximal spacing. Then, by enlarging this initial value iteratively, we define an increasing sequence of radii and check whether any of them satisfies the conditions to define the maximal spacing.

This algorithm could be generalized to $d > 2$. At the end of the following paragraph we explain how to adapt the first stage for $d = 3$. As for the iterative "enlarging" stage, the indicated methodology should also work for $d > 2$ although the computationally efficient implementation is not straightforward.

The Voronoi diagram and the Delaunay triangulation

The Voronoi diagram of a finite sample of points $\aleph_n$ in $\mathbb{R}^2$ is a covering of the plane by n regions V_i where, for each $i \in \{1, \ldots, n\}$, the cell V_i consists of all points in $\mathbb{R}^2$ which have X_i as nearest sample point. That is, $V_i = \{x \in \mathbb{R}^2 : \ \|x - X_i\| \leq \|x - X_j\| \text{ for all } X_j \in \aleph_n\}$. Two sample points X_i and X_j are said to be Voronoi neighbours if the cells V_i and V_j share a common point. We denote the Voronoi Diagram of $\aleph_n$ by $VD(\aleph_n)$. The Delaunay triangulation of $\aleph_n$, denoted by $DT(\aleph_n)$, is defined as the partition of $\mathbb{R}^2$ by triangles which are delimited by the segments (called Delaunay edges) connecting the Voronoi neighbours. Thus, there exists a Delaunay edge connecting the sample points X_i and X_j if and only if they are Voronoi neighbours. The Delaunay triangulation was introduced by Voronoi (1907) and extended by Delaunay (1934) who proved the following *empty circle property*: No vertex in a Delaunay triangulation is included inside a circumcircle drawn through the vertices of a Delaunay triangle. This property will be used below in order to get a radius to initialize our algorithm. For a survey of these topics we refer to Aurenhammer (1991). Figure 2 shows a Voronoi diagram, the corresponding Delaunay triangulation and the empty circle property. The circumference in gray represents the largest circumcircle contained in the set.

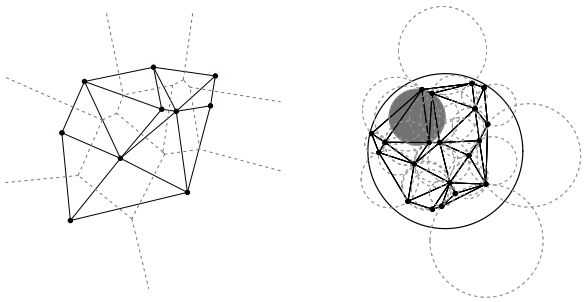

Fig. 2 Voronoi diagram and Delaunay triangulation (left). The empty circle property (right); the circumference in gray corresponds to the first step of the algorithm to calculate Δ_n, being S the ball in solid line.

The Delaunay triangulation and its duality to the Voronoi diagrams generalize to higher dimensions. The Delaunay triangulation of a set of points in $\mathbb{R}^d$ is a tessellation of the convex hull of the points such that no d-sphere defined by the d-triangles contains any other points from the set. Figure 3 shows the Delaunay triangulation of a set of points in $\mathbb{R}^3$. The sphere represents the largest circumsphere contained in the set $S = [0, 1]^3$. Regarding the computation of the Delaunay triangulation in general dimension, Barber *et al.* (1996) present the implementation of the Quickhull algorithm, that computes convex hulls, Delaunay triangulations, Voronoi vertices, furthest-site Voronoi vertices, and halfspace intersections in $\mathbb{R}^d$. This algorithm is available for computing languages such as R, MATLAB, and Mathematica. See, for the former, the package `geometry`, by Grasman and Gramacy (2010).

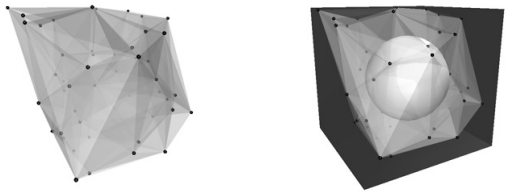

Fig. 3 Delaunay triangulation of a sample of points in $\mathbb{R}^3$ (left). The sphere corresponds to the first step of the algorithm to calculate Δ_n, being S the unit cube (right).

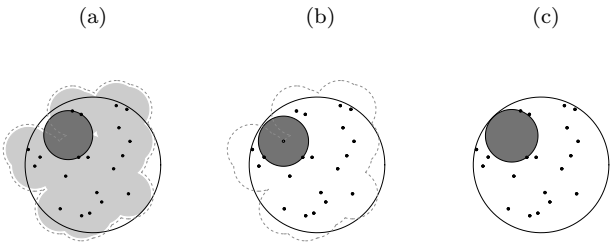

Fig. 4 Maximal spacing algorithm with known support S. (a) In gray, ball with radius $\Delta_n^{(0)}$. In light gray, $B(\aleph_n, \Delta_n^{(0)})$. In dashed line $\partial B(\aleph_n, r^{(1)})$. (b) Ball with radius $\Delta_n^{(1)}$. (c) Ball with radius $\Delta_n^{(N)}$.

The case of known support

Let us first assume that the support S is known. Note that, even though they may not be contained in S, the open circumdiscs defined by the Delaunay triangulation do not intersect the sample. Therefore $\Delta_n(S;P) \geq \Delta_n^{(0)}$, where $\Delta_n^{(0)} = \max\{r : \exists B(x,r) \subset S,\ B(x,r) \text{ circumdisc defined by } DT(\aleph_n)\}$; this is the radius of the circumference in gray in Figure 2. The decision on whether a circumdisc, $B(x,r)$, is contained in S is made based on $d(x, \partial S)$ which is not difficult to calculate for certain sets S.

