Boletín de Estadística e Investigación Operativa
Vol. 25, No. 2, Junio 2009, pp. 71-85

Estadística

Set estimation: Another bridge between statistics and geometry

Antonio Cuevas
Departamento de Matemáticas, Facultad de Ciencias
Universidad Autónoma de Madrid
✉ antonio.cuevas@uam.es

Abstract

Set estimation has to do with the statistical reconstruction of sets from random set of points. This theory is closely related with nonparametric functional estimation as well as with stochastic geometry. A non-exhaustive expository overview of set estimation theory is given. The aim is to present the basic ideas, some typical tools involved in the theory and a few applications. Most technicalities are omitted or summarized.

Keywords: Support estimation, level set estimation, boundary estimation.
AMS Subject classifications: 62G07, 62G20.

1. Introduction

The use of geometry tools is customary in different areas of statistics (multivariate regression, classification, clustering, principal component analysis, spatial statistics, stereology, study of parametric families, etc.). This paper is devoted to *set estimation*, a relatively young chapter of the modern mathematical statistics where geometry plays a essential role, not only as a source of auxiliary tools but also as a major motivation. Roughly speaking, this theory deals with the estimation of a set $S \subset \mathbb{R}^d$ in the Euclidean space from a random sample of points $X_1 \ldots, X_n$.

In a way, set estimation is the geometric counterpart of the classical theory of nonparametric functional estimation (e.g., Simonoff, 1996). In both theories the estimators typically depend on a sequence of smoothing parameters, the theoretical results make special emphasis on asymptotic properties, especially consistency and convergence rates, and the overall aim is to get results as general as possible in the sense that they typically hold under very general conditions (for example, no normality assumption is made) for a general dimension d. The main differences with respect to nonparametric functional estimation are related to the much stronger geometrical motivation behind set estimation. Since in this theory the target is estimating sets, rather than functions, it is natural that the distances between sets, as well as the geometrical conditions concerning their shapes, play here an important role. As a consequence, set estimation is in the intersection of nonparametric statistics, stochastic geometry and geometric measure theory.

The aim of this work. Our purpose here is to provide a short overview, by no means exhaustive, of the state of the art in set estimation theory. The style will be mainly expository, in order to convey the main ideas and some applications. Therefore, most technical aspects are omitted or briefly summarized and no attempt is made to give a complete bibliography. For a much more detailed survey the reader is referred to Cuevas and Fraiman (2009).

2. What is set estimation about?

We will outline here some typical problems of interest in set estimation, in particular those concerning the estimation of the support, the level sets, the support boundary and some related functionals.

Some notation. In what follows $X_1, \ldots, X_n$ will denote a sample of a $\mathbb{R}^d$-valued random variable X, whose distribution will be denoted by P_X. When X is assumed to be absolutely continuous, f will stand for the corresponding density. The Lebesgue measure will be represented by μ. The closed ball with center x and radius r will be denoted by $B(x,r)$ and, for $S \subset \mathbb{R}^d$, $B(S,\epsilon)$ will stand for the parallel set $B(S,\epsilon) = \cup_{x \in S} B(x,\epsilon)$. When convenient, the set of sample points will be denoted by $\aleph_n$.

Support estimation. Some simple estimators. This is perhaps the simplest, more direct, set estimation problem: Here the target S is the support of the distribution P_X in $\mathbb{R}^d$. We want to approximate S from a sample $X_1, \ldots, X_n$ of random observations drawn inside S. The question of how to construct a suitable estimator of S has a quite natural response if S is assumed to be convex. Then the *convex hull* of the sample, $S_n = \text{conv}(\aleph_n)$ provides a simple estimator. This is just the intersection of all convex sets including $\aleph_n$. It provides always an estimator "from the inside" but this bias is not important from an asymptotical point of view.

The estimate $S_n = \text{conv}(\aleph_n)$ has received attention in the literature since at least fifty years ago. See e.g., Schneider (1988) and references therein. See also Dümbgen and Walther (1996) and Reitzner (2003) for more recent references. Clearly, the assumption of convexity is very restrictive for many purposes but still the convex hull is an important set estimator not only for historical reasons. As we will see below some of the key ideas in set estimation are borrowed from the field of convex analysis.

If S is not convex there are still several simple estimators which do the job under very general conditions on S. For example the Devroye and Wise (1980) estimator (see also Chevalier, 1976, Korostelev and Tsybakov, 1993 and Cuevas and Rodríguez-Casal, 2004),

$$S_n = \bigcup_{i=1}^{n} B(X_i, \epsilon_n), \tag{2.1}$$

is just a sort of "dilated" version of the sample $\aleph_n$, where ϵ_n is a sequence of smoothing parameters which must tend to zero, but not too quickly ($n\epsilon_n^d \to \infty$) in order to get a consistent estimation as $n \to \infty$ (see Devroye and Wise, 1980).

