



**UNIVERSIDADE DE
SANTIAGO DE COMPOSTELA
DEPARTAMENTO DE
ESTADÍSTICA E INVESTIGACIÓN OPERATIVA**

**Multivariate uniformity tests: the distance to boundary method
for the case of unknown support**

J. R. Berrendero, A. Cuevas, B. Pateiro-López

Report 11-03

Reports in Statistics and Operations Research

Testing uniformity for the case of a planar unknown support

Abstract

A new test is proposed for the hypothesis of uniformity on bi-dimensional supports. The procedure is an adaptation of the “distance to boundary test” (DB test) proposed in Berrendero, Cuevas and Vázquez-Grande (2006). This new version of the DB test, called DBU test, allows us (as a novel, interesting feature) to deal with the case where the support S of the underlying distribution is unknown. This means that S is not specified in the null hypothesis so that, in fact, we test the null hypothesis that the underlying distribution is uniform on some support S belonging to a given class \mathcal{C} . We pay special attention to the case that \mathcal{C} is either the class of compact convex supports or the (broader) class of compact λ -convex supports. The basic idea is to apply the DB test in a sort of plug-in version, where the support S is approximated by using methods of set estimation. The DBU method is analyzed from both the theoretical and practical point of view, via some asymptotic results and a simulation study, respectively.

Keywords: convexity; distance to boundary; λ -convexity; set estimation; uniformity test.

AMS 2000 subject classifications: Primary 62G05; secondary 62G20.

1 Introduction

We are concerned with the problem of testing the null hypothesis

H_0 : the random variable X has a uniform distribution on some support S .

We assume throughout that the available information is given by an iid sample X_1, \dots, X_n drawn from the d -dimensional random variable X .

The vast majority of theoretical developments and applications for this problem deal with either the univariate case $d = 1$ or the bivariate models with $d = 2$. The motivations for both situations are quite different. While the univariate uniformity tests are often motivated by the need of having good “random number generators”, the bivariate uniformity problems arise usually in the setting of spatial statistics. Anyway, the bivariate problem is considerably harder in several senses. A first obvious difficulty for $d = 2$ is the lack of a distribution-free procedure (such as the univariate Kolmogorov-Smirnov test) based on the empirical distribution. Also, the choice of the support is not an issue in most univariate uniformity problems, as they are naturally set out in a known interval $S = [a, b]$ which can be reduced to the standard case $[a, b] = [0, 1]$; on the contrary, when we are dealing with bivariate data there is no good reason for restricting us to a fixed

support as, for example, $S = [0, 1]^2$. Of course, $S = [0, 1]^2$ is a relevant case, but there are many other conceivable interesting supports (such as polygons, ellipses, etc.) and one might even consider the case where S is not known in advance and only a generic regularity assumption on its structure is imposed. In other words, the class $\mathcal{U}(\mathbb{R}^2)$ of uniform distributions with connected support in \mathbb{R}^2 is much more complicated than its one-dimensional analog, $\mathcal{U}(\mathbb{R})$. The latter is a parametric family so that, even if the support S were unknown, its estimation is a simple matter based on standard methods. This is not at all the case with $\mathcal{U}(\mathbb{R}^2)$. Thus, it is clear that the goodness-of-fit problem to the non-parametric family $\mathcal{U}(\mathbb{R}^2)$ (or to appropriate sub-families of it) involves non-trivial geometric and statistical issues which lead us to the main point of this work.

The purpose of this paper

We specifically aim at developing a new uniformity test, based on an iid sample of size n , for the null hypothesis

$$H_0 : \text{the random variable } X \text{ has a uniform distribution belonging to the class } \mathcal{U}_{\mathcal{C}}, \quad (1)$$

where $\mathcal{U}_{\mathcal{C}}$ is the class of bivariate uniform distributions whose support S belongs to a given class \mathcal{C} of compact supports in \mathbb{R}^2 . As we will see, the natural assumption of connectedness for S can be incorporated to our approach but it is not strictly needed.

Our test will consist of an adaptation of the *Distance-to-Boundary Method* (DB) which was proposed by Berrendero, Cuevas and Vázquez-Grande (2006) for the simplest, usual case that the support S is completely known and specified in the null hypothesis; in the notation (1), this would amount to take a class $\mathcal{C} = \{S\}$ with a unique member. The DB-method was based on calculating the distances Y_i from the sampling observations X_i to the boundary of the support S . The test checks the fit of the empirical distribution of this variables to that corresponding to the case where H_0 is true (a more detailed account will be given below). The purpose of this paper is to show that this method can be adapted to the case where the support S is unknown so that we deal in fact with a general problem of type (1). Our extension of the DB procedure, which we will denote DBU test, relies on methods of set estimation (see Cuevas and Fraiman (2009) for a survey of this topic). The basic idea is a sort of plug-in device: we apply the DB test presented in Berrendero, Cuevas and Vázquez-Grande (2006) replacing the support S by a suitable estimator S_n . If the estimated boundary ∂S_n approaches fast enough to the population counterpart ∂S , the respective critical regions in both tests (with the tests statistics calculated from S and S_n , respectively) will be asymptotically equivalent.

There are many possible different choices for $\mathcal{U}_{\mathcal{C}}$ in (1). We will pay especial attention to the cases where \mathcal{C} is either the class of compact convex supports or the class of compact λ -convex supports. The notion of λ -convexity arises as a natural generalization of convexity, so every convex set is also λ -convex for all $\lambda > 0$. In short a set is λ -convex if it can be expressed as the intersection of the complements of a family of open balls with radii λ ; see Perkal (1956), Walther (1997, 1999), Cuevas, Fraiman and Pateiro-López (2011) and references therein.

Some related literature

In the recent paper by Berrendero, Cuevas and Pateiro-López (2011) a further uniformity test is proposed for the problem (1), when \mathcal{C} is also the class of compact supports which are either convex or λ -convex. However the idea behind this test is completely different from that developed here as it is based on the size of the estimated maximal bivariate spacing (so we call it EMS test), as defined in Janson (1987).

As we will see in the simulations below, the EMS procedure is, in some sense, complementary of the DBU test. While the former is particularly suitable for alternative hypothesis of Neyman-Scott type, e.g., for departures from uniformity which lead to “clustered observations”, the DBU test turns out to be more powerful for “contamination models”, prone to give more observations close to (or far away from) ∂S than expected under uniformity.

In the work by Jain et al. (2002) it is analyzed (especially from the practical and computational point of view) another method for the uniformity testing problem with unknown support. It is based on ideas of graph theory.

Let us finally mention the interesting proposal by Liang et al. (2001). These authors propose a uniformity test, easy to implement even for very large dimensional data. However their method is designed for the specific case that $S = [0, 1]^d$.

This paper is organized as follows. In Section 2, the basic ideas of the DB test proposed in Berrendero, Cuevas and Vázquez-Grande (2006) are summarized. Then the corresponding DBU test (for the case of unknown support) is defined. Also, some notions on λ -convex sets and their estimation are recalled. In Section 3 we show that the test-statistic D_n of the DB-method and its counterpart D_n^* in the new DBU procedure satisfy $|D_n - D_n^*| \rightarrow 0$, in probability, so that both tests are asymptotically equivalent regarding their properties of consistency and asymptotic preservation of the significance level. Section 4 is devoted to empirical results. Some geometric and computational issues are discussed in Section 5.

