The Annals of Statistics
1984, Vol. 12, No. 4, 1263–1284

ON A CLASS OF NONPARAMETRIC DENSITY AND REGRESSION ESTIMATORS[1]

BY V. K. KLONIAS

The Johns Hopkins University

A class of maximum penalized likelihood estimators (MPLE) of the density function f is constructed, through the use of a rather general roughness-penalty functional. This class contains all the density estimates in the literature that arise as solutions to MPLE problems with penalties on $f^{1/2}$. In addition, the flexibility of the penalty functional permits the construction of new spline estimates with improved performance at the peaks and valleys of the density curves. The consistency of the estimators in probability and a.s., in the $L_p(\mathbb{R})$ – norms, $p = 1, 2, \infty$, in the Hellinger metric and Sobolev norms is established in a unified manner. A class of penalty functionals is identified which leads to estimators which approach the optimal rates of convergence predicted in Farrell (1972). Based on the above estimates, a class of MPLE regression estimators is introduced which has the appealing property of reducing to the classical nonparametric regression estimates when a smoothing parameter goes to zero. Finally, a theoretically justifiable and numerically efficient method for a data based choice of the smoothing parameter is proposed for further study. A number of numerical examples and graphs are presented.

1. Introduction. Let $X_1, \cdots, X_n$ be independent observations from a distribution function F with density f over $\mathbb{R}^p$, $p \in \mathbb{Z}_+$ and let F_n denote the associated empirical distribution function. The nonparametric maximum penalized likelihood method of density estimation (MPLE) introduced by Good and Gaskins (1971) and (1980), produces as the estimator of f the maximizer of the log-likelihood minus a "roughness" penalty functional $\Phi(v)$ which is usually expressed in terms of the root density $v = f^{1/2}$, e.g., $\Phi_1(v) = \alpha \int (v')^2$, $\Phi_2(v) = \beta \int (v')^2 + \alpha \int (v'')^2$ with $\alpha > 0$, $\beta \geq 0$. In DeMontricher, Tapia and Thompson (1975), the existence and uniqueness of the MPLE's were rigorously established within the framework of Sobolev spaces $W^{2,\ell} = \{u \in L_2(\mathbb{R})\colon \| u^{(\ell)} \|_2 < +\infty\}$, with ℓ a positive integer; $L_2(\mathbb{R})$ denotes the space of square integrable functions and $\| u \|_2^2 = \int u^2$. They also introduced a class of MPLE with finite support and penalties expressed in terms of f. For a discretized version of the MPLE problem see Tapia and Thompson (1978) and Scott, Tapia and Thompson (1980). In Silverman (1982) the MPLE problem is extensively studied in the case where the "roughness" penalty is imposed on $\log f$ rather than $f^{1/2}$, e.g. $\Phi_2(\log f)$. For

Received July 1983; revised June 1984.

[1] Research supported in part by the Office of Naval Research Contract No. N00014-79-C-0801.

AMS 1970 *subject classifications.* Primary 62G05, 62G10, 60F15, 41A25; secondary 60F25, 40A30, 62E10.

Key words and phrases. Nonparametric density estimation, nonparametric regression estimation, maximum penalized likelihood method, strong consistency, rates of convergence, likelihood functional, spline function, data based smoothing parameters.

another penalty method for the estimation of the score function (and hence the density), see Cox (1983).

The MPLE's u of $f^{1/2}$ we consider here, are solutions to the following optimization problem:

$$\max\left\{\textstyle\sum_{i=1}^{n} \log u(X_i)^2 - \lambda(2\pi)^{-p} \displaystyle\int |\tilde{u}|^2 \, d\mu,\ u \in H\right\} \tag{1.1}$$
$$\text{subject to } u(X_i) \geq 0, \quad i = 1, \cdots, n,$$

where $\tilde{u}$ denote the Fourier transform of u, $H = \{u \in L_2(R^p), \int |\tilde{u}|^2 \, d\mu < +\infty\}$ and $\lambda > 0$ is such that $\int u^2 = 1$. A unique MPLE u corresponds to each positive measure μ, dominated by the Lebesgue measure with density $m(t)$, and is a spline function, given implicitly by

$$u(x) = \lambda^{-1} \textstyle\sum_{i=1}^{n} u(X_i)^{-1} \kappa_\mu(x_1 - X_{i1}, \cdots, x_p - X_{ip}), \quad x \in \mathbb{R}^p, \tag{1.2}$$

with $m\tilde{\kappa}_\mu = 1$. We then estimate f by $f_n = u^2$. In Sections 3 and 4, where the consistency of these estimators is discussed, we will let μ depend on a parameter $h \in \mathbb{R}^p$, so that $m(t) = m_0(h_1 t_1, \cdots, h_p t_p)$, $t \in \mathbb{R}^p$, with $h_j > 0$, $j = 1, \cdots, p$. Then,

$$\kappa_\mu(z) = (h_1 \cdots h_p)^{-1} k(z_1/h_1, \cdots, z_p/h_p), \quad z \in \mathbb{R}^p,$$

where k is such that $\tilde{k} m_0 = 1$.

Note that if we set

$$m(t) = \textstyle\sum_{j=0}^{\ell} a_j t^{2j}, \tag{1.3}$$

with a_0, $a_\ell > 0$ and $a_j \geq 0$, $j = 1, \cdots, \ell - 1$ for some positive integer ℓ; by an application of Parseval's Theorem (see, e.g., Yosida, 1970, page 154) problem (1.1) is seen to be equivalent to that treated in DeMontricher et al. (1975), giving for $\ell = 1, 2$ the "first and second MPLE of Good and Gaskins", corresponding to Φ_1, Φ_2 and being generated by kernels κ_μ of the form

$$(a/2)\exp\{-a \,|\, x \,|\}$$

and

$$[4 \,|\, ab \,|\, (a^2 + b^2)]^{-1} \exp\{- \,|\, ax \,|\}[\,|\, b \,|\cos|\, bx \,| + |\, a \,|\, \sin \,|\, bx \,|\,]$$

respectively. For $\ell > 2$, κ_μ is a convolution of these two kernels.