Figure 1 shows the the convex hull (left) and the Devroye-Wise estimator for the same random sample of size $n = 40$. Not surprisingly, both estimators have very different appearances as the first one incorporates the assumption of convexity.

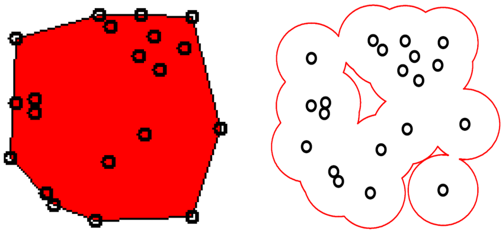

Figure 1.- Support estimation using the convex hull and Devroye-Wise estimator

Other more sophisticated estimators are considered, e.g., in Cuevas and Fraiman (1997) relying on the use of an auxiliary non-parametric density estimator (e.g., of kernel type; see Simonoff, 1996). If f is the underlying density, the support coincides essentially (at least in regular cases) with the set $\{f > 0\}$. So one could think of estimating the support by using a sequence $S_n = \{f_n > c_n\}$ where c_n is a suitable numerical sequence $c_n \downarrow 0$. Note that if the underlying support is compact and the auxiliary estimators (as it is often the case) have an unbounded support the simplest choice $c_n = 0$ would be in general inappropriate.

Level set estimation. The plug-in approach. In many cases the support S of the underlying density is not of special interest since an important part of S is "almost empty" from the probability point of view. In other words if the distribution is absolutely continuous with density f, the areas of the support where f is very close to zero are usually of lesser interest for many practical purposes since the probability of finding points there is extremely low. In these situations it could make sense to consider c-level sets of type $L(c) = \{f \geq c\}$ (where $c > 0$ is a given constant) that might be considered as the "substantial support". Thus, the estimation of density level sets (or even regression level sets where f is replaced for a regression function) is another typical concern in set estimation.

The estimators of *plug-in type* are a natural choice in this problem. They have the general form $\{f_n \geq c\}$, where f_n is a nonparametric estimator of f. See, e.g., Cadre (2006) for a recent deep study on this class of estimators. Another approach, especially suitable for "smooth" level sets is proposed in the interesting paper by Walther (1997).

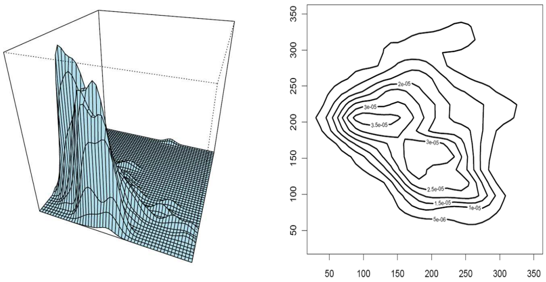

Figure 2.- A bivariate density and its level sets.

Boundary estimation. To estimate a set is not exactly the same as estimating its boundary. This is a more difficult task. To see this note that, depending on the considered criterion (see subsection 3.1 below), the sample $\aleph_n = \{X_1, \ldots, X_n\}$ itself is a consistent estimator of the support S but in general it is not a satisfactory boundary estimator in the sense that the boundary of $\aleph_n$ will not approach that of S. However, the estimation of the boundary is perhaps the most interesting task in many instances, for example in problems related to image analysis. It is intuitively clear that the boundary ∂S_n of the Devroye-Wise estimator (2.1) could provide a good estimation of ∂S provided that S is not too complicate and the sequence of smoothing parameters $\{\epsilon_n\}$ converges to zero slowly enough. This problem has been considered in Cuevas and Rodríguez-Casal (2004); see subsection 3.3 below.

Estimation of support functionals. In some cases the interest is focussed on a specific functional $\Phi(S)$ of the support S (e.g., the Lebesgue measure, the gravity center, the boundary measure of S,...). A recent example is the work by Cuevas, Fraiman and Rodríguez-Casal (2007) on the estimation of the measure of the boundary of S, as defined by the so-called Minkowski content. See also Cuevas and Fraiman (2009) for further details and references.

3. The mathematical setup

We briefly summarize, in a few schematic features, the main mathematical tools and results involved in set estimation.

