Journal of Multivariate Analysis 147 (2016) 82–101

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Local convex hull support and boundary estimation

C. Aaron *, O. Bodart

Laboratoire de Mathématiques CNRS (UMR 6620) Université Blaise Pascal (Clermont-Ferrand 2), 63177 Aubière cedex, France

ARTICLE INFO

Article history:
Received 10 June 2015
Available online 23 January 2016

AMS 2000 subject classifications:
62G05
62G20
62H99

Keywords:
Convex-Hull
Polyhedron
Support estimation
Topological data analysis
Geometric inference

ABSTRACT

In this paper we introduce a new estimator for the support of a multivariate density. It is defined as a union of convex hulls of observations contained in balls of fixed radius. We study the asymptotic behavior of this "local convex hull" for the estimation of the support and its boundary. When the support is smooth enough, the proposed estimator is proved to be, eventually almost surely, homeomorphic to the support. Numerical simulations on simulated data illustrate the performance of our estimator.

1. Introduction

Let $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ be a set of n independent and identically distributed (i.i.d.) random variables with probability density f defined on $\mathbb{R}^d$. Let

$$S = \overline{\{x \in \mathbb{R}^d, f(x) > 0\}} \subset \mathbb{R}^d,$$

be the support of the probability density f. We aim at building an estimator of S and of its boundary ∂S, based on the set of observations $\mathcal{X}_n$.

Throughout this article, we will make use of the classical topological notations: $\overline{A}$, $\mathring{A}$, $\partial A = A \setminus \mathring{A}$ and A^c will denote respectively the closure, the interior, the boundary and the complementary set of a set $A \subset \mathbb{R}^d$. Moreover, $\mathcal{H}(A)$ and $|A|$ will denote respectively the convex hull and the Lebesgue measure of A.

Given an estimator $\hat{S}_n$ of the support S, its quality can be quantified by the study of the following errors:

1. the measure of the symmetric difference $|\hat{S}_n \,\Delta\, S|$, where:

$$|A \,\Delta\, B| = |(A \cap B^c) \cup (A^c \cap B)|,$$

2. the Hausdorff distance between $\hat{S}_n$ and S, $d_H(\hat{S}_n, S)$, where:

$$d_H(A, B) = \max(\max_{a\in A}(\min_{b\in B} \|a - b\|), \max_{b\in B}(\min_{a\in A} \|a - b\|)),$$

where $\|a - b\|$ denotes the Euclidean distance between a and b.

* Corresponding author.
E-mail addresses: catherine.aaron@math.univ-bpclermont.fr (C. Aaron), olivier.bodart@math.univ-bpclermont.fr (O. Bodart).

http://dx.doi.org/10.1016/j.jmva.2016.01.003

The first criterion is the most commonly used for the estimation of supports. However, when it comes to evaluate the quality of the estimation of ∂S, only the Hausdorff distance $d_H(\partial \hat{S}_n, \partial S)$ between $\partial \hat{S}_n$ and ∂S is relevant.

The most intuitive and simple way to estimate S was introduced by Devroye and Wise. In [13] (see also [9]), the following estimator is proposed:

$$\hat{S}_r = \bigcup_1^n \mathcal{B}(X_i, r), \tag{1}$$

where $\mathcal{B}(X, r)$ denotes the closed ball centered in x and of radius $r > 0$. The properties of this estimator have been widely studied in the context of the estimation of both S and ∂S.

In [13], for a sequence of radii $r_n \to 0$ such that $nr_n^d \to \infty$, $\hat{S}_{r_n}$ has been proved to be universally consistent with respect to the symmetric difference.

In [1], assuming standardness on S and f (see Definition 2), a rate of convergence of order $(\ln n/n)^{1/d}$ for the symmetric difference criterion is established. With additional assumptions on f, central limit type theorems type are obtained in [3]. In [12], the authors study the estimation of the boundary of S in terms of the Hausdorff distance.

