Simple Estimation of the Mode of a Multivariate Density

Christophe ABRAHAM [a], Gérard BIAU [b,*] and Benoît CADRE [c]

[a] ENSAM-INRA, UMR Biométrie et Analyse des Systèmes,
2, Place Pierre Viala, 34060 Montpellier Cedex 1, France;
[b] Laboratoire de Statistique Théorique et Appliquée,
Université Pierre et Marie Curie – Paris VI,
Boîte 158, 175 rue du Chevaleret, 75013 Paris, France;
[c] Laboratoire de Probabilités et Statistique, Université Montpellier II,
Cc 051, Place Eugène Bataillon, 34095 Montpellier Cedex 5, France

Abstract

We consider an estimate of the mode θ of a multivariate probability density f using a kernel estimate f_n drawn from a sample $X_1, \ldots, X_n$. The estimate θ_n is defined as any x in $\{X_1, \ldots, X_n\}$ such that $f_n(x) = \max_{i=1,\ldots,n} f_n(X_i)$. The strong consistency of θ_n towards θ is shown and an almost sure rate of convergence is provided. This rate relies on the sharpness of the density near θ, which is measured by a peak index.

Index Terms **— Multivariate probability density, mode, kernel estimate, rate of convergence.**

AMS 2000 Classification**: 62G05.**

1 Introduction

The problem of estimating the mode of a probability density has received considerable attention in the literature. For a historical and mathematical survey, we refer the reader to Sager [13]. Recent years have witnessed a renewal of interest in estimating the mode and in related multivariate mapping problems. The main reason is the recent advent of powerful mathematical tools and computational machinery that render these problems much more

*Corresponding author. Email: biau@ccr.jussieu.fr .

tractable. One of the most recent application of mode estimation is in unsupervised *cluster analysis*, where one tries to break a complex data set into a series of piecewise similar groups or structures, each of which may then be regarded as a separate data state, thus reducing overall data complexity. Cluster analysis has a long and rich history and excellent reviews of many methods may be found in Everitt [3], Hartigan [4], Jain and Chandrasekaran [5] and Jain and Dubes [6]. Among these methods, the nonparametric approach is based on the premise that groups correspond to modes of a density. The goal is then to estimate the modes and assign each observation to the "domain of attraction" of a mode. But there are many other fields where the knowledge of the mode is of great interest. For example, statistical mapping has points of contact with modal estimation. Contours, or isopleths, connect points of equal (probability) density and therefore bound modal regions, the smallest of which is the mode. So the estimation of isopleths and modal regions is a natural extension of the estimation of modal points. Nevertheless, the statistical properties of most isopleth mapping techniques remain largely unknown.

In this paper, we consider the problem of estimating the mode θ of a multivariate unimodal probability density f with support in $\mathbb{R}^d$ from independent random variables $X_1, \ldots, X_n$ with density f. Formally, a mode of a density is a value that maximizes the density. Since the density may be arbitrarily redefined on any set of measure zero, this definition is somewhat unsatisfactory. For example, with $\mathbf{1}_A$ denoting the indicator function of the set A,

$$f(x) = \frac{1}{2} \exp\big(- |x - \theta|\big) + \mathbf{1}_{\{\theta+1\}}(x)$$

would have a mode at $\theta + 1$, whereas the mode should be θ. This technical difficulty is usually overcome by selecting a particular representative from the equivalence class or by imposing additional constraints on f such as continuity. Therefore it will be assumed throughout the paper that the density f is continuous on a neighborhood (with respect to the support) of its unique mode θ.

We will define an estimate of the mode using a kernel density estimate $\big($Rosenblatt [12], Parzen [9], Devroye [2]$\big)$. Given a kernel K (*i.e.*, a probability density) and a bandwidth $h_n > 0$ such that $h_n \to 0$ as n grows to infinity, the kernel estimate is given by

$$f_n(x) = \frac{1}{nh_n^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right), \quad x \in \mathbb{R}^d.$$

The estimation of the mode has received a great deal of attention, see for example Parzen [9], Konakov [7], Samanta [14], Devroye [1], Romano [11], Vieu [17], Leclerc and Pierre-Loti-Viaud [8] and the references therein. Mostly, the estimate $\hat{\theta}_n$ of θ is defined as any maximizer of f_n, *i.e.*,

$$\hat{\theta}_n \in \operatorname{argmax}_{\mathbb{R}^d} f_n \,. \tag{1.1}$$

In the univariate case ($d = 1$) Romano [11] proved, under mild conditions on f and K, the strong consistency of $\hat{\theta}_n$ to θ. Assuming at least that f is twice continuously differentiable, Konakov [7], Samanta [14], Romano [11], Vieu [17] and Leclerc and Pierre-Loti-Viaud [8] also studied the rate of convergence of the estimate. Although interest in the mode is currently returning, there is still much to be done, particularly in defining estimates that will be computationally efficient and have good statistical properties. Indeed, estimating the mode using (1.1) has several severe drawbacks. In addition to the calculation of f_n, it involves a numerical step for the computation of the argmax. As noticed by Devroye [1], classical search methods of the argmax perform satisfactorily only when f_n is sufficiently regular (continuous, unimodal, etc.) Thus, in practice, the argmax is usually computed over a finite grid. This failing is seldom discussed by authors, although it may affect the asymptotic properties of the estimate (we refer the reader to the discussion of Section 3 and more particularly to Figure 1). Moreover, when the dimension of the sample space is large, or when accurate estimation is needed, the grid size (which exponentially increases with the dimension) leads to time-consuming computations. Finally, the search grid should be located around high density areas. In high dimension, this is a difficult task and the search grid usually includes low density areas.

