A TAILOR–MADE NONPARAMETRIC DENSITY ESTIMATE

DANIEL CARANDO, RICARDO FRAIMAN, AND PABLO GROISMAN

Abstract. We consider a problem of nonparametric density estimation under shape restrictions. The first relevant result in this direction was the case of monotone (decreasing) densities considered by Grenander (1956). In spite of what happens under no restrictions, in this context the maximum likelihood estimate turns out to be a strongly consistent estimate. In our case, the shape restriction condition is that the density belongs to a class of Lipschitz functions with a uniformly bounded Lipschitz constant, a quite natural shape condition. Devroye (1987) considered these classes of estimates as tailor-made estimates, in opposite in some way to universally consistent estimates. In our framework the maximum likelihood estimate can be easily characterized but it is not easy to compute. Some simpler approximations are also considered.

1. Introduction.

It is well known that the maximum likelihood estimation method fails in the non–parametric setting of density estimation. This is because we consider as the parameter space the class of all density functions, which is too large. However, there are some smaller classes of densities, that are still a non–parametric family, where this is not the case. A relevant result in this direction is the case of monotone (decreasing) densities. For this problem, Grenander (1956) introduced an estimate defined as the derivative of the least concave majorant (concave envelope) of the empirical cumulative distribution function of the data. It turns out, that this estimate is the maximum likelihood estimate (MLE) restricted to the class of decreasing densities on R^+ (for a proof see, for instance, Grenander (1981) or Barlow, et al (1972)). The asymptotic behavior of Grenander's estimate has been studied by several authors, (see for instance Groeneboom (1985))

The first author is partially supported by CONICET Resolución N°1584-04, Universidad de Buenos Aires under grant X058 and ANPCyT PICT 03-15033. The third author is partially supported by Universidad de Buenos Aires under grant X066, Fundación Antorchas Project 13900-5, and by ANPCyT PICT No. 03-13719.

AMS 2000 Subject Classification. Primary 62G07; Secondary 62F30, 62G20.

Key words and phrases. Density Estimation, Maximum Likelihood, Tailor–made Estimates.

and in particular, it provides a simple strongly consistent estimate of the unknown decreasing density f.

An additional important property of this estimate is that in order to build it up, it does not require of any additional parameter like a bandwidth h or a number of nearest neighbors k. Of course the estimate will not be consistent if the true underlying density f is not monotone. In this sense we can consider Grenander's estimate as a tailor–made estimate (an expression coined by Devroye (1987)), in opposite in some way to universally consistent estimates. The extra information about the density function, allows to do the search for the estimate in a smaller class of functions sharing the extra properties we have assumed.

In what follows we will consider the case where the unknown density f is a Lipschitz function on its support. This restriction is quite natural in some real problems. For instance we often are dealing with situations in which we know that the speed at which the density can grow is clearly bounded. The density f is allowed to be discontinuous in the boundary of its support. In this context we will also obtain a "bandwidth free estimate", which turns out to be the maximum likelihood estimate. We also propose another simple consistent approximation. The "bandwidth free estimate" property is particularly important in high dimensions, since it is a way to avoid the well known problem of the curse of dimensionality that universally consistent estimates shares when we want to estimate a density function in high dimensions. However, it should be pointed that we need to know a bound for the Lipschitz constant C, which might be thought as a rough penalty parameter.

To be more precise, the problem will be to estimate a density function f on $\mathbb{R}^d$, from which we know in advance that has a bounded convex support $S(f)$ (but we do not know which the support is), and that is a Lipschitz function on its support with Lipschitz constant C, from a sample of i.i.d. random vectors $\{X_i : i \geq 1\}$ in $\mathbb{R}^d$, with density f. We start in Section 2 with the case of the maximum likelihood estimate in this setting, which we call the cone estimate. There we show existence, uniqueness and strong consistency of the estimate. If the support of f, $S(f)$, is known, then we do not need to require the support to be convex (see Theorem 2.2 below). For example, this will be the case if we are interested on uniform convergence over a fixed compact set K. In this case, we estimate the conditional density given that the

data are in K. Take into account that also outside a big compact set, the usual kernel density estimates, just reproduce the kernel shape. In sub–section 2.1 we provide an algorithm to calculate the estimate, and we give a few examples. In Section 3 we introduce a simpler approximation to the MLE estimate, which also turns out to be strongly consistent. We also include some examples that illustrate the behavior of the estimates.

In what follows, for each $n \geq 1$, $L(g)$ will stand for the log–likelihood function

$$L(g) = L_n(g) = \frac{1}{n} \sum_{j=1}^{n} \log\ g(X_i).$$

Also, we denote by $I(g)$ the integral of the function g over $\mathbb{R}^d$ with respect to the Lebesgue measure:

$$I(g) = \int_{\mathbb{R}^d} g\ d\mu$$

2. The cone estimate

Let $\{X_i\}_{i\geq 1}$ be independent and identically distributed random variables in $\mathbb{R}^d$ with common density f, defined on a probability space $(\Omega, \mathcal{A}, P)$. Through out we will assume that:

H1 The density f is supported in a convex compact set $S(f)$, and $f|_{S(f)}$ is a Lipschitz function with Lipschitz constant C ($f \in \mathbb{L}(C, S(f))$).

Having this knowledge of f one can look for a maximum likelihood estimate in this small class.