2 The DBU test

Let $S \subset \mathbb{R}^2$ be a compact set. Let us also consider a two-dimensional random variable X taking values in S and denote by $\mathcal{X}_n = \{X_1, \dots, X_n\}$ a sample drawn from X . As a first step in the development of our DBU test we briefly describe below the implementation of the original distance-to-boundary test with known support (DB test) proposed by Berrendero, Cuevas and Vázquez-Grande (2006).

The DB test: The support S is known

The target is to test the null hypothesis

$$H_0 : \text{the distribution of } X \text{ is uniform with support } S.$$

Some notation: $D(x, y)$ denotes the Euclidean distance between points x and y ; for $A \subset \mathbb{R}^2$, $D(x, A) = \inf_{y \in A} D(x, y)$. The distribution function of the random variable $Y = D(X, \partial S)$ under H_0 will be denoted by F and \mathbb{F}_n is the empirical distribution function corresponding to Y_1, \dots, Y_n , where $Y_i = D(X_i, \partial S)$. The usual Kolmogorov-Smirnov statistic is denoted by D_n , so that $D_n = \sqrt{n} \|F - \mathbb{F}_n\|$, where $\|\cdot\|$ stands for the sup-norm. The closed and open balls with center y and radius r will be denoted respectively by $B(y, r)$ and $\overset{\circ}{B}(y, r)$.

The implementation of the DB test can be summarized as follows.

1. Given the original sample X_1, \dots, X_n , compute the distances to the boundary $Y_i = D(X_i, \partial S)$, $i = 1, \dots, n$.
2. Compute the “maximum depth” $R = \max\{D(x, \partial S), x \in S\}$ and define the “normalized distances” $Y_i^R = Y_i/R$, for $i = 1, \dots, n$.
3. If the set S is “invariant by erosion upon an homothecy” (see Berrendero, Cuevas and Vázquez-Grande (2006) for details and Pegden (2011) for closely related ideas) it can be proved that the distribution function F^R of the Y_i^R , under H_0 , is beta with parameters $a = 1$ and $b = 2$ (regardless of the support S). Then the DB test would reject H_0 , at a level α , if the Kolmogorov-Smirnov statistic based on the normalized distances $D_n^R = \sqrt{n} \|\mathbb{F}_n^R - F^R\|$ is greater than the corresponding critical value $D_{n, \alpha}$.
4. Otherwise (i.e., when S does not fulfill the mentioned shape assumption), the distribution of the Y_i^R will depend, in general, on S . So the normalization by R indicated in the second step above is not particularly useful. In this case the test is performed, as indicated in the previous step, using in the Kolmogorov-Smirnov statistic D_n calculated from the (non-normalized) distances Y_i . If the distribution under H_0 of the Y_i is difficult to calculate in a closed form it can be approximated by a Monte Carlo procedure by just drawing a large number of artificial iid observations \hat{X}_i , $i = 1, \dots, m$ from the uniform distribution on S and taking the corresponding empirical distribution associated with $\hat{Y}_i = D(\hat{X}_i, \partial S)$ as an approximation to F .

We next present the adaptation of the DB method for the case that the support S is not specified in the null hypothesis. So, we will deal with the general problem (1) stated in the introduction. As commented above, the crucial idea is to replace the support S with an appropriate estimator $S_n = S_n(X_1, \dots, X_n)$. There are all purpose set estimators which provide consistency properties (and even known convergence rates) under very general conditions on S ; see Cuevas and Fraiman (2009) for details. However, given the special role of ∂S in the DB test, it is important for the plug-in estimator ∂S_n to approximate the population counterpart at a fast enough rate. This will lead us to impose some restriction on the class \mathcal{C} of possible supports. We will further comment on this below. Now, let us formally state the implementation of the DBU test.

DBU test: The support S is unknown

1. Construct S_n , an estimator of S based on the sample \mathcal{X}_n .
2. Define $\mathcal{X}_n^* = \{X_1^*, \dots, X_n^*\} = \{X_i \in \mathcal{X}_n, X_i \notin \partial S_n, i = 1, \dots, n\}$.
3. Compute $Y_i^* = D(X_i^*, \partial S_n), i = 1, \dots, n^*$.
4. Generate an artificial sample $\hat{\mathcal{X}}_m = \{\hat{X}_1, \dots, \hat{X}_m\}$, from a uniform distribution on S_n .
5. Compute $\hat{Y}_i = D(\hat{X}_i, \partial S_n), i = 1, \dots, m$.
6. Perform a two-sample Kolmogorov-Smirnov test of the null hypothesis that Y_i^* and \hat{Y}_i were drawn from the same continuous distribution. Thus, the test is based on the statistic $D_n^* = \sqrt{n} \|\mathbb{F}_n^* - \hat{F}\|$, where \mathbb{F}_n^* is the empirical distribution of $Y_i^* = D(X_i^*, \partial S_n)$ and \hat{F} is the distribution function of \hat{Y}_i . Since this distribution under H_0 is difficult to calculate it is approximated by the corresponding empirical distribution.

The choice of the support estimator S_n

Keeping in mind that S_n must provide an efficient, easy-to-compute estimator for both S and (via ∂S_n) for ∂S , a natural choice for the class \mathcal{C} in (1) would be

$$\mathcal{C} = \{\text{class of compact convex supports in } \mathbb{R}^2\}.$$

In this case the natural estimator of S is the convex hull of \mathcal{X}_n ,

$$S_n = \text{conv}(\mathcal{X}_n). \quad (2)$$

The properties of this estimator have been extensively analyzed since the early sixties; see Reitzner (2009). We will need here the consistency properties established by Dümbgen and Walther (1996). In particular, these authors show that (for the two-dimensional case $d = 2$), with probability 1 (a.s.),

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

where $d_H(A, B)$ stands for the Hausdorff distance between two compact non-empty sets A and B .

As we will see, this convergence rate is not fast enough for our purposes. Under additional smoothness assumptions on S (see Walther (1997, 1999), Rodríguez-Casal (2007)) we have

$$d_H(S_n, S) = O\left(\left(\frac{\log n}{n}\right)^{2/3}\right), \text{ a.s.} \quad (3)$$

and, more importantly,

$$d_H(\partial S_n, \partial S) = O\left(\left(\frac{\log n}{n}\right)^{2/3}\right), \text{ a.s.} \quad (4)$$

Whereas convexity is a simple, natural and well-studied assumption to be imposed on S , it is quite restrictive for many practical purposes. For example, when analyzing spatial patterns in ecological data, it is not always reasonable to assume that the habitat of a certain plant species is a convex domain. Hence we will also consider a second (much less popular) condition called λ -convexity which allows for a much more flexible class of possible supports. For another recent application of this condition to the problem of testing uniformity see Berrendero, Cuevas and Pateiro-López (2011).

A closed set $S \subset \mathbb{R}^2$ is said to be λ -convex for some $\lambda > 0$ if S coincides with its λ -convex hull, that is $S = C_\lambda(S)$, where

$$C_\lambda(S) = \bigcap_{\mathring{B}(y, \lambda) \cap S = \emptyset} \mathring{B}(y, \lambda)^c. \quad (5)$$

The origin of this notion goes back to Perkal (1956). See Walther (1997), Rodríguez-Casal (2007), Berrendero, Cuevas and Pateiro-López (2011), and references therein, for additional insights as well as statistical applications.