The aim of the paper is to study an estimate of the mode which eliminates these problems. This estimate has been first considered by Devroye [1]. Denoting by S_n the set $\{X_1, \ldots, X_n\}$, we let the estimate θ_n be defined as

$$\theta_n \in \operatorname{argmax}_{S_n} f_n \,,$$

i.e.,

$$\theta_n \in \big\{x \in S_n : f_n(x) = \max_{i=1,\ldots,n} f_n(X_i)\big\}.$$

Since the sample points are naturally concentrated in high density areas, the set S_n can be regarded as the most natural (random) grid for approximating the mode. Clearly, the more the density is sharp around the mode, the more the data will concentrate around it, and the more θ_n will perform. We emphasize that the main advantage of using θ_n instead of the argmax estimate

(1.1) is that the former is easily computed in a finite number of operations. Our motivation is to derive asymptotic properties of this estimate under the weakest assumptions on the underlying density. For instance, we will not assume any differentiability condition on f around the mode.

The paper is organized as follows. Section 2 is devoted to the main results. Here, we assert the strong consistency of θ_n towards θ as well as the existence of a positive sequence $(\varphi_n)_{n\geq 1}$ tending to zero such that, for all $p > 0$, $\mathbf{P}\big(\|\theta - \theta_n\| \geq \varphi_n\big) = \mathrm{o}(1/n^p)$. From this, we deduce an almost sure rate of convergence of θ_n towards θ. Our results are valid for a large class of densities f not necessarily differentiable around the mode. The results of Section 2 strongly rely on the sharpness of the density near θ. This sharpness will be measured by a *peak index*, which is discussed in the appendix. Section 3 is dedicated to a discussion. Proofs are gathered in Section 4.

2 Main results

We equip $\mathbb{R}^d$ with the Euclidean norm $\|.\|$. For convenience, we shall assume throughout the paper that the support of K is compact and we denote by a a positive real number such that supp K is contained in the closed ball with center at the origin and radius a. Moreover, as we will use Pollard's results [10], K will be assumed to be of the form $M\big(\|.\|\big)$, where M is a monotone nonincreasing function on $\mathbb{R}^+$.

We consider, for all $\varepsilon > 0$, the *level sets*

$$A(\varepsilon) = \big\{x \in \mathbb{R}^d : f(x) > f(\theta) - \varepsilon\big\},$$

which will play a crucial role throughout. We denote by $V(\delta)$ the open ball with center at θ and radius $\delta > 0$. We shall assume that there exists $\delta_0 > 0$ such that f is continuous on $V := V(\delta_0)$. Without loss of generality, we assume throughout the paper that δ_0 is small enough to ensure $\inf_V f > 0$. The notation diam $A(\varepsilon)$ stands for the diameter of $A(\varepsilon)$ $\big($*i.e.*, $\sup\big\{\|x-y\| : x \in A(\varepsilon), y \in A(\varepsilon)\big\}\big)$. Finally, for any function g, we denote by $\|g\|_\infty$ (*resp.* $\|g\|_V$) the supremum norm of g on $\mathbb{R}^d$ (*resp.* on V).

2.1 Consistency of θ_n

Theorem 2.1 *Assume the density f is such that* diam $A(\varepsilon) \to 0$ *as $\varepsilon \to 0$. Then, under the conditions $h_n \to 0$ and $nh_n^d/\log n \to \infty$, one has $\theta_n \to \theta$ a.s.*

The condition on diam $A(\varepsilon)$ has been introduced to avoid high density areas arbitrarily far from θ. Note that an equivalent condition is given in Lemma 4.1. To illustrate the usefulness of this condition, consider the continuous univariate density

$$f(x) = c\,(1 - \mathbf{1}_{[2,\infty[}(x)/k)(x-k+1)(k+1-x)\exp\big(-(x-k)^2k^4\big)$$

for $k-1 \leq x \leq k+1$, $k = 1, 2, \ldots$ where c is a normalization constant. This density has its maximum at $\theta = 1$ but it has spikes in its tails which approach the maximum value c arbitrarily closely. As the sample size increases, it is possible that θ_n will occasionally fall into a tail spike.

2.2 Rate of convergence of θ_n

Let us assume that for some real number $\alpha > 0$ the following inequalities hold:

$$0 < \liminf_{\varepsilon \to 0} \frac{\text{diam } A(\varepsilon)}{\varepsilon^\alpha} \leq \limsup_{\varepsilon \to 0} \frac{\text{diam } A(\varepsilon)}{\varepsilon^\alpha} < \infty. \tag{2.1}$$

When α does exist, attention shows that it is uniquely defined. In this case, we take the liberty to call it the *peak index* of f. This peak index measures the sharpness of the density around the mode. Roughly, the more the density is sharp around θ, the more the peak index is large. As an illustration, consider the family of univariate densities $f(x) = N\exp(-|(x-m)/\sigma|^\gamma)$, $\gamma > 0$, where $m \in \mathbb{R}$ and $\sigma > 0$ are respectively position and scale parameters, and where N is a normalization constant. Each member of this family admits a peak index $\alpha = 1/\gamma$. This is for instance the case for normal densities ($\gamma = 2$) and Laplace densities ($\gamma = 1$). In the appendix, we provide a classification of the values of α, depending on the local smoothness of f.