Let $\mathcal{F}_C$ be the class of densities $g : \mathbb{R}^d \to \mathbb{R}$ with convex compact support that verifies

$$|g(x) - g(y)| \leq C\|x - y\|, \qquad x, y \in S(g).$$

That is, $\mathcal{F}_C$ is the class of Lipschitz densities with prescribed Lipschitz constant C. We allow g to be discontinuous in the boundary of its support.

Without loss of generality we will assume throughout that $C = 1$. Otherwise, we consider the variables $Y_i = \sqrt{C}X_i$. These new variables have a density with Lipschitz constant 1. We denote $\mathcal{F} = \mathcal{F}_1$. If E is a closed subset of $\mathbb{R}^d$, we denote by $\mathcal{F}(E)$ the family of functions in $\mathcal{F}$ whose support is exactly E and $\bar{\mathcal{F}}(E)$ the family of functions that are Lipschitz (with Lipschitz constant 1) in E with support contained in E, but possibly smaller.

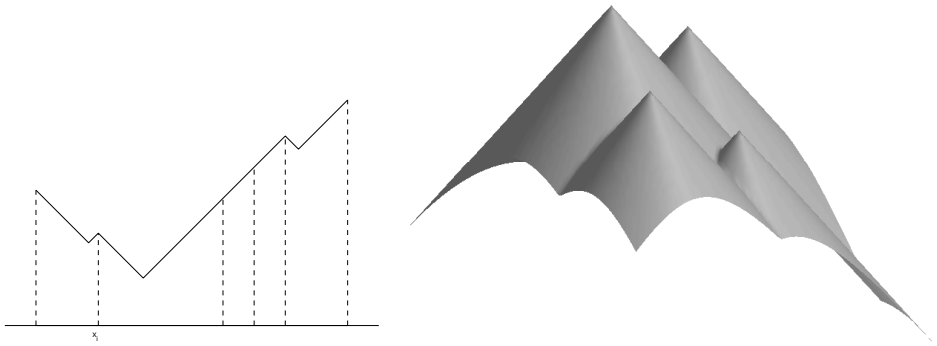

FIGURE 2.1. The maximizer in dimension one and two.

Although we are in a nonparametric setting, the following theorem proves that the maximum likelihood estimate is well defined and shows how it looks.

Theorem 2.1. *Under H1 we have that*

(i) *There exists a unique maximizer $\hat{f}_n$ of $L(g)$ in $\mathcal{F}$. Moreover, $\hat{f}_n$ is supported in $\mathcal{C}_n$, the convex hull of $\{X_1, \dots, X_n\}$, and its value there is given by the maximum of n "cone functions", i.e.*

$$\hat{f}_n(x) = \max_{1 \le i \le n} \left(\hat{f}_n(X_i) - \|x - X_i\| \right)^+. \tag{2.1}$$

(ii) *$\hat{f}_n$ is consistent in the following sense: for every compact set $K \subset S(f)^\circ$ (the interior of $S(f)$),*

$$\lim_{n \to \infty} \|\hat{f}_n - f\|_{L^\infty(K)} \to 0 \qquad a.s.$$

Hence $\hat{f}_n$ is determined by its values at the sample points and takes the form of a cone around each of them. If $d = 1$, $\hat{f}$ is piecewise linear, with slopes 1 or -1.

Proof. **Existence, uniqueness and characterization of the maximizer.**

First we show that for any $g \in \mathcal{F}$ there exists a function of the form (2.1) with at least the same likelihood. So let $g \in \mathcal{F}$ and consider $\overline{g}(x)$ supported in $\mathcal{C}_n$ and given by

$$\overline{g}(x) = \max_{1 \le i \le n} \left(g(X_i) - \|x - X_i\| \right)^+ \quad \text{for } x \in \mathcal{C}_n.$$

Observe that $L(\overline{g}) = L(g)$ but, since $g \in \mathbb{L}(1, \mathcal{C}_n)$, we have $g(x) \geq \overline{g}(x)$. Then $\int \overline{g} \leq 1$ and hence we can augment $\overline{g}$ uniformly in order to achieve $\int \overline{g} = 1$. The augmented version of $\overline{g}$ belongs to $\mathcal{F}$ and verifies $L(\overline{g}) \geq L(g)$.

Hence, the maximizer of L among functions of the form (2.1) (which exists since these functions form a compact class) is a global maximizer.

Uniqueness follows from the fact that any maximizer must lie in $\mathbb{L}(1, \mathcal{C}_n)$. This class is convex and L is a concave functional.

Consistency. The proof of (ii) is based on Theorem 1 in Huber (1967). In this direction it is desirable to look for the maximizer in a compact class (see Lemma 1 in Huber (1967)). Unfortunately $\mathcal{F}$ is not compact and it is not clear that there exists a compact set in which $\hat{f}_n$ almost surely ultimately stays. Hence we introduce an auxiliary statistic f_n that lies in the compact class $\bar{\mathcal{F}}(S(f))$ for every $n \geq 1$. Let

$$f_n := A_n \max_{1 \leq i \leq n} \left(\hat{f}_n(X_i) - \|x - X_i\| \right)^+, \qquad \text{for all } x \in S(f).$$

The constant A_n is chosen to guarantee $I(f_n) = 1$. Observe that the difference between equation (2.1) and the above formula is that the latter holds for all $x \in S(f)$, while (2.1) gives the value of $\hat{f}_n$ only for $x \in \mathcal{C}_n$.

This statistic cannot actually be computed since $S(f)$ in unknown, but we are going to prove that is asymptotically equivalent to $\hat{f}_n$ and consistent. This will prove the consistency of $\hat{f}_n$.