Once $\Delta_n^{(0)}$ is determined, we proceed iteratively. The goal is to find a disc, $B(x,r)$, such that $\text{int}(B(x,r)) \subset S \setminus \aleph_n$, with $r > \Delta_n^{(0)}$. If such a disc does exist then $x \notin B(\aleph_n, \Delta_n^{(0)})$, where

$$B(\aleph_n, \Delta_n^{(0)}) = \bigcup_{X_i \in \aleph_n} B(X_i, \Delta_n^{(0)})$$

is the dilation of radius $\Delta_n^{(0)}$ of the sample. This means that the centres of the possible maximal balls necessarily lie outside $B(\aleph_n, \Delta_n^{(0)})$. Now, fix a maximum number of iterations N and $\varepsilon > 0$. Set $r^{(0)} = \Delta_n^{(0)}$ and, for each $k = 1, \ldots, N$, proceed as follows:

1. Set $r^{(k)} = r^{(k-1)} + \varepsilon$.
2. Determine the set $D^{(k)} = S \cap \partial B(\aleph_n, r^{(k)})$.
3. If $d(x, \partial S \cup \aleph_n) > \Delta_n^{(k-1)}$ for any $x \in D^{(k)}$, set $\Delta_n^{(k)} = \max\{d(x, \partial S \cup \aleph_n) : d(x, \partial S \cup \aleph_n) > \Delta_n^{(k-1)},\ x \in D^{(k)}\}$ and $r^{(k)} = \Delta_n^{(k)}$. Else, set $\Delta_n^{(k)} = \Delta_n^{(k-1)}$.

The radius $\Delta_n^{(N)}$ approximates the maximal spacing $\Delta_n(S;P)$. The implementation of this algorithm is visualized in Figure 4.

The case of unknown support

Assume now that the support S is unknown. As indicated above, in this situation the maximal spacing $\Delta_n(S;P)$ is estimated by $\Delta_n(S_n, P)$, where S_n is a suitable estimator of S. Thus, the first step in this case is to compute S_n based on the sample. If we assume that S is λ-convex for some $\lambda > 0$ we will take $S_n = C_\lambda(\aleph_n)$, the λ-convex hull of the sample. The algorithm for computing $C_\lambda(\aleph_n)$ is described in Edelsbrunner *et al.* (1983). Let $\widehat{\Delta}_n^{(0)} = \max\{r : B(x,r) \subset C_\lambda(\aleph_n),\ B(x,r) \text{ circumdisc defined by } DT(\aleph_n)\}$.

The structure of the λ-convex hull can be used to decide whether a circumdisc, $B(x,r)$, is contained in $C_\lambda(\aleph_n)$. By DeMorgan's law, the complement of $C_\lambda(\aleph_n)$ can be written as the union of all open balls of radius λ which contain no point of $\aleph_n$. As a consequence, $\partial C_\lambda(\aleph_n)$ is formed by arcs of balls of radius λ (besides possible isolated sample points). These arcs are determined by the intersections of some of the balls that define the complement of the λ-convex hull, see Figure 5. Therefore, for points $x \in C_\lambda(\aleph_n)$, we can compute $d(x, \partial C_\lambda(\aleph_n))$ from the minimum distance between x and the centers of the balls that define $\partial C_\lambda(\aleph_n)$.

Once $\widehat{\Delta}_n^{(0)}$ is determined the algorithm follows the same steps as in the case of known support.

One issue that needs to be addressed is the selection of the parameter λ which, in practice, may be unknown. Different values of λ result in different sets and, consequently, the choice of this parameter may affect the value of $\widehat{\Delta}_n^{(N)}$. For sufficiently large λ, the λ-convex hull tends to the convex hull of the sample whereas, as λ decreases, the estimator shrinks until that, for sufficiently small λ, the λ-convex hull reduces to the sample points. Therefore, if the original set S is known to be convex, the estimator works reasonably well for large values of

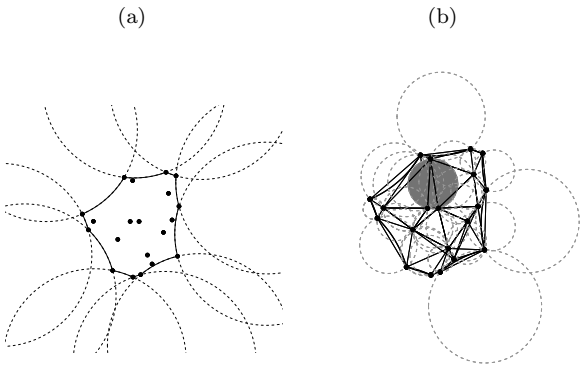

Fig. 5 Random sample of size $n = 20$ on $S = B(0, 1/\sqrt{\pi})$. (a) In solid black line, $\partial C_\lambda(\aleph_n)$ for $\lambda > 0$. The boundary of $C_\lambda(\aleph_n)$ is formed by arcs of balls of radius λ (in dashed line). (b) In dashed line, circumcircles defined by the Delaunay triangulation of the sample. In gray, circumdisc with radius $\widehat{\Delta}_n^{(0)}$.

λ. However, if S is not convex and λ is too large, the estimator may not be able to identify the cavities of the set. In the absence of an automatic parameter selection method, small values of λ result in more conservative estimations.

In any case, as we will see in the simulation experiments of the next section, the numerical outputs (in terms of significance level) are quite robust against small and moderate changes in λ.

5 Simulation study

We offer here a small simulation study in order to check the practical performance of our proposals. The details are as follows.

Tests under study

Our main target is to study the behaviour of our maximal spacing tests (denoted by MS and EMS) whose critical region are (6) or (9) corresponding, respectively, to the case where the support S is known (with $\mu(S) = 1$) and estimated.