We shall assume in this section that f admits a peak index $\alpha > 0$ and that

$$\liminf_{\varepsilon \to 0} \frac{\lambda\big(A(\varepsilon)\big)}{\big(\text{diam } A(\varepsilon)\big)^d} > 0. \tag{2.2}$$

Roughly speaking, the above inequality means that the Lebesgue measure of each level set $A(\varepsilon)$ is of the same order as the volume of a sphere with the same diameter as $A(\varepsilon)$. Without loss of generality, one can also assume that ε_0 is small enough so that the following property holds:

(P_1) There exists $L > 0$ such that diam $A(\varepsilon) \leq L\,\varepsilon^\alpha$ for $\varepsilon \leq \varepsilon_0$.

Using (P_1), we observe that $\theta_n \to \theta$ a.s. in virtue of Theorem 2.1.

Theorem 2.2 *Assume that (2.1) holds for some* $\alpha > 0$ *and (2.2) holds. Assume there exist two real numbers* $c > 0$ *and* $\beta > 0$ *such that* $\alpha\beta \leq 1$ *and, for all* $n \geq 1$, $\|\mathbf{E}f_n - f\|_V \leq ch_n^\beta$, *with*

$$h_n = \frac{(\log n)^{2/(2\beta+d)}}{n^{1/(2\beta+d)}}.$$

Then, for all $p > 0$,

$$\mathbf{P}\Big(\|\theta - \theta_n\| \geq \big(a + (16c)^\alpha L\big)\frac{(\log n)^{2/(2\beta+d)}}{n^{\alpha\beta/(2\beta+d)}}\Big) = \mathrm{o}\Big(\frac{1}{n^p}\Big),$$

where L *is the constant defined in* (P_1).

Observe that the condition $\|\mathbf{E}f_n - f\|_V \leq ch_n^\beta$ is automatically fulfilled for any h_n in the important case where f is Hölder of order $\beta > 0$ on some neighborhood of V. In this case, the inequality $\alpha\beta \leq 1$ also holds, see Proposition A.1. The interest of Theorem 2.2 is essentially theoretical. From a practical point of view, it is worth pointing out that it is useless, just because the parameters α, β, c and L depend upon the unknown density f (recall that the parameter a has been defined in Section 2). To solve this problem, a natural idea consists in estimating these four parameters. For simplicity, consider the case $d = 1$ and f twice continuously differentiable with $f''(\theta) \neq 0$. In that case, $\alpha = 1/2$, $L = 2\sqrt{2/|f''(\theta)|}$ (see Corollary A.1), $\beta = 2$ and $c = 1/2\|f''\|_V \int_{-a}^{a} t^2 K(t)\,\mathrm{d}t$ (use the fact that K is even). The estimation of f'' is easily done using the almost sure consistency of the second derivative of a smooth kernel estimate of f towards f'' (see for example Silverman [15]). In the non-smooth case, however, more work is needed, at least for the estimation of α. A possible approach could rely on the estimation of level sets, see Tsybakov [16].

We can as well deduce from Theorem 2.2 and the Borel-Cantelli lemma an almost sure rate of convergence of θ_n towards θ.

Corollary 2.1 *Under the assumptions of Theorem 2.2, we have a.s.*

$$\|\theta - \theta_n\| = \mathrm{O}\Big(\frac{(\log n)^{2/(2\beta+d)}}{n^{\alpha\beta/(2\beta+d)}}\Big).$$

As expected, we note from Corollary 2.1 that the rate of convergence of θ_n increases with α.

Remark When f is Hölder of order β, $\mathbf{E}X = \theta$ (*e.g.*, if f is symmetric around θ) and $\alpha\beta > \beta + d/2$, Corollary 2.1 provides us with an estimate that improves the standard empirical mean estimate.

Finally, one easily shows that, under the assumptions of Theorem 2.2,

$$\mathbf{P}\Big(\frac{1}{(n\log n)^{1/d}} \leq \|\theta - \theta_n\| \leq \big(a + (16c)^\alpha L\big)\frac{(\log n)^{2/(2\beta+d)}}{n^{\alpha\beta/(2\beta+d)}}\Big) \to 1 \text{ as } n \to \infty. \tag{2.3}$$

As an illustration, let us consider again the family of univariate probability densities $f(x) = N\exp\big(-|(x-m)/\sigma|^\gamma\big)$. It is clear that f is locally Hölder of order $\beta = \min(1,\gamma)$ with a peak index $\alpha = 1/\gamma$. Moreover $\theta = m$, and (2.3) tells us that, for some constant κ,

$$\mathbf{P}\Big(\frac{1}{n\log n} \leq |m - \theta_n| \leq \kappa\,\frac{(\log n)^{2/(2\beta+1)}}{n^{\beta/(\gamma(2\beta+1))}}\Big) \to 1 \text{ as } n \to \infty.$$

For γ near 0, the upper and lower bounds are of the same order (up to a logarithm), making this result not far from optimality.