Recall that, since the support of $\hat{f}_n$ is the convex hull of the sample points $\{X_1, \ldots, X_n\}$, the Hausdorff distance $dist_H(S(\hat{f}_n), S(f)) \to 0$ (see for instance Rényi and Sulanke (1963, 1964) or Dumbgen and Walther (1996) for rates of convergence). This means, on the one hand, that given any compact subset $K \subset S(f)^\circ$, K is contained in $S(\hat{f}_n)$ for n large enough a.s. On the other hand, we have that the Lebesgue measure $\mu(S(f) \setminus S(\hat{f}_n)) \to 0$ as $n \to \infty$. From these two observations we have that $A_n \to 1$ and for large n

$$\|f_n - \hat{f}_n\|_{L^\infty(K)} \leq |A_n - 1| \, \|\hat{f}_n\|_{L^\infty(K)} \to 0, \tag{2.2}$$

since $(\|\hat{f}_n\|_{L^\infty(K)})_n$ is bounded a.s.

Next we define $T_n \colon \mathbb{R}^n \to \bar{\mathcal{F}}(S(f))$ by

$$T_n(X_1, \ldots, X_n) = f_n,$$

and we check that T_n is a sequence of maximum likelihood estimates in the more general sense that,

$$-L(T_n) + \inf_{g \in \bar{\mathcal{F}}(S(f))} L(g) \to 0, \quad \text{a.s.}, \tag{2.3}$$

as defined in Huber (1967). Indeed

$$0 \leq -L(T_n) + \inf_{g \in \bar{\mathcal{F}}(S(f))} L(g) \leq -L(T_n) + \inf_{g \in \mathcal{F}(\mathcal{C}_n))} L(g) = -L(T_n) + L(\hat{f}_n),$$

which converges to zero a.s. since

$$-L(T_n) + L(\hat{f}_n) = -\log A_n \to 0.$$

It remains to proof assumptions (A-1), (A-2'), (A-3) and (A-4) of Huber (1967), namely

(A-1) $\rho(x, g) := -\log(g(x))$ *is separable in the sense of Doob.*

(A-2') *Consider a family of neighborhoods* U *of* f *that shrinks to* $\{f\}$*, then*

$$\inf_{g \in U} -\log(g(X)) \to -\log(f(X)), \qquad (U \to \{f\}) \text{ a.s.}$$

(A-3) $E((-\log(g(X)))^-) < \infty$ *for every* $g \in \bar{\mathcal{F}}(S(f))$ *and*
$E((-\log(g(X)))^+) < \infty$ *for some* $g \in \bar{\mathcal{F}}(S(f))$

(A-4) $E(-\log(g(X)) > E(-\log(f(X))$ *for all* $g \in \mathcal{F}(S(f)), \quad g \neq f$.

Assumption (A-1) holds since $\Theta := \bar{\mathcal{F}}(S(f))$ is compact and separable. (A-2') is immediate.

Since $E((-\log(g(X)))^-) = E((\log(g(X)))^+)$ and $S(f)$ is compact, the first statement of (A-3) holds. For the second, one can take any function g strictly positive in $S(f)$.

To prove (A-4) define $Y = \log(g(X)) - \log(f(X))$ and recall that if Y is not a constant (a.s.) and $E(|Y|) < \infty$, we have by Jensen's inequality

$$E(Y) < \log E(e^Y).$$

Since in our case $E(e^Y) = I(g) = 1$, we have

$$E(\log(g(X)) - E(\log(f(X)) = E(Y) < 0.$$

Details of this argument can be found in Wald (1949).

Therefore, we can apply Huber's Theorem to conclude that f_n is consistent in $L^\infty(S(f))$ and hence it is also consistent in the topology of uniform convergence on compact subsets of $S(f)$. From (2.2) we get the consistency of $\hat{f}_n$. □

We want to remark that (2.3) means that T_n is a maximum likelihood estimate in the sense described by Huber (1967). These estimates are in a more general setup, and allows that different estimates fall in this framework. In particular, some asymptotically equivalent estimates verify (2.3). This approach has been particularly fruitful in the robust literature, where M-estimates can be considered as generalized maximum likelihood estimates under non standard conditions (i.e. when the true underlying distribution is not exactly that of the considered parametric model).

One of the strengths of the theorem is that the support $S(f)$ of the density f is unknown (in fact, the estimate $\hat{f}_n$ involves an estimation of $S(f)$). This is the main reason of the hypothesis of convexity imposed to $S(f)$. If the support $S(f)$ is known in advance, we do not need the convexity assumption. In this case, we consider $\mathcal{F}(S(f))$, the set of densities with support $S(f)$ that are Lipschitz (with constant 1) on $S(f)$. The set $\mathcal{F}(S(f))$ is convex and, by the Arzela-Ascoli theorem, it is also compact. This places us in a very good situation for both the maximization of L and the consistency of the maximizer (the hypotheses of Huber's theorem are easier to verify when there is a fixed compact set containing all possible estimates). Therefore, the proof of the previous theorem can be considerably simplified to obtain:

Theorem 2.2. *Suppose $S(f)$ is known (not necessarily convex). Then under H1 we have that:*

(i) *There exists a unique maximizer $\hat{f}_n$ of $L(g)$ in $\mathcal{F}(S(f))$, which verifies*

$$\hat{f}_n(x) = \max_{1\leq i\leq n} \left(\hat{f}_n(X_i) - \|x - X_i\|\right)^+.$$

(ii) *$\hat{f}_n$ is consistent*

As pointed in the Introduction, there are many situations where is reasonable to assume that the support is known. On the other hand, the case of non-convex and unknown support can be also dealt with in practice. One should first estimate the support of f by a set estimation method (see, for instance, Cuevas and Rodriguez-Casal (2003) for a review on this field) and then define the cone estimate on the estimated support. This will be the subject of a future work.