In the case of unknown support, we also need to compute the area of the estimator, see the numerator of (9). The area of the λ-convex hull estimator can be exactly determined. Since the λ-convex hull is defined as the complement of the intersection of open balls of radius λ, its boundary is formed by the union of arcs of balls of radius λ besides possible isolated points. In view of this observation, the area of the λ-convex hull can be computed by adding and subtracting (if there are holes) the areas of the polygons that result from joining the extreme points of adjacent arcs and the areas of the circle segments determined by those arcs. Despite the fact that we are able to compute the area of the estimator, there is still one problem we have to face with. Since, with probability one, $C_\lambda(\aleph_n) \subset S$, the area of the estimator systematically underestimates the area of the support and, as a consequence, the test tends to reject the null hypothesis more readily than one would wish. Finding an appropriate dilation of the estimator in order to correct the bias is an open problem which was already pointed out by Ripley and Rasson (1977) for the case of the convex hull estimator. As in their paper, we propose to estimate $\mu(S)$ by means of

$$a_n^* = \frac{n}{n - v_n}\mu(C_\lambda(\aleph_n)),$$

where v_n is the number of vertices of the λ-convex hull. Thus, we reject the null hypothesis of uniformity on an unknown support whenever,

$$\widehat{V}_n > \frac{a_n^*(u_\alpha + \log n + (d-1)\log\log n + \log\beta)}{n}.$$

As competitor procedures we will consider several tests proposed by Liang *et al.* (2001). These tests are based on different statistics (we will consider here those denoted A_n and T_n) as well as on different discrepancy measures (we will use here those called *symmetric*, *centered*, and *star*). We will also consider the one-sided (o-s) "distance-to-boundary" (DB) test proposed by Berrendero *et al.* (2006). Let us recall that all these competitors require the knowledge of the support S. Therefore the comparison is not fair for our EMS method (as the competitors incorporate extra information) but still it will allow us, at least, to assess the loss of efficiency entailed on the estimation of the support. The test by Smith and Jain (1984) addresses the problem of testing the uniformity of multidimensional data over some compact set, called the sampling window. It uses the Friedman-Rafsky test to determine if two samples come from the same population. One of the samples is the given data and the other one is generated uniformly over the sampling window when it is known. For unknown sampling windows, the second sample is generated uniformly over the convex hull of the given data.

Considered supports

We will use the following supports:

(a) $S = [0, 1]^2$.
(b) $S = \left\{(x, y) \in \mathbb{R}^2 : \ x^2 + y^2 \leq \frac{1}{\pi}\right\}$.

(c) $S = \left\{(x,y) \in \mathbb{R}^2 : \ \left(\frac{x}{0.8}\right)^2 + \left(\frac{y}{0.39}\right)^2 \leq 1\right\}$.
(d) $S = \left\{(x,y) \in \mathbb{R}^2 : \ 0.32^2 \leq x^2 + y^2 \leq 0.65^2\right\}$.

As mentioned before, most existing methods in the literature for testing uniformity in the multidimensional context are limited to the unit hypercube. We have chosen the support (a) in order to compare our proposal with other tests although, for the case of unknown support, $S = [0,1]^2$ does not fulfill property (CS) in Theorem 2. Note that (CS) is used only as a sufficient condition to guarantee a fast enough rate in the estimation of S (see the proof of Theorem 2). Nevertheless, the test will work whenever this rate holds, even if (CS) is not fulfilled.

Outputs

The numerical results given below have been obtained using the R software, R Development Core Team (2008), and the `alphahull` package. See Pateiro-López and Rodríguez-Casal (2010) for a description of the library.

Table 1 gives the outputs corresponding to the empirical significance level obtained (as an average over 10000 independent runs) with the different tests intended for a nominal significance level $\alpha = 0.05$. Sample sizes are $n = 100, 200, 500$. Since the tests proposed by Liang *et al.* (2001) are designed for the unit hypercube, they are only evaluated for that support. The test by Smith and Jain (1984) is evaluated for compact convex supports, since it has been developed under this assumption.

Table 2 gives the empirical powers of the different procedures under the so-called Neyman-Scott model. This is a typical deviation from the uniformity assumption, often considered in the theory of point processes; see, e.g., Moller and Waagepetersen (2004), chapter 5 for details. This kind of non-uniformity pattern appears as quite natural in our case, as it can be defined, and easily interpreted, irrespectively of the structure of the support S. Under this model the sample tends to provide "clustered" observations. Figure 6 shows two examples of data sets generated under a Neyman-Scott process.

Table 3 shows the empirical significance levels attained with different choices of λ for the EMS test under the models (a)-(c). The results suggest a remarkable stability with respect to the values of this shape parameter.

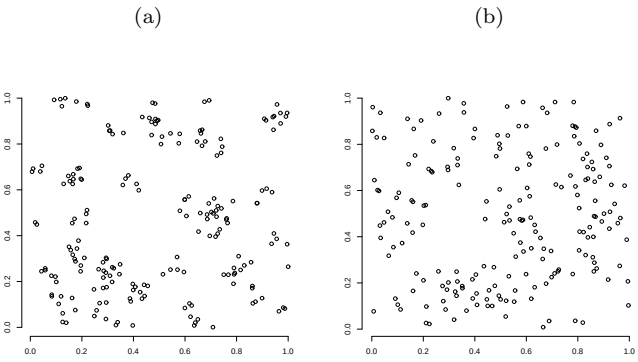

Fig. 6 Realization from the Neyman-Scott process over the unit square with size $n = 200$. Each cluster consist of 5 points, generated from the uniform distribution on a disc of radius r. (a) $r = 0.05$. (b) $r = 0.1$.