3 Discussion

In the univariate setting, Vieu [17] showed, under the additional assumption that f is of class C^k ($k \geq 2$) on a neighborhood of θ with $f^{(2)}(\theta) < 0$ and $f^{(k)}(\theta) \neq 0$, that the argmax estimate defined in (1.1) achieves the almost sure rate $\mathrm{O}\big((\log n/n)^{(k-1)/(2k+1)}\big)$. This result was recently improved by Leclerc and Pierre-Loti-Viaud [8], who obtained a $\mathrm{O}\big((\log\log n/n)^{(k-1)/(2k+1)}\big)$ rate. Under the same smoothness conditions on f, Corollary 2.1 applies with $\alpha = 1/2$ (by Corollary A.1) and $\beta = 2$ (recall the kernel K is assumed to be even). Thus a.s., $|\theta - \theta_n| = \mathrm{O}\big((\log n)^{2/5}/n^{1/5}\big)$. The obtained almost sure rate of convergence of our estimate is slightly less than the rate obtained for the estimate (1.1). Actually, as enlightened by the results given in appendix, the less the density under study is smooth around θ the better the rate of θ_n will be. As pointed out by a referee, an interesting question is then to know how much is lost by our estimate when f is not smooth. In other words, which of the argmax estimate $\hat{\theta}_n$ defined in (1.1) and the estimate θ_n is superior in the case of non-smooth density? This is a difficult question, just because, as far as we know, there are no available results about the rate of convergence of the argmax estimate in cases where the density f has bad local behavior (*e.g.*, it has infinite derivative). Much remains to be done in this domain. However, we insist on the fact that the main interest of the estimate θ_n is numerical.

Indeed, modal estimation using the traditional argmax estimate induces a bias due to the computation over a finite grid. This drawback is enlightened by Figure 1, which presents typical trajectories of θ_n and $\hat{\theta}_n$ for n ranging from 1 to 600. Here, we estimate the mode of a standard Laplace density $f(x) = 1/2 \exp\big(-|x|\big)$ using one unit apart grid points $\{\ldots, -1/2, 1/2, \ldots\}$. We intentionally chose a wide grid in order to underline the main features of the gridding error. Whereas the estimate θ_n approaches the true value of the mode ($\theta = 0$), the argmax estimate takes the values of the grid points which are the closest from θ. Thus, as illustrated by Figure 1, the estimation error $\theta - \hat{\theta}_n$ will directly depend on the grid step. Attention shows that the number of grid points needed in d dimensions to make gridding error be of the same order as the error of θ_n is $\mathrm{O}\big(n^{1/3}/(\log n)^{2/3}\big)^d$ for a multivariate Laplace density. This enlightens the fact that the more the dimension is large, the more our estimate is computer-time competitive.

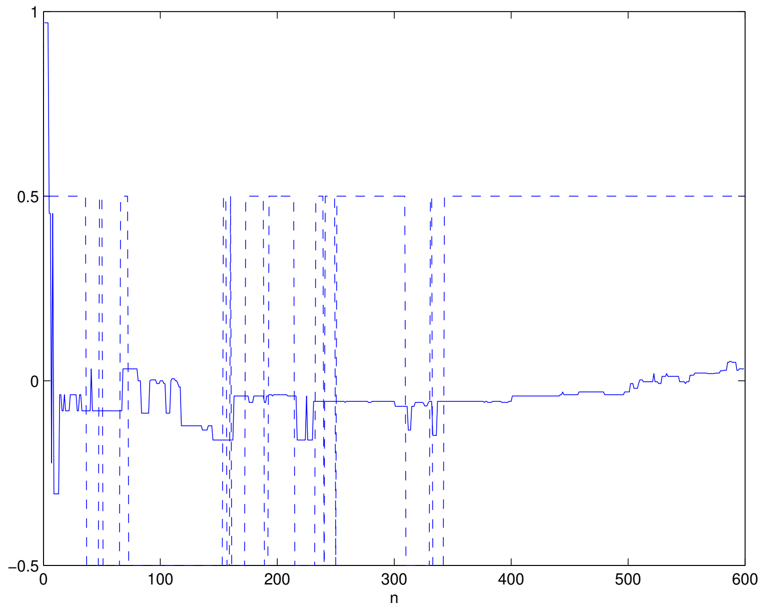

Figure 1: Typical trajectories of $\theta_{\mathbf{n}}$ (continuous line) and $\hat{\theta}_{\mathbf{n}}$ (dashed line) for estimating the mode ($\theta = 0$) of a standard Laplace density.

Let us finally mention that the choice of the bandwidth in the non-smooth case presented on Figure 1 is delicate. Namely, most of available methods for selecting local bandwidths pertain to sufficiently smooth target densities, which is not the case here. As a first benchmark, we have taken h_n as in Theorem 2.2. We realize however that more work is needed. An interesting approach could use the results of Proposition A.2. The idea is to substitute the asymptotic development of Proposition A.2 for classical Taylor's series expansion in the derivation of the error criteria.

4 Proofs

4.1 Proof of Theorem 2.1

The proof of Theorem 2.1 will rely on the following two lemmas.

Lemma 4.1 *If* diam $A(\varepsilon) \to 0$ *as* $\varepsilon \to 0$ *then, for any* $\delta > 0$, $\sup_{V(\delta)^c} f < f(\theta)$.

As a matter of fact, one can easily proves the equivalence between the two assertions of the lemma.