In what follows, we will deal with the situation of a convex and unknown support $S(f)$. However, most of the results can be adapted to handle the case of (non-convex) known support. The proofs in the known-support setting are generally simpler (as it happens with Theorems 2.1 and 2.2).

2.1. **How to compute it.** Although the theorem above characterizes the estimator, we do not have an explicit formula for it based in the sample points. By equation (2.1), it remains to determine the value of $\hat{f}_n(X_i)$ for each $i = 1, \dots, n$. But at this stage, we have a finite dimensional optimization problem. So we define the following space

$$\mathcal{W} = \mathcal{W}(X_1, \dots, X_n) = \left\{ g \in \mathbb{L}(1, \mathcal{C}_n) : g(x) = \max_{1 \le i \le n} \Big(g(X_i) - \|x - X_i\| \Big)^+ \mathbf{1}_{C_n}(x) \right\}.$$

In view of the observations that we made in the proof of Theorem 2.1, the unique maximizer of L in $\mathcal{F}$ must belong to $\mathcal{W}$. Hence $\hat{f}_n$ solves the following optimization problem:

$$\text{maximize } \prod_{i=1}^{n} g(X_i) \quad \text{subject to } g \in \mathcal{W} \text{ and } I(g) = 1. \tag{2.4}$$

Since this is a finite dimensional problem, we restate it as an optimization problem in $\mathbb{R}^n$. Let $y = (y_1, \dots, y_n) \in \mathbb{R}^n$ be such that $|y_i - y_j| \le \|X_i - X_j\|$. We define g_y as:

$$\hat{g}_y(x) = \max_{1 \le i \le n} \Big(\hat{y}_i - \|x - X_i\| \Big)^+ \qquad (x \in \mathcal{C}_n),$$

in other words, g_y is the only function in $\mathcal{W}$ that takes the value y_i at X_i. Therefore, (2.4) can be stated as follows:

$$\text{maximize } \ell(y) = \prod_{i=1}^{n} y_i, \quad y \in \Omega \tag{2.5}$$

$$\text{where } \Omega = \{ y \in \mathbb{R}^n : y_i > 0 \text{ and } |y_i - y_j| \le \|X_i - X_j\| \text{ for } i \ne j \text{ and } I(g_y) = 1 \}.$$

Numerical solutions to this problem can be obtained with most of the numerical methods for optimization problems. We have used the `fmincon` routine provided by MATLAB. Convergence to the optimum is guaranteed since ℓ is concave and Ω is a convex subset of $\mathbb{R}^n$. To show that Ω is a convex set, consider the mapping $T : \{ g \in \mathcal{W} : I(g) = 1 \} \longrightarrow \Omega$ given by $T(g) = (g(X_1), \dots, g(X_n))$. T is a bijective mapping and verifies

$$T(\alpha g_1 + (1 - \alpha) g_2) = \alpha T(g_1) + (1 - \alpha) T(g_2) \quad \text{ for } 0 \le \alpha \le 1.$$

Since $\{ g \in \mathcal{W} : I(g) = 1 \}$ is convex, so is Ω.

In the one dimensional case $d = 1$, given $y \in \Omega$, the integral $I(g_y)$ can be easily computed. In fact, if $X^{(1)}, \dots, X^{(n)}$ stands for the order statistics of the vector $(X_1, \dots, X_n)$, we have

$$I(g_y) = \frac{1}{4} \sum -(X^{(i+1)} - X^{(i)})^2 + 2(y_{i+1} + y_i)(X^{(i+1)} - X^{(i)}) + (y_{i+1} - y_i)^2$$

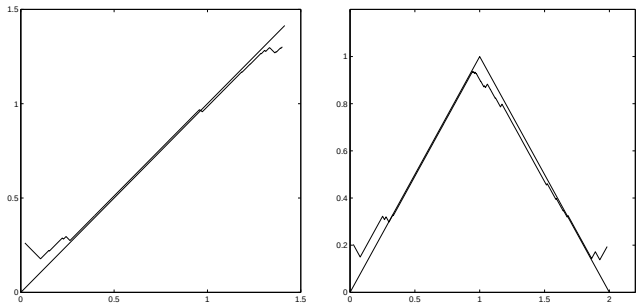

FIGURE 2.2. Two realizations of the cone estimator and the underlying densities. Left: a sample from the density $f(x) = x\mathbf{1}_{[0,\sqrt{2}]}$. Right: a sample from the sum of two uniform random variables. Sample sizes: n=100.

Also, the Lipschitz condition is simply

$$-(X^{(i+1)} - X^{(i)}) \leq y_{i+1} - y_i \leq X^{(i+1)} - X^{(i)}, \quad \text{for } i = 1, \ldots, n-1.$$

Figure 2.2 shows the realization of the cone estimator for two different distributions.