In general, density estimators tend to underestimate the "peaks" and overestimate the "valleys" of the curves. This issue has been addressed in a number of papers, see, e.g., Wahba (1976), Breiman, Meisel and Purcell (1977), Hall (1983). The generality of the measure μ we are allowed in Problem (1.1) and the fact that $f_n = u^2$ is nonnegative even if κ_μ is not, permits the consideration of the following approach to the construction of the "roughness" penalty, which seems to improve the performance of the estimates at the "peaks" and "valleys" of the density curve (compare e.g., Figures 1C and 3B).

Since we use global penalties for "roughness" and since the square root is a variance stabilizing transformation for the probability density estimation prob-

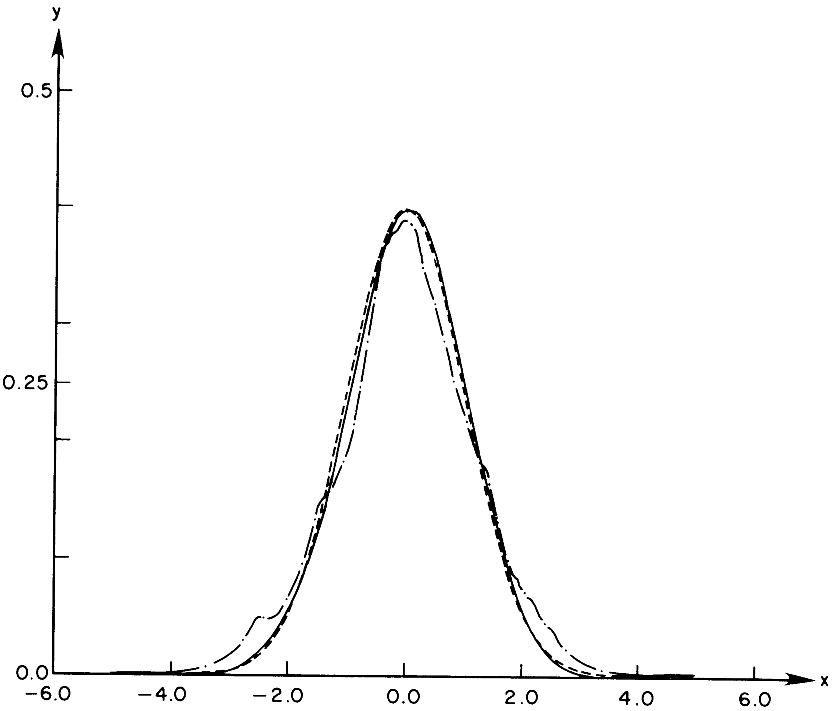

FIG. 1. A: --- *standard normal density.* B: -·- *kernel estimator,* $h = .450$. C: —— *MPLE with* $k = \phi - \phi''$, $h = 1.275$.

lem, it seems preferable to control the "smoothness" of the root-density, as in Good and Gaskins (1971) rather than the density estimator. Also, in view of the over-and-under-estimation problems mentioned above it is desirable to penalize lighter for "roughness" near the "peaks" and "valleys" of the curve, the location of which is unknown. To this end, note that if we convolute u with a symmetric around zero density $\nu_h = h^{-1}\nu(\cdot/h)$, $h > 0$, we have that

$$(u^*\nu_h)(x) - u(x) = h^2(m_2/2)v''(x) + O(h^4), \quad x \in \mathbb{R}, \tag{1.4}$$

where m_i denotes the ith moment of ν and $O(h^4)$ is bounded in absolute value by $(m_4/4!)\,\|u^{(4)}\|_2 h^4$. Since the dominant term on the RHS of (1.4) is negative at the concave parts and positive at the convex part of u, at least for h small, the "peaks" of $u^*\nu_h$ will tend to lie below and the "dips" above those of u. Then, $u^*\nu_h$ will be "flatter" than u at the places we wish to penalize lighter, and we propose to penalize $u^*\nu_h$ rather than u directly. This simply means using $|\tilde{\nu}_h|^2 m$ for the weighted function in (1.1). To avoid the introduction of new modes we can use a strongly unimodal ν (see, e.g., Lukacs, 1970). Note that in the special case that $m(t) = \exp\{h^2t/2\}$, $h > 0$ and ν_h is the log-concave density (and hence strongly unimodal) with $\tilde{\nu}_h(t) = (1 + h^2t^2)^{-1/2}$ (see Ibragimov, 1956), the MPLE given by (1.2) is generated by $\kappa_\mu(\cdot) = h^{-1}k(\cdot/h)$, with $k = \phi - \phi''$ where ϕ denotes the standard normal density. So that, in effect, the estimate is corrected by