The condition of λ -convexity is clearly reminiscent of the plain notion of convexity, as it can be seen by replacing the balls in (5) by half-spaces. In fact, every closed convex set is also λ -convex for all $\lambda > 0$. It is also apparent that λ -convexity is a much more flexible condition which allows the set to have inlands (as long as they are not too sharp) or holes and even to be disconnected.

From a statistical point of view, the most important feature of definition (5) is the fact that a λ -convex support S has a natural estimator from a random sample \mathcal{X}_n which is the λ -convex hull of the sample points,

$$S_n = C_\lambda(\mathcal{X}_n). \quad (6)$$

This estimator turns out to be computationally feasible; the R-package `alphahull` developed by Pateiro-López and Rodríguez-Casal (2010) provides an efficient calculation of (6) in the two-dimensional case. Moreover, under appropriate smoothness conditions, this estimator exhibits also the fast convergence rates (3) and (4). This will be important in the theoretical developments of the following section.

3 Asymptotic properties

The aim of this section is to show that, under suitable shape restrictions on the class \mathcal{C} in (1), the DBU test is asymptotically equivalent to the DB test proposed in Berrendero, Cuevas and Vázquez-Grande (2006) for the case of a known support. According to the notation introduced in Section 2, this amounts to show that $|D_n - D_n^*| \rightarrow 0$, in probability

as $n \rightarrow \infty$. We will establish this in two results (Theorems 1 and 2 below), obtained under two different assumptions for \mathcal{C} .

In what follows, $\mu(A)$ will stand for the Lebesgue measure of a set A and the cardinal of the set $\{i : X_i \in \partial S_n\}$ will be denoted by N_n (that is, $N_n = n - n^*$). All the convergence results below correspond to limits as $n \rightarrow \infty$.

First, we will establish two lemmas giving the conditions under which \hat{F} is close enough to F , and \mathbb{F}_n^* is close enough to \mathbb{F}_n , respectively.

Lemma 1. *Assume that the set S and the estimator S_n are such that F is Lipschitz continuous, $S_n \subset S$ with probability one, $\sqrt{n} \mu(S \setminus S_n) \xrightarrow{P} 0$ and $\sqrt{n} d_H(\partial S, \partial S_n) \xrightarrow{P} 0$. Then, $\sqrt{n} \|\hat{F} - F\| \xrightarrow{P} 0$.*

PROOF: Since both F and \hat{F} have compact support, there exists $K > 0$ (not depending on n) such that $\|\hat{F} - F\| = \sup_{t \in [0, K]} |\hat{F}(t) - F(t)|$. Let B be the closed unit ball in \mathbb{R}^2 and denote by $C \ominus D = \{x : x + D \subset C\}$ the Minkowski difference of two sets C and D . Observe that $Y \geq t$ if and only if $X \in S \ominus tB$, and $\hat{Y} \geq t$ if and only if $\hat{X} \in S_n \ominus tB$. Then,

$$|\hat{F}(t) - F(t)| = |\mathbb{P}(\hat{X} \in S_n \ominus tB) - \mathbb{P}(X \in S \ominus tB)|,$$

and, using the triangle inequality,

$$\begin{aligned} |\hat{F}(t) - F(t)| &\leq |\mathbb{P}(\hat{X} \in S_n \ominus tB) - \mathbb{P}(X \in S_n \ominus tB)| \\ &\quad + |\mathbb{P}(X \in S_n \ominus tB) - \mathbb{P}(X \in S \ominus tB)|. \end{aligned} \quad (7)$$

Regarding the first term in the right-hand side of inequality (7), observe that, for all $t \in [0, K]$,

$$\begin{aligned} |\mathbb{P}(\hat{X} \in S_n \ominus tB) - \mathbb{P}(X \in S_n \ominus tB)| &= \frac{\mu(S_n \ominus tB)}{\mu(S_n)} - \frac{\mu(S_n \ominus tB)}{\mu(S)} \\ &\leq \frac{\mu(S_n \ominus tB)}{\mu(S_n)} \left(1 - \frac{\mu(S_n)}{\mu(S)}\right) \leq 1 - \frac{\mu(S_n)}{\mu(S)} = \frac{\mu(S \setminus S_n)}{\mu(S)}. \end{aligned}$$

Since, by assumption, $\sqrt{n} \mu(S \setminus S_n) \xrightarrow{P} 0$, we also have

$$\sqrt{n} \sup_{t \in [0, K]} |\mathbb{P}(\hat{X} \in S_n \ominus tB) - \mathbb{P}(X \in S_n \ominus tB)| \xrightarrow{P} 0.$$

For the second term in the right-hand side of inequality (7), observe that, for all $t \in [0, K]$,

$$\mathbb{P}(X \in (S \ominus tB) \setminus (S_n \ominus tB)) \leq \mathbb{P}(Y \geq t, D(X, \partial S_n) < t) + \frac{\mu(S \setminus S_n)}{\mu(S)},$$

since $X \in S \ominus tB$ amounts to $Y \geq t$, and $X \notin S_n \ominus tB$ implies that $D(X, \partial S_n) < t$ or $X \in S \setminus S_n$.

Also, $D(X, \partial S_n) < t$ implies $Y < t + \epsilon_n$, where $\epsilon_n = d_H(\partial S, \partial S_n)$. Indeed, since $D(X, \partial S_n) < t$, there exists $z_n \in \partial S_n$ such that $D(X, z_n) < t$. By definition of Hausdorff distance, there exists $z \in \partial S$ with $D(z, z_n) \leq \epsilon_n$. Hence,

$$Y = D(X, \partial S) \leq D(X, z) \leq D(X, z_n) + D(z_n, z) < t + \epsilon_n.$$

As a consequence,

$$\mathbb{P}(X \in (S \ominus tB) \setminus (S_n \ominus tB)) \leq \mathbb{P}(t \leq Y < t + \epsilon_n) + \frac{\mu(S \setminus S_n)}{\mu(S)}. \quad (8)$$

Since F is Lipschitz continuous, there exists $M > 0$ such that $\mathbb{P}(t \leq Y < t + \epsilon_n) = F(t + \epsilon_n) - F(t) \leq M\epsilon_n$. From this bound, (8) and taking into account the assumptions we deduce

$$\sqrt{n} \sup_{t \in [0, K]} \mathbb{P}(X \in (S \ominus tB) \setminus (S_n \ominus tB)) \leq M\sqrt{n}\epsilon_n + \frac{\sqrt{n}\mu(S \setminus S_n)}{\mu(S)} \xrightarrow{P} 0.$$

□

Lemma 2. *Assume that the set S and the estimator S_n are such that F is Lipschitz continuous, $S_n \subset S$ with probability one, $N_n/\sqrt{n} \xrightarrow{P} 0$ and $n^{1/2+\delta} d_H(\partial S, \partial S_n) \xrightarrow{a.s.} 0$, for some $\delta > 0$. Then, $\sqrt{n}\|\mathbb{F}_n^* - \mathbb{F}_n\| \xrightarrow{P} 0$.*

PROOF: For $i = 1, \dots, n$ define $\tilde{Y}_i = D(X_i, \partial S_n)$ and let $\tilde{\mathbb{F}}_n$ be the empirical distribution function corresponding to $\tilde{Y}_1, \dots, \tilde{Y}_n$. Since

$$\sqrt{n}\|\mathbb{F}_n^* - \mathbb{F}_n\| \leq \sqrt{n}\|\mathbb{F}_n^* - \tilde{\mathbb{F}}_n\| + \sqrt{n}\|\tilde{\mathbb{F}}_n - \mathbb{F}_n\|, \quad (9)$$

it is enough to prove that both terms in the right-hand side of the last inequality go to zero in probability. Since there exists $K > 0$ such that all the involved distributions have supports included in $[0, K]$, the sup-norms can always be computed on a compact interval $[0, K]$ instead of \mathbb{R} .

