Testing multivariate uniformity: the distance-to-boundary method

José R. BERRENDERO, Antonio CUEVAS and Francisco VÁZQUEZ–GRANDE

Abstract: Given a random sample taken on a compact domain $\mathcal{S} \subset \mathbb{R}^d$, the authors propose a new method for testing the hypothesis of uniformity of the underlying distribution. The test statistic is based on the distance of every observation to the boundary of $\mathcal{S}$. The proposed test has a number of interesting properties, namely: Unlike most available methods, it is feasible and particularly suitable for high dimensional data, it is distribution-free for a wide range of choices of $\mathcal{S}$, it can be adapted to the case that the support $\mathcal{S}$ is unknown and also allows for one-sided versions. Moreover, the results suggest that, in some cases, this procedure does not suffer from the well-known "curse of dimensionality". The authors study the properties of this test from both a theoretical and practical point of view. In particular, an extensive simulation study is given in order to compare the performance of our methods with some recent alternative procedures. The conclusions suggest that the proposed test provides quite a satisfactory balance between statistical power, computational simplicity, and flexibility of application for different dimensions and supports.

Title in French: we can supply this

Résumé : Given a random sample taken on a compact domain $\mathcal{S} \subset \mathbb{R}^d$, the authors propose a new method for testing the hypothesis of uniformity of the underlying distribution. The test statistic is based on the distance of every observation to the boundary of $\mathcal{S}$. The proposed test has a number of interesting properties, namely: Unlike most available methods, it is feasible and particularly suitable for high dimensional data, it is distribution-free for a wide range of choices of $\mathcal{S}$, it can be adapted to the case that the support $\mathcal{S}$ is unknown and also allows for one-sided versions. Moreover, the results suggest that, in some cases, this procedure does not suffer from the well-known "curse of dimensionality". The authors study the properties of this test from both a theoretical and practical point of view. In particular, an extensive simulation study is given in order to compare the performance of our methods with some recent alternative procedures. The conclusions suggest that the proposed test provides quite a satisfactory balance between statistical power, computational simplicity, and flexibility of application for different dimensions and supports.

1. INTRODUCTION

The problem of testing whether a pattern of points in the Euclidean space $\mathbb{R}^d$ is distributed uniformly (or "at random", using this term in an improper restricted way) arises in different fields, including biology, forestry, astronomy and computer science. The univariate case, where the observed points are numbers in the real line, has received much more attention in the literature than the multivariate case. This paper is concerned with the latter. Whereas this multivariate extension has an obvious practical interest, it is far from trivial as it involves a number of additional difficulties. These range from the huge variety of potential supports to be considered (unlike the

univariate case where the obvious choice for the sample space is a compact interval), to the so-called "'curse of dimensionality" (which imposes high sample size requirements in order to preserve the power) as well as other technical reasons, such as the lack of the distribution-free property for some standard statistics (Kolmogorov-Smirnov, Cramer-von Mises,...) used often in the univariate goodness of fit tests.

The uniformity tests have been tackled in the literature from different approaches, depending on the assumed underlying model for the observed points. There are three usual choices for this model. Under the *point process model* the sample observations are assumed to be a realization of a point process on $\mathbb{R}^d$ (usually for $d = 2$). The null hypothesis is that such process is of Poisson type with a constant intensity function. In practice, the observation is always restricted to a given area (typically, a rectangle in the considered space) which entails the presence of a boundary effect. Different procedures have been proposed for the testing problem. The most popular are maybe those based on nearest neighbours statistics (see, e.g., Ripley 1979; Diggle 2003; Moller and Waagepetersen 2004). In *the binary sequence model* the problem is testing "randomness" (i.e. equiprobability and independence) from a given string of binary digits. This problem arises in several fields, including the assessment of different random number generators. A fairly comprehensive survey of existing methods and some theoretical insights can be found in Rukhin (2001). Finally, in *the iid model* the sample information consists of a fixed number of independent identically distributed (iid) random variables. This is the usual framework in the classical goodness of fit theory as well as the model to be considered hereafter throughout this work.

The study of multivariate uniformity tests under the iid model is less developed than the analogous problem (e.g., see Székely and Rizzo 2005) of testing multivariate normality. Let us mention however the paper by Justel, Peña and Zamar (1997) where a uniformity test for bi-variate observations on $[0, 1]^2$ is proposed using the corresponding bi-dimensional version of the Kolmogorov-Smirnov statistic. An algorithm is given for the non-trivial problem of exactly calculating this statistic in the bivariate case $d = 2$. An approximate algorithm valid for $d > 2$ is also discussed briefly but, as the authors point out, *"An extension of the computing algorithm* [for $p > 2$] *is maybe possible but still the numerical complications seem considerable"*.

A different approach is followed in the paper by Liang, Fang, Hickernell and Li (2001) where several tests are proposed for testing uniformity on $[0, 1]^d$. These tests are based on non-trivial number-theoretic considerations and can be used even in high-dimensional cases. The basic idea is to use some "discrepancy measures" in order to evaluate the distance between expectations of type $E(f(X))$, f being a real function on $[0, 1]^d$ and their Monte Carlo approximations. The asymptotic behaviour of these discrepancies under the uniformity assumption is derived in Liang *et al.* (2001) and used to define several asymptotic uniformity tests. They can be implemented in any dimension d, with a reasonable performance in terms of both type I (significance) and type II (power) errors. In the Monte Carlo outputs provided by these authors it can be observed a remarkable increase of the power (for a given sample size n) as the dimension increases. This fact, which is also observed in our test, can be of some interest in connection with the current increasing attention on high dimensional problems (those with $d >> n$; see, e.g., Hall, Marron and Neeman 2005).

Let us finally recall that the notion of "maximal spacing" (i.e. the maximum distance between consecutive points of the sample) is also a classical tool for testing uniformity on $[0, 1]$; see Reschen-hofer and Bomze (1991). The multivariate generalization of the concept of spacing proposed by Deheuvels (1983) and the asymptotic results obtained by Janson (1987) could be used to derive an (asymptotic) multivariate uniformity test.

The paper is organized as follows: In Section 2 the new uniformity test is presented and its main theoretical properties are discussed. Section 3 is devoted to a Monte Carlo comparative study. Some conclusions and final remarks are included in Section 4. The proofs are given in an Appendix.

