Available online at www.sciencedirect.com

ScienceDirect

Journal of Multivariate Analysis 99 (2008) 2185–2207

Journal of Multivariate Analysis

www.elsevier.com/locate/jmva

Exact rates in density support estimation

Gérard Biau[a], Benoît Cadre[b], Bruno Pelletier[c,*]

[a] *LSTA & LPMA, Université Pierre et Marie Curie – Paris VI, Boîte 158, 175 rue du Chevaleret, 75013 Paris, France*
[b] *IRMAR, ENS Cachan Bretagne, CNRS, UEB, Campus de Ker Lann, Avenue Robert Schuman, 35170 Bruz, France*
[c] *Institut de Mathématiques et de Modélisation de Montpellier, UMR CNRS 5149, Equipe de Probabilités et Statistique, Université Montpellier II, CC 051, Place Eugène Bataillon, 34095 Montpellier Cedex 5, France*

Received 23 May 2007
Available online 14 February 2008

Abstract

Let f be an unknown multivariate probability density with compact support S_f. Given n independent observations $X_1, \ldots, X_n$ drawn from f, this paper is devoted to the study of the estimator $\hat{S}_n$ of S_f defined as unions of balls centered at the X_i and of common radius r_n. To measure the proximity between $\hat{S}_n$ and S_f, we employ a general criterion d_g, based on some function g, which encompasses many statistical situations of interest. Under mild assumptions on the sequence (r_n) and some analytic conditions on f and g, the exact rates of convergence of $d_g(\hat{S}_n, S_f)$ are obtained using tools from Riemannian geometry. The conditions on the radius sequence are found to be sharp and consequences of the results are discussed from a statistical perspective.

AMS 2000 subject classifications: 62G05; 62G20

1. Introduction

Let f be an unknown probability density function defined with respect to the Lebesgue measure on $\mathbb{R}^d$. This paper is concerned with the problem of estimating the support of f, i.e., the closed set

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

given a random sample $X_1, \ldots, X_n$ drawn from f. Here and later, $\overline{A}$ means the closure of the set A. Since the earlier works of Rényi and Sulanke [23,24] and Geffroy [12], the problem of

[*] Corresponding author.
E-mail addresses: gerard.biau@upmc.fr (G. Biau), Benoit.Cadre@bretagne.ens-cachan.fr (B. Cadre), pelletier@math.univ-montp2.fr (B. Pelletier).

0047-259X/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2008.02.021

support estimation has been considered by several authors [see, e.g., [5,9,14,6,19,20,15,18,21,7, 11,1,16]]. The scope of application is vast, as support estimation is routinely employed across the entire and diverse range of applied statistics, including problems in medical diagnosis, machine condition monitoring, marketing or econometrics [see the discussion in [1] and the references therein]. In close connection with the related topic of estimating a density level set [22,26,27,3], the problem of support estimation has been also addressed via unsupervised learning methods, such as the one-class kernel Support Vector Machines algorithm presented in [25].

Among the various approaches that have been proposed to date to estimate S_f, the probably most simple and intuitive one has been considered in [9]. The estimator is defined as

$$\hat{S}_n = \bigcup_{i=1}^{n} \mathcal{B}(X_i, r_n), \tag{1.1}$$

where $\mathcal{B}(x, r)$ denotes the closed Euclidean ball centered at x and of radius r, and where (r_n) is an appropriately chosen sequence of positive smoothing parameters. Note that this approach amounts to estimating the support of the density by the support of a kernel estimate, the kernel of which has a ball-shaped support. The sequence (r_n) then plays a role analogous to that of the kernel bandwidth. The practical properties of the support estimator (1.1) are explored in [1], who argue that this estimator is a good generalist when no *a priori* information is available on S_f. Moreover, from a practical perspective, the relative simplicity of the naive strategy (1.1) arises as a major advantage in comparison with competing multidimensional set estimation techniques, that are faced with severe difficulties owing to a heavy computational burden.

To measure the performance of the support estimator, i.e., the closeness of $\hat{S}_n$ to S_f, a standard choice is to use the distance $d_1(\hat{S}_n, S_f)$ defined by

$$d_1(\hat{S}_n, S_f) = \lambda(\hat{S}_n \triangle S_f),$$

where $\triangle$ denotes the symmetric difference and λ is the Lebesgue measure on $\mathbb{R}^d$. This criterion of proximity between sets, which is *geometric* by essence, has been successfully employed for example by Korostelev and Tsybakov [20], Härdle, Park and Tsybakov [15], and Mammen and Tsybakov [21] who have considered maximum-likelihood-type estimators and have derived minimax rates of convergence under various assumptions on the boundary sharpness of f, that is, the behavior of f near the boundary of the support S_f.

The distance d_1 may be easily extended to the much more general measure-based distance d_μ defined by

$$d_\mu(\hat{S}_n, S_f) = \mu(\hat{S}_n \triangle S_f),$$

where μ is any measure on the Borel sets of $\mathbb{R}^d$. In this context, Cuevas and Fraiman [7] discuss the d_μ-asymptotic properties of a plug-in estimator of S_f of the form $\{f_n > \alpha_n\}$, where f_n is a non-parametric density estimator of f, and where α_n is a tuning parameter converging to zero. These authors establish also asymptotic results in terms of the Hausdorff metric, which is another natural criterion of proximity between sets [6,20,18,8].

Assuming for convenience that μ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$, with a density g, the criterion d_μ may be written as

$$d_g(\hat{S}_n, S_f) = \int_{\mathbb{R}^d} \mathbf{1}_{\hat{S}_n \triangle S_f}(x) g(x) \mathrm{d}x. \tag{1.2}$$