3.1. Distances between sets

So far, most typical results in set estimation are of consistency type:

$$D(S_n, S) \to 0, \text{ as } n \to \infty, \text{ either in probability or almost surely,}$$

where D is a suitable distance between sets. Even more interesting are the results on the rate (or speed) of convergence of S_n towards S, of type

$$D(S_n, S) = \mathcal{O}_P(R_n), \text{ or } D(S_n, S) = \mathcal{O}(R_n), \text{with probability one,}$$

where R_n is a numerical sequence $R_n \uparrow \infty$ and, as usual, the notation $\mathcal{O}$ is used to represent the convergence order while $\mathcal{O}_P$ denotes "order in probability".

There are two main distances used for this type of result. The first one is the distance "in measure" defined, for bounded Borelian sets, by

$$d_\mu(S_n, S) = \mu(S_n \Delta S),$$

where the symbol Δ denotes the usual symmetric difference and the Lebesgue measure could be replaced with another measure of interest in the problem.

Another typical distance is the *Hausdorff metric*, defined by

$$\begin{aligned} d_H(S_n, S) &= \inf\{\epsilon > 0, S \subset B(S_n, \epsilon), S_n \subset B(S, \epsilon)\} \\ &= \max\left\{\sup_{x \in S_n} \inf_{y \in S} \|x - y\|, \sup_{y \in S} \inf_{x \in S_n} \|x - y\|\right\}. \end{aligned}$$

This distance reflects a different, more "visual" notion of closeness between compact sets. It has been often used in image analysis and fractals theory.

3.2. Geometric conditions that define nice sets

From the geometrical point of view, the family of bounded Borelian sets in $\mathbb{R}^d$ is a "monters parade" which includes extremely complicate hard-to-imagine sets. There is little hope to properly reconstruct most of these strange sets using statistical methods, unless we are willing to accept very rough approximations which do not identify many relevant features. So, there is a natural need to impose some conditions oriented to identify different classes of "nice" sets for which the statistical approximations are better suited.

In this section we list and briefly comment some of these properties. All of them have a clear geometric and intuitive character together with some deep mathematical implications. All of them have, we think, some independent interest besides its application to set estimation.

Standardness. This restriction typically arises (sometimes under slightly different versions and names) in set estimation and stochastic geometry. In intuitive terms it establishes that every ball of small enough radius centered at a point of S has at least a fixed proportion of it inside S. Formally, S is μ-standard if there exist $\epsilon_0 > 0$ and $\delta > 0$ such that for all $x \in S$ and $\epsilon \leq \epsilon_0$

$$\mu(B(x,\epsilon) \cap S) \geq \delta\mu(B(x,\epsilon)). \tag{3.1}$$

There are different versions of this property: For example it can be imposed "from the outside", by replacing S with the complement S^c in (3.1); also, the measure μ in the left-hand side of (3.1) could be replaced with another suitable measure (e.g., the distribution P_X of the random variable which generates the sample). The "house-shaped" set of Figure 3 (left) is standard; the set shown in the right is not standard as the condition fails in the sharp "non-linear" peaks.

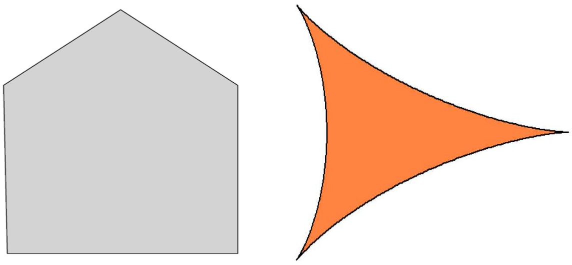

Figure 3.- The "house" fulfills standardness assumption, the "spiky" set on the right doesn't

The assumption of standardness appears often in set estimation. For example, in order to study the d_H-convergence rate at which the d-variate sample $\aleph_n = \{X_1, \ldots, X_n\}$ tends to the (compact) support S, let us take any $\epsilon > 0$ and assume that S is P_X-standard. Consider a minimal covering $\mathcal{C}(\epsilon) = \{B_1, B_2, \ldots\}$ of S (with cardinality $N(\epsilon)$) by closed balls of radius ϵ with centers in S. Then,

$$\begin{aligned}
&\mathbb{P}\{d_H(\aleph_n, S) > 2\epsilon\} \leq \mathbb{P}\{\text{there is a ball in } \mathcal{C}(\epsilon) \text{ with no sample point inside}\} \\
&= \mathbb{P}\left\{\cup_{k=1}^{N(\epsilon)} \cap_{i=1}^{n} \{X_i \notin B_k\}\right\} \leq N(\epsilon) \max_k (1 - \mathbb{P}\{X_1 \in B_k\})^n \\
&= N(\epsilon)(1 - \min_k \mathbb{P}\{X_1 \in B_k\})^n \overset{(*)}{\leq} N(\epsilon) \exp(-n\delta\omega_d\epsilon^d) \\
&\overset{(**)}{\leq} C\epsilon^{-d} \exp(-n\delta\omega_d\epsilon^d).
\end{aligned} \tag{3.2}$$

Inequality (*) follows from the standardness property $\mathbb{P}\{X \in B_k\} = P_X(B_k) \geq \delta\mu(B_k) = \delta\omega_d\epsilon^d$, where $\omega_d = \mu(B(0,1))$, together with the inequality $(1-x)^n \leq \exp(-nx)$, for $0 \leq x \leq 1$. For (**) we have used the fact that $N(\epsilon) \leq C\epsilon^{-d}$, where C is a constant which depends on the set S.