Other estimators have also been proposed by various authors. In [19], the authors study the case when S is a subset of the unit square in $\mathbb{R}^2$. A piecewise polynomial estimator of ∂S is proposed. When f is an α- decreasing function (see Definition 5), it is proved to have an optimal convergence rate (which depends on α and the regularity of ∂S).

When S is a convex set, the convex hull of the observations is a natural estimator of S which has been widely studied (see e.g. [2,25,27,14,16]). There also exist generalizations of the convex hull which can be used to estimate non-convex supports. For example, one can consider, for $r > 0$, the r-convex hull:

$$C_r(A) = \bigcap_{x\in\mathbb{R}^d, A\subset(\mathring{\mathcal{B}}(x,r))^c} (\mathring{\mathcal{B}}(x, r))^c.$$

Notice that $C_\infty(A) = \mathcal{H}(A)$.

In [26], the authors study $C_r(\mathcal{X}_n)$ when the support S is r_0-convex (i.e. $C_{r_0}(S) = S$ and $C_{r_0}(S^c) = S^c$) for some $r_0 > 0$. They also assume the density f to be bounded from below. In this case, $C_r(\mathcal{X}_n)$ is proved to be consistent, with a convergence rate of order $n^{2/(d+1)}$. This order is the same as the one obtained with the convex hull when the support S is convex.

When S is a manifold of dimension $d' < d$, this estimator may degenerate. For example, when S is a d'-manifold in $\mathbb{R}^d$ with a positive reach r' (see [17]), one can easily prove that $C_r(\mathcal{X}_n) = \mathcal{X}_n$ for all $r < r'$. In [10], the authors study a generalization of $C_r(\mathcal{X}_n)$, which allows to weaken the assumptions on S. However the estimator possesses the same degeneracy drawback.

In this work, we also propose a generalization of the convex hull, defined as follows:

$$\hat{H}_r = \bigcup_{X\in\mathcal{X}_n} \mathcal{H}(\mathcal{B}(X, r) \cap \mathcal{X}_n). \tag{2}$$

Let us first notice that, while $C_r(\mathcal{X}_n)$ is an intersection of subsets, the local convex hull estimator $\hat{H}_r$ is built as a union of subsets. This prevents from degenerate behaviors.

Moreover, one can expect the convergence rate of the estimator $\hat{H}_r$ to be closer to the convergence rate of the convex hull estimator than the one obtained with the Devroye–Wise estimator. This actually the case (see Theorem 2). Indeed, the Devroye–Wise estimator is built by setting the same volume around each observation point, when our method is based on a different idea. Obviously, we have

$$S = \bigcup_{x\in S}(S \cap \mathcal{B}(x, r_n)).$$

Therefore, when the sample size n tends to infinity, the following set

$$\bigcup_i(S \cap \mathcal{B}(X_i, r_n))$$

is a good approximation of S. Eventually, when ∂S is smooth enough, as $r_n \to 0$, the set $(S \cap \mathcal{B}(X_i, r_n))$ can be approximated by the convex set $\mathcal{H}(\mathcal{X}_n \cap \mathcal{B}(X_i, r))$.

In order to estimate the boundary of the support by $\partial\hat{S}_r$, the inclusion condition $S \subset \hat{S}_r$ has to be imposed (see [12]). This causes an overfilling phenomenon. More precisely it has been numerically shown (see Section 6) the following: on the one hand, the best estimator (in the sense of the previous distance criteria) for the support is not consistent for boundary estimation, and does not preserve the topological properties of the support. On the other hand the best estimator for the boundary, while giving good results in terms of topological properties, drastically overfills the support.

The local convex hull estimator $\hat{H}_r$ is designed to overcome these drawbacks. It can be used, with a unique choice of radius, to estimate the support and its boundary simultaneously, while preserving the topological properties of the support.

Let us note that the first idea of using a local convex hull estimator has been introduced in [18], using nearest-neighbors instead of fixed radius. This estimator has good performances in different applications as home-range estimation, zoology