2. THE DISTANCE-TO-BOUNDARY (DB) TEST

If $X = (\xi_1, \ldots, \xi_d)$ is a random variable with compact (known) support $\mathcal{S} \subset \mathbb{R}^d$, define the distance-to-the-boundary $D(X, \partial \mathcal{S}) := \min\{x \in \partial \mathcal{S} : \|x - X\|\}$, $\|.\|$ being the Euclidean norm on $\mathbb{R}^d$ and $\partial \mathcal{S}$ the topological boundary of $\mathcal{S}$. Denote by R the maximum distance to the boundary that can be attained on $\mathcal{S}$, $R := \max\{D(x, \partial \mathcal{S}) : x \in \mathcal{S}\}$. Finally, let $Y := D(X, \partial \mathcal{S})/R$. We will sometimes refer to Y using the term "relative depth" which is motivated by obvious intuitive considerations. It is well-known that the term "depth" is also used in the statistical literature with different purposes, mostly oriented to generalize the notion of order statistics to multivariate frameworks; see, e.g., Zuo and Serfling (2000). Our target is to test the null hypothesis

H_0: X is uniformly distributed on $\mathcal{S}$,

from an iid random sample $X_1, \ldots, X_n$ of X.

The test we propose here is based on the real variables $Y_i = D(X_i, \partial S)/R$, $i = 1, \ldots, n$. Let $G(y)$, $G_n(y)$ be the distribution function of Y and the empirical distribution function associated with the sample $Y_1, \ldots, Y_n$, respectively. Denote by $H(y)$ the distribution of Y under the null hypothesis.

2.1 The two- and one-sided DB test

The basic idea behind our DB uniformity test is as follows: We reduce the problem of testing H_0 to that of testing $H_0^* : G = H$ with the classical Kolmogorov-Smirnov test, that is, we reject the null hypothesis, at a significance level α, whenever

$$\sup_y \sqrt{n}|G_n(y) - H(y)| > C_{n,\alpha},$$

where the critical value $C_{n,\alpha}$ is obtained in the usual way from the distribution of the Kolmogorov-Smirnov statistic (i.e., $P_{H_0}\{\sup_y \sqrt{n}|G_n(y) - H(y)| > C_{n,\alpha}\} = \alpha$).

An obvious advantage of this procedure is the fact that the problem, originally proposed in dimension d, has been reduced to a much simpler one-dimensional version. Of course, this entails some loss in the discrimination capacity, as some different underlying distributions G could lead to the same value of the distance $\sup |G(x) - H(x)|$ (see the Subsection 2.2 below). However, there are at least five important aspects which make this idea feasible and worth considering: First, the distribution H can be in many cases explicitly known; second, this distribution does not depend on $\mathcal{S}$ provided that this set fulfills a fairly general condition; third, the resulting test is extremely simple to implement and turns out to be competitive when compared with other available alternative tests for the same problem; fourth, the test is exact (in the sense that no asymptotical approximation is involved); fifth, the test does not depend on the chosen coordinate axes, it is invariant under translations and rotations.

A further interesting feature of the proposed test is that it admits of directional (one-sided) versions valid for those cases where the alternative hypothesis is either $H_1^* : G \leq H$ or $H_1^* : G \geq H$. Both cases have an easy intuitive interpretation: For example, $G \leq H$ corresponds to those distributions which tend to provide points "deeper" (that is, further away from the support boundary) than those in the uniform case. Of course, for these one-sided alternatives we would use the corresponding one-sided versions of the Kolmogorov-Smirnov test.

2.2 Basic properties

The distribution H of the relative depth variable Y under the null hypothesis of uniformity remains invariant for a wide class of sets $\mathcal{S}$. Before proving this (see Theorem 1 below), we need to introduce some notation. Let B denote the unit closed ball in $\mathbb{R}^d$ and $B(x, r)$ the closed ball centred at x with radius r. Given $r > 0$ we define the r–erosion of $\mathcal{S}$ by

$$\mathcal{S} \ominus rB = \{x : B(x, r) \subset \mathcal{S}\}.$$

We will say that a set $\mathcal{S}$ is "invariant by erosion" upon an homothecy, if for every $r \in (0, R)$ there exists a constant λ_r such that

$$\mathcal{S} \ominus rB = \lambda_r \mathcal{S}. \tag{1}$$

It can be easily seen that (1) holds for a square of type $\mathcal{S} = [-M, M]^d$ or for a ball $\mathcal{S} = B(0, M)$. It can be also seen that (1) entails that 0 is the deepest point (so $R = D(0, \partial \mathcal{S})$) but this is not essential in what follows as the results below are also valid, with trivial changes, when (1) holds after a translation.

The following theorem establishes a distribution-free property, which makes easier the use of the proposed test, provided that $\mathcal{S}$ is invariant by erosion upon an homothecy:

THEOREM 1. *(a) If the (compact) support $\mathcal{S}$ fulfills (1) and X is uniform on $\mathcal{S}$, then the distribution function of the relative depth variable $Y = D(X, \mathcal{S})/R$ is $H(y) = 1 - (1 - y)^d$, for $0 \le y \le 1$, that is, H is a beta distribution with parameters $a = 1$ and $b = d$. In particular, the limit distribution of dY when d tends to infinity is exponential with parameter 1.*
(b) Assume that the support of the uniform variable X is $\mathcal{S} = \mathcal{S}_1 \setminus \mathcal{S}_0$, where $\mathcal{S}_1$ is a set fulfilling (1) with $R = D(0, \partial \mathcal{S}_1)$ and $\mathcal{S}_0 = (r/R)\mathcal{S}_1$, for some $r \in (0, R)$. Then the distribution of the relative depth variable Y is

$$H(y) = 1 - \frac{[2R - (R - r)y]^d - [2r + (R - r)y]^d}{2^d(R^d - r^d)}, \quad 0 \le y \le 1.$$

We next show that, besides the squares and the balls, condition (1) is also fulfilled for a wide class of convex polyhedra. See Apostol and Mnatsakanian (2004) for closely related geometrical ideas.

THEOREM 2. *Every convex polyhedron (with deepest point at 0) circumscribed to a ball fulfills condition (1)*

Let us conclude this subsection with two important remarks regarding the practical use of the DB-test:

First note that, even if condition (1) is not fulfilled, the distribution H of the test statistic is easy to evaluate in many other instances. For example, if the support S is a rectangle $\mathcal{S} = \Pi_{i=1}^{d}[0, 2R_i]$, then $H(y) = 1 - \Pi_{i=1}^{d}(1 - \kappa_i y)$, for all $0 \le y \le 1$, where $\kappa_i = R/R_i$ and $R = \min\{R_1 \ldots, R_d\}$ is the maximum depth.