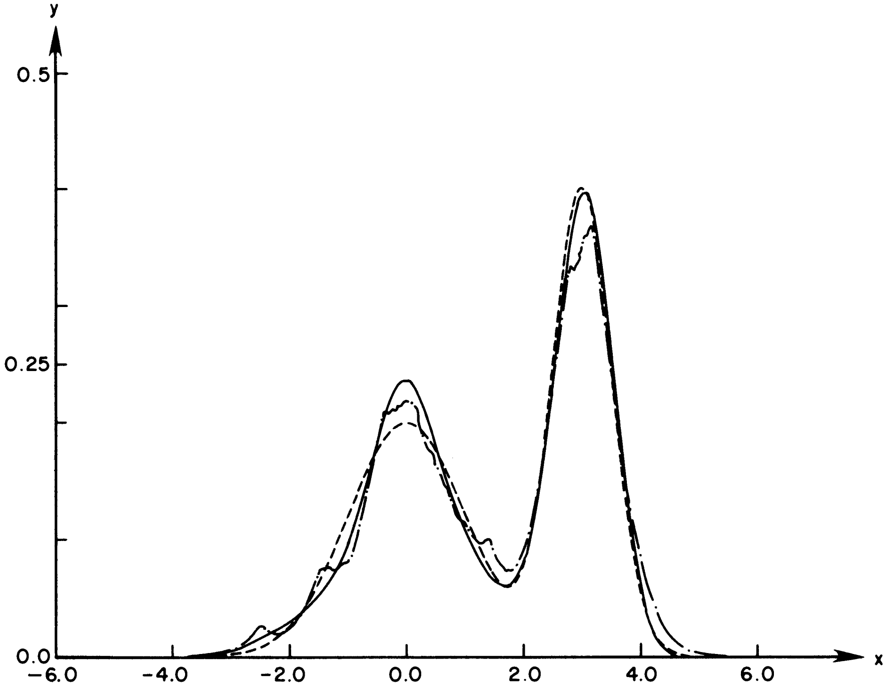

FIG. 2. A: --- *bimodal density* $[\phi(x) + 2\phi(2(x - 3))]/2$. B: -·- *kernel estimator*, $h = .30$. C: —— MPLE *with* $k = \phi - (1/2)\phi''$, $h = .65$.

subtracting its second derivative. This seems to improve the performance of the estimates at the "peaks" and "valleys" of the curve; see, e.g., Figures 1, 2. Clearly any kernel of the form

$$k = \phi - c\phi'', \quad c > 0$$

makes sense in the context above (the choice of $c = 1$ is suggested by an ad hoc approximation argument). In fact for $c = 1/2$, $\int x^2 k(x)\, dx = 0$, so that the resulting MPLE attains a higher rate of convergence (see, e.g., Proposition 4.1, and also Fig. 2C). In Section 4 we show that the estimators of f based on kernels with s zero even moments, $s = 0, 1, \cdots$, attain the enhanced rates of convergence of kernel estimates based on such kernels. It should be noted that although the MPLE u of $f^{1/2}$ based on such a kernel may assume negative values, the estimator u^2 of f will be a nonnegative density estimate.

Also, note that the added flexibility in the choice of μ allows us to consider as kernels κ_μ in (1.2) a wide variety of symmetric densities as long as $\tilde{\kappa}_\mu > 0$, including the family

$$(1.5) \qquad \kappa_\mu(x) = [2\Gamma(1 + \gamma^{-1})]^{-1}\exp\{-|x|^\gamma\}, \quad x \in \mathbb{R}, \quad \gamma \in [1, 2],$$

as well as kernels with finite support, e.g. $\kappa_\mu(x) = 1 - |x|$, $x \in (-1, 1)$ (see Klonias and Nash, 1983a), which should require less numerical effort, a consid-

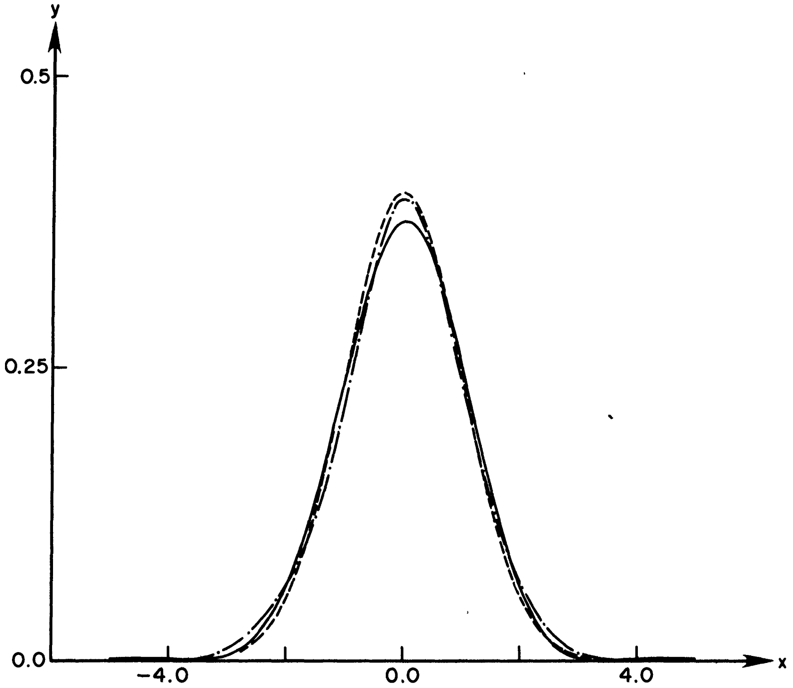

FIG. 3. A: --- *standard normal density*, B: -·- MPLE *with* $k = \phi$, $h = .75$. C: —— MPLE *with* $k = \phi - \phi''$, *data based* $h = 1.525$.

eration of some importance for the algorithm giving the data based band-width h (see Figures 3C and 4C) and for the MPLE's of multivariate densities.