Conclusions

We observe that, in which concerns the significance level, both tests (MS and EMS) perform reasonably well for large sample sizes, except perhaps in the case (d) of the circular crown where larger sample sizes are needed. In the remaining cases, even in the case of unknown support, the estimated maximal spacing test does a good job in preserving the nominal value. The reason is that, whenever the support estimator is able to the recover faithfully the boundary of the original set, the maximal spacing with unknown support is not much smaller than the maximal spacing with known support. This is illustrated in Figure 7, where the balls corresponding to the maximal spacing and estimated maximal spacing are represented. For the the unit square in (a), the choice of a large value of λ ensures that the support estimator works well. For the supports (b), (c) and (d), the parameter λ is chosen as the largest value for which the corresponding support is λ-convex.

We should say that other uniformity tests reported in literature, such as the proposals by Liang *et al.* (2001), show a better performance for small sample sizes. Recall, however, that these tests are designed for the case $S = [0,1]^d$ so that, in principle, they are not applicable to the case where S is a support different from the unit square or to the case where S is unknown. We should also keep this in mind when comparing the power outputs of the different tests in the case of unknown support. However, even when S is unknown, the test EMS based on the maximal spacing shows a clear superiority in the case of Neyman-Scott clustering alternatives over the proposals by Liang *et al.* (2001) and

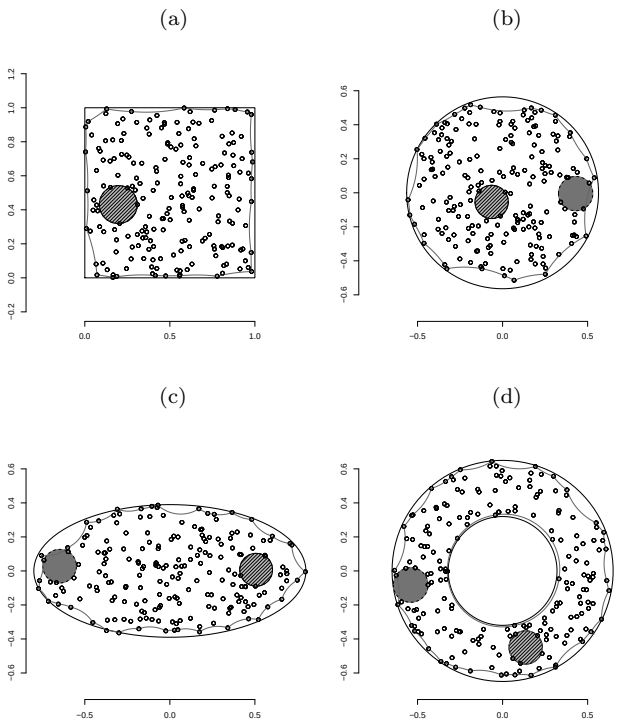

Fig. 7 Uniform samples of size $n = 200$ on different supports S. In dark gray, maximal ball with known support. In light gray and dashed, maximal ball with unknown support. The support is estimated through $C_\lambda(\aleph_n)$, whose boundary is represented in gray. The values of λ are (a) $\lambda = 1$, (b) $\lambda = 0.56$, (c) $\lambda = 0.19$, and (d) $\lambda = 0.32$.

the DB test by Berrendero *et al.* (2006). The reason is that under the Neyman-Scott processes empty spaces or holes tend to arise between the clusters of points and, consequently, the maximal balls are significantly larger than those in the uniform case; see Figure 8. Regarding Smith and Jain's (1984) test for convex supports, the power outputs are satisfactory and slightly better than those of their competitors. However, the observed significance levels are always larger than the theoretical ones, especially for large sample sizes. This could be due to the fact that we are using, as proposed in Smith and Jain (1984), the conditional version of the test statistic given in Friedman and Rafsky (1979, p. 702). This "conditional version" (easier to compute) is obtained by replacing the variance in the denominator of the test statistic by a conditional variance; see equation (14) in Friedman and Rafsky (1979).

We should also mention that the performance of the EMS test (and even that of the MS) is much worse under a further class of "natural" deviations from uniformity, namely the "ϵ-contamination models" as those

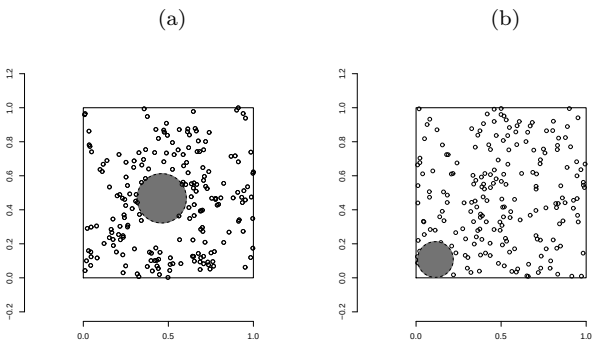

Fig. 8 Maximal ball with known support. (a) Neyman-Scott process over the unit square of size $n = 200$. (b) Uniform sample on the unit square of size $n = 200$.

considered in Berrendero *et al.* (2006). In these models the sample includes some contamination given by an extra proportion ϵ of "outliers" (observations close to the boundary of S) or "inliers" (observations close to the middle of S).

As a final conclusion, we think that the EMS test is to be especially recommended whenever there are some reasons to suspect departures from uniformity via "clustered" observations. The fact that (unlike the remaining standard tests) the EMS procedure does not require the knowledge of the underlying support, provides an extra flexibility. Our simulation results show that, if the shape of S is somewhat involved, the EMS procedure could be rather conservative. Still, we have performed other numerical experiments, not reported in Tables 1-3, which suggest a good behavior with respect to power even in this case.