As a second remarkable feature we should mention that, under many usual alternative distributions, the power of our uniformity test tends to increase as the dimension d grows (for a given sample size). This phenomenon is clearly observed in the simulations of Section 3 below as well as in those given by Liang *et al.* (2001). At first sight, this could seem a bit surprising as it contradicts the well-known "curse of dimensionality" phenomenon which affects many statistical procedures. However, after a more careful consideration, some kind of "dimensionality blessing" could be expected for a multivariate uniformity test (at least for some alternatives) since a higher number of components in the data X_i provides additional chances to detect deviations from uniformity. Obviously, in the simple χ^2 uniformity test the phenomenon of dimensionality blessing does not appear; rather the opposite is true since this procedure is based only on frequency counts which suffer from the curse of dimensionality associated with the "empty spaces" phenomenon (see, e.g., Silverman 1986).

Figure 1 (top) shows a typical behaviour for the absolute difference between $G := G(.; d)$ (the true distribution of the distance-to-boundary variable Y) and $H := H(.; d)$ (the distribution of Y when the X_i are iid uniform). Intuitively, as the maximum of this absolute difference increases we expect the power of the test to be greater. These graphics correspond to the case where the
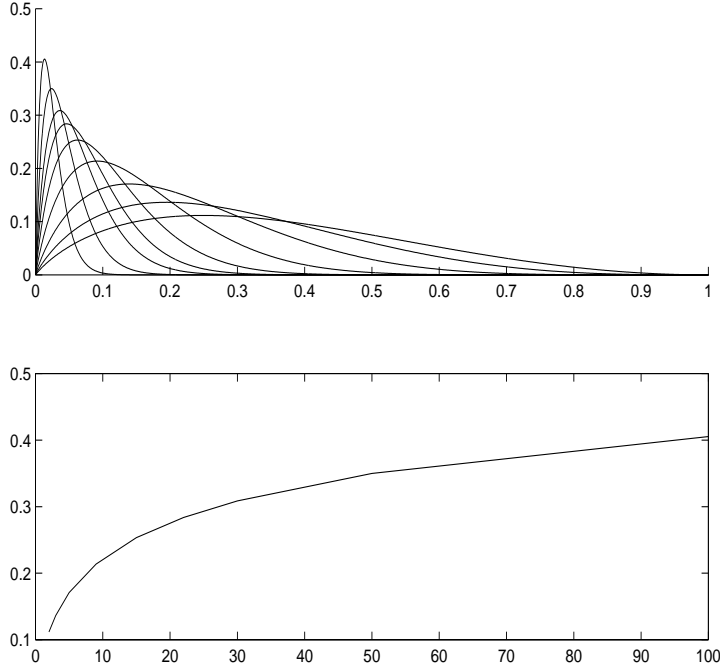
Figure 1: The upper panel shows the absolute difference between the distributions of Y under independent beta marginals and those under the null hypothesis, for several values of d. In the lower panel the maximum difference is represented as a function of d.

support is $\mathcal{S} = [0,1]^d$ and the marginals ξ_i of X are iid with distribution F in the beta family (with parameters $a = b = 1.3$). Thus the curves at the top of Figure 1 show the differences

$$|G(y) - H(y)| = |(1-y)^d - [F(1-y/2) - F(y/2)]^d|$$

for $0 \le y \le 1$ and for different values of the dimension d ranging from $d = 2$ to $d = 100$. It can be seen that the quantity $\sup_y |G(y) - H(y)|$ increases as d grows. Also, the point $y_d \in [0,1]$ where the supremum $\sup_y |G(y) - H(y)|$ is attained tends to be smaller (and closer to 0) as d grows. The curve at the bottom of Figure 1 shows the plot of the function $\Psi(d) = |G(y_d) - H(y_d)|$.

It is not easy to give a formal account of this improvement in high dimensions through a general theorem. The reason is the huge number of essentially different supports and departures from uniformity to be considered. However, the "asymptotic-dimensional" behaviour of the relative depth variable Y when the sample size n is fixed and the dimension d grows to infinity can be explicitly derived in some cases. For example, assume that the considered support is $\mathcal{S} = [0,1]^d$ and the components of the observed random variable $X = (\xi_1, \ldots, \xi_d)$ are iid with density function f. A straightforward calculation gives

$$\lim_{d \to \infty} \mathrm{P}\left(dY \le y\right) = 1 - e^{-[f(0)+f(1)]y/2}, \quad \text{for } y \ge 0.$$

This means that dY has, in the limit, an exponencial distribution with parameter $[f(0) + f(1)]/2$. In particular, under the null hypothesis that the ξ_i are uniform we would get that, for d large, dY

is approximately exponential with parameter 1. So, if d is large enough (keeping n constant) the simple observation of the relative depth variable Y should allow us to detect the non-uniformity of f provided that $f(0) + f(1)$ (which represents, in some sense, "the weight of the boundary") is different from 2 (the corresponding uniform value).

2.3 Crown-uniformity and consistency

Let us note that the original two-sided DB-test should be seen itself as a directional test, tailor-designed to detect some specific alternatives. The point is that the distribution G of the depth variable Y associated with some non-uniform random variables X, could coincide with that of the uniform case, H. For example, assume that the distribution P_X of X fulfills $P_X([y, 1-y]^d) = (1-2y)^d$ for all $y \in [0, 1/2]$. This would be the case when the marginals ξ_i in $X = (\xi_1, \ldots, \xi_d)$ are iid random variables taking values on $[0, 1]$ with absolutely continuous distribution function F such that $F(1-y) - F(y) = 1 - 2y$, for all $y \in [0, 1/2]$ (for example $F(y) = y^2$). Thus, in strict sense, our procedure is intended to consistently test the "crown-uniformity" hypothesis that the underlying distribution gives the same mass as the uniform one to every outside crown $[0, 1]^d \setminus [y, 1-y]^d$, that is $P_X([0, 1]^d \setminus [y, 1-y]^d) = 1 - (1-2y)^d$. A similar comment holds for a general support $\mathcal{S}$; in this case the crown-uniformity is defined by the condition $P_X(\mathcal{S} \ominus \epsilon B) = \mu_L(\mathcal{S} \ominus \epsilon B)/\mu_L(\mathcal{S})$, where μ_L is the Lebesgue measure on $\mathbb{R}^d$. Therefore if the DB method is used to test uniformity, the consistency is not guaranteed unless we can exclude beforehand the crown-uniformity. In fact, this lack of universal consistency is quite common in the statistical theory and practice. We could mention for instance the randomness tests based of the average nearest-neighbour distance considered by Ripley (1979), the normality tests based on skewness and kurtosis (see Mardia 1970) and even the classical χ^2-test for goodness of fit (although in this case the consistency can be achieved asymptotically). The following proposition shows that crown-uniformity is equivalent to uniformity in the class of spherically symmetric distributions. This has some practical relevance, as the symmetric sphericity could also be tested in an efficient way (see, e.g., Baringhaus 1991, and references therein) as a previous step to the use of the DB-test.