This conclusion is useful in several aspects. First, it provides a qualitative assessment for the proximity between $\aleph_n$ to S. Second (and more importantly), it helps to get convergence rates, (in probability or almost surely) for $d_H(\aleph_n, S)$. If we show $R_n d_H(\aleph_n, S) \to 0$ (either almost surely or in probability) for a numerical sequence $0 < R_n \uparrow \infty$ we may ensure that $d_H(\aleph_n, S)$ goes to zero (a.s. or in prob.) at a rate faster than R_n^{-1}. Thus, under the assumed standardness condition, the bound (3.2) allows us to directly show that $R_n d_H(\aleph_n, S) \to 0$, in probability, for any $R_n = n^q$ with $0 < q < 1/d$. Finally, a bound of type (3.2), based on a similar reasoning, is also involved in the obtention of convergence rates for the simple estimator (2.1) and other usual set estimators.

Rolling property. This is a smoothness property which is established in purely geometrical terms, without any explicit use of differentiation concepts. A closed set S is said to fulfill the (inner) rolling property if "a ball can roll freely along the boundary, from inside, having contact with all the boundary points". In formal terms: There exists $r > 0$ such that for all $x \in \partial S$ there is a ball $B(a, r)$ such that $x \in B(a, r)$ and $B(a, r) \subset S$. Of course, an "outer rolling property" could be similarly defined by imposing the above condition on $\overline{S^c}$.

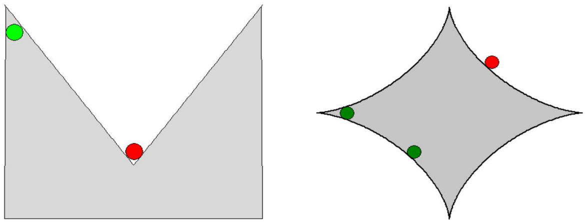

Figure 4.- A set with no rolling property (left) and with the outer rolling property (right).

It is clear that the rolling property must have an interpretation in terms of differentiability properties. This is done in Walther (1999). As for its use in set estimation we refer to Walther (1997) which provides fast convergence rates for the estimation of level sets under a rolling property, and Cuevas and Rodríguez-Casal (2004) where this property is used to establish an inequality of type $d_H(\partial S, \partial B(S, \epsilon)) \leq \epsilon$, for ϵ small enough, which is particularly useful in the problem of boundary estimation.

Reach condition. The reach of a closed set $S \subset \mathbb{R}^d$ is defined as the largest (possibly ∞) value r_0 such that if $x \in \mathbb{R}^d \setminus S$ and the distance from x to S is smaller than r_0, then S contains a unique point nearest to x. Every compact convex set $S \subset \mathbb{R}^d$ has positive reach, in fact for these sets reach$(S) = \infty$. So the positive reach condition can be seen as a generalization of convexity. This situation is quite common in set estimation and, more generally, in stochastic geometry: Although convexity is a too restrictive property for many purposes, it is also an extremely rich and natural condition which can be useful in different ways just identifying some key features of convex sets and defining the respective classes of sets which fulfill these properties. This leads to define different types of generalizations of convex sets (star-shaped sets, sets with positive reach, etc...) which are still "intuitive" and easy to handle.

This condition is closely related to the rolling property as well as with the α-convexity considered below. See Pateiro-López (2008, Appendix A) for details.

The reach condition was introduced by Federer (1959) to get an interesting generalization of Steiner's theorem: This result establishes that for a compact convex set, the "volume function" $V(\epsilon) = \mu(B(S, \epsilon))$ is a polynomial in ϵ of order d. Federer (1959) has proved that if S is a bounded closed set with $\text{reach}(S) = r_0 > 0$ then $V(\epsilon)$ is a polynomial of order d for $\epsilon \in [0, r_0)$. This assumption on the structure of the parallel set is useful sometimes in set estimation (see Cuevas, Fraiman and Rodríguez-Casal, 2007).

The set on the left of Figure 4 does not satisfy the positive reach condition since all the points on an upper vertical half-line with origin in the middle-vertex have two projections on the set.