Proof of Lemma 4.1 Set $\delta_\varepsilon = \operatorname{diam} A(\varepsilon)$ for any $\varepsilon > 0$. Then $A(\varepsilon) \subset V(2\delta_\varepsilon)$. Let $x \in V(2\delta_\varepsilon)^c$. Then $x \in A(\varepsilon)^c$ and hence $f(x) \leq f(\theta) - \varepsilon$. Consequently, $\sup_{V(2\delta_\varepsilon)^c} f < f(\theta)$. As δ_ε tends to 0 as ε goes to 0, the proof is complete. ■

Lemma 4.2 *Assume f satisfies the condition of Theorem 2.1. Then, a.s., for any* $\delta > 0$, $\max_{S_n \cap V(\delta)} f \to f(\theta)$ *as* $n \to \infty$.

Proof of Lemma 4.2 Let $0 < \varepsilon < f(\theta)$. Then

$$
\begin{aligned}
&\mathbf{P}\Big(f(\theta) - \max_{S_n \cap V(\delta)} f \geq \varepsilon\Big) \\
&\quad = \mathbf{P}\Big(\max_{S_n \cap V(\delta)} f \leq f(\theta) - \varepsilon\Big) \\
&\quad = \mathbf{P}\Big(\bigcap_{i=1}^{n} \big(f(X_i) \leq f(\theta) - \varepsilon,\, X_i \in V(\delta)\big) \cup \big(X_i \in V(\delta)^c\big)\Big) \\
&\quad = \Big[\mathbf{P}\Big(f(X) \leq f(\theta) - \varepsilon,\, X \in V(\delta)\Big) + \mathbf{P}\Big(X \in V(\delta)^c\Big)\Big]^n \\
&\quad = \Big[1 - \mathbf{P}\Big(f(X) > f(\theta) - \varepsilon,\, X \in V(\delta)\Big)\Big]^n \\
&\quad = \Big[1 - \mathbf{P}\Big(X \in A(\varepsilon) \cap V(\delta)\Big)\Big]^n.
\end{aligned}
$$

By the Borel-Cantelli lemma, we only need to show that $\mathbf{P}\big(X \in A(\varepsilon) \cap V(\delta)\big) > 0$. Using the condition on f, it is easy to see that there exists $\eta > 0$ such that $A(\eta) \subset V(\delta)$. Distinguishing between the values of ε

leads to $A(\varepsilon) \cap V(\delta) = A(\varepsilon)$ if $\varepsilon \leq \eta$ and $A(\varepsilon) \supset A(\eta)$ otherwise. In this last case, we obtain $A(\varepsilon) \cap V(\delta) \supset A(\eta) \cap V(\delta) = A(\eta)$. Therefore, $\mathbf{P}\big(X \in A(\varepsilon) \cap V(\delta)\big) \geq \min\big[\mathbf{P}\big(X \in A(\varepsilon)\big), \mathbf{P}\big(X \in A(\eta)\big)\big]$. It is thus enough to show that for any $\gamma > 0$, $\mathbf{P}\big(X \in A(\gamma)\big) > 0$.

Let $\gamma > 0$. By continuity of f on V, there exists $0 < h_0 \leq \delta_0$ such that for $h \in \mathbb{R}^d$ with $\|h\| \leq h_0$, $f(\theta) - f(\theta + h) < \gamma$. This implies $V(h_0) \subset A(\gamma)$ and consequently $\mathbf{P}\big(X \in A(\gamma)\big) \geq \mathbf{P}\big(X \in V(h_0)\big)$. Denoting by λ the Lebesgue measure on $\mathbb{R}^d$, one sees that the rightmost term of this inequality is positive since $\lambda\big(V(h_0)\big) > 0$ and $\inf_{V(h_0)} f > 0$. ■

We are now ready to prove Theorem 2.1. The proof is inspired from Romano [11], who considers the univariate case.

Proof of Theorem 2.1 Observe first that, for any $x \in \mathbb{R}^d$ and $n \geq 1$,

$$\mathbf{E} f_n(x) = \int_{\mathbb{R}^d} K(t) f(x - h_n t)\, \mathrm{d}t.$$

Thus,

$$\sup_{V(\delta)^c} \mathbf{E} f_n \leq \sup_{V(\delta - a h_n)^c} f,$$

and consequently, as $h_n \to 0$ as $n \to \infty$, for all n large enough,

$$\sup_{V(\delta)^c} \mathbf{E} f_n \leq \sup_{V(\delta/2)} f.$$

Therefore, using Lemma 4.1, it is deduced that for any $\delta > 0$,

$$\limsup_n \sup_{V(\delta)^c} \mathbf{E} f_n < f(\theta).$$

Moreover, using the hypotheses on the kernel K, we have according to Pollard [10] (Theorem 37 and Problem 28 of Chapter II),

$$\|\mathbf{E} f_n - f_n\|_\infty \to 0 \quad \text{a.s.} \tag{4.1}$$

Consequently, a.s. and for all $\delta > 0$,

$$\limsup_n \sup_{V(\delta)^c} f_n < f(\theta),$$

so that

$$\limsup_n \max_{S_n \cap V(\delta)^c} f_n < f(\theta).$$

By (4.1) and Bochner's lemma, one also has $\|f_n - f\|_V \to 0$ a.s. In accordance with Lemma 4.2, we obtain a.s., for all $\delta \leq \delta_0$,

$$\limsup_n \max_{S_n \cap V(\delta)^c} f_n < \lim_n \max_{S_n \cap V(\delta)} f_n.$$

Observing finally that δ is as small as desired, the last inequality shows that $\theta_n \to \theta$ a.s. ■

4.2 Proof of Theorem 2.2

In addition to property (P_1), we assume in this paragraph, without loss of generality, that ε_0 is small enough so that the following two properties hold:

(P_2) There exists $l > 0$ such that $\lambda\big(A(\varepsilon)\big) \geq l\,\varepsilon^{d\alpha}$ for $\varepsilon \leq \varepsilon_0$.
(P_3) The inequality $\inf_{A(\varepsilon_0)} f > 0$ holds.