PROPOSITION 1. *Assume that the underlying distribution P_X has a density f, whose support is the unit ball B on $\mathbb{R}^d$, such that $f(x) = g(\|x\|)$ for $x \in B$, where $\|.\|$ is the Euclidean norm and $g : [0, 1] \to [0, \infty)$ is a continuous function. Assume also that P_X fulfills the "crown-uniformity" condition $P_X(B(0, \epsilon)) = \mu_L(B(0, \epsilon))/\mu_L(B)$. Then, P_X is necessarily uniform on B.*

Note that if the ball B is replaced for a polyhedron S circumscribed to B, an analogous conclusion can be obtained but in this case the sphericity assumption $f(x) = g(\|x\|)$ must be imposed on the distribution obtained by transforming the original S into B in the following way: the image of $x \in S$ is the point $T(x) \in B$ such that $T(x)$ is in the segment joining 0 with x and the distance from $T(x)$ to the boundary of B equals the distance from x to the boundary of S. Observe that T takes each ϵ-crown $S \setminus (S \ominus \epsilon B)$ of S into the circular crown $B \setminus (B \ominus \epsilon B)$. Thus if X is a S-supported random variable with a crown-uniform distribution, then $T(X)$ is crown-uniform on B and conversely. Also X is uniform on S if and only if $T(X)$ is uniform on C. Finally, from the Proposition above, if X is crown-uniform on S and the distribution of $T(X)$ is spherical, X must necessarily be uniform.

2.4 The case of unknown support

Our method can be adapted to the case where the null hypothesis is
    H_0: The distribution of X is uniform on some (unknown) support $\mathcal{S}$.
    Thus the framework would be analogous to that of the problem of goodness of fit to a parametric family. The natural solution here is analogous to that of the classical Lilliefors version of Kolmogorov-Smirnov test: the unknown "parameter" $\mathcal{S}$ is replaced by a suitable estimator S_n

and the original test is performed on S_n. However, the fact that $\mathcal{S}$ is not a member of a finite-dimensional space entails some additional difficulties. Typically, the shape of S_n will not be nice enough to provide a simple form for the distribution of the depth variable (let us denote it by $Y^{(n)}$) defined from uniform samples drawn on S_n. Hence this distribution will be derived by a Monte Carlo mechanism: a large number, B, of Monte Carlo (artificial) samples of size n consisting of iid uniform observations are drawn on S_n. The corresponding empirical distribution is used as an approximation for the distribution of $Y^{(n)}$.

As the "empirical" depth variable $Y^{(n)}$ is defined in terms of ∂S_n, the whole idea would not work unless ∂S_n provides a consistent estimator of $\partial \mathcal{S}$ in the sense that

$$d_H(\partial S_n, \partial \mathcal{S}) \to 0, \text{ as } n \to \infty, \tag{2}$$

d_H being the Hausdorff distance between compact sets. This is fulfilled, under some regularity conditions (see, e,g., Cuevas and Rodríguez-Casal (2004), Theorems 1 and 4) by the Devroye-Wise's (1980) estimator

$$\widehat{S}_n(\epsilon_n) = \bigcup_{i=1}^{n} B(X_i, \epsilon_n). \tag{3}$$

It is worth noting that the boundary of this estimator is easily identified, since $\partial \widehat{S}_n = \{x : \min_{i=1,\ldots,n} \|x - X_i\| = \epsilon_n\}$.

An important practical question is how to choose the sequence ϵ_n in order to achieve (2). Theorem 4 in Cuevas and Rodríguez-Casal provides some hints in this respect: If ω_d is the Lebesgue measure of the unit ball in $\mathbb{R}^d$, and δ denotes the *standardness index* of $\mathcal{S}$ (that is the minimum portion of any (small enough) ball centred at a point of $\mathcal{S}$ which is included in $\mathcal{S}$; for example $\delta = 1/4$ for an square, $\delta = 1/6$ for an equilateral triangle), then a choice of ϵ_n of type

$$\epsilon_n = C \left(\frac{\log n}{n} \right)^{\frac{1}{d}}, \tag{4}$$

where C is a constant such that $C > \left(\frac{2}{\delta \omega_d} \right)^{\frac{1}{d}}$, leads to (2). According to the mentioned theorem, it also turns out that the sequence (4) satisfies $\mathcal{S} \subset \widehat{S}_n$ eventually almost surely. So, the estimator (3) with ϵ_n given by (4) provides an overestimation of $\mathcal{S}$. For this reason, we use $S_n = \widehat{S}_n \ominus \epsilon_n B$, a "eroded" version of $\widehat{S}_n$ to draw the Monte Carlo samples required for approximating the null distribution of the depth statistic. Note that some original sample points X_i must necessarily be on the boundary of S_n. These are excluded in the evaluation of the empirical distribution G_n. Thus in fact we work with a "censored" version of the original sample, which makes the procedure conservative with respect to deviations from uniformity tending to give many points close to the boundary. This loss of efficiency should be considered as a price to be paid in order to work with an unknown support. However, it is easy to see that the proportion of original sample observations in the boundary of S_n tends to 0 as n tends to infinity (provided that the natural condition $\epsilon_n \to 0$ holds). Thus the loss of efficiency associated with this boundary effect tends to disappear for large values of n.

In some cases, additional geometrical criteria can also be used in the choice of ϵ_n. For example, if the unknown support may be assumed to be connected, we could select ϵ_n in order to get a connected estimated support S_n.

3. MONTE CARLO RESULTS

In the above sections we have pointed out the simplicity and flexibility of the DB test. Obviously, these properties would be of little interest unless this test proves to be competitive in terms of both, power and preservation of the nominal confidence level.