Observe that, for $t \geq 0$, $\tilde{\mathbb{F}}_n(t) = (1 - N_n/n)\mathbb{F}_n^*(t) + N_n/n$. Therefore,

$$\sqrt{n}\|\mathbb{F}_n^* - \tilde{\mathbb{F}}_n\| = \frac{N_n}{\sqrt{n}} \sup_{t \in [0, K]} (1 - \mathbb{F}_n^*(t)) \leq \frac{N_n}{\sqrt{n}} \xrightarrow{P} 0,$$

by assumption.

Regarding the second term of the right-hand side of (9), notice that $\tilde{Y}_i \leq Y_i \leq \tilde{Y}_i + \epsilon_n$, where $\epsilon_n = d_H(\partial S, \partial S_n)$. Then,

$$\sqrt{n}\|\tilde{\mathbb{F}}_n - \mathbb{F}_n\| \leq \sup_{t \in [0, K]} \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbb{I}_{\{t < Y_i \leq t + \epsilon_n\}} \quad (10)$$

Define the sequence $b_n = n^{-1/2-\delta}$, where $\delta > 0$ is given in the assumptions of the lemma. Notice that, from the assumption on $d_H(\partial S, \partial S_n)$, we have $b_n > \epsilon_n$ eventually with probability 1. Then,

$$\sup_{t \in [0, K]} \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbb{I}_{\{t < Y_i \leq t + \epsilon_n\}} \leq \sup_{t \in [0, K]} \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbb{I}_{\{t < Y_i \leq t + b_n\}}, \text{ eventually with probability 1.} \quad (11)$$

Now, denote by C_n , for each n , the minimal covering of $(0, K]$ by intervals of the form $I_j^n = (jb_n, (j+1)b_n]$, $j = 1, 2, \dots$. Clearly, the cardinality of C_n is $O(b_n^{-1}) = O(n^\gamma)$ with $\gamma = 1/2 + \delta$. Also, since F , the distribution of the Y_i 's, is Lipschitz continuous, there exists M such that

$$\max_{I \in C_n} \mathbb{P}_F(I) \leq Mb_n = o(n^{-1/2}).$$

Therefore, the sequence of coverings C_n fulfills the assumptions in Lemma 2.2 of Fernholz (1991). It follows that $T_n/\sqrt{n} \xrightarrow{\text{a.s.}} 0$, where T_n is the maximum number of Y_i 's with values in any $I \in C_n$. Then,

$$\sup_{t \in [0, K]} \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbb{I}_{\{t < Y_i \leq t + b_n\}} \leq \frac{2T_n}{\sqrt{n}} \xrightarrow{\text{a.s.}} 0. \quad (12)$$

From (10), (11) and (12) we get $\sqrt{n} \|\tilde{\mathbb{F}}_n - \mathbb{F}_n\| \xrightarrow{\text{a.s.}} 0$. \square

The following result is a straightforward consequence of the two previous lemmas.

Lemma 3. *Assume that the set S and the estimator S_n are such that F is Lipschitz continuous, $S_n \subset S$ with probability one, $\sqrt{n} \mu(S \setminus S_n) \xrightarrow{P} 0$, $N_n/\sqrt{n} \xrightarrow{P} 0$, and $n^{1/2+\delta} d_H(\partial S, \partial S_n) \xrightarrow{\text{a.s.}} 0$, for some $\delta > 0$. Then, $|D_n - D_n^*| \xrightarrow{P} 0$.*

PROOF: Applying the triangle inequality,

$$\|\mathbb{F}_n^* - \hat{F}\| \leq \|\mathbb{F}_n^* - \mathbb{F}_n\| + \|\mathbb{F}_n - F\| + \|F - \hat{F}\|,$$

and

$$\|\mathbb{F}_n - F\| \leq \|\mathbb{F}_n - \mathbb{F}_n^*\| + \|\mathbb{F}_n^* - \hat{F}\| + \|\hat{F} - F\|.$$

Hence,

$$D_n - \sqrt{n} \|\mathbb{F}_n^* - \mathbb{F}_n\| - \sqrt{n} \|F - \hat{F}\| \leq D_n^* \leq D_n + \sqrt{n} \|\mathbb{F}_n^* - \mathbb{F}_n\| + \sqrt{n} \|F - \hat{F}\|$$

and the result follows from Lemmas 1 and 2. \square

Now we apply Lemma 3 to the cases when we can assume that S is convex and λ -convex respectively. We will also need the following smoothness condition: A ball of radius r is said to *roll freely inside* a closed set $A \subset \mathbb{R}^d$ if for each point $a \in \partial A$ there exists $x \in \mathbb{R}^d$ such that $a \in B(x, r) \subset A$.

Theorem 1. *Let $S \subset \mathbb{R}^2$ be a compact convex set with nonempty interior such that F is Lipschitz continuous and such that a ball of radius $r > 0$ rolls freely inside S for some $r > 0$. Let S_n be the convex hull of \mathcal{X}_n . It holds $|D_n - D_n^*| \xrightarrow{P} 0$.*

PROOF: We are going to check the assumptions of Lemma 3. By Theorem 3 and Remark 3 in Rodríguez-Casal (2007), it holds

$$d_H(\partial S, \partial S_n) = O\left(\frac{\log n}{n}\right)^{2/3}$$

with probability 1. Observe that, if S is convex and $S_n, C_\lambda(\mathcal{X}_n)$ stand for the convex hull and the λ -convex hull, respectively, of \mathcal{X}_n , then $C_\lambda(\mathcal{X}_n) \subset S_n \subset S$ for any $\lambda > 0$. Thus, $n^{1/2+\delta} d_H(\partial S, \partial S_n) \xrightarrow{\text{a.s.}} 0$, for $0 < \delta < 1/6$.

Theorem 1 in Schütt (1994) ensures that for any convex body $S \subset \mathbb{R}^d$, $\mathbb{E}[\mu(S \setminus S_n)] = O(n^{-2/(d+1)})$. In particular, for $d = 2$ and using Markov inequality we have $n^\beta \mu(S \setminus S_n) \xrightarrow{P} 0$, for $0 \leq \beta < 2/3$.