In higher dimensions ($d > 1$), it is not so simple to obtain a formula for $\int g_y$. However, Monte-Carlo methods can be employed to compute this integral. It is important to note that in high dimensions, more effort will be needed to compute the integral, but that the number of restrictions does not depend on d, it only depends on the number of sample points. For $d \geq 2$, the number of restrictions can be roughly bounded by $n(n-1)/2$, the number of pairs of sample points which should verify the Lipschitz condition. In the next section, an alternative estimator is presented, that simplifies the computation of the integral and decreases the number of restrictions.

3. An alternative Maximum likelihood-type estimate

As we observed in the previous section, many different estimators can be viewed as maximum likelihood-type estimates. In this section we consider an alternative estimator to the one described above.

This new estimator is smoother and cheaper (in terms of amount of computations) than the one described in the previous section, $\hat{f}_n$. It can be viewed basically as a modification of it.

This new estimator will lead us to a simpler optimization problem: the integral $I(g)$ will be easier to compute and the number of restrictions will be substantially lower for $d \geq 2$. For the sake of simplicity we first analyze the one dimensional case.

3.1. **Dimension one: Piecewise linear maximum likelihood.** We observed that the MLE described in the previous section on the one hand has too many peaks and on the other hand a nonlinear problem has to be solved in order to compute it. In order to avoid these two problems we propose to look for a maximum likelihood estimate just in the class of piecewise linear densities with knots at the sample points. Let $X^{(1)}, \dots, X^{(n)}$ stand for the order statistics of the vector $(X_1, \dots, X_n)$. Now consider

$$\mathcal{V} = \mathcal{V}(X_1, \dots, X_n) = \{g \in \mathcal{F} : g|_{[X^{(i)}, X^{(i+1)}]} \text{ is linear } \},$$

and let $\tilde{f}_n$ be the maximum of L over $\mathcal{V}(X_1, \dots, X_n)$. We will call this estimator the PLMLE. Existence and uniqueness of this estimator is guaranteed since $\mathcal{V}$ is a finite dimensional compact and convex subset of $\mathcal{F}$.

Although $\tilde{f}_n$ has lower likelihood than $\hat{f}_n$, it has some nice properties that $\hat{f}_n$ does not posses. For example, in order to compute $\tilde{f}_n$ we only need to solve a linear problem, which means faster algorithms and lower errors. In addition, this estimator presents less oscilations. We will also show that $\tilde{f}_n$ is a maximum likelihood-type estimator too in the sense of (2.3). If we define $V(g)$ as the linear interpolant of the points $(X^{(i)}, g(X^{(i)}))$, we have

$$L\left(\frac{\hat{f}_n}{I(V(\hat{f}_n))}\right) \leq L(\tilde{f}_n) \leq L(\hat{f}_n).$$

The first inequality holds since

$$L\left(\frac{\hat{f}_n}{I(V(\hat{f}_n))}\right) = L\left(V\left(\frac{\hat{f}_n}{I(V(\hat{f}_n))}\right)\right),$$

$V(\hat{f}_n / I(V(\hat{f}_n))$ belongs to $\mathcal{V}$ and therefore has lower likelihood than $\tilde{f}_n$.

We observe that for any $g \in \mathcal{W}$,

$$I(g) + \sum_{i=1}^{n-1} \left(X^{(i+1)} - X^{(i)}\right)^2 \geq I(V(g)) \geq 1. \tag{3.1}$$

Note that

$$\sum_{i=1}^{n-1} \left(X^{(i+1)} - X^{(i)}\right)^2 \leq \mu(S(g)) \max_{1 \leq i \leq n-1} \left(X^{(i+1)} - X^{(i)}\right).$$

Since the maximal spacing converges almost surely to 0 – see for instance Devroye (1981), Deheuvels (1983) – so does $\sum_{i=1}^{n-1} \left(X^{(i+1)} - X^{(i)}\right)^2$. Then, $I(V(\hat{f}_n)) \xrightarrow{a.s.} 1$ and

$$L\left(\frac{\hat{f}_n}{I(V(\hat{f}_n))}\right) - L(\hat{f}_n) = -L(I(V(\hat{f}_n))) \xrightarrow{a.s.} 0.$$

In consequence,

$$0 \geq L(\tilde{f}_n) - \max_{g \in \mathcal{F}} L(g) \geq L\left(\frac{\hat{f}_n}{I(V(\hat{f}_n))}\right) - L(\hat{f}_n) \to 0$$

holds almost surely. So $\tilde{f}_n$ verifies equation (1) and therefore is in the context of Huber's Theorem.

In order to compute the linear estimator $\tilde{f}_n$ we observe that if the sample takes the values $(x_1, \ldots, x_n)$ (assume that they are sorted), then we have to solve the following optimization problem

maximize $\prod_{i=1}^{n} y_i$; subject to

$$-a \leq Ay \leq a, \qquad By = 1.$$

The matrices A, a and B read as

$$A = \begin{pmatrix} -1 & 1 & 0 & \cdots & & 0 \\ 0 & -1 & 1 & & & \\ \vdots & & \ddots & \ddots & & \vdots \\ & & & -1 & 1 & 0 \\ 0 & \cdots & & 0 & -1 & 1 \end{pmatrix}, \qquad a = \begin{pmatrix} x_2 - x_1 \\ \vdots \\ x_{i+1} - x_i \\ \vdots \\ x_n - x_{n-1} \end{pmatrix},$$

$$B = \frac{1}{2}\left(x_2 - x_1, x_3 - x_1, \ldots, x_{i+1} - x_{i-1}, \ldots, x_n - x_{n-2}, x_n - x_{n-1}\right)$$

The equation $-a \leq Ay \leq a$ guarantees the Lipschitz condition and $By = 1$ represents the restriction $I(\tilde{f}) = 1$.