In the present section we check the performance of our test through a simulation study. It is clear that a distribution can be non-uniform in a potentially infinite number of different meaningful ways. So any Monte Carlo study is necessarily limited in scope, given the huge amount of interesting alternative hypotheses to be checked. We think that, in our case, a plausible criterion is to design the study in order to make its results comparable with those in Liang *et al.* (2001). The reason is that the uniformity tests proposed by these authors are, as our DB test, applicable to iid high--dimensional data. Thus all the models considered in our Monte Carlo study have support $[0,1]^d$ since the Liang *et al.* (2001) tests are all exclusively designed for this situation.

3.1 Description of the study

Five tests have been considered in our simulations: the two-sided DB test, the appropriate DB one-sided version, the χ^2 test (which is is included here as a benchmark for comparisons) and the tests proposed by Liang *et al.* (2001) based on the statistics A_n and T_n (we follow here the notation of these authors).

The cells used for the χ^2 test are obtained from equispaced partitions of size $m = \lfloor (n/5)^{1/d} \rfloor$ in every axis. It is well-known that the sample size must be large enough for a proper use of χ^2 test given the asymptotic nature of this procedure. This in practice leads to some requirements on the expected number of observations, e_i, in each considered cell. The usual condition is $e_i \geq 5$ for all i (or, at least, for most i) which is not fulfilled in high dimensions unless unrealistically large sample sizes are used. The choice of m indicated above is aimed at fulfilling these requirements. In the tables below there are some missing χ^2 outputs, corresponding to some situations in which the sample size turned out to be insufficient for a reliable use of this test.

The use of the one-side DB test requires some prior knowledge on the nature of the underlying distribution: We must know if it tends to provide observations either closer or further away from the boundary than those in the uniform case. Although it seems quite reasonable to have some prior guess in this respect in many practical situations, this information could also be unavailable. Of course in such situations the standard two-sided version should be used.

Our null hypothesis is in all cases

H_0: The underlying distribution is uniform on $[0,1]^d$ ($U([0,1]^d)$).

Four values have been considered for the dimension d: 2, 3, 5, 20. The nominal significance level is 0.05. In order to assess how this value is kept in practice by the different tests we have evaluated their rejection percentages along 10000 independent runs based on samples drawn from the null hypothesis. The results are given in Table 1.

As for the power study, we have checked three different models in the choice of the alternative distribution:

A) *The beta model*: All the marginals are independent, identically distributed according to a beta distribution $Beta(a,b)$. Two different values have been taken for the parameter (a,b): (1.3, 1.3), (0.8, 0.8) (Table 2).

   The beta family looks like a very natural choice to check in a compact-supported case, as it is a very rich and flexible class which includes distributions of almost every conceivable shape. Note that the case considered here of independent marginals should be the most difficult to detect, since it is also the case with the uniform distribution on the unit square.

B) *Meta-type uniform distributions*: These are the distributions of those $[0,1]^d$-supported random variables $X = (\xi_1, \ldots, \xi_d)$ obtained from some $\mathbb{R}^d$-supported elliptically distributed random variables $Y = (\eta_1, \ldots, \eta_d)$, via a transformation of type $\xi_i = F_i(\eta_i)$, where F_i denotes the (continuous) distribution function of η_i. Note that though every ξ_i is uniform on $[0,1]$, X need not be uniform on $[0,1]^d$, since the ξ_i are not independent in general. Thus,

this family is useful in order to assess the capacity of our uniformity test for detecting departures from independence. Specifically, we have considered the following meta-type uniform distributions for X (see Liang *et al.* (2001) for more details):

> MNU: obtained from $Y \sim N(0, \Sigma)$, with $\Sigma = (\sigma_{ij})$, $\sigma_{ii} = 1$, $\sigma_{ij} = 0.5$ for $i \neq j$. See Table 3.

> MTU: the transformation is defined from a multivariate Student's-t random variable Y with 5 degrees of freedom. See Table 3.

> MPVIIU: Y is a Pearson-type VII distribution with parameters 11 and 2. See Table 4.

> MCU: obtained from a Cauchy variable Y. See Table 4.

C) *The contamination model*: The observed points are a sample drawn from a random variable $Y = (\eta_1, \ldots, \eta_d)$ whose distribution is a contaminated uniform, given by a mixture of type $(1 - \varepsilon)U([0,1]^d) + \varepsilon U(C_d)$, where C_d denotes a cube with the same centre as $[0,1]^d$ and measure $1/2$. We have taken $\varepsilon = 0.1$, 0.2. The corresponding outputs are given in Table 5.

It can be observed that the distributions described in the paragraphs A and B above are complementary in the sense that they take into account dual aspects of non-uniformity: Whereas the distributions in A account for marginal deviations from uniformity (keeping the independence of the ξ_i), those in B consider deviations from independence (keeping the marginals' uniformity). The model in C has maybe a more direct intuitive interpretation. It takes into account a typical deviation from the theoretical model (quite common in the classical robustness theory) associated with the presence of "inliers".

3.2 The numerical outputs

The simulation outputs are summarized in Tables 1-5 which provide the rejection proportions for the null (uniformity) hypothesis on 10000 independent runs.

Regarding the several proposals of Liang *et al.* (2001) we should say that, in which concerns power, the best of them (called T_n *with symmetric discrepancy* in that paper) performs better than ours in the meta-type uniform distributions. In the beta model, the test based on T_n shows also a better performance in general, though not uniformly. Finally, in the contamination model the DB test shows a clear superiority. It could be noted that the power of the different tests (except the χ^2) improves as the dimension d increases in the models considered in Tables 2 to 4. In the contamination model (Table 5) no such phenomenon is observed, as a constant proportion of "inliers" is present for all considered values of d. However, interestingly, no "curse of dimensionality" is observed in this case.

As for the proposed significance level Table 1 shows that in general the DB test does a better job in preserving the nominal value. In fact, Liang *et al.* (2001) report relative average deviations with respect to the nominal significance level of 21% in the case of $\alpha = 0.05$, 17% when $\alpha = 0.1$ and 81% if $\alpha = 0.01$.

4. CONCLUSIONS

These are, in our view, the most relevant features of the distance-to-boundary approach that has been introduced in this article:

a) Since the test statistic is evaluated from the distances of the sample observations to the boundary of the support, the method is easy to implement. Moreover, it seems a suitable choice for high dimensional data as it is not particularly affected by the curse of dimensionality.