α-convexity. As mentioned above, the wealth of interesting properties of the convex sets is a continuous source of inspiration in order to identify "nice" classes of sets for different purposes. Thus a useful class of sets is obtained by recalling that a closed convex set S can be obtained as the intersection of the halfspaces which contain S. Now, if we replace the halfspaces with the complements of balls of radius α we get the following definition: A closed set A is said to be α-convex if $A = C_\alpha(A)$, where

$$C_\alpha(A) = \bigcap_{\text{int}(B(x,\alpha)) \cap A = \emptyset} (\text{int}(B(x, \alpha)))^c$$

is called the *α-convex hull* of A. It can be seen that a closed convex set is α-convex for all $\alpha > 0$. The reciprocal is true when A has non-empty interior (see Walther, 1999).

The interesting fact is that an α-convex support S has a natural hull-type estimator from a sample $\aleph_n = \{X_1, \ldots, X_n\}$ which is just the *α-convex hull* of the sample points, $S_n = C_\alpha(\aleph_n)$. The properties of this estimator have been recently studied by Rodríguez-Casal (2007) and Pateiro-López (2008).

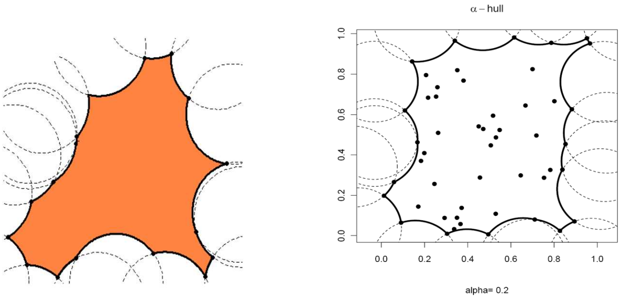

Figure 5.- α-convex set (left), α-convex hull of a random sample (right)

3.3. Some results

We will summarize here a few typical results and methods which fit the scope of this limited survey. They have not been selected according to any hierarchy of importance, as other similar results could have been selected instead. The only purpose is to convey, through a few flashes, a general idea on the atmosphere in set estimation. Many technical details will be omitted.

On convergence rates for particular classes. To our knowledge, the first monograph on set estimation with a nonparametric approach is the book by Korostelev and Tsybakov (1993) which includes a survey of the theory and a compilation of the authors' work on the subject. In particular, the estimator (2.1) is studied in that book under the title "A simple and rough support estimator" (section 7.2). It is shown there that if the X_i are uniform on S and S belongs to a certain class $\mathcal{G}$ of sets with piecewise Lipschitz boundaries, a suitable choice of the smoothing parameters ϵ_n gives a uniform convergence rate of type

$$\sup_{S\in\mathcal{G}} \mathbf{E} d_\mu(S_n, S) = \mathcal{O}\left(\frac{\log n}{n}\right)^{1/d}. \tag{3.3}$$

This kind of results are typical in set estimation: If the target set is nice enough we can guarantee a certain convergence rate. That obtained in (3.3) is also quite usual: If we ignore the logarithm factor, the rate is essentially $n^{-1/d}$, which deteriorates quickly as d increases (this is the so-called "curse of dimensionality").

A result of consistency with rates for boundary estimation. The Devroye-Wise estimator $S_n = \cup_i B(X_i, \epsilon_n)$ is also studied in Cuevas and Rodríguez-Casal (2004). This authors consider the natural problem of finding conditions under which the boundary of S_n will approximate the boundary of S. It is clear that a condition on the minimum size of ϵ_n is required. The precise answer given in that paper (Prop. 1 and Th. 4) is as follows:

Assume that S is standard with respect to P_X (i.e., $P_X(B(x,\epsilon)) \geq \delta\mu(B(x,\epsilon))$ for $x \in S$ and ϵ small enough). Assume also that S fulfills the (outer) rolling property, and

$$\epsilon_n = C\left(\frac{\log n}{n}\right)^{\frac{1}{d}}, \text{ for some } C > \left(\frac{2}{\delta\omega_d}\right)^{\frac{1}{d}}, \ \omega_d = \mu(B(0,1)).$$

Then, with probability one, $d_H(\partial S_n, \partial S) \leq \epsilon_n$ and $d_H\left(S_n, S\right) \leq \epsilon_n$, eventually.

An image analysis model with auxiliary variables. Mammen and Tsybakov (1995) consider a model with auxiliary variables for $S \subset [0,1]^d$. The sample data are of type (X_i, Y_i), $i = 1, \ldots, n$ with $Y_i = (2\mathbf{1}_{\{X_i \in S\}} - 1)\xi_i$, where the ξ_i are i.i.d. random variables, independent from the X_i, taking values 1 and -1 with probabilities $1/2 + a_n$ and $1/2 - a_n$, respectively, a_n being a sequence with $0 < a_n < 1/2$. The value Y_i is interpreted as the image level (in the grayscale) at the point X_i; the black is codified by $Y_i = 1$ and the white for $Y_i = -1$; $\mathbf{1}_A$ stands for the indicator function of A.