Figure 3.1 shows the PLMLE in dimension one for two samples: the first was obtained from the sum of two uniform random variables and the second from the maximum of two uniform random variables. The estimated densities are plotted together with the real densities and the estimation is compared to the kernel estimation of the same samples. Opposite to the kernel estimation, the PLMLE estimator does not assume any particular behavior of the density near the boundary of its support. This is more apparent in the case when the density is not zero in the boundary.

Remark Observe that we take $\mathcal{V}$ to be a piecewise linear function space, but is also possible to take $\mathcal{V}$ as a space of spline functions of higher order. In this case we loss the linear essence of the optimization problem but we gain in regularity.

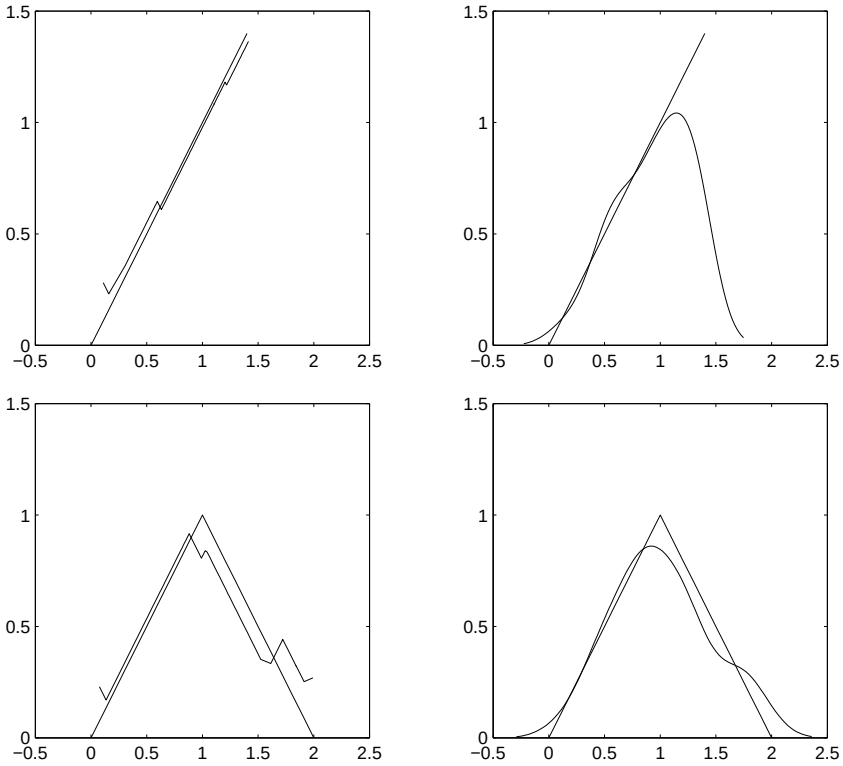

FIGURE 3.1. The PLMLE and the underlying density (left) compared with the kernel estimation (right) for the same sample of size 100.

We also observe that this method gives another spline approach to nonparametric estimation that, in general, do not coincide with the well known penalized maximum likelihood estimate.

3.2. **Higher dimensions.** Now we introduce the $d-$dimensional version of the estimator described above. We get back to the case $f : \mathbb{R}^d \to \mathbb{R}$ with Lipschitz constant 1 when restricted to its support. Let $\{X_i\}_{i\geq 1}$ be independent and identically distributed random vectors with common density f.

We consider the Delaunay tessellation T of the points $X_1, \ldots, X_n$. This tessellation consists of a set of simplices $\{\tau_1, \ldots, \tau_N\}$ such that none of the data points is contained in any circumspheres of the simplices. The construction of the Delaunay tessellation is based on Voronoi diagrams (see for example the book by George and Borouchaki (1998)).

These tessellations have many desirable properties among which we want to stress the following:

For any $i \neq j$, $\tau_i \cap \tau_j$ is either a point, a $(d-1)-$dimensional face, or the empty set.

If we consider now the class

$$\mathcal{V} = \mathcal{V}(X_1, \ldots, X_n) = \{g \in \mathcal{F} : g|_{\tau_i} \text{ is linear}\},$$

we can define as in the previous section, the PLMLE $\tilde{f}_n$ as the argument that maximizes L over $\mathcal{V}(X_1, \ldots, X_n)$.

As a consequence of the above property, functions of $\mathcal{V}$ are univocally determined by their value at the sample points. In fact, $\mathcal{V}$ is a compact subset of the finite dimensional linear space $\tilde{\mathcal{V}}$ of continuous functions g defined on $\cup_k \tau_k$ which are linear in each τ_k. A basis of $\tilde{\mathcal{V}}$ is given by the functions φ_i, $1 \leq i \leq n$ defined by

$$\varphi_i(X_j) = \delta_{i,j}.$$

A function $g \in \tilde{\mathcal{V}}$ has the representation

$$g(x) = \sum_i y_i \varphi_i(x),$$

where $y_i = g(X_i)$. These spaces and bases are frequently used when applying the well known finite element method for the numerical treatment of partial differential equations (see for example Ciarlet (1978)).