In Section 2 we show the existence and uniqueness of the estimators and indicate the approach employed for their numerical evaluation. Details on the algorithms will be reported elsewhere jointly with Stephen G. Nash. After some preliminary lemmas in Section 3, we derive in Section 4, in a unified manner, the consistency of the MPLE's of $f^{1/2}$ and f in probability and a.s. in the $L_p(\mathbb{R})$-norms, $p = 1, 2, \infty$. The rates of convergence of the MPLE's given in Proposition 4.1 (ii), approach the optimal predicted by Farrell (1972), page 172, and Stone (1980). For the MPLE's of Good and Gaskins and those in DeMontricher et al. (1975) with support on $\mathbb{R}$, we show in addition, that they also converge in Sobolev norms. Similar results are obtained for the MPLE u of $f^{1/2}$. In Section 5 we present a number of numerical examples and propose an approach for a data based choice of the smoothing parameters of the MPLE's which seems to perform well but needs further study. An efficient algorithm for the numerical evaluation of h has been constructed jointly with Stephen G. Nash, details of which will be reported elsewhere. In Section 6 we present a nonparametric regression estimator based on the MPLE's, see (6.1), which has the appealing property of reducing to the classical nonparametric regression estimators when a smoothing parameter

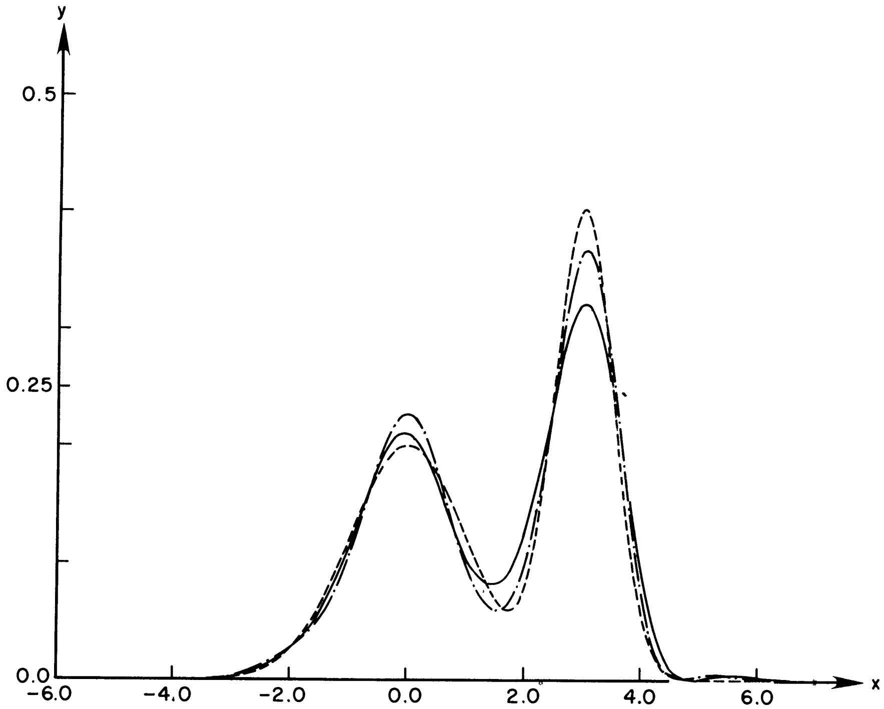

FIG. 4. A: --- *bimodal density* $[\phi(x) + 2\phi(2(x-3))]/2$. B: -·- MPLE *with* $k = \phi - \phi''$, $h = .925$. C: —— MPLE *with* $k = \phi - \phi''$, *data based* $h = 1.1$.

goes to zero. For other regression estimators based on penalty methods see Reinsch (1967), Wahba (1975), Bartoszyński, Brown, McBride and Thompson, (1981), Anderson and Blair (1982), and Rosenblatt (1983).

2. On the existence and numerical evaluation of the estimators.

In this section we obtain the spline function (1.2) as the unique solution to problem (1.1), and indicate the approach employed for the numerical evaluation of the estimators.

When the measure of $L_2(\mathbb{R}^p)$, $p \in \mathbb{Z}_+$, is different from the Lebesgue measure we denote the space by $L_2(\mu)$, μ a positive measure on $\mathbb{R}^p$ dominated by the Lebesgue measure with density $m(t)$. Let $H \equiv \{u \in L_2(\mathbb{R}^p)\colon \tilde{u} \in L_2(\mu)\}$—a Hilbert space—where $\tilde{u}(t) = \int_{\mathbb{R}^p} e^{-it^Tx} u(x)\, dx$ denotes the Fourier transform of u, with corresponding inversion formula $u(x) = (2\pi)^{-p} \int_{\mathbb{R}^p} e^{it^Tx} \tilde{u}(t)\, dt$, $t, x \in \mathbb{R}^p$. The measures μ we consider here are such that there exists a symmetric around zero function $k_\mu \in L_2(\mathbb{R}^p)$ with $\tilde{k}_\mu(t) m(t) = 1$, e.g., $m(t) = \exp\{h^2t^2/2\}$ with $k_\mu(\cdot) = h^{-1}\phi(\cdot/h)$, where ϕ denotes the standard normal density. To see that H is a reproducing kernel Hilbert space (RKHS) with kernel $\kappa^*(x, y) = \kappa_\mu(x-y)$, let us denote the inner product of H by

$$\langle u, v\rangle \equiv (2\pi)^{-p} \int \tilde{u}\tilde{v}\, d\mu,$$

and the induced norm by

$$\| u \| \equiv \langle u, u \rangle^{1/2} = (2\pi)^{-p/2} \| m^{1/2}\tilde{u} \|_2.$$

Then, setting $\kappa_x^*(\cdot) = \kappa^*(\cdot, x) \in H$, $x \in \mathbb{R}^p$, we have that

$$u(x) = \langle \kappa_x^*, u \rangle \ \forall\, u \in H \Leftrightarrow (2\pi)^{-p} \int_{\mathbb{R}^p} e^{-it^Tx}\bar{\tilde{u}}(t)\, dt$$

$$= (2\pi)^{-p} \int_{\mathbb{R}^p} \tilde{\kappa}_x^*(t)\bar{\tilde{u}}(t)m(t)\, dt,\ \forall\, u \in H \Leftrightarrow \tilde{\kappa}_x^*(t) = e^{-it^Tx}/m(t)$$

$$\Leftrightarrow \tilde{\kappa}_x^*(t) = e^{-it^Tx}\tilde{\kappa}_\mu(t),$$

$x, t \in \mathbb{R}^p$, which by the inversion formula gives the result.