9

Table 1: Empirical significance level (over 10000 runs) of the considered uniformity tests. The nominal value is 0.05

d=2	25	50	100	200
DB	0.0508	0.0507	0.049	0.0467
DB o-s	0.0497	0.0489	0.0503	0.0493
DB o-s	0.0511	0.0519	0.0471	0.0509
χ^2	0.0389	0.0541	0.0476	0.0512
A_n	0.0664	0.0543	0.0495	0.0482
T_n	0.0565	0.0578	0.0518	0.0563
d=3				
DB	0.0532	0.0524	0.0516	0.0493
DB o-s	0.0503	0.0514	0.0541	0.0516
DB o-s	0.0532	0.0514	0.0492	0.0485
χ^2	-	0.0432	0.0517	0.0488
A_n	0.0649	0.0565	0.0528	0.0547
T_n	0.0557	0.0535	0.0519	0.0554
d=5				
DB	0.0472	0.0501	0.0536	0.0503
DB o-s	0.0491	0.0487	0.0537	0.0511
DB o-s	0.046	0.0516	0.0495	0.0515
χ^2	-	-	-	0.051
A_n	0.0662	0.056	0.0533	0.0523
T_n	0.0523	0.0539	0.0557	0.0542
d=20				
DB	0.0476	0.0558	0.0529	0.0517
DB o-s	0.0488	0.05	0.0528	0.0518
DB o-s	0.05	0.0526	0.0512	0.0517
χ^2	-	-	-	-
A_n	0.0885	0.0743	0.0607	0.0545
T_n	0.0523	0.0531	0.0455	0.0522

Table 2: Empirical powers (over 10000 runs) of the considered uniformity tests. The true underlying distributions are of Beta type with the indicated parameters (a, b). The significance level is 0.05.

Beta	(1.3,1.3)				(0,8,0.8)			
d=2	25	50	100	200	25	50	100	200
DB	0.2069	0.3839	0.678	0.9409	0.1853	0.3118	0.548	0.8436
DB o-s	0.3182	0.5205	0.7973	0.9725	0.2776	0.4336	0.6705	0.9127
χ^2	0.041	0.0959	0.1709	0.3277	0.0375	0.0795	0.1222	0.2458
A_n	0.292	0.4976	0.7917	0.9787	0.2227	0.3833	0.6334	0.9127
T_n	0.1745	0.3416	0.6476	0.94	0.1862	0.316	0.5408	0.8511
d=3								
DB	0.3036	0.5545	0.8466	0.9895	0.2552	0.4349	0.7243	0.9533
DB o-s	0.4298	0.6883	0.9133	0.9962	0.3575	0.5654	0.8249	0.9763
χ^2	-	0.0455	0.0509	0.2643	-	0.0432	0.0528	0.1563
A_n	0.4101	0.6723	0.9238	0.9979	0.2935	0.5057	0.8025	0.9768
T_n	0.2663	0.5199	0.8491	0.9924	0.2404	0.4253	0.7277	0.9557
d=5								
DB	0.4564	0.7536	0.967	0.9998	0.3644	0.6282	0.9011	0.9958
DB o-s	0.5922	0.8441	0.9855	1	0.4932	0.7413	0.9463	0.9984
χ^2	-	-	-	0.0487	-	-	-	0.0469
A_n	0.6123	0.8749	0.9921	1	0.421	0.7013	0.9446	0.9996
T_n	0.4676	0.7852	0.9796	1	0.3486	0.6213	0.9076	0.9971
d=20								
DB	0.8639	0.9918	1	1	0.7586	0.9653	0.9996	1
DB o-s	0.9228	0.9966	1	1	0.8493	0.9832	0.9998	1
χ^2	-	-	-	-	-	-	-	-
A_n	0.9819	0.9997	1	1	0.8349	0.9921	1	1
T_n	0.9806	0.9998	1	1	0.8162	0.992	1	1

Table 3: Empirical powers (over 10000 runs) of the uniformity tests under study. The underlying distributions are metauniforms of type MNU and MTU. The significance level is 0.05.

		MNU				MTU			
d=2		25	50	100	200	25	50	100	200
	DB	0.0644	0.0752	0.1068	0.1543	0.0852	0.1128	0.1813	0.3133
	DB o-s	0.0997	0.1193	0.1649	0.2443	0.1323	0.1831	0.2775	0.4267
	χ^2	0.228	0.5016	0.8329	0.9876	0.2306	0.513	0.846	0.9917
	A_n	0.0765	0.0841	0.1043	0.1533	0.0878	0.0928	0.1328	0.1886
	T_n	0.1342	0.2586	0.6715	0.9914	0.1375	0.2565	0.6499	0.9898
d=3									
	DB	0.0948	0.1442	0.2292	0.4145	0.1761	0.2869	0.4995	0.7961
	DB o-s	0.1514	0.2261	0.3325	0.5416	0.2586	0.398	0.6321	0.8758
	χ^2	-	0.8197	0.9908	1	-	0.8184	0.9892	1
	A_n	0.1283	0.1852	0.3114	0.563	0.1705	0.2392	0.3885	0.6357
	T_n	0.2975	0.7111	0.9948	1	0.3091	0.72	0.9912	1
d=5									
	DB	0.1834	0.307	0.5518	0.8457	0.4037	0.6714	0.929	0.9982
	DB o-s	0.2731	0.4203	0.6728	0.9107	0.5279	0.775	0.9621	0.9996
	χ^2	-	-	-	1	-	-	-	1
	A_n	0.4064	0.672	0.941	0.9995	0.4823	0.7299	0.952	0.9989
	T_n	0.7514	0.9954	1	1	0.7647	0.9937	1	1
d=20									
	DB	0.6986	0.9368	0.9988	1	0.983	1	1	1
	DB o-s	0.7963	0.9678	0.9994	1	0.9921	1	1	1
	χ^2	-	-	-	-	-	-	-	-
	A_n	1	1	1	1	1	1	1	1
	T_n	1	1	1	1	1	1	1	1

Table 4: Empirical powers (over 10000 runs) of the uniformity tests under study. The underlying distributions are meta-uniforms MPVII (generated from independent Pearson-type variables), and MCU (generated from independent Cauchy random variables). The significance level is 0.05