Finally, according to a well-known result due to Efron (1965) and using again Schütt's Theorem,

$$\mathbb{E}(N_n) = n \frac{\mathbb{E}[\mu(S \setminus S_{n-1})]}{\mu(S)} = O(n^{(d-1)/(d+1)}).$$

In particular, for $d = 2$ and using Markov's inequality we have $N_n/n^\beta \xrightarrow{P} 0$, for $\beta > 1/3$. \square

Theorem 2. *Let $S \subset \mathbb{R}^2$ be a compact λ -convex set with nonempty interior such that $\overline{S^c}$ is also λ -convex and $\text{int}(S_i) \neq \emptyset$ for each path-connected component $S_i \subset S$. Assume that F is Lipschitz continuous. Let S_n be the λ -convex hull of \mathcal{X}_n . It holds $|D_n - D_n^*| \xrightarrow{P} 0$.*

PROOF: We are going to check the assumptions of Lemma 3. By Theorem 3 in Rodríguez-Casal (2007), with probability 1

$$d_H(\partial S, \partial S_n) = O\left(\frac{\log n}{n}\right)^{2/3}$$

and the same rate holds for $\mu(S \setminus S_n)$. Then, $n^\beta \mu(S \setminus S_n) \xrightarrow{\text{a.s.}} 0$, for $0 \leq \beta < 2/3$, and $n^{1/2+\delta} d_H(\partial S, \partial S_n) \xrightarrow{\text{a.s.}} 0$, for $0 < \delta < 1/6$. Finally, by Theorem 3 in Pateiro-López and Rodríguez-Casal (2011), we have $\mathbb{E}(N_n) = O(n^{1/3})$ and using Markov inequality we obtain $N_n/n^\beta \xrightarrow{P} 0$ for $\beta > 1/3$. \square

As a consequence of these results the DBU test inherits the properties of the DB test studied in Berrendero, Cuevas and Vázquez-Grande (2006), in particular, it asymptotically preserves the prescribed significance level and it is consistent to detect non-uniform alternatives provided that they do not include “crown-uniform” distributions.

4 Empirical results

4.1 Empirical significance level

We have checked the performance of the DBU test in terms of preservation of the nominal confidence level. The numerical results given below have been obtained using the R

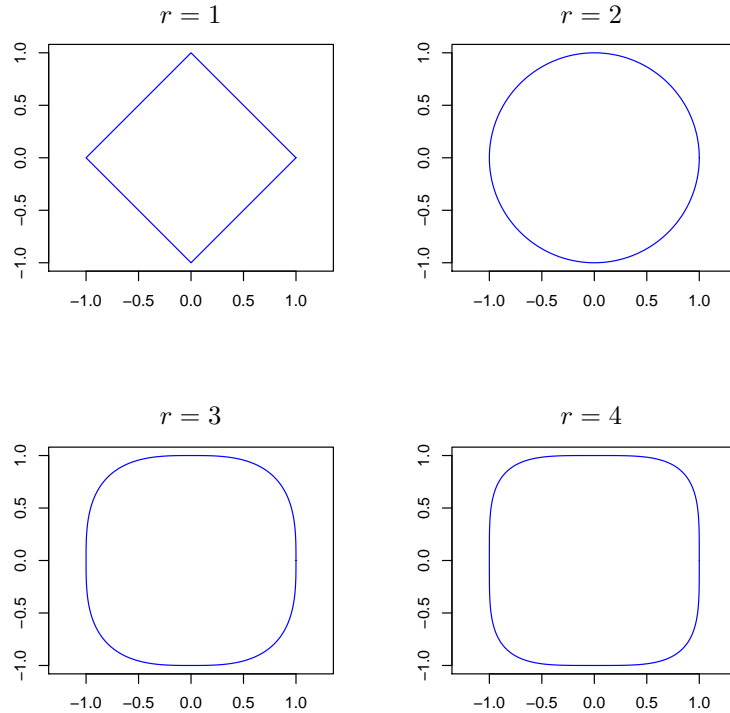


Figure 1: Lamé curves $|x|^r + |y|^r = 1$ and different values of r .

software, see R Development Core Team (2008).

A simulation example: the “unknown” support S is a set limited by a Lamé curve. The possible supports in the null hypothesis are either convex or λ -convex

Table 1 gives the outputs corresponding to the empirical significance level obtained (as an average over 5000 independent runs) with the DBU test and the DB test intended for nominal significance levels $\alpha = 0.05, 0.1$. Sample sizes are $n = 50, 100, 200$. The considered supports are sets limited by different Lamé curves (also called superellipses), that is, sets of the form $S = \{(x, y) \in \mathbb{R}^2 : |x|^r + |y|^r \leq 1\}$ for different values of r , see Figure 1. Note that for $r = 1$ and $r = 2$ the equation of the Lamé curve describes a square and a circle, respectively. We refer to Jaklič, Leonardis and Solina (2000) for further discussion of superellipses and their properties.

The supports limited by these curves for $r = 1$ and $r = 2$ are invariant by erosion upon an homothecy (the sets fulfilling this condition are called “resilient to erosion” in Pegden (2011)). Thus we are under the assumptions of Theorem 1 in Berrendero, Cuevas and Vázquez-Grande (2006) for the DB test, so that the distribution of $Y^R = D(X, \partial S)/R$ under the null hypothesis is totally known (it is a beta distribution $\beta(1, 2)$) and we

may perform a classical one-sample Kolmogorov-Smirnov test of goodness of fit to that distribution.

For other values of r , the set S does not fulfill the mentioned shape restriction and the distribution of Y is derived in practice by a Monte Carlo mechanism; see the description of the implementation of the test in Section 2. Moreover, the non-normalized distances Y_i in the DB test are approximated numerically, since there is not a close form solution for the distance to the Lamé curve when $r = 3$ or $r = 4$, see Rosin and West (1995). For the DBU test we use as estimator S_n both the convex hull of the sample $\mathcal{H}(\mathcal{X}_n)$ and the λ -convex hull of the sample $C_\lambda(\mathcal{X}_n)$ (with $\lambda = 1$). This corresponds to take \mathcal{C} in the the null hypothesis (1) to be the class of compact convex sets or the class of compact λ -convex sets, respectively.

Table 1: Empirical significance level of the DBU test and DB test over 5000 uniform samples of size $n = 50, 100, 200$ on the supports $S = \{(x, y) \in \mathbb{R}^2 : |x|^r + |y|^r \leq 1\}$ for different values of r . The nominal values are 0.05, 0.1. For the DBU test, we consider $S_n = \mathcal{H}(\mathcal{X}_n)$ and $S_n = C_\lambda(\mathcal{X}_n)$ with $\lambda = 1$.

		DBU test $S_n = \mathcal{H}(\mathcal{X}_n)$		DBU test $S_n = C_\lambda(\mathcal{X}_n)$		DB test		
		α	0.05	0.1	0.05	0.1	0.05	0.1
$r = 1$	$n = 50$		0.0436	0.0868	0.0422	0.0860	0.0460	0.0896
	$n = 100$		0.0414	0.0888	0.0406	0.0858	0.0408	0.0834
	$n = 200$		0.0492	0.0962	0.0476	0.0960	0.0418	0.0864
$r = 2$	$n = 50$		0.0474	0.0906	0.0400	0.0866	0.0472	0.0940
	$n = 100$		0.0416	0.0828	0.0452	0.0920	0.0450	0.0934
	$n = 200$		0.0510	0.0934	0.0490	0.0966	0.0522	0.1018
$r = 3$	$n = 50$		0.0502	0.0974	0.0482	0.0932	0.0500	0.0954
	$n = 100$		0.0468	0.0902	0.0442	0.0902	0.0472	0.0938
	$n = 200$		0.0480	0.0998	0.0510	0.1016	0.0444	0.0890
$r = 4$	$n = 50$		0.0416	0.0820	0.0376	0.0790	0.0414	0.0834
	$n = 100$		0.0448	0.0920	0.0432	0.0880	0.0428	0.0846
	$n = 200$		0.0470	0.0946	0.0468	0.0908	0.0344	0.0718

A case with non-connected support

Let S be the set in Figure 2, which is not convex but λ -convex for $\lambda = 2$. Table 2 gives the outputs corresponding to the empirical significance level obtained (as an average

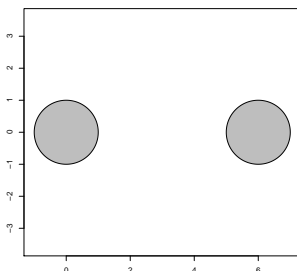


Figure 2: Non-convex support $S = B(x, 1) \cup B(y, 1)$, with $x = (0, 0)$ and $y = (6, 0)$. The set S is not convex but λ -convex for $\lambda = 2$.

over 10000 independent runs) with the DBU test and the DB test intended for nominal significance levels $\alpha = 0.05, 0.1$.