We can now obtain the spline function (1.2) as the unique solution to the optimization problem (1.1).

PROPOSITION 2.1. *The optimization problem*

$$\max\{[\textstyle\prod_{i=1}^n u^2(X_i)]\exp\{-\lambda \ \| u \|^2\},\quad u \in H\} \tag{2.1}$$

$$\textit{subject to: } u(X_i) \geq 0,\quad i = 1, 2, \cdots, n,$$

has a unique solution, given by (1.2) *with* $\tilde{\kappa}_\mu m = 1$, $\lambda > 0$.

PROOF. In view of the fact that H is a RKHS, the existence and uniqueness of the solution is a direct consequence of Propositions 2.1 in DeMontricher et al. (1975). The constraints cannot be active at the maximum and hence the stationary point of the Lagrangian of the problem, after taking logarithms, is given by:

$$2 \textstyle\sum_{j=1}^n u(X_i)^{-1}\eta(X_i) - \lambda(2\pi)^{-p}\left\{\displaystyle\int m\tilde{u}\bar{\tilde{\eta}} + \int m\bar{\tilde{u}}\tilde{\eta}\right\} = 0 \quad \forall\ \eta \in H$$

$$\Leftrightarrow \left\{\textstyle\sum_{j=1}^n u(X_j)^{-1} \displaystyle\int e^{-it^TX_j}\bar{\tilde{\eta}}(t)\, dt - \lambda \int m\tilde{u}\bar{\tilde{\eta}}\right\}$$

$$+ \left\{\textstyle\sum_{j=1}^n u(X_j)^{-1} \displaystyle\int e^{it^TX_j}\tilde{\eta}(t)\, dt - \lambda \int m\bar{\tilde{u}}\tilde{\eta}\right\} = 0 \quad \forall\ \eta \in H$$

$$\Leftrightarrow Re\left\{\int [\textstyle\sum_{j=1}^n u(X_j)^{-1}e^{-it^TX_j} - \lambda\, m(t)\tilde{u}(t)]\bar{\tilde{\eta}}(t)\, dt\right\} = 0 \quad \forall\ \eta \in H.$$

For the validity of this last identity it is sufficient to have:

$$\int [\textstyle\sum_{j=1}^n u(X_j)^{-1}e^{-it^TX_j} - \lambda m(t)\tilde{u}(t)]\bar{\tilde{\eta}}(t)\, dt = 0 \quad \forall\ \eta \in H,$$

$$\Leftrightarrow \textstyle\sum_{j=1}^n u(X_i)^{-1}e^{-it^TX_j} - \lambda m(t)\tilde{u}(t) \equiv 0$$

$$\Leftrightarrow \tilde{u}(t) = \lambda^{-1} \textstyle\sum_{j=1}^n u(X_j)^{-1}e^{-it^TX_j}\tilde{\kappa}_\mu(t),$$

which by the inversion formula gives (1.2). That this is the only solution to problem (2.1) follows from Proposition 2.1 in DeMontricher et al. (1975) and is a consequence of the fact that the second differential of the Lagrangian, given by

$$-2\left\{\sum_{j=1}^{n} u(X_i)^{-2}\eta(X_i)^2 + \lambda(2\pi)^{-p}\int m\,|\,\tilde{\eta}\,|^2\right\},$$

is negative definite everywhere. □

The parameter λ is chosen so that $\int u^2 = 1$. To obtain λ, note that $q \equiv \lambda^{1/2}u$ does not depend on λ and $\lambda = \| q \|_2^2$.

For the numerical evaluation of the estimate, note that setting $x = X_j$, $j = 1, \cdots, n$ in (1.2) we obtain the system

$$(2.2)\qquad q(X_j) = \sum_{i=1}^{n} q(X_i)^{-1}\kappa_\mu(X_j - X_i), \quad j = 1, \cdots, n.$$

To solve this system we equivalently minimize the convex function

$$\mathbf{g}^T \textstyle\sum \mathbf{g} - \sum_{i=1}^{n} \log g_i^2 \text{ over } \{\mathbf{g} \in \mathbb{R}^n\colon g_i \geq 0,\, i = 1, \cdots, n\},$$

where

$$g_i \equiv q(X_i)^{-1},\, i = 1, \cdots, n \text{ and } \textstyle\sum \equiv [\kappa_\mu(X_i - X_j)] \in \mathbb{R}^{2n}.$$

The algorithm we use is based on a truncated-Newton method, described in Nash (1982). Then, the parameter $\lambda = \sum_{i=1}^{n}\sum_{j=1}^{n} g_i g_j(\kappa_\mu * \kappa_\mu)(X_i - X_j)$. For details on the numerical evaluation of the estimators see Klonias and Nash (1983b) and for a summary of the techniques see Klonias and Nash (1983a).

3. Preliminary lemmas.

In this section we present four lemmas on which the proofs of the consistency of the estimators—appearing in Section 4—are based.

Let us denote by $\|\cdot\|_p$ the $L_p(\mathbb{R})$ norms $p = 1, 2, \infty$ and let $\|\cdot\|$ denote the norm induced by the inner product of H. The proofs of the consistency of the estimators are based on the following lemma.