		MPVII (N=11, m=2)				MCU			
d=2		25	50	100	200	25	50	100	200
	DB	0.059	0.058	0.0675	0.0935	0.2195	0.3854	0.6474	0.9176
	DB o-s	0.0732	0.0878	0.1033	0.1452	0.3179	0.5122	0.7585	0.958
	χ^2	0.2171	0.5094	0.8474	0.9872	0.223	0.6494	0.9728	1
	A_n	0.0646	0.0598	0.054	0.0655	0.1363	0.1623	0.2129	0.3145
	T_n	0.1425	0.2684	0.6545	0.9889	0.1496	0.2785	0.6394	0.9739
d=3									
	DB	0.0678	0.0767	0.0947	0.1346	0.5525	0.8487	0.9899	1
	DB o-s	0.09	0.1089	0.1404	0.2065	0.6755	0.911	0.9963	1
	χ^2	-	0.8171	0.9894	1	-	0.8216	0.9922	1
	A_n	0.0701	0.0738	0.0814	0.1117	0.2746	0.3936	0.5677	0.8069
	T_n	0.3308	0.6965	0.9884	1	0.3557	0.7305	0.9884	1
d=5									
	DB	0.0737	0.0838	0.1116	0.1833	0.9323	0.9982	1	1
	DB o-s	0.0685	0.0822	0.1102	0.1661	0.966	0.9996	1	1
	χ^2	-	-	-	1	-	-	-	1
	A_n	0.0824	0.0928	0.1348	0.2311	0.6568	0.86	0.9803	0.9997
	T_n	0.6976	0.9779	1	1	0.8133	0.9957	1	1
d=20									
	DB	1	1	1	1	1	1	1	1
	DB o-s	1	1	1	1	1	1	1	1
	χ^2	-	-	-	-	-	-	-	-
	A_n	0.8972	0.9593	0.9956	0.9999	1	1	1	1
	T_n	1	1	1	1	1	1	1	1

Table 5: Empirical powers (over 10000 runs) of the uniformity tests under study. The underlying distributions are contaminated uniforms. The significance level is 0.05.

		$0.9U([0,1]^d) + 0.1U(C_d)$				$0.8U([0,1]^d) + 0.2U(C_d)$			
d=2		25	50	100	200	25	50	100	200
	DB	0.0737	0.0967	0.1425	0.2464	0.1431	0.2362	0.4386	0.7359
	DB o-s	0.113	0.1478	0.2207	0.351	0.2203	0.341	0.5633	0.8243
	χ^2	0.0421	0.0526	0.0588	0.08	0.0432	0.062	0.0848	0.1869
	A_n	0.0892	0.1043	0.1391	0.2322	0.1591	0.2332	0.3885	0.6553
	T_n	0.0633	0.0751	0.0981	0.16	0.0973	0.1426	0.267	0.5102
d=3									
	DB	0.0735	0.0951	0.1414	0.2386	0.1386	0.2369	0.4381	0.7378
	DB o-s	0.115	0.1521	0.2175	0.341	0.2142	0.3412	0.5566	0.8269
	χ^2		0.0458	0.0478	0.0516	0.0465	0.0479	0.0701	
	A_n	0.0873	0.0953	0.1253	0.1943	0.1378	0.1979	0.3377	0.576
	T_n	0.0626	0.0715	0.0912	0.1367	0.085	0.129	0.2316	0.4396
d=5									
	DB	0.0768	0.0926	0.1413	0.2434	0.1385	0.2461	0.4359	0.7368
	DB o-s	0.1215	0.1505	0.219	0.3439	0.2159	0.3489	0.5577	0.8267
	χ^2	-	-	-	0.0508	-	-	-	0.0528
	A_n	0.0913	0.0876	0.1101	0.1534	0.1201	0.1677	0.2617	0.4501
	T_n	0.0638	0.065	0.0837	0.1097	0.0826	0.1182	0.1855	0.342
d=20									
	DB	0.073	0.097	0.1394	0.2451	0.1352	0.248	0.4356	0.73
	DB o-s	0.1145	0.1536	0.2189	0.3478	0.2148	0.3502	0.5588	0.82
	χ^2	-	-	-	-	-	-	-	-
	A_n	0.1026	0.0891	0.0839	0.0912	0.1178	0.1187	0.1255	0.1831
	T_n	0.062	0.0651	0.0639	0.0773	0.0743	0.0876	0.0969	0.142

b) The DB test does not involve any asymptotic approximation as it rests on the well-established Kolmogorov-Smirnov test. The underlying ideas are closely related to some concepts borrowed from computational geometry, mathematical morphology or set estimation theory. In fact, this geometrical approach is on the basis of the remarkable simplicity and the distribution-free property of the DB procedure.

c) From an applied point of view, one-sided versions of the test could be helpful to incorporate some prior knowledge about the particular problem at hand.

d) The simulation study that we have carried out shows that the test has an acceptable behaviour in terms of power and significance level when compared with competing alternative tests. In particular, the DB test is outperformed (in terms of power) by an asymptotic test proposed by Liang *et al.* (2001) for some alternatives (including those of beta or meta-uniform type) but it is more powerful for the case of contaminated (mixture) models and it is also better with regard to the preservation of the nominal significance level.

e) Using non-parametric set estimation techniques, it is possible to specify a method for the difficult problem of testing multivariate uniformity on an unknown compact support.

APPENDIX

Proof of Theorem 1. (a) Let us first note that condition (1) entails $\lambda_r = 1 - r/R$. To see this let us note that $\lambda_r R$ and $R - r$ are the distances from 0 to the boundaries of $\lambda_r \mathcal{S}$ and $\mathcal{S} \ominus rB$, respectively. Now, the result follows from a direct calculation: If μ_L stands for the Lebesgue measure on $\mathbb{R}^d$,

$$H(y) = 1 - \mathrm{P}(D(X, \partial \mathcal{S}) > Ry) = 1 - \frac{\mu_L(\lambda_{Ry}\mathcal{S})}{\mu_L(\mathcal{S})} = 1 - \lambda_{Ry}^d = 1 - (1-y)^d.$$

(b) In this case we have that, for all $0 \leq x \leq (R-r)/2$,

$$\mathrm{P}(D(X, \partial \mathcal{S}) > x) = \frac{\mu_L \left(\lambda_x \mathcal{S}_1 \right) - \mu_L \left(\lambda_{R-(r+x)} \mathcal{S}_1 \right)}{\mu_L(\mathcal{S})} = \frac{(R-x)^d - (r+x)^d}{R^d - r^d}. \qquad (5)$$

Now, for all $0 \leq y \leq 1$, $\mathrm{P}(Y > y) = \mathrm{P}(D(X, \partial \mathcal{S}) > (R-r)y/2)$. Then, the result follows easily from (5) with $x = (R-r)y/2$.