The paper is mainly concerned with the best attainable rates rather than with the explicit construction of the optimal estimates. However, the information on optimal rates is still interesting for practitioners as it sheds some light on the intrinsic nature of the problem at hand. More specifically, assuming that ∂S has a smooth parametrization, Mammen and Tsybakov (1995) derive the expression of the asymptotically optimal d_μ-rates for the estimation of S, which depend on the smoothness degree of the boundary parametrization. As these authors point out, the requirement of smooth parametrization does not entail that the boundary itself is smooth. So the results are quite general. They extend those obtained in Korostelev and Tsybakov (1993) for classes of sets defined as the hypograph of a smooth function (these sets are sometimes called "boundary fragments").

The excess-mass methodology. This is an important idea in level set estimation, useful to incorporate into the estimators shape restrictions on the nature of the target (density) level set $L(c) = \{f \geq c\}$. It relies in the observation (see Hartigan, 1987) that the functional $H_c(B) = \int_B (f(x) - c)dx = P(B) - c\mu(B)$ is maximized by the level set $L(c)$. Then if $\mathcal{B}$ is a given class of sets, a natural estimator $L_n(c)$ of $L(c)$ under the shape restriction $L(c) \in \mathcal{B}$ would be the maximizer on $\mathcal{B}$ of the *empirical excess mass* $H_{c,n}(B) = P_n(B) - c\mu(B)$. Hartigan (1987) considered the case where $\mathcal{B}$ is the class of convex sets and proposed and algorithm involving $\mathcal{O}(n^3)$ steps to obtain $L_n(c)$.

Asymptotic results for the estimator restricted to much more general classes $\mathcal{B}$ are given by Polonik (1995) using the empirical process theory.

4. Some connections with other topics

Set estimation has an "indirect" application in the problem of supervised classification where the optimal solution is essentially given by a (regression) level set; see e.g., the survey by Cuevas and Fraiman (2009) for more details and references. However, we will limit ourselves to some situations in which the estimation of a set is a first natural step in the data analysis.

Exploratory data analysis. Clustering. In some cases, set estimation can be helpful in the visualization of a large mass of data. For example, in Cuevas, González-Manteiga and Rodríguez-Casal (2006) it is considered the analysis of a large data set consisting of more than 11000 double stars identified during the period 1989–1993 by the satellite Hipparcos of the European Space Agency. The positions of these stars are projected on the unit sphere of the tri-dimensional space so that in fact their distances are not considered, only the directions are relevant. Thus, we deal with a sample of 11749 points on the unit sphere (this is an example of the so-called "spherical" or "directional" data). Several interesting questions could be raised in connection with this data set. For example, one could ask whether the distribution of the stars is similar in both northern and southern hemispheres. If we want to have a visual idea of the structure of this data set, the simple representation of these points projected on the sphere is not very useful as we only can see a huge mass of points which covers almost completely the sphere surface. Figure 6 (left) shows such a representation for a sample of 500 randomly chosen stars. It can be seen that, even for this reduced data set, such a direct representation does not allow us to draw any precise idea. Level set estimation provides a reasonable alternative. For example one could estimate the level sets $L(c) = \{f \geq c\}$, f being the "true" underlying density on the sphere, for different values of c. The estimates could be of type $L_n(c) = \{f_n \geq c\}$ where f_n denotes a suitable density estimator for spherical data. Figure 6 (right) shows such a level set (for $c = 0.2$). It appears as a few close small areas in the southern hemisphere. No element is found in the northern hemisphere for this value of c. This suggests that the distribution is not identical in both hemispheres (as it is indeed the case). Of course, a much more complete analysis would be needed for a better understanding of this phenomenon but these few hints provide an idea of the usefulness of level set estimation techniques.

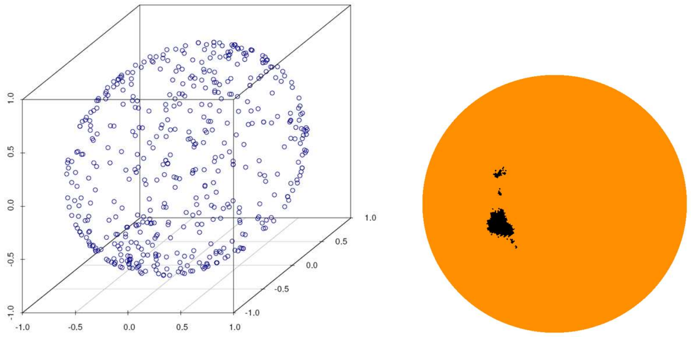

Figure 6.- Positions of 500 double stars in the unit sphere (left) and estimated level set in the southern hemisphere (right).