6 A real data example

There are different contexts in which to deal with a set of points distributed within a region. One example is the study of spatial patterns in natural vegetation. In order to understand the establishment or the growth of plant communities, it is important to analyze their spatial characteristics. Given a plant population, a good starting point is to test whether the point locations of the individuals are distributed completely at random. A positive answer would dissuade us from attempting to model any kind of structure in the data. Usually the randomness assumption is formalized in terms of the validity of an homogeneous Poisson process. This could be translated into our approach by conditioning to a fixed sample size.

The data we have studied, due to Gerrard (1969), come from an investigation of a 19.6 acre square plot in

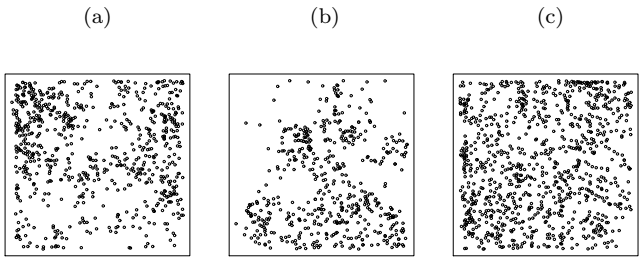

Fig. 9 Location of trees in Lansing Woods, Clinton County, Michigan, USA. (a) Hickories. (b) Mapples. (c) Oaks.

Lansing Woods, Clinton County, Michigan, USA. The complete data set give the locations of 2251 trees. It is available from the R package `spatstat`, see Baddeley and Turner (2005). We have considered for this study the locations corresponding to the three major species (hickory, maple and oak). Figure 9 shows the three spatial point patterns. The original plot size has been rescaled to the unit square. As discussed in Diggle (2003), the visual examination of the data suggests evidence of aggregation for the hickories and especially for the maples. For the oaks, the plot shows no obvious structure and might be considered as a realization from a uniform distribution. This intuition is supported by the results of several complete spatial randomness tests in Diggle (2003). We have applied the uniformity tests discussed in this work to the Lansing Woods data. Results are given in Table 4. The tests based on maximal spacings (MS and EMS) accept the null hypothesis of uniformity for the oaks, at a significance level of 0.05, and reject in favour of the alternative hypothesis for both the hickories and the maples. For the EMS test we have chosen λ=1. As a conclusion, we cannot assume that the locations of hickories and maples are distributed completely at random. The next natural step in a further study of the data would be to find groups in these two species and to estimate the number of clusters. We refer to Cuevas *et al.* (2000) and Tibshirani *et al.* (2001) for different approaches to this major problem.

7 Proofs

In this section we give the proofs of the stated theorems.

Proof of Theorem 1: If ξ_n and η_n are positive sequences of random variables, the notations

$$\xi_n \lesssim \eta_n,\ \xi_n \lessapprox \eta_n \text{ and } \xi_n \approx \eta_n$$

will mean, respectively that, with probability one (that is, a.s.),

$$\limsup \frac{\xi_n}{\eta_n} < 1,\ \limsup \frac{\xi_n}{\eta_n} \leq 1 \text{ and } \lim \frac{\xi_n}{\eta_n} = 1.$$

For the sake of clarity we will divide the proof in several steps:

Step 1: It will suffice to prove that

$$V_n(S) \lesssim V_n(S;P), \tag{11}$$

where $V_n(S)$ and $V_n(S;P)$ are based on independent samples drawn from the uniform on S and the distribution P, respectively.

Indeed, note that the null hypothesis, under uniformity would be rejected whenever

$$nV_n(S) - \log n - (d-1)\log\log n - \log\beta > u_\alpha \tag{12}$$

and, under P the critical region is

$$nV_n(S;P) - \log n - (d-1)\log\log n - \log\beta > u_\alpha. \tag{13}$$

The difference between the statistics of the left-hand sides is

$$n[V_n(S;P) - V_n(S)] = n\left[\frac{V_n(S;P)}{V_n(S)} - 1\right] V_n(S).$$

If (11) holds, then for all $\epsilon > 0$ small enough we have that, with probability one, there exists n_0 such that

$$\frac{V_n(S;P)}{V_n(S)} > \frac{1}{1-\epsilon},\ \forall n \geq n_0.$$

On the other hand, from Equation (5) $nV_n(S) = \infty$, a.s. Then we conclude $n[V_n(S;P) - V_n(S)] = \infty$ a.s. and, since the rejection probability under uniformity in (12) is α, the rejection probability in (13) tends to 1 as $n \to \infty$.

Step 2: Let B be the ball indicated in the definition of hypothesis H_1. Note that, in order to prove (11) it is in turn sufficient to prove that for any $c_1 \in (c,1)$,

$$V_n(c_1^{-1/d}S) \lesssim V_n(S,B;P) \tag{14}$$

since, from the definition of the spacings and Equation (5), we have that

$$V_n(c_1^{-1/d}S) \approx c_1^{-1}V_n(S) > V_n(S)$$

and

$$V_n(S,B;P) \leq V_n(S;P).$$

Step 3: To see (14) we will find a sequence $V_m(B)$, with appropriately chosen $m = m(n)$, such that

$$V_n(c_1^{-1/d}S) \lesssim V_m(B) \tag{15}$$

and

$$V_m(B) \lessapprox V_n(S, B; P). \tag{16}$$

More precisely, we will take

$$m = \left\lfloor \frac{\mu(B)n}{\mu(c^{-1/d}S)} \right\rfloor = \lfloor c\mu(B)n \rfloor$$

where $\lfloor k \rfloor$ denotes the integer part of k.