Table 2: Empirical significance level of the DBU test and DB test over 10000 uniform samples of size $n = 50, 100, 200, 500$ on S in Figure 2. The nominal values are 0.05, 0.1. For the DBU test, we consider $S_n = C_\lambda(\mathcal{X}_n)$ with $\lambda = 2$.

	DBU test		DB test		
	$S_n = C_\lambda(\mathcal{X}_n)$				
	α	0.05	0.1	0.05	0.1
$n = 50$		0.0433	0.0860	0.0480	0.0981
$n = 100$		0.0472	0.0933	0.0437	0.0922
$n = 200$		0.0439	0.0934	0.0417	0.0889
$n = 500$		0.0466	0.0967	0.0495	0.0948

Some results in \mathbb{R}^3

We have also studied the behavior in terms of significance level of the DB test and DBU test in \mathbb{R}^3 . The algorithms are essentially the same as those described in Section 2. Table 3 gives the outputs corresponding to the empirical significance level obtained (as an average over 10000 independent runs) with the DBU test and the DB test intended for nominal significance levels $\alpha = 0.05, 0.1$. Sample sizes are $n = 50, 100, 200, 500$. The considered supports are the unit cube $S = [0, 1]^3$ and unit ball $S = B(0, 1)$ in \mathbb{R}^3 . Since both supports are invariant by erosion upon an homothecy we perform for the DB test a classical one-sample Kolmogorov-Smirnov test of the null that the distribution function

of the random variable $Y^R = D(X, \partial S)/R$ is a beta distribution with parameters $a = 1$ and $b = 3$. For the DBU test, we restrict ourselves to the case where the support S is assumed to be convex and is estimated through the convex hull of the sample $\mathcal{H}(\mathcal{X}_n)$.

Table 3: Empirical significance level of the DBU test and DB test over 10000 uniform samples of size $n = 50, 100, 200, 500$ on $S = [0, 1]^3$ and $S = B(0, 1)$ in \mathbb{R}^3 . The nominal values are 0.05, 0.1. For the DBU test, we consider $S_n = \mathcal{H}(\mathcal{X}_n)$.

		DBU test $S_n = \mathcal{H}(\mathcal{X}_n)$			DB test	
		α	0.05	0.1	0.05	0.1
$S = [0, 1]^3$	$n = 50$		0.0399	0.0831	0.0483	0.0969
	$n = 100$		0.0434	0.0886	0.0423	0.0872
	$n = 200$		0.0468	0.0916	0.0487	0.0950
	$n = 500$		0.0470	0.0938	0.0483	0.0956
$S = B(0, 1)$	$n = 50$		0.0378	0.0798	0.0480	0.0972
	$n = 100$		0.0453	0.0886	0.0488	0.0968
	$n = 200$		0.0449	0.0909	0.0433	0.0872
	$n = 500$		0.0510	0.0970	0.0507	0.0948

4.2 Power study

As for the power study we have considered two different models in the choice of the alternative distribution.

Contamination model

The sample points are drawn from a random variable whose distribution is given by a mixture of type $(1 - \epsilon)U(S) + \epsilon U(S \setminus S_0)$, where $S = \{(x, y) \in \mathbb{R}^2 : |x|^r + |y|^r \leq 1\}$ for $r = 3$ and S_0 denotes a set like S with the same centre and area $\mu(S)/2$. We have taken $\epsilon = 0.1, 0.2, 0.3$, see Figure 3. We have compared the performance of the DBU test to that of the EMS test (based on multivariate spacings) by Berrendero, Cuevas and Pateiro-López (2011). The corresponding outputs are summarized in Table 4.

Neyman-Scott clustering alternatives

This is a typical deviation from the uniformity assumption, often considered in the theory of point processes. Under this model the sample tends to provide “clustered”

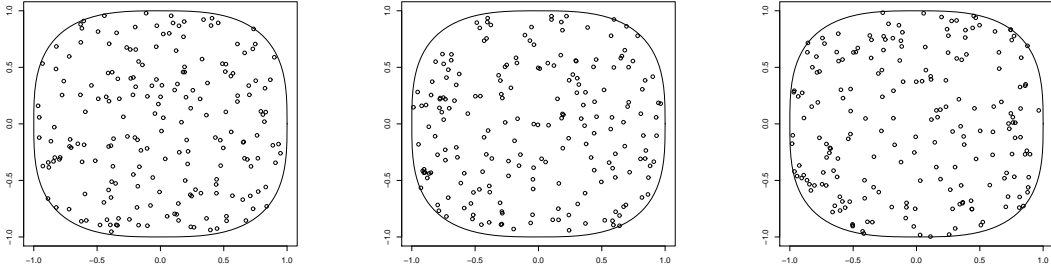


Figure 3: Random samples of size $n = 200$ from mixtures of type $(1 - \epsilon)U(S) + \epsilon U(S \setminus S_0)$, where $S = \{(x, y) \in \mathbb{R}^2 : |x|^r + |y|^r \leq 1\}$ for $r = 3$ and S_0 denotes a set like S with the same centre and area $\mu(S)/2$. Left, $\epsilon = 0.1$. Center, $\epsilon = 0.2$. Right, $\epsilon = 0.3$.

observations. For the simulated samples each cluster consist of m points, generated from the uniform distribution on a disc of radius r . The corresponding outputs are summarized in Table 5.

The support estimator used in the second column of Table 4 and in Table 5 is $S_n = C_\lambda(\mathcal{X}_n)$ with $\lambda = 1$.

4.3 Conclusions

1. The results in Tables 1 and 2 show that the DBU test succeeds in preserving the significance level (though it tends to be slightly conservative). The cost of estimating the support (pointed out by the difference observed with the DB test, where the support is known) turns out to be moderate and quite affordable in statistical terms.
2. Note that the asymptotic validity of the DBU test in the tri-dimensional case is not covered by our theoretical results in Section 3 (which apply only for $d = 2$). However, the outputs in Table 3 suggest that the method could work even in this case. A new, quite different, theoretical approach would be needed in this case, as the arguments in Section 3 rely essentially on the assumption $d = 2$.
3. The power results in Tables 4 and 5 show also a foreseeable behavior: the procedure works efficiently for detecting “contaminated” distributions but it is much less powerful for Neyman-Scott alternatives. Again, the loss of efficiency associated with the estimation of the support is surprisingly low. As mentioned in the Introduction, the “spacing-based” EMS procedure (see Berrendero, Cuevas and Pateiro-López (2011) for details) can be thought as complementary to the DBU test. The EMS

Table 4: Empirical powers over 5000 runs of the DBU test, EMS test and DB test. The underlying distributions are contaminated uniforms $(1 - \epsilon)U(S) + \epsilon U(S \setminus S_0)$, where $S = \{(x, y) \in \mathbb{R}^2 : |x|^r + |y|^r \leq 1\}$ for $r = 3$ and S_0 denotes a set like S with the same centre and area $\mu(S)/2$. The significance level is 0.05.