LEMMA 3.1. *Let u_λ denote the solution to problem* (1.1) *for some $\lambda > 0$. If $v = f^{1/2} \in H$, then*

$$\| u_\lambda - v \|^2 \leq \lambda^{-2} \sum_{i=1}^{n}\sum_{j=1}^{n} v(X_i)^{-1} v(X_j)^{-1}\kappa_\mu(X_i - X_j) - 2(n/\lambda) + \| v \|^2.$$

PROOF. Note that

$$\begin{aligned} &\sum_{i=1}^{n} u_\lambda(X_i)^{-1}v(X_i)^{-1}(u_\lambda(X_i) - v(X_i))^2 \geq 0 \\ &\Leftrightarrow \langle \lambda^{-1}\sum_{i=1}^{n} v(X_i)^{-1}\kappa_\mu(\cdot\, - X_i) - u_\lambda,\, u_\lambda - v\rangle \geq 0 \\ &\Leftrightarrow \langle \lambda^{-1}\sum_{i=1}^{n} v(X_i)^{-1}\kappa_\mu(\cdot\, - X_i) - v,\, u_\lambda - v\rangle \geq \| u_\lambda - v \|^2. \end{aligned}$$

Then, by the Cauchy-Schwartz inequality we have,

$$\begin{aligned} \| u_\lambda - v \|^2 &\leq \| \lambda^{-1}\sum_{i=1}^{n} v(X_i)^{-1}\kappa_\mu(\cdot\, - X_i) - v \|^2 \\ &= \lambda^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n} v(X_i)^{-1}v(X_j)^{-1}\kappa_\mu(X_i - X_j) - 2(n/\lambda) + \| v \|^2. \qquad □ \end{aligned}$$

Next, let λ_n denote the value of λ such that $\int u_{\lambda_n}^2 = 1$. The following lemma describes the probabilistic behavior of λ_n/n.

LEMMA 3.2. *If $v \in H$, then*

$$|(\lambda_n/n)^{1/2} - 1|$$
$$\leq \|\kappa_\mu\|_1^{1/2}\{n^{-2} \textstyle\sum_{i=1}^n \sum_{j=1}^n v(X_i)^{-1}v(X_j)^{-1}\kappa_\mu(X_i - X_j) - 2 + \|v\|^2\}^{1/2}.$$

PROOF. From Lemma 3.1, with $\lambda = n$, we have that

$$\|u_n - v\|^2 \leq n^{-2} \textstyle\sum_{i=1}^n \sum_{j=1}^n v(X_i)^{-1}v(X_j)^{-1}\kappa_\mu(X_i - X_j) - 2 + \|v\|^2. \tag{3.1}$$

Also, from (2.2) we see that $\lambda^{1/2}u(x_i)$, $i = 1, \cdots, n$ and hence $\lambda^{1/2}u$ do not depend on λ. Then, $\lambda_n^{1/2}u_{\lambda_n} = n^{1/2}u_n$, so that $1 = \|u_{\lambda_n}\|_2 = (n/\lambda_n)^{1/2}\|u_n\|_2$, i.e.,

$$\lambda_n/n = \|u_n\|_2^2. \tag{3.2}$$

Also, note that $1 = m(t)\tilde{\kappa}_\mu(t) \leq m(t)\|\kappa_\mu\|_1$, so that

$$\|\cdot\|_2^2 \leq \|\kappa_\mu\|_1\|\cdot\|^2 \tag{3.3}$$

and hence, using also (3.2), we obtain

$$|(\lambda_n/n)^{1/2} - 1| = |\|u_n\|_2 - \|v\|_2| \leq \|u_n - v\|_2 \leq \|\kappa_\mu\|_1\|u_n - v\|,$$

which along with (3.1) give the result. □

Henceforth we consider RKHS H with kernel $\kappa^*(x, y) = \kappa_\mu(x - y)$, with $\kappa_\mu(\cdot) = h^{-1}k(\cdot/h)$, $h > 0$ and take $h = h_n = O(n^{-t})$ for some $t \in (0, ½)$, so that (1.2) takes the form

$$u(x) = \lambda_n^{-1} \textstyle\sum_{i=1}^n u(X_i)^{-1}h^{-1}k((x - X_i)/h), \quad x \in \mathbb{R}. \tag{3.4}$$

LEMMA 3.3. *Under the assumptions*:

A$_1$: *$E|X|^\tau < +\infty$ for some $\tau > (1 - t)^{-1}$, $t \in (0, ½)$,*

A$_2$: *$\|v^{(s+2)}\|_2 < +\infty$, where $s = 0, 1, 2, \cdots$ denotes the number of even moments of the kernel k which are zero, we have that*

$$|n^{-2} \textstyle\sum_{i=1}^n \sum_{j=1}^n v(X_i)^{-1}v(X_j)^{-1}h^{-1}k((X_i - X_j)/h) - 1|$$
$$= O_p(n^{-(1-\delta-\tau^{-1}-t)}) + O_p(n^{-(1/2+(2+s)t)}) + O(n^{-2(1+s)t})$$

for $\delta > 0$.

PROOF. Let $Y_{ij} \equiv v^{-1}(X_i)v^{-1}(X_j)h^{-1}k((X_i - X_j)/h)$, $i, j = 1, \cdots, n$. Note that $k(0)^{-1}hY_{ii} = f^{-1}(X_i)$, $i = 1, \cdots, n$, are i.i.d. random variables and under our moment assumption $E\{f^{-r}(X)\} < +\infty$ for $r < (1 + \tau^{-1})^{-1}$. Then, from the Marcinkiewicz's version of the SLLN (see, e.g. Loève, 1977, page 254), we have

$$k(0)h^{-1}n^{-2} \textstyle\sum_{i=1}^n f^{-1}(X_i) = O(n^{(1/r)+t-2}) \quad \text{a.s.}, \tag{3.5}$$

so that we need only deal with the term $\sum \sum_{i\neq j} Y_{ij}$.