Proof of Theorem 2. Assume first that $d = 2$. Let $S \subset \mathbb{R}^2$ be a polygon circumscribing a circle of centre 0 and radius R, and let $r \in (0, R)$. Then, $S \ominus rB$ also circumscribes a circle with centre 0 but now with radius $R - r$. We will show that for any $x \in \partial(S \ominus rB)$ there exists $x_0 \in \partial S$ such that $x = \lambda_r x_0$, where $\lambda_r = 1 - r/R$. This fact implies that $\partial(S \ominus rB) = \lambda_r \partial S$ and hence the statement holds.

Consider the ray ℓ with origin 0 that meets the point x, and let x_0 be the intersection between ℓ and ∂S. Take the ray ℓ' with origin 0 that intersects both ∂S and $\partial(S \ominus rB)$ in the same sides as ℓ does but containing the tangency points (see Figure 2).

Denote by y and y_0 these tangency points, that is, the intersections of ℓ' with $\partial(S \ominus rB)$ and ∂S respectively. Observe that the triangles $\triangle x0y$ and $\triangle x_0 0 y_0$ are similar since their three angles are equal. Therefore, we have the following relationship:

$$\frac{\|y\|}{\|y_0\|} = \frac{\|x\|}{\|x_0\|}. \qquad (6)$$

Since $\|y_0\| = R$ and $\|y\| = R - r$, equation (6) implies $\|x\| = \lambda_r \|x_0\|$. Since both x and x_0 belong to ℓ, it follows that $x = \lambda_r x_0$.

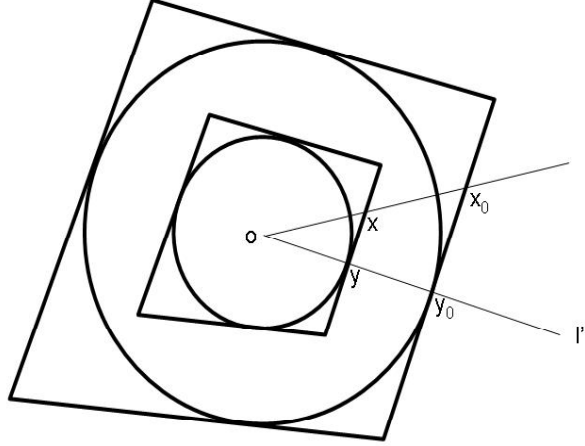For $d > 2$, the result follows from an analogous argument, but now considering the hyperplane defined by the rays ℓ and ℓ'.

Figure 2: The similarity argument in the proof of Theorem 2

Proof of Proposition 1. Let c be the value of the uniform density at $\mathcal{S}$. Assume for instance that there exists $x_0 \in \mathcal{S}$ such that $g(\|x_0\|) < c$. If we denote $\epsilon = c - g(\|x_0\|)$, by the continuity of g we

have that there is a neighbourhood of $\|x_0\|$ where the values of g are, say, smaller than $c - \epsilon/2$. Therefore, there exists a "circular crown" containing x_0 whose probability is lesser than that corresponding to the uniform distribution, which contradicts the assumption of crown-uniformity.

ACKNOWLEDGEMENTS

REFERENCES

Apostol, T. M. and Mnatsakanian, M. A. (2004). Figures circumscribing circles. *American Mathematical Monthly*, 111, 853–863.

Baringhaus, L. (1991). Testing for spherical symmetry of a multivariate distribution. *The Annals of Statistics*, 19, 899–917.

Cuevas, A. and Rodríguez-Casal, A. (2004). On boundary estimation. *Advances in Applied Probability*, 36, 340–354.

Deheuvels, P. (1983). Strong bounds for multidimensional spacings. *Z. Wahrsch. verw. Gebiete*, 62, 465–474.

Devroye, L. and Wise, G. L. (1980). Detection of abnormal behavior via nonparametric estimation of the support. *SIAM Journal of Applied Mathematics*, 38, 480–488.

Diggle, P. J. (2003). *Statistical Analysis of Spatial Point Patterns, 2nd edition*, Edward Arnold, London.

Hall, P., Marron, J. S. and Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society Series B*, 67, 427–444.

Janson, S. (1987). Maximal spacings in several dimensions. *The Annals of Probability*, 15, 274-280.

Justel, A., Peña, D. and Zamar, R. (1997). A multivariate Kolmogorov-Smirnov test of goodness of fit. *Statistics & Probability Letters*, 35, 251–259.

Liang, J. J., Fang, K. T., Hickernell, F. J. and Li, R. (2001). Testing multivariate uniformity and its applications. *Mathematics of Computation*, 70 , 337–355.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57, 519–530.

Moller, J. and Waagepetersen, R. P. (2004). *Statistical Inference and Simulation for Spatial Point Processes*. Chapman & Hall/CRC, Boca Raton.

Reschehofer, E. and Bomze, I. M. (1991). Length tests for goodness of fit. *Biometrika*, 78, 207–216.

Ripley, B. D.. (1979). Tests of 'randomness' for spatial point patterns. *Journal of the Royal Statistical Society Series B*, 39, 172–212.

Rukhin, A. L. (2001). Testing randomness: a suite of statistical procedures. *Theory of Probability and its applications*, 45, 111–132.

Silverman, B. (1986). *Density estimation for statistics and data analysis*. Chapman & Hall, London.

Székely, G. J. and Rizzo, M. L. (2005). A new test for multivariate normality. *Journal of Multivariate Analysis*, 93, 58–80.

Zuo, Y. and Serfling, R. (2000). General notions of statistical depth function. *The Annals of Statistics*, 28, 461–482.

José R. BERRENDERO: `joser.berrendero@uam.es`
Departamento de Matemáticas
Universidad Autónoma de Madrid
28049 Madrid, Spain

Antonio CUEVAS: `antonio.cuevas@uam.es`
Departamento de Matemáticas
Universidad Autónoma de Madrid
28049 Madrid, Spain

Francisco VÁZQUEZ-GRANDE: `vazquez-grande@uchicago.edu`
Graduate School of Business
University of Chicago
Chicago, Illinois 60637 USA