To prove consistency of the estimator we proceed as in the one-dimensional case, the only modification being in equation (3.1), replacing

$$\sum_{i=1}^{n-1} \left(X^{(i+1)} - X^{(i)}\right)^2$$

by

$$\sum_{i=1}^{N-1} |\tau_i| \text{diam}(\tau_i).$$

The result that ensures that maximal spacing

$$\max_{1 \leq i \leq n-1} \left(X^{(i+1)} - X^{(i)}\right) \to 0 \text{ a.s.}$$

must be replaced by a similar result for a notion of multivariate spacings. This will be done using the notion of multivariate spacings introduced by Deheuvels (1983) (see also Deheuvels et al (1988) for a more general setting). The maximal k-spacing $M_{k,n}$ with respect to a family of regular subsets $\mathbf{C}$ (which in our case will be the Euclidean balls) is defined in Deheuvels et al (1988) as

$$M_{k,n} = \sup\{\mu(C) : C \in \mathbf{C} \text{ and } nP_n(C) < k\},$$

where $P_n(.)$ stands for the empirical measure associated with $X_1, \dots, X_n$ and μ is the Lebesgue measure on $\mathbb{R}^d$. We use the results in Deheuvels et al (1988), on the asymptotic behaviour of the second spacing $M_{2,n}$ for the class $\mathbf{C}$ of Euclidean balls (Theorem 1) to obtain

$$\sum_{i=1}^{N-1} |\tau_i| \mathrm{diam}(\tau_i) \leq \mu(S(g)) \max_{1\leq i\leq N} \mathrm{diam}(\tau_i) \leq \mu(S(g)) M_{2,n} \xrightarrow{a.s.} 0.$$

So, $\tilde{f}_n$ verifies (2.3) and hence we are ready to prove the following

Theorem 3.1. *Assume H1, then for every compact set $K \subset S(f)$ we have*

$$\|\tilde{f}_n - f\|_{L^\infty(K)} \to 0 \qquad a.s.$$

Proof. It is immediate once we have proved (2.3). The only point to be careful is in the fact that, as in Theorem 2.1, $\tilde{f}_n$ is not Lipschitz in the whole $S(f)$. To avoid this problem, we proceed as in the proof of that theorem by considering an auxiliary statistic asymptotically equivalent to $\tilde{f}_n$. This statistic can be constructed extending $\tilde{f}_n$ from $\mathcal{C}_n$ to the hole $S(f)$ by any function that preserves the Lipchitzianity (and the norm) and the positivity of $\tilde{f}_n$. The fact that this auxiliary statistic is asymptotically equivalent to $\tilde{f}_n$ can be proved exactly as in Theorem 2.1. Likewise, (2.3) and assumptions (A-1), (A-2'), (A-3), (A-4) hold. Therefore, it is consistent and so is $\tilde{f}_n$. □

3.3. **Computation.** To compute this estimator we observe that if the density $g\colon \mathbb{R}^d \to \mathbb{R}$ is regular, the Lipschitz condition

$$|g(x) - g(y)| \leq \|x - y\|$$

is equivalent to

$$\sup_x \|\nabla g(x)\|^* \leq 1,$$

where

$$\|\nabla g(x)\|^* = \sup_y \frac{|\nabla g(x) y|}{\|y\|}$$

is the norm induced by the norm considered in $\mathbb{R}^d$. It is well known that, for $1 \leq p \leq \infty$ we have $\|\cdot\|_p^* = \|\cdot\|_q$, where $\frac{1}{p} + \frac{1}{q} = 1$.

In order to have the regularity required in the above paragraph, it is enough, for example, to have a tessellation $T = \{\tau_1, \dots, \tau_N\}$ such that $g|_{\tau_k}$ is differentiable and $g|_{\cup \tau_k}$ is continuous.

If $g \in \tilde{\mathcal{V}}$, then $\nabla g|_{\tau_k} \equiv \nabla_k g$ is constant for all k and hence $g \in \mathcal{V}$ if and only if

$$\|\nabla_k g\|^* \leq 1, \qquad \text{for all } 1 \leq k \leq N$$

and

$$I(g) = 1.$$

Note that, if the sample takes the values $(x_1, \ldots, x_n)$, then

$$I(g) = I\left(\sum_{i=1}^{n} g(x_i)\varphi_i\right) = By,$$

where $B = (I(\varphi_1), \ldots, I(\varphi_n))$ and $y = (g(x_1), \ldots, g(x_n))$. We also have

$$\nabla_k g = \sum_i y_i \nabla_k \varphi_i = A_k y,$$

where A_k is the matrix whose $i-$th column is the gradient of the $i-$th basis function φ_i restricted to the simplex τ_k (the gradients are constant on each τ_k). That is:

$$A_k = (\nabla_k^t \varphi_1 | \ldots | \nabla_k^t \varphi_n).$$

Hence, our optimization problem reads as follows.

maximize $\prod_{i=1}^{n} y_i$; subject to

$$\|A_k y\|^* \leq 1, \quad 1 \leq k \leq N, \qquad By = 1.$$

Observe that if $\|\cdot\|^* = \|\cdot\|_\infty$, the above problem has linear restrictions. That is the case when $\|\cdot\|_1$ is considered in $\mathbb{R}^d$.

Remark. Observe that all the optimization problems treated above have the form:

minimize $\alpha(x)$; subject to

$$h_1(x) \leq 0$$

$$h_2(x) = 0.$$

where α is concave and h_1 and h_2 are convex functions. Hence, standard algorithms for convex programming problems can be applied to compute the estimator. The concavity/convexity ensures convergence in all of our situations. We have used the `fmincon` routine provided by MATLAB. For a description of the algorithm and further references see

`http://www.mathworks.com/access/helpdesk/help/toolbox/optim/fmincon.html`

Figure 3.3 shows the bidimensional PLMLE (left) from a sample of size 250 together with the underlying density (right). Finally, Figure 3.3 shows the estimation of a uniform random variable with just 200 observations. Observe that these are rather small samples for two-dimensional problems.