Since $EY_{ij}^2 = +\infty$, in order to study the convergence of the series we will use a truncation technique. For i, j with $i \neq j$, let $Z_{ij} \equiv X_i - X_j$,

$$Z_{ij}^c = \begin{cases} +\infty & \text{if} \quad |X_i| > c_i \quad \text{and} \quad |X_j| > c_j \\ Z_{ij} & \text{otherwise}, \end{cases}$$

with $c_i \equiv i^\zeta$ for $\zeta > \tau^{-1}$, and define $Y_{ij}^c \equiv v^{-1}(X_i)v^{-1}(X_j)h^{-1}k(Z_{ij}^c/h)$. Note that $Y_{ij} \neq Y_{ij}^c$ if and only if $Z_{ij} \neq Z_{ij}^c$. Hence $\forall\ h > 0$,

$$\begin{aligned} P(Y_{ij} \neq Y_{ij}^c \text{ i.o.}) &= P(Z_{ij} \neq Z_{ij}^c \text{ i.o.}) \\ &\leq \lim_m \textstyle\sum_{i=m}^\infty \sum_{j=m}^\infty P(|X_i| > c_i)P(|X_j| > c_j) \\ &= \{\lim_m \textstyle\sum_{i=m}^\infty P(|X_i| > c_i)\}^2. \end{aligned}$$

But from Chebyshev's inequality (see, e.g., Chung, 1974, page 48), we have that $P(|X_i| > c_i) \leq E|X|^\tau |c_i|^{-\tau} = E|X|^\tau i^{-\zeta\tau}$ and hence the series $\sum_{i=1}^\infty P(|X_1| > c_i)$ converges. Then, $\forall\ h > 0$, $P(Y_{ij} \neq Y_{ij}^c \text{ i.o.}) = 0$. Then, setting $S_n = \sum^n \sum_{i\neq j}^n v^{-1}(X_i)v^{-1}(X_j)k(Z_{ij}/h_n)$ and defining S_n^c similarly through Z_{ij}^c, note that for all positive integers n

$$|S_n - S_n^c| \leq \|k\|_\infty \{\textstyle\sum_{i\in A} v^{-1}(X_i)\}^2,$$

where A remains a finite set and $P(v(X_i) = 0) = 0$ since f is continuous. Hence, for any sequence $a_n \to 0$ as $n \to \infty$

$$a_n|S_n - S_n^c| \to 0 \text{ a.s. as } n \to \infty,$$

i.e., the series a_nS_n, $a_nS_n^c$ are convergence equivalent and hence it is enough to show the result for the truncated random variables.

Next note that,

$$EY_{ij}^c = EY_{ij} - \left[\int_{c_i}^{+\infty}\int_{c_j}^{+\infty} + \int_{c_i}^{+\infty}\int_{-\infty}^{-c_j} + \int_{-\infty}^{-c_i}\int_{-\infty}^{-c_j} + \int_{-\infty}^{-c_i}\int_{c_j}^{+\infty}\right]$$
$$\cdot[h^{-1}k((x-y)/h)f^{1/2}(x)f^{1/2}(y)]\,dx\,dy.$$

But,

$$\begin{aligned} &\left|\int_{c_j}^{+\infty}\int_{c_j}^{+\infty} h^{-1}k\left(\frac{x-y}{h}\right)f^{1/2}(x)f^{1/2}(y)\,dx\,dy\right| \\ &\leq \left\{\int_{c_i}^{+\infty}\int_{c_j}^{+\infty} h^{-1}\left|k\left(\frac{x-y}{h}\right)\right| f(x)\,dx\,dy\right\}^{1/2} \\ &\quad\cdot\left\{\int_{c_i}^{+\infty}\int_{c_j}^{+\infty} h^{-1}\left|k\left(\frac{x-y}{h}\right)\right| f(y)\,dx\,dy\right\}^{1/2} \\ &\leq \|k\|_1[1 - F(c_i)]^{1/2}[1 - F(c_j)]^{1/2} \\ &\leq \|k\|_1 P(|X_i| > c_i)^{1/2}P(|X_j| > c_j)^{1/2} \\ &\leq \|k\|_1 E|X|^\tau(ij)^{-\zeta\tau/2} \end{aligned}$$

and the same bound is similarly derived for the other three integrals, so that

$$|EY_{ij}^c - EY_{12}| \leq 4 \, \|k\|_1 E|X|^\tau (ij)^{-\zeta\tau/2}, \tag{3.6}$$

and hence we have that

$$[n(n-1)]^{-1} E\{\sum \sum_{i \neq j} Y_{ij}^c\} = EY_{12} + O(n^{-\zeta\tau}). \tag{3.7}$$

Then, in order to obtain the rate at which the bias vanishes it is enough to check the following term:

$$\begin{aligned}
|EY_{12} - 1| &= \left| \int \int h^{-1} k\left(\frac{x-y}{h}\right) v(x) v(y) \, dx \, dy - 1 \right| \\
&= \left| \int \int k(z) v(x) v(x + hz) \, dx \, dz - 1 \right| \\
&= \left(\frac{h^2}{2}\right) \left| \int z^2 k(z) \int v'(x) v'(x + \bar{h}z) dx \, dz \right|, \quad \bar{h} \in (0, h) \\
&\leq \left(\frac{h^2}{2}\right) \|v'\|_2^2 \int z^2 |k(z)| \, dz.
\end{aligned}$$

Also, note that if $\int z^{2s} k(z) \, dz = 0$ for some positive integer s, we can carry the Taylor expansion above further to obtain

$$|EY_{12} - 1| \leq \frac{h^{2(s+1)}}{(2s+1)!} \|v^{(s+1)}\|_2^2 \int z^{2(s+1)} |k(z)| \, dz.$$