Step 4: Now the proof is reduced to check (15) and (16). In order to prove (16) we will first show that

$$V_n(c^{-1/d}S, B_1) \approx V_m(B), \tag{17}$$

where B_1 denotes a ball such that $B_1 \subset c^{-1/d}S$ and $\mu(B_1) = \mu(B)$. Indeed, relation (17) follows from the fact that, according to the Strong Law of Large Numbers, if Y_i are i.i.d. uniformly distributed on $c^{-1/d}S$, then the random variable

$$N_n = \#\{i : Y_i \in B_1\}$$

is such that, for all $\epsilon > 0$,

$$m_1 := \lfloor (1-\epsilon)nc\mu(B) \rfloor \leq N_n \leq \lceil (1+\epsilon)nc\mu(B) \rceil := m_2,$$

eventually with probability 1, where $\lceil r \rceil = \lfloor r \rfloor + 1$. Then,

$$V_{m_2}(B) \lessapprox V_n(c^{-1/d}S, B_1) \lessapprox V_{m_1}(B),$$

but, from (5),

$$V_{m_1}(B) \approx \frac{\log m_1}{m_1}\mu(B) \approx \frac{\log n}{(1-\epsilon)nc},$$

$$V_{m_2}(B) \approx \frac{\log m_2}{m_2}\mu(B) \approx \frac{\log n}{(1+\epsilon)nc},$$

for all $\epsilon > 0$. Therefore

$$V_n(c^{-1/d}S, B_1) \approx \frac{\log n}{nc}$$

which proves (17) since, using again (5)

$$V_m(B) \approx \frac{\log n}{nc}.$$

Step 5: Since the uniform density on $c^{-1/d}S$ is equal to c on the set B_1 and the density f of P is such that $f < c$ on B,

$$V_n(c^{-1/d}S, B_1) \lessapprox V_n(S, B; P).$$

This, together with (17), proves (16).

Step 6: Likewise, (15) follows from

$$V_n(c_1^{-1/d}S) \approx c_1^{-1}V_n(S) \lesssim \frac{\log n}{nc} \approx V_m(B).$$

Finally, as indicated in Step 3, we conclude (14) since we have proved (15) and (16).

□

Proof of Theorem 2: To prove the first statement (concerning the asymptotic preservation of the level) the crucial point is to show that, under H_0, the weak convergence (4) holds also when V_n is replaced with $\widehat{V}_n$.

To see this denote by $d_\mu(S_n, S)$ and $d_H(S_n, S)$ the symmetric difference distance and the Hausdorff distance between S_n and S defined, respectively, by

$$d_\mu(S_n, S) = \mu(S_n \Delta S) = \mu\left((S_n \setminus S) \cup (S \setminus S_n)\right)$$
$$d_H(S_n, S) = \max\{\sup_{x \in S_n} d(x, S), \sup_{y \in S} d(y, S_n)\},$$

where, for any $A \subset \mathbb{R}^d$ and $x \in \mathbb{R}^d$,

$$d(x, A) = \inf\{\|x - y\| : y \in A\}.$$

Under the assumption (CS) it has been proven (see Rodríguez-Casal 2007, Theorem 3) that, with probability one,

$$d_H(\partial S_n, \partial S) = O\left(\left(\frac{\log n}{n}\right)^{2/(d+1)}\right),$$

and the same holds for $d_H(S_n, S)$ and $d_\mu(S_n, S)$.

Define

$$\epsilon_n = \max\{d_H(\partial S_n, \partial S), d_H(S_n, S)\},$$

$$T_n = S \setminus \{x \in S : \ d(x, \partial S) \leq 2\epsilon_n\}.$$

We now see that $T_n \subset S_n$ or, equivalently, $S_n^c \subset T_n^c$. Indeed, take $x \notin S_n$ such that $x \in S$ (the case $x \notin S$ is trivial). Since $d_H(S, S_n) \leq \epsilon_n$, there exists $y \in S_n$ such that $d(x, y) \leq \epsilon_n$. Also, since the projection of x on S_n must belong to ∂S_n, we may take $y \in \partial S_n$. On the other hand, since $d_H(\partial S, \partial S_n) \leq \epsilon_n$, there exists $z \in \partial S$ such that $d(y, z) \leq \epsilon_n$. From the triangle inequality, $d(x, z) \leq 2\epsilon_n$ so that $x \notin T_n$. Therefore $S_n^c \subset T_n^c$.

Now, let $B(x', \Delta_n)$ be any "maximal" ball realizing the definition (3) for some $x' \in S$. We have

$$B(x', \Delta_n - 2\epsilon_n) \subset T_n \subset S_n.$$

and therefore $\widehat{\Delta}_n \geq \Delta_n - 2\epsilon_n$, where $\widehat{\Delta}_n$ is the estimated maximal radius defined in (8). This, together with $\widehat{\Delta}_n \leq \Delta_n$ a.s. (which follows from $S_n \subset S$ a.s.) proves that, with probability one,

$$\widehat{\Delta}_n - \Delta_n = O(\log n/n)^{2/(d+1)}. \tag{18}$$

Now denote $\widehat{\Delta}_n - \Delta_n = r_n$. We want to prove that (4) holds also when nV_n is replaced with $n\widehat{V}_n$. As $\widehat{V}_n = \mu(B(0,1))\widehat{\Delta}_n^d$, it will suffice to prove that, with probability one,

$$n\widehat{\Delta}_n^d - n\Delta_n^d \to 0 \tag{19}$$

To this end put

$$n\widehat{\Delta}_n^d = n(\Delta_n + r_n)^d = n\sum_{k=0}^{d}\binom{d}{k}\Delta_n^{d-k}r_n^k$$

and recall that from (5), Δ_n^{d-k} is (with probability one) of exact order $(\log n/n)^{(d-k)/d}$ for $0 \le k \le d$. Thus, from (18), with probability one,

$$n\Delta_n^{d-k}r_n^k = nO\left(\left(\frac{\log n}{n}\right)^{(d-k)/d}\left(\frac{\log n}{n}\right)^{2k/(d+1)}\right) \tag{20}$$

which proves $n\Delta_n^{d-k}r_n^k \to 0$ a.s., since (up to $\log n$ terms) the right hand side of (20) is of order n^q with

$$q = 1 - \frac{d^2+d-kd-k+2kd}{d^2+d} < 0.$$

We have thus proved (19) and therefore, from (4), we also conclude

$$n\widehat{V}_n - \log n - (d-1)\log\log n - \log\beta \xrightarrow{w} U$$

with $\mathbb{P}(U \le u) = \exp(-\exp(-u))$ for $u \in \mathbb{R}$.