		DBU test	DBU test	DB test	EMS test
		$S_n = \mathcal{H}(\mathcal{X}_n)$	$S_n = C_\lambda(\mathcal{X}_n)$		
$\epsilon = 0.1$	$n = 50$	0.0646	0.0598	0.1078	0.0130
	$n = 100$	0.1042	0.0974	0.1716	0.0404
	$n = 200$	0.2028	0.1934	0.3112	0.0566
$\epsilon = 0.2$	$n = 50$	0.1438	0.1168	0.2584	0.0212
	$n = 100$	0.3346	0.2990	0.4786	0.0638
	$n = 200$	0.6110	0.5844	0.7778	0.0992
$\epsilon = 0.3$	$n = 50$	0.3218	0.2570	0.5118	0.0374
	$n = 100$	0.6598	0.6178	0.8176	0.1126
	$n = 200$	0.9456	0.9338	0.9852	0.1910

Table 5: Empirical powers of the uniformity tests under study over 5000 runs of sample size $n = 100$ and $n = 200$ from Neyman-Scott clustering alternatives. Each cluster consist of m points, generated from the uniform distribution on a disc of radius r .

			DBU test	DB test	EMS test
$r = 0.05$	$m = 5$	$n = 100$	0.6174	0.4946	0.9790
		$n = 200$	0.5608	0.4976	0.9976
$r = 0.05$	$m = 10$	$n = 100$	0.9030	0.7556	0.9994
		$n = 200$	0.8504	0.7560	1.0000
$r = 0.1$	$m = 5$	$n = 100$	0.3458	0.3668	0.7952
		$n = 200$	0.3338	0.3670	0.8442
$r = 0.1$	$m = 10$	$n = 100$	0.5784	0.5822	0.9828
		$n = 200$	0.5404	0.5746	0.9970

test is suitable for alternative hypothesis that provide “clustered” observations but it is less powerful for “contamination models”, where the DBU test shows a clear superiority. The slight loss of power observed when increasing the sample size in some cases in Table 5 may be explained by the dependence of the observations generated from the Neyman-Scott model.

5 Appendix: Generation of uniform samples on S_n

The uniformity test for the case of an unknown support S is based on the statistic

$$D_n^* = \sqrt{n} \|\mathbb{F}_n^* - \hat{F}\|,$$

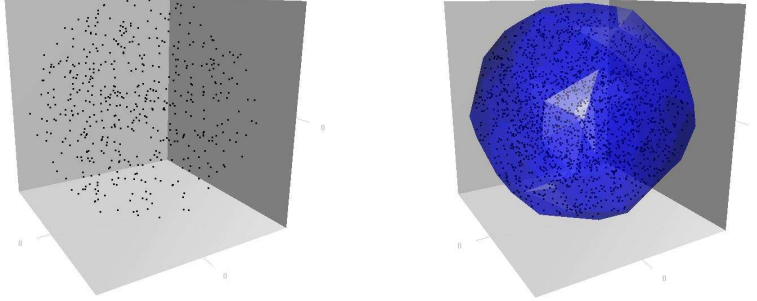


Figure 4: Uniform sample \mathcal{X}_n in $B(0,1)$ in \mathbb{R}^3 of size $n = 500$ (left). Convex hull $\mathcal{H}(\mathcal{X}_n)$ and uniform sample generated on $\mathcal{H}(\mathcal{X}_n)$ of size $m = 2000$ (right).

being \hat{F} the distribution of the random variable $\hat{Y} = D(\hat{X}, \partial S_n)$, where \hat{X} is uniform on S_n . Since \hat{F} is unknown, this distribution is derived in practice by a Monte Carlo mechanism. A large number of iid uniform observations $\hat{X}_i, i = 1, \dots, m$ are drawn on S_n . The empirical distribution corresponding to the sample $\hat{Y}_i = D(\hat{X}_i, \partial S_n), i = 1, \dots, m$ is used as an approximation for \hat{F} .

Uniform samples on $\mathcal{H}(\mathcal{X}_n)$.

Assume that we choose as estimator $S_n = \mathcal{H}(\mathcal{X}_n)$. The problem of how to generate uniform random vectors on the convex hull of a set of points in \mathbb{R}^2 is well-known. Note that this is a particular case of uniform random generation on a convex polygon in the plane, which is solved by means of triangulation. See Devroye (1986) for a description of the algorithm. The procedure in \mathbb{R}^3 is similar. In this case, we partition the convex hull of the sample into tetrahedra by means of the Delaunay triangulation of the sample, which can be computed in R by means of the library `geometry`, see Grasman and Gramacy (2010). To generate a point uniformly in the triangulated polyhedron, we first sample one of the tetrahedra with probabilities proportional to their volumes and then we sample a point uniformly in the selected tetrahedron. The generation of uniform random vectors in a tetrahedron is a particular case of the generation of uniform random vectors in a simplex for dimension $d = 3$. See Figure 4.

Uniform samples on $C_\lambda(\mathcal{X}_n)$.

Assume now that $S_n = C_\lambda(\mathcal{X}_n)$. In order to generate uniform samples on $C_\lambda(\mathcal{X}_n)$ we proceed as follows: first, we generate a large sample of uniform observations in the convex hull $\mathcal{H}(\mathcal{X}_n)$. Note that the λ -convex hull is contained in the convex hull. Then, we remove the points that belong to any of the balls defining the complement of the alpha convex hull. The resulting sample is uniform in $C_\lambda(\mathcal{X}_n)$, see Figure 5.

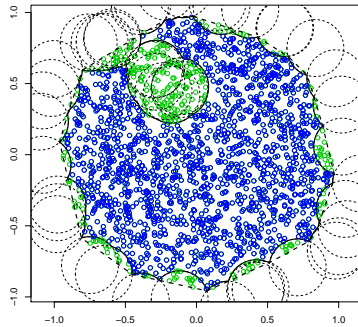


Figure 5: In blue, uniform sample in $C_\lambda(\mathcal{X}_n)$. The sample is obtained from uniform observations (in green) in the convex hull $\mathcal{H}(\mathcal{X}_n)$ after removing the sample points that belong to any of the balls defining the complement of the λ -convex hull (balls in dashed line).

Acknowledgements

This work has been partially supported by Spanish Grants MTM2010-17366 and CCG10-UAM/ESP-5494 (first and second author) and by Spanish Grant MTM2008-03010 and the IAP research network grant no. P6/03 from the Belgian government (third author).

References

- Berrendero, J.R., Cuevas, A. and Vázquez-Grande, F. (2006). Testing multivariate uniformity: The distance-to-boundary method. *Canad. J. Statist.* **34** 693–707.
- Berrendero, J.R., Cuevas, A., and Pateiro-López, B. (2011). A multivariate uniformity test for the case of unknown support. *Stat. Comput.* <http://dx.doi.org/10.1007/s11222-010-9222-z>
- Cuevas, A. and Fraiman, R. (2009). *Set estimation*. In *New Perspectives on Stochastic Geometry*. W.S. Kendall and I. Molchanov, eds. Oxford University Press, 374–397.
- Cuevas, A., Fraiman, R. and Pateiro-López, B. (2011). On statistical properties of sets fulfilling rolling-type conditions. *Preprint*.
- Devroye, L. (1986). *Nonuniform random variate generation*. Springer-Verlag, New York.
- Diggle, P.J. (2003). *Statistical Analysis of Spatial Point Patterns, 2nd edition*. Edward Arnold, London.