Another related more specific application is *cluster analysis*. The purpose of this wide methodology is to classify a (usually large) data set into several subsets (clusters) which are relatively different from each other but internally "homogeneous" in the sense that they are made of "similar" observations. The most popular clustering procedure is maybe the k-means algorithm which is based in a distance criterion which tends to provide "globular" clusters defined around suitably chosen "centroids".

Sometimes the nature of the data under study leads to consider clusters of very different shapes, not necessarily globular. This is the case, for example, in astronomy where the analysis of clusters of galaxies is an interesting subject and the shape of the clusters is a relevant fact. Set estimation appears in some "shape oriented" cluster techniques where the starting point is Hartigan's definition of the (population) c-clusters as the connected components of the level set $L(c) = \{f \geq c\}$. These population clusters can be estimated with their empirical counterparts in $L_n(c) = \{f_n \geq c\}$. Finally, the data could be grouped according to the empirical clusters they belong. See Cuevas, Febrero and Fraiman (2000) for a development of these ideas. See also Jang and Hendry (2007) for an interesting application in astronomy.

An application in economics: The efficient frontier problem. One can think that a company transforms some inputs x (capital investments, human resources, etc.) into an output y (capital gains). The efficiency of this company can be measured by the difference $g(x) - y$ between y and the "best attainable output" associated with the input x, which is defined by some function $g(x)$. In practice, $g(x)$ is not exactly known so that it must be estimated from a sample $(x_1, y_1), \ldots, (x_n, y_n)$ corresponding to the performances of n randomly selected companies.

This problem amounts to the estimation of the "upper boundary" of the *hypograph set*

$$S = \{(x, y) : x \in S_0, \, y \leq g(x)\} ,$$

where S_0 is the set of all possible inputs. Note that the sample data are taken in S.

This is a relevant problem in the field of productivity analysis which has motivated a considerable amount of literature since the pioneering paper by Farrell (1957). It is often tackled assuming different monotonicity-type or concavity assumptions on g or convexity on S. See e.g., Simar and Wilson (2000) for more details and references.

Statistical Quality Control. The basic idea is as follows: Assume that we sequentially get independent observations $X_1, X_2, \ldots$ from a d-dimensional random variable X, for example, certain quality characteristics of a manufactured item. This process will be monitored in order to detect a potential change in the distribution of X.

Typically the process is "in control" along a initial period where the observations follow a distribution F, with density f. So we may assume that we have a "pilot" sample $X_1, \ldots, X_n$ of F from a monitoring (in control) period. At some stage, the process may run out of control and the distribution of the X_i's changes to G. The aim is to detect a real change in the distribution of subsequent observations X_{n+k}, $k \geq 1$, as quickly as possible. This is a relevant problem in applied statistics which has received constant attention in the literature. Many tools have been developed for this and related problems, starting from the well-known Shewhart charts. The possible usefulness of set estimation in these quality control or detection problems was first pointed out in Devroye and Wise (1980). In short the idea is to use a level set estimator $S_n = \{f_n \geq c_n\}$, based on the observations $X_1, \ldots, X_n$ and to raise an alarm for X_{n+k} when $X_{n+k} \notin \{f_n \geq c_n\}$. The constant c_n can be chosen in order to approximately get a given probability of false alarm α. Thus, set estimation offers a sort of nonparametric multivariate alternative to the classical Shewhart charts.

Further references and details can be found in Baíllo and Cuevas (2006).

Image analysis. The reconstruction of a set from a random sample of points has some obvious reminiscences from image analysis. For example, one could think of estimating the habitat of a plant or animal species from a sample of elements; see, e.g., De Haan and Resnick (1994). A more recent reference with an image analysis motivation is Ray Chaudhuri et al. (2004).

Acknowledgements

This work has been partially supported with Spanish grant MTM2007-66632.

I am indebted to all my co-authors in set estimation issues and especially to Ricardo Fraiman. I am also grateful to Beatriz Pateiro-López for her help with some figures. The R-package *alphahull* by Pateiro-López and Rodríguez-Casal (2009) has been used in the obtention of Figures 1 and 5.

References

[1] Baíllo, A. and Cuevas, A. (2006). Parametric versus nonparametric tolerance regions in detection problems. *Comp. Stat.*, **21**, 523–536.

[2] Cadre, B. (2006). Kernel estimation of density level sets. *J. Multivariate Anal.*, **97**, 999–1023.