We start with a technical proposition.

Proposition 4.1 *Let $(\gamma_n)_{n\geq 1}$ be a sequence of positive real numbers tending to 0. Then, for all n large enough,*

$$\mathbf{P}\Big(\|\theta - \theta_n\| \geq ah_n + 8^\alpha L\gamma_n^\alpha\Big) \leq \Big(1 - l\gamma_n^{d\alpha} \inf_{A(\varepsilon_0)} f\Big)^n + \mathbf{P}\Big(\|\mathbf{E}f_n - f_n\|_\infty \geq \gamma_n\Big) + \mathbf{P}\Big(\|\mathbf{E}f_n - f\|_V \geq \gamma_n\Big).$$

Proof of Proposition 4.1 Set, for $n \geq 1$, $k_n = ah_n + 8^\alpha L\gamma_n^\alpha$. Then, for all n large enough (to ensure that $V(k_n) \subset V$),

$$\begin{aligned}
&\mathbf{P}\Big(\|\theta - \theta_n\| \geq k_n\Big)\\
&\quad\leq \mathbf{P}\Big(\max_{S_n \cap V(k_n)} f_n \leq \max_{S_n \cap V(k_n)^c} f_n\Big)\\
&\quad\leq \mathbf{P}\Big(-\|\mathbf{E}f_n - f_n\|_\infty + \max_{S_n \cap V(k_n)} \mathbf{E}f_n \leq \max_{S_n \cap V(k_n)^c} \mathbf{E}f_n + \|\mathbf{E}f_n - f_n\|_\infty\Big)\\
&\quad= \mathbf{P}\Big(\max_{S_n \cap V(k_n)} \mathbf{E}f_n \leq \max_{S_n \cap V(k_n)^c} \mathbf{E}f_n + 2\,\|\mathbf{E}f_n - f_n\|_\infty\Big)\\
&\quad\leq \mathbf{P}\Big(\max_{S_n \cap V(k_n)} f \leq \sup_{V(k_n)^c} \mathbf{E}f_n + 2\,\|\mathbf{E}f_n - f_n\|_\infty + \|\mathbf{E}f_n - f\|_V\Big)\\
&\quad\leq \mathbf{P}\Big(\max_{S_n \cap V(k_n)} f \leq \sup_{V(k_n)^c} \mathbf{E}f_n + 3\gamma_n\Big) + \mathbf{P}\Big(\|\mathbf{E}f_n - f_n\|_\infty \geq \gamma_n\Big)\\
&\qquad + \mathbf{P}\Big(\|\mathbf{E}f_n - f\|_V \geq \gamma_n\Big).
\end{aligned}$$

To show the result, it suffices to bound the first term of the right hand side of the last inequality. Recall that we assume that supp K is contained in the closed ball with center at the origin and radius $a > 0$. We can write, for all $n \geq 1$ and $x \in V(k_n)^c$,

$$\mathbf{E} f_n(x) = \int_{\mathbb{R}^d} K(t) f(x - h_n t)\,\mathrm{d}t \leq \sup_{V(k_n - a h_n)^c} f.$$

Thus,

$$\begin{aligned}
&\mathbf{P}\Big(\max_{S_n \cap V(k_n)} f \leq \sup_{V(k_n)^c} \mathbf{E} f_n + 3\gamma_n\Big) \\
&\quad \leq \mathbf{P}\Big(\max_{S_n \cap V(k_n)} f \leq \sup_{V(k_n - a h_n)^c} f + 3\gamma_n\Big) \\
&\quad = \Big[\mathbf{P}\Big(f(X) \leq \sup_{V(k_n - a h_n)^c} f + 3\gamma_n,\ X \in V(k_n)\Big) + \mathbf{P}\Big(X \in V(k_n)^c\Big)\Big]^n \\
&\quad = \Big[1 - \mathbf{P}\Big(f(X) > \sup_{V(k_n - a h_n)^c} f + 3\gamma_n,\ X \in V(k_n)\Big)\Big]^n. \qquad (4.2)
\end{aligned}$$

Choosing n large enough, we may assume that $8\gamma_n \leq \varepsilon_0$, that is

$$\frac{(k_n - a h_n)^{1/\alpha}}{2L^{1/\alpha}} \leq \frac{\varepsilon_0}{2}.$$

Fix t in $V(k_n - a h_n)^c$. If t meets the condition $f(\theta) - f(t) \leq \varepsilon_0/2$, then

$$\|\theta - t\| \leq \operatorname{diam} A\Big(2\big(f(\theta) - f(t)\big)\Big) \leq 2^\alpha L\big(f(\theta) - f(t)\big)^\alpha,$$

by Property (P_1). Consequently,

$$f(t) \leq f(\theta) - \frac{(k_n - a h_n)^{1/\alpha}}{2L^{1/\alpha}}. \qquad (4.3)$$