To show (10) we only need to prove $a_n \to a$ with probability one. But this follows from Theorem 3 in Rodríguez-Casal (2007) which, in particular, states that $d_\mu(S_n, S) \to 0$, a.s.

Finally, the consistency result against an alternative hypothesis of type of those considered in Theorem 1 follows from (5) and (19) which together imply the validity of the "estimated analog" of expression (5),

$$\frac{n\widehat{V}_n}{\log n} \to 1, \text{ a.s..} \tag{21}$$

Now the proof of consistency follows the same lines as the proof of Theorem 1, replacing (5) with (21). □

Acknowledgements This work has been partially supported by Spanish Grant MTM2010-17366 (first and second author) and by Spanish Grant MTM2008-03010, PGIDIT06PXIB207009PR and the IAP research network grant no. P6/03 from the Belgian government (third author).

References

1. Aurenhammer, F.: Voronoi diagrams. A survey of a fundamental geometric data structure. ACM Computing Surveys. **23**, 345–405 (1991)
2. Baddeley, A. and Turner, R.: Spatstat: an R package for analyzing spatial point patterns. J. Stat. Softw. **6**,1–42 (2005)
3. Baringhaus, L. and Henze, N.: A test for uniformity with unknown limits based on D'Agostino's D. Statist. Probab. Lett. **9**, 299–304 (1990)
4. Berrendero, J.R., Cuevas, A. and Vázquez-Grande, F.: Testing multivariate uniformity: The distance-to-boundary method. Canad. J. Statist. **34**, 693–707 (2006)
5. Barber. B. C., Dobkin D. P. and Huhdanpaa, H.: The Quickhull Algorithm for Convex Hulls. ACM Transactions on Mathematical Software**22**, 469–483 (1996)
6. Cuevas, A., Febrero, M. and Fraiman, R.: Estimating the number of clusters. Canad. J. Statist. **28**, 367–382 (2000)
7. Cuevas, A. and Fraiman, R, F.: Set estimation. In New Perspectives on Stochastic Geometry. W.S. Kendall and I. Molchanov, eds. Oxford University Press, 366–389 (2009)
8. Deheuvels, P.: Strong bounds for multidimensional spacings. Z. Wahrsch. Verw. Gebiete **64**, 411–424 (1983)
9. Delaunay, B.: Sur la sphere vide. Bull. Acad. Sci. USSR **7**, 793–800 (1934)
10. Diggle, P. J.: Statistical Analysis of Spatial Point Patterns, 2nd edition. Edward Arnold, London (2003)
11. Edelsbrunner, H., Kirkpatrick D.G. and Seidel, R.: On the shape of a set of points in the plane. IEEE Trans. Inform. Theory **29**, 551–559 (1983)
12. Friedman, J.H. and Rafsky, L.C.: Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. Ann. Statist. **7**, 697–717 (1979)
13. Gerrard, D.: Competition Quotient: A New Measure of the Competition Affecting Individual Forest Trees. Research Bulletin 20, Agricultural Experiment Station, Michigan State University. (1969)
14. Grasman, R. and Gramacy, R. B.: Geometry: Mesh generation and surface tesselation. R package version 0.1-7, `http://CRAN.R-project.org/package=geometry`. (2010)
15. Jain, A., Xu, X., Ho, T. and Xiao, F.: Uniformity testing using minimal spanning tree. Proceedings of the 16th International Conference on Pattern Recognition, **4**, 281–284 (2002)
16. Jammalamadaka, S.R. and Goria, M.N.: A test of goodness-of-fit based on Gini's index of spacings. Statist. Probab. Lett. **68**, 177–187 (2004)
17. Janson, S.: Maximal spacings in several dimensions. Ann. Probab. **15**, 274–280 (1987)
18. Liang, J. J., Fang, K. T., Hickernell, F. J. and Li, R.: Testing multivariate uniformity and its applications. Math. Comp. **70**, 337–355 (2001)
19. Marhuenda, Y., Morales, D. and Pardo, M.C.: A comparison of uniformity tests. Statistics **39**, 315–328 (2005)
20. Moller, J. and Waagepetersen, R. P.: Inference and Simulation for Spatial Point Processes. Chapman & Hall/CRC, Boca Raton (2004)
21. Pateiro-López, B.: Set estimation under convexity-type restrictions. Ph. D. Thesis. Universidad de Santiago de Compostela (2008)
22. Pateiro-López, B. and Rodríguez-Casal, A.: Generalizing the convex hull of a sample: The R package alphahull. J. Stat. Softw. **5**, 1–28 (2010)
23. R Development Core Team.: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna. `http://www.R-project.org` (2008)
24. Ripley, B. D.: Tests of 'randomness' for spatial point patterns. J. Roy. Statist. Soc. Ser. A **39**, 172–212 (1979)