[3] Chevalier, J. (1976). Estimation du support et du contour de support d'une loi de probabilité. *Ann. Inst. H. Poincaré B*, **12**, 339–364.

[4] Cuevas, A., Febrero, M. and Fraiman, R. (2000). Estimating the number of clusters. *Canad. J. Statist.*, **28**, 367–382.

[5] Cuevas, A. and Fraiman, R. (1997). A plug-in approach to support estimation. *Ann. Statist.*, **25**, 2300–2312.

[6] Cuevas, A. and Fraiman, R. (2009). Set estimation. In: *New Perspectives in Stochastic Geometry*, W. Kendall and I. Molchanov, eds. Oxford University Press, Oxford (UK).

[7] Cuevas, A., Fraiman, R. and Rodríguez-Casal, A. (2007). A nonparametric approach to the estimation of lengths and surface areas. *Ann. Statist.*, **35**, 1031–1051.

[8] Cuevas, A., González-Manteiga, W. and Rodríguez-Casal, A. (2006). Plug-in estimation of general level sets. *Aust. N. Z. J. Stat.*, **48**, 7–19.

[9] Cuevas, A. and Rodríguez-Casal, A. (2004). On boundary estimation. *Adv. in Appl. Probab.*, **36**, 340–354.

[10] Devroye, L. and Wise, G. (1980). Detection of abnormal behavior via nonparametric, estimation of the support. *SIAM J. Appl. Math.*, **3**, 480–488.

[11] De Haan, L. and Resnick, S. (1994). Estimating the home range. *J. Appl. Probab.*, **31**, 700–720.

[12] Dümbgen, L. and Walther, G. (1996). Rates of convergence for random approximations of convex sets. *Adv. in Appl. Probab.*, **28**, 384–393.

[13] Farrell, M.J. (1957). The measurement of productive efficiency. *J. Roy. Statist. Soc. Ser. A*, **120**, 253-281.

[14] Federer, H. (1959). Curvature measures. *Trans. Amer. Math. Soc.*, **93**, 418–491.

[15] Hartigan, J.A. (1987). Estimation of a convex density contour in two dimensions. *J. Amer. Statist. Assoc.*, **82**, 267–270.

[16] Jang, W. and Hendry, M. (2007). Cluster analysis of massive datasets in astronomy. *Stat. Comput.*, **17**, 253–262.

[17] Korostelev, A.P. and Tsybakov, A.B. (1993). *Minimax Theory of Image Reconstruction*. Lecture Notes in Statistics 82. Springer-Verlag, New York.

[18] Mammen, E. and Tsybakov, A.B. (1995). Asymptotical minimax recovery of sets with smooth boundaries. *Ann. Statist.*, **23**, 502–524.

[19] Pateiro-López, B. (2008). Set estimation under convexity-type restrictions. Ph. D. Thesis. Universidad de Santiago de Compostela.

[20] Pateiro-López, B. and Rodríguez-Casal, A. (2009). *alphahull*: Generalization of the convex hull of a sample of points in the plane. R package version 0.1.

[21] Polonik, W. (1995). Measuring mass concentration and estimating density contour clusters-an excess mass approach. *Ann. Statist.*, **23**, 855–881.

[22] Ray Chaudhuri, A., Basu, Tan, K., Bhandari, S.K. and Chaudhuri, B.B. (2004). An efficient set estimator in high dimensions: consistency and applications to fast data visualization. *Comput. Vis. Image Und.*, **93**, 260–287.

[23] Reitzner, M. (2003). Random polytopes and the Efron–Stein jackknife inequality *Ann. Prob.*, **31**, 2136–2166.

[24] Rodríguez-Casal, A. (2007). Set estimation under convexity-type assumptions. *Ann. Inst. H. Poincaré Probab. Statist.*, **43**, 763–774.

[25] Schneider, R. (1988). Random approximation of convex sets. *J. Microsc.*, **151**, 211–227.

[26] Simar, L. and Wilson, P. (2000). Statistical inference in nonparametric frontier models: The state of the art. *J. Prod. Anal.*, **13**, 49–78.

[27] Simonoff, J. S. (1996). *Smoothing Methods in Statistics.* Springer-Verlag, New York.

[28] Walther, G. (1997). Granulometric smoothing. *Ann. Statist.*, **25**, 2273–2299.

[29] Walther, G. (1999). On a generalization of Blaschke's Rolling Theorem and the smoothing of surfaces. *Math. Methods Appl. Sci.*, **22**, 301–316.

About the author

Antonio Cuevas is a Professor in the Mathematics Department of Universidad Autónoma de Madrid. His research interests include non-parametric statistical methodology (in particular, set estimation), statistics with functional data and classification methods.