On the other hand, if t meets the condition $f(\theta) - f(t) > \varepsilon_0/2$, inequality (4.3) is obviously valid in this case. Therefore,

$$\sup_{V(k_n - a h_n)^c} f \leq f(\theta) - \frac{(k_n - a h_n)^{1/\alpha}}{2L^{1/\alpha}},$$

i.e.,

$$\sup_{V(k_n - a h_n)^c} f \leq f(\theta) - 4\gamma_n.$$

Thus, by (4.2), we obtain, for all n large enough,

$$\begin{aligned}
&\mathbf{P}\Big(\max_{S_n\cap V(k_n)} f \leq \sup_{V(k_n)^c} \mathbf{E}f_n + 3\gamma_n\Big) \\
&\quad \leq \Big[1-\mathbf{P}\Big(f(X) > f(\theta) - 4\gamma_n + 3\gamma_n,\, X \in V(k_n)\Big)\Big]^n \\
&\quad = \Big[1-\mathbf{P}\Big(X \in A(\gamma_n)\cap V(k_n)\Big)\Big]^n.
\end{aligned}$$

Observing now (see (4.3) for similar computations) that

$$A(\gamma_n) \subset V(k_n),$$

we deduce that

$$\begin{aligned}
&\mathbf{P}\Big(\max_{S_n\cap V(k_n)} f \leq \sup_{V(k_n)^c} \mathbf{E}f_n + 3\gamma_n\Big) \\
&\quad \leq \Big[1-\mathbf{P}\Big(X \in A(\gamma_n)\Big)\Big]^n \\
&\quad \leq \Big(1 - l\gamma_n^{d\alpha} \inf_{A(\varepsilon_0)} f\Big)^n,
\end{aligned}$$

where the last inequality follows from Property (P_2). Since $\inf_{A(\varepsilon_0)} f > 0$ (by Property (P_3)), we finally obtain, for all n large enough,

$$\begin{aligned}
\mathbf{P}\Big(\|\theta-\theta_n\| \geq k_n\Big) \leq \Big(1 - l\gamma_n^{d\alpha} \inf_{A(\varepsilon_0)} f\Big)^n + \mathbf{P}\Big(\|\mathbf{E}f_n - f_n\|_\infty \geq \gamma_n\Big) \\
+ \mathbf{P}\Big(\|\mathbf{E}f_n - f\|_V \geq \gamma_n\Big).
\end{aligned}$$

■

Proof of Theorem 2.2 Set, for $n \geq 1$, $\gamma_n = 2ch_n^\beta$. According to the exponential inequality in the proof of Theorem 37 and Problem 28 of Chapter II in Pollard [10], and since $\alpha\beta \leq 1$, an easy computation shows that for all $p > 0$,

$$\mathbf{P}\Big(\|\mathbf{E}f_n - f_n\|_\infty \geq \gamma_n\Big) = \mathrm{o}\Big(\frac{1}{n^p}\Big). \tag{4.4}$$

Furthermore, the condition $\|\mathbf{E}f_n - f\|_V \leq ch_n^\beta$ leads, for all $n \geq 1$, to

$$\mathbf{P}\Big(\|\mathbf{E}f_n - f\|_V \geq \gamma_n\Big) = 0. \tag{4.5}$$

Finally, by the very definition of γ_n, and using the fact that $\alpha\beta \le 1$, we also have for all $p > 0$,

$$\Big(1 - l\gamma_n^{d\alpha} \inf_{A(\varepsilon_0)} f\Big)^n = \mathrm{o}\Big(\frac{1}{n^p}\Big). \tag{4.6}$$

Using (4.4)-(4.6) and Proposition 4.1, it is deduced that for all $p > 0$,

$$\mathbf{P}\Big(\|\theta - \theta_n\| \ge ah_n + 8^\alpha L\gamma_n^\alpha\Big) = \mathrm{o}\Big(\frac{1}{n^p}\Big).$$

Using $\alpha\beta \le 1$ leads to the desired inequality. ■

A Appendix: on the peak index of f

Proposition A.1 *Assume f is Hölder of order $\beta > 0$ on V with a peak index $\alpha > 0$. Then $\alpha\beta \le 1$.*

Proof of Proposition A.1 Let ε_0 and L be positive real numbers such that

$$\text{diam } A(\varepsilon) \le L\,\varepsilon^\alpha \quad \text{for } \varepsilon \le \varepsilon_0.$$

Observe that, by continuity of f on V, there exists a sequence $(t_k)_{k\ge1}$ in V tending towards θ such that for all $k \ge 1$, $f(\theta) - f(t_k) \le \varepsilon_0/2$. Thus, for all $k \ge 1$,

$$\|\theta - t_k\| \le \text{diam } A\Big(2\big(f(\theta) - f(t_k)\big)\Big) \le 2^\alpha L\big(f(\theta) - f(t_k)\big)^\alpha.$$

Consequently, using the fact that f is Hölder of order β on V, there exists $C > 0$ such that

$$\|\theta - t_k\| \le C\,\|\theta - t_k\|^{\alpha\beta}, \quad \forall k \ge 1.$$

As $t_k \to \theta$ as k grows to infinity, we must have $\alpha\beta \le 1$. ■

Let us now focus on the case $d = 1$. Using similar arguments, it can be proved that for $\alpha > 1$ f is not differentiable at the point θ. The following result holds:

Proposition A.2 *Assume f admits an asymptotic expansion around θ of the form*

$$f(x) = f(\theta) + a_p\,|x - \theta|^p + \mathrm{o}\big(|x - \theta|^p\big),$$

where $p > 0$ and a_p are real numbers. Assume further that diam $A(\varepsilon) \to 0$ *as $\varepsilon \to 0$. Then*

$$\lim_{\varepsilon\to0} \frac{\text{diam } A(\varepsilon)}{\varepsilon^{1/p}} = \frac{2}{|a_p|^{1/p}},$$

and hence f has a peak index $\alpha = 1/p$.