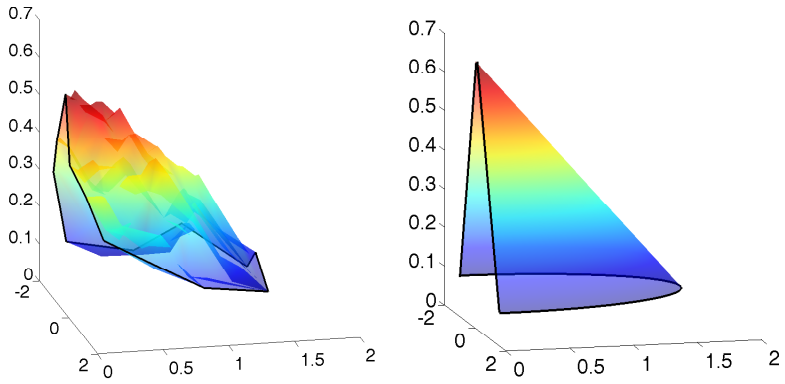

FIGURE 3.2. The PLMLE (left) for a sample of size 250 and the underlying desnsity (right).

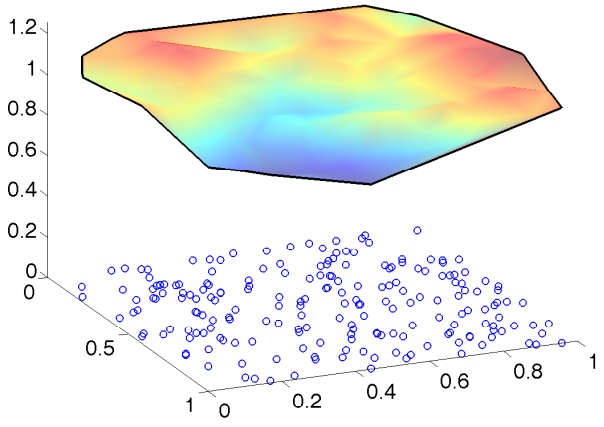

FIGURE 3.3. The PLMLE for a sample of size 200 of a uniform variable over the unit square and the sample points.

ACKNOWLEDGMENTS

We want to thank Ricardo Durán for very helpful comments about Delaunay triangulations.

REFERENCES

BARLOW R. E., BARTHOLOMEW D. J., BREMNER J. M., AND BRUNK H. D. (1972) *Statistical inference under order restrictions. The theory and application of isotonic regression.* John Wiley & Sons, London-New York-Sydney. Wiley Series in Probability and Mathematical Statistics.

Ciarlet, P. G. (1978) *The finite element method for elliptic problems. Studies in Mathematics and its Applications, Vol. 4* North-Holland Publishing Co.,Amsterdam.

Cuevas, A. and Rodríguez-Casal, A. (2003) *Set estimation: an overview and some recent developements.* In *Recent advances and trends in nonparametric statistics.* Arkitas, M. and Politis, D., editors. Elsevier.

Deheuvels P. (1983) Upper bounds for kth maximal spacings. *Z. Wahrsch. Verw. Gebiete*, 62, 465–474.

Deheuvels P., Einmahl J. H. J., Mason D. M., and Ruymgaart , Frits H. (1988) The almost sure behavior of maximal and minimal multivariate k_n-spacings. *J. Multivariate Anal.*, 24, 155–176.

Devroye L. (1981) Laws of the iterated logarithm for order statistics of uniform spacings. *Ann. Probab.*, 9, 860–867.

Devroye L. (1987) *A course in density estimation*, volume 14 of *Progress in Probability and Statistics*. Birkhäuser Boston Inc., Boston, MA.

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

George, P.L. and Borouchaki, H., *Delaunay triangulation and meshing.* Editions Hermès, Paris, 1998.

Grenander U. (1956) On the theory of mortality measurement. II. *Skand. Aktuarietidskr.*, 39, 125–153.

Grenander U. (1981) *Abstract inference*. John Wiley & Sons Inc., New York. Wiley Series in Probability and Mathematical Statistics.

Groeneboom P. (1985) Estimating a monotone density. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II (Berkeley, Calif., 1983)*, Wadsworth Statist./Probab. Ser., 539–555, Belmont, CA. Wadsworth.

Huber P. J. (1967) The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66), Vol. I: Statistics*, 221–233. Univ. California Press, Berkeley, Calif.

Rényi, A. and Sulanke, R. (1963) Uber die konvexe Hülle von n zufällig gewählten Punkten. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 2, 75–84.

Rényi, A. and Sulanke, R. (1964) Uber die konvexe Hülle von n zufällig gewählten Punkten II. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 3, 138–147.

Wald, A. (1949) Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statistics*, 20, 595–601.

Departamento de Matemática, Universidad de San Andrés, Vito Dumas 284 (1644) Victoria, Pcia. de Buenos Aires, Argentina

E-mail address: daniel@udesa.edu.ar, rfraiman@udesa.edu.ar

Instituto de Cálculo, FCEyN, UBA, (1428) Buenos Aires, Argentina.

E-mail address: pgroisma@dm.uba.ar