25. Ripley, B. D. and Rasson, J. P.: Finding the Edge of a Poisson Forest. J. Appl. Probab. **14**, 483–491 (1977)
26. Rodríguez-Casal, A.: Set estimation under convexity-type assumptions. Ann. Inst. H. Poincare Probab. Statist. **43**, 763–774 (2007)
27. Smith, S. P. and Jain, A. K.: Testing for uniformity in multidimensional data. IEEE Trans. Pattern Anal. Machine Intell. **PAMI-6**, 73–81 (1984)
28. Tenreiro, C.: On the finite sample behavior of fixed bandwidth Bickel-Rosenblatt test for univariate and multivariate uniformity. Comm. Statist. Simulation Comput. **36**, 827–846 (2007)
29. Tibshirani, R., Guenther, W. and Hastie, T.: Estimating the number of clusters in a data set via the gap statistic. J. R. Statist. Soc. B **63**, 411–423 (2001)
30. Voronoi, G.: Nouvelles applications des paramètres continus à la théorie des formes quadratiques. J. Reine Angew. Math. **133**, 97–178 (1907)
31. Walther, G.: Granulometric smoothing. Ann. Statist. **25**, 2273–2299 (1997)
32. Walther, G.: On a generalization of Blaschke's Rolling Theorem and the smoothing of surfaces. Math. Methods Appl. Sci. **22**, 301–316 (1999)

Table 1 Empirical significance level over 10000 uniform samples of size $n = 100, 200, 500$ on different supports S. The uniformity tests considered for the case of known support are the maximal spacing test (MS), the (SJ) test by Smith and Jain (1984), the tests by Liang *et al.* (2001) and the one-sided DB test by Berrendero *et al.* (2006). The uniformity tests considered for the case of unknown support are the estimated maximal spacing test (EMS) and the (SJ) test by Smith and Jain (1984). The nominal value is 0.05.

			n		
			100	200	500
Support (a)	known	MS	0.0334	0.0449	0.0504
		SJ	0.0558	0.0581	0.0746
		A_n symmetric	0.0501	0.0578	0.0457
		centered	0.0519	0.0520	0.0438
		star	0.0510	0.0496	0.0505
		T_n symmetric	0.0604	0.0566	0.0527
		centered	0.0604	0.0556	0.0519
		star	0.0674	0.0636	0.0619
		DB o-s	0.0441	0.0491	0.0493
	unknown	EMS	0.0231	0.0388	0.0471
		SJ	0.0552	0.0591	0.0709
Support (b)	known	MS	0.0369	0.0449	0.0548
		SJ	0.0553	0.0592	0.0758
	unknown	EMS	0.0247	0.0356	0.0491
		SJ	0.0579	0.0557	0.0686
Support (c)	known	MS	0.0326	0.0385	0.0515
		SJ	0.0563	0.0528	0.0766
	unknown	EMS	0.0138	0.0281	0.0463
		SJ	0.0598	0.0523	0.0724
Support (d)	known	MS	0.0000	0.0177	0.0380
	unknown	EMS	0.0000	0.0062	0.0324

Table 2 Empirical powers of the uniformity tests under study over 5000 runs of sample size $n = 100$ from Neyman-Scott clustering alternatives. Each cluster consist of m points, generated from the uniform distribution on a disc of radius r.

			$r = 0.05$		$r = 0.1$	
			$m = 5$	$m = 10$	$m = 5$	$m = 10$
Support (a)	known	MS	0.9852	1.0000	0.8582	0.9930
		SJ	1.0000	1.0000	0.9006	0.9998
		A_n symmetric	0.4072	0.5974	0.3608	0.5466
		centered	0.4220	0.5954	0.3900	0.5634
		star	0.3478	0.5156	0.3398	0.4904
		T_n symmetric	0.8772	0.9954	0.8036	0.9692
		centered	0.8206	0.9842	0.7528	0.9448
		star	0.7614	0.9700	0.6906	0.9198
		DB o-s	0.3022	0.4536	0.2394	0.3416
	unknown	EMS	0.9804	0.9992	0.7978	0.9794
		SJ	0.9998	1.0000	0.8444	0.9936

Table 3 Empirical significance level over 10000 uniform samples of size $n = 100, 200, 500$ on different supports S. The uniformity test considered is the estimated maximal spacing test (EMS). The support estimator $C_\lambda(\aleph_n)$ is constructed for different values of λ. The nominal value is 0.05.

				n		
				100	200	500
Support (a)	unknown	EMS	$\lambda = 0.9$	0.0233	0.02381	0.0469
			$\lambda = 1$	0.0231	0.0388	0.0471
			$\lambda = 1.1$	0.0233	0.0392	0.0475
Support (b)	unknown	EMS	$\lambda = 0.5$	0.0233	0.0349	0.0491
			$\lambda = 0.56$	0.0247	0.0356	0.0491
			$\lambda = 0.6$	0.0245	0.0362	0.0491
Support (c)	unknown	EMS	$\lambda = 0.15$	0.0000	0.0265	0.0441
			$\lambda = 0.19$	0.0138	0.0281	0.0463
			$\lambda = 0.25$	0.0169	0.0297	0.0462

Table 4 Analysis of Lansing Woods data. For each species, p-values of the associated uniformity tests.

		p-value		
		Hickory	Maple	Oak
Known support	MS	< 0.001	< 0.001	0.350
	SJ	< 0.001	< 0.001	0.273
	A_n symmetric	0.004	< 0.001	0.018
	centered	< 0.001	0.774	0.004
	star	< 0.001	< 0.001	0.087
	T_n symmetric	< 0.001	< 0.001	< 0.001
	centered	< 0.001	< 0.001	< 0.001
	star	< 0.001	< 0.001	< 0.001
	DB o-s	< 0.001	< 0.001	0.002
Unknown support	EMS	< 0.001	< 0.001	0.313
	SJ	< 0.001	< 0.001	0.165