Proof of Proposition A.2 Let $0 < \delta < 1$. Denote by I_δ any open interval centered around θ such that

$$I_\delta \subset \Big\{ x \in \mathbb{R} : \big| \mathrm{o}\big(|x-\theta|^p\big) \big| \leq \delta |a_p| \, |x-\theta|^p \Big\}.$$

The following two properties hold for ε small enough:

$$\begin{array}{ll} \text{(i)} & A(\varepsilon) \subset I_\delta \\ \text{(ii)} & \left]\theta - \left(\dfrac{\varepsilon}{(1+\delta)|a_p|}\right)^{1/p}, \theta + \left(\dfrac{\varepsilon}{(1+\delta)|a_p|}\right)^{1/p}\right[\subset I_\delta. \end{array}$$

Thus, since

$$A(\varepsilon) = \big\{ x \in \mathbb{R} : a_p \, |x-\theta|^p + \mathrm{o}\big(|x-\theta|^p\big) > -\varepsilon \big\},$$

we are led to

$$A(\varepsilon) \subset \big\{ x \in \mathbb{R} : |a_p| \, |x-\theta|^p (1-\delta) < \varepsilon \big\} \quad \big(\text{using (i)}\big)$$

and

$$A(\varepsilon) \supset \big\{ x \in \mathbb{R} : |a_p| \, |x-\theta|^p (1+\delta) < \varepsilon \big\} \quad \big(\text{using (ii)}\big).$$

We obtain

$$2\left(\frac{1}{(1+\delta)|a_p|}\right)^{1/p} \leq \frac{\operatorname{diam} A(\varepsilon)}{\varepsilon^{1/p}} \leq 2\left(\frac{1}{(1-\delta)|a_p|}\right)^{1/p}.$$

Observing that the inequalities above are valid for any $0 < \delta < 1$ leads to the desired result. ∎

Corollary A.1 *Assume the density f is of class C^p ($p \geq 2$, necessarily even) on a neighborhood of θ, with $f'(\theta) = \ldots = f^{(p-1)}(\theta) = 0$ and $f^{(p)}(\theta) < 0$ (necessarily). Assume further that* $\operatorname{diam} A(\varepsilon) \to 0$ *as $\varepsilon \to 0$. Then*

$$\lim_{\varepsilon \to 0} \frac{\operatorname{diam} A(\varepsilon)}{\varepsilon^{1/p}} = 2\left(\frac{p!}{|f^{(p)}(\theta)|}\right)^{1/p},$$

and hence f has a peak index $\alpha = 1/p$.

Distinguishing left and right derivatives of f at θ, a similar result valid for all $p \geq 1$ can be obtained.

Acknowledgments. The authors greatly thank an Associate Editor and an anonymous referee for a careful reading of the paper. They also thank Luc Devroye and Nicolas Hengartner for many helpful comments.

References

[1] Devroye, L. (1979). Recursive estimation of the mode of a multivariate density. *The Canadian Journal of Statistics.* **7** 159–167.

[2] Devroye, L. (1987). *A Course in Density Estimation.* Birkhäuser, Boston.

[3] Everitt, B. (1974). *Cluster Analysis.* Wiley, New York.

[4] Hartigan, J. (1975). *Clustering Algorithms.* Wiley, New York.

[5] Jain, A.K. and Chandrasekaran, B. (1982). *Dimensionality and Sample Size Considerations in Pattern Recognition Practice*, in *Handbook of Statistics, Volume II*, ed. P.R. Krishnaiah and L.N. Kanal, **18** 835–855, North-Holland, Amsterdam.

[6] Jain, A.K. and Dubes, R.C. (1988). *Algorithms for Clustering Data.* Prentice-Haal, New Jersey.

[7] Konakov, V.D. (1973). On asymptotic normality of the sample mode of multivariate distributions. *Theory of Probability and its Applications.* **18** 836–842.

[8] Leclerc, J. and Pierre-Loti-Viaud, D. (2000). Vitesse de convergence presque sûre de l'estimateur à noyau du mode. *Comptes Rendus de l'Académie des Sciences de Paris.* **331** 637–640.

[9] Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics.* **33** 1065–1076.

[10] Pollard, D. (1984). *Convergence of Stochastic Processes.* Springer–Verlag, New York.

[11] Romano, J.P. (1988). On weak convergence and optimality of kernel density estimates of the mode. *The Annals of Statistics.* **16** 629–647.

[12] Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics.* **27** 832–837.

[13] Sager, T.W. (1983). Estimating modes and isopleths. *Communications in Statistics – Theory and Methods.* **12(5)** 529-557.

[14] Samanta, M. (1973). Nonparametric estimation of the mode of a multivariate density. *South African Statistical Journal.* **7** 109–117.

[15] Silverman, B. (1978). Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *The Annals of Statistics.* **6** 177–184.

[16] Tsybakov, A.B. (1997). On nonparametric estimation of density level sets. *The Annals of Statistics.* **25** 948–969.

[17] Vieu, P. (1996). A note on density mode estimation. *Statistics and Probability Letters.* **26** 297–307.