- Dümbgen, L. and Walther, G. (1996). Rates of convergence for random approximation of convex sets. *Advances in Applied Probability* **28** 384–386.
- Fernholz, L.T. (1991). Almost sure convergence of smoothed empirical distribution functions. *Scandinavian Journal of Statistics*. **18** 255–262.
- Efron, B. (1965). The convex hull of a random set of points. *Biometrika*, **52** 331–343.
- Grasman, R. and Gramacy, R. B. (2010). *geometry: Mesh generation and surface tessellation*. R package version 0.1-7. <http://CRAN.R-project.org/package=geometry>.
- Jain, A., Xu, X., Ho, T. and Xiao, F.(2002) Uniformity testing using minimal spanning tree. *Proceedings of the 16th International Conference on Pattern Recognition*, **4** 281–284.
- Jaklič, A., Leonardis, A. and Solina, F. (2000). *Segmentation and Recovery of Superquadrics: Computational imaging and vision*. Kluwer.
- Janson, S. (1987). Maximal spacings in several dimensions. *Ann. Probab* **15** 274–280.
- Liang, J. J., Fang, K. T., Hickernell, F. J. and Li, R. (2001). Testing multivariate uniformity and its applications. *Math. Comp* **70** 337–355.
- Moller, J. and Waagepetersen, R. P. (2004). *Inference and Simulation for Spatial Point Processes*. Chapman & Hall/CRC, Boca Raton.
- Pateiro-López, B. (2008). *Set estimation under convexity-type restrictions*. Ph. D. Thesis. Universidad de Santiago de Compostela.
- Pateiro-López, B. and Rodríguez-Casal, A. (2010). Generalizing the convex hull of a sample: The R package alphahull. *J. Stat. Softw* **5** 1–28.
- Pateiro-López, B. and Rodríguez-Casal, A. (2011). Recovering the shape of a point cloud in the plane. *arXiv:1105.5945v1*.
- Pegden, W. (2011). Sets resilient to erosion. *Adv. Geom.* **11** 201–224.
- Perkal, J. (1956). Sur les ensembles ϵ -convexes. *Colloq. Math.* **4** 1–10.
- R Development Core Team. (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. <http://www.R-project.org>.
- Reitzner, M. (2009). *Random polytopes*. In *New Perspectives on Stochastic Geometry*. W.S. Kendall and I. Molchanov, eds. Oxford University Press, 45–76.
- Ripley, B. D.(1979). Tests of ‘randomness’ for spatial point patterns. *J. Roy. Statist. Soc. Ser. A* **39** 172–212.
- Ripley, B. D. and Rasson, J. P. (1977). Finding the Edge of a Poisson Forest. *J. Appl. Probab* **14** 483–491.

- Smith, S. P. and Jain, A. K. (1984). Testing for uniformity in multidimensional data. *IEEE Trans. Pattern Anal. Machine Intell.* **PAMI-6** 73–81.
- Rodríguez-Casal, A. (2007). Set estimation under convexity type assumptions, *Ann. Inst. H. Poincaré Prob. Statist.* **43** 763–774.
- Rosin, P.L. and West, G.A.W. (1995). Curve Segmentation and Representation by Superellipses. *Proc. IEE: Vision, Image, and Signal Processing* **142** 280–288.
- Schütt (1994). Random polytopes and affine surface area. *Math. Nachr.*, **170** 227–249.
- Walther, G. (1997). Granulometric smoothing. *Ann. Statist.* **25** 2273–2299.
- Walther, G. (1999). On a generalization of Blaschke’s rolling theorem and the smoothing. *Math. Methods Appl. Sci.* **22** 301–316.

Reports in Statistics and Operations Research

2005

- 05-01 SiZer Map for Evaluating a Bootstrap Local Bandwidth Selector in Nonparametric Additive Models. M. D. Martínez-Miranda, R. Raya-Miranda, W. González-Manteiga and A. González-Carmona.
- 05-02 The Role of Commitment in Repeated Games. I. García Jurado, Julio González Díaz.
- 05-03 Project Games. A. Estévez Fernández, P. Borm, H. Hamers.
- 05-04 Semiparametric Inference in Generalized Mixed Effects Models. M. J. Lombardía, S. Sperlich.

2006

- 06-01 A unifying model for contests: effort-prize games. J. González Díaz.
- 06-02 The Harsanyi paradox and the "right to talk" in bargaining among coalitions. J. J. Vidal Puga.
- 06-03 A functional analysis of NO_x levels: location and scale estimation and outlier detection. M. Febrero, P. Galeano, W. González-Manteiga.
- 06-04 Comparing spatial dependence structures. R. M. Crujeiras, R. Fernández-Casal, W. González-Manteiga.
- 06-05 On the spectral simulation of spatial dependence structures. R. M. Crujeiras, R. Fernández-Casal.
- 06-06 An L₂-test for comparing spatial spectral densities. R. M. Crujeiras, R. Fernández-Casal, W. González-Manteiga.

2007

- 07-01 Goodness-of-fit tests for the spatial spectral density. R. M. Crujeiras, R. Fernández-Casal, W. González-Manteiga.
- 07-02 Presmoothed estimation with left truncated and right censored data. M. A. Jácome, M. C. Iglesias-Pérez.
- 07-03 Robust nonparametric estimation with missing data. G. Boente, W. González-Manteiga, A. Pérez-González.
- 07-04 k-Sample test based on the common area of kernel density estimators. P. Martínez-Cambor, J. de Uña Álvarez, N. Corral-Blanco.

07-05 A bootstrap based model checking for selection-biased data. J. L. Ojeda, W. González-Manteiga, J. A. Cristobal.

07-06 The Gaussian mixture dynamic conditional correlation model: Bayesian estimation, value at risk calculation and portfolio selection. P. Galeano, M. C. Ausín.

2008

08-01 ROC curves in nonparametric location-scale regression models. W. González-Manteiga, J. C. Pardo Fernández, I. Van Keilegom.

08-02 On the estimation of α -convex sets. B. Pateiro-López, A. Rodríguez-Casal.

2009

09-01 Lasso Logistic Regression, GSoft and the Cyclyc Coordinate Descent Algorithm. Application to Gene Expression Data. M. García-Magariños, A. Antoniadis, R. Cao, W. González-Manteiga.

2010

10-01 Asymptotic behaviour of robust estimators in partially linear models with missing responses: The effect of estimating the missing probability on simplified marginal estimators. A. Bianco, G. Boente, W. González-Manteiga, A. Pérez-González.

10-02 First-Price Winner-Takes-All Contents. J. González-Díaz.

10-03 Goodness of Fit Test for Interest Rate Models: an approach based on Empirical Process. A. E. Monsalve-Cobis, W. González-Manteiga, M. Febrero-Bande.

2011

11-01 Exploring wind direction and SO₂ concentration by circular-linear density estimation. E. García-Portugués, R.M. Crujeiras, W. González-Manteiga.

11-02 Utilities for Statistical Computing in Functional Data Analysis: The R Package fda.usc. M. Oviedo de la Fuente, M. Febrero-Bande

11-03 Multivariate uniformity tests: the distance to boundary method for the case of unknown support. J. R. Berrendero, A. Cuevas, B. Pateiro-López

Previous issues (2001 – 2003):

<http://eio.usc.es/reports.php>