Hence we have that

$$|EY_{12} - 1| = O(h^{2(s+1)}). \tag{3.8}$$

Then, in view of (3.7) and (3.8), to conclude the proof we need only show that

$$[n(n-1)]^{-1} \sum \sum_{i \neq j} (Y_{ij}^c - EY_{ij}^c) = O_p(n^{-1/2} h^{2+s}) + O_p(n^{-1+(\zeta/2)} h^{-1/2}), \tag{3.9}$$

for which, by the Chebyshev inequality, it is enough to show that

$$E\{[n(n-1)]^{-1} \sum \sum_{i \neq j} (Y_{ij}^c - EY_{ij}^c)\}^2 = O(n^{-1} j^{4+2s}) + O(n^{-2+\zeta} h^{-1}). \tag{3.10}$$

First, note that

$$\begin{aligned}
&\sum \sum_{i \neq j} E\{Y_{ij}^c\}^2 \\
&\leq \sum \sum_{i \neq j} \left\{ \int_{-c_j}^{c_j} \int_{-\infty}^{+\infty} h^{-2} k^2\left(\frac{x-y}{h}\right) dx \, dy + \int_{-c_i}^{c_i} \int_{-\infty}^{+\infty} h^2 k^2\left(\frac{x-y}{h}\right) dx \, dy \right\} \\
&= 2 \, \|k\|_2^2 h^{-1} \sum \sum_{i \neq j} (c_i + c_j) \leq 4 \, \|k\|_2^2 h^{-1} n^{2+\zeta},
\end{aligned}$$

so that,

$$[n(n-1)]^{-2} \sum \sum_{i \neq j} E(Y_{ij}^c - EY_{ij}^c)^2 = O(h^{-1} n^{-2+\zeta}). \tag{3.11}$$

The remaining nonzero term of (3.10) is a triple sum of the terms we bound below:

$$
\begin{aligned}
&|EY^c_{ij}Y^c_{j\ell} - EY^c_{ij}EY^c_{j\ell}| \\
(3.12)\qquad &\le |EY^c_{ij}Y^c_{j\ell} - EY_{ij}Y_{j\ell}| + |EY_{ij}Y_{j\ell} - EY_{ij}EY_{j\ell}| \\
&\quad + |EY^c_{j\ell}|\,|EY^c_{ij} - EY_{12}| + |EY_{12}|\,|EY^c_{j\ell} - EY_{12}|.
\end{aligned}
$$

But, after setting $I_i = [-c_i, c_i]$, we have that

$$
\begin{aligned}
&|EY^c_{ij}Y^c_{j\ell} - EY_{ij}Y_{j\ell}| \\
&= \Bigg| \left\{\int_{I_i}\int_{I^c_j}\int_{I^c_\ell} + \int_{I^c_i}\int_{I^c_j}\int_{-\infty}^{+\infty}\right\} \\
&\quad \cdot\left[h^{-2}k\left(\frac{x-y}{h}\right)k\left(\frac{y-z}{h}\right)v(x)v(z)\right] dx\,dy\,dz \Bigg| \\
&\le \left\{\int_{-\infty}^{+\infty}\int_{I^c_\ell} + \int_{I^c_i}\int_{-\infty}^{+\infty}\right\}\left[h^{-1}(|k| * |k|)\left(\frac{x-y}{h}\right)v(x)v(z)\right] dx\,dz \\
&\le \|k\|_1^2[P(|X| > c_\ell)^{1/2} + P(|X| > c_i)^{1/2}],
\end{aligned}
$$

and hence

$$
(3.13)\qquad |EY^c_{ij}Y^c_{j\ell} - EY_{ij}Y_{j\ell}| \le \|k\|_1^2 E|x|^\tau(\ell^{-\zeta\tau/2} + i^{-\zeta\tau/2}).
$$

Next, note that for a kernel k with $s = 0, 1, 2, \cdots$ zero even moments, $k * k$ has s zero moments and

$$
m^*_{2s+2} = 2m_{2s+2},
$$

where m_r, m^*_r denote the rth moment of k and $k * k$ respectively. Then,

$$
\begin{aligned}
&EY_{ij}Y_{j\ell} - EY_{ij}EY_{j\ell} \\
&= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} h^{-1}(k * k)\left(\frac{x-y}{h}\right)v(x)v(y)\,dx\,dy \\
&\quad - \left\{\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} h^{-1}k\left(\frac{x-y}{h}\right)v(x)v(y)\,dx\,dy\right\}^2 \\
&= 1 + (-1)^{s+1}\frac{h^{2s+2}}{(2s+2)!}m^*_{2s+2}\|v^{(s+1)}\|_2^2 + O(h^{2s+4}) \\
&\quad - \left\{1 + (-1)^{s+1}\frac{h^{2s+2}}{(2s+2)!}m_{2s+2}\|v^{(s+1)}\|_2^2 + O(h^{2s+4})\right\}^2,
\end{aligned}
$$

so that,

$$
(3.14)\qquad EY_{ij}Y_{j\ell} - EY_{ij}EY_{j\ell} = O(h^{4+2s}).
$$

Then, (3.6) and (3.12) through (3.14) with $\zeta\tau > 2$, imply

$$[n(n-1)]^{-2} \textstyle\sum_i \sum_{\substack{j \\ i\neq j\neq \ell}} \sum_{\ell} E(Y_{ij}^c Y_{j\ell}^c - EY_{ij}^c EY_{j\ell}^c) = O(n^{-2}) + O(n^{-1}h^{4+2s}), \tag{3.15}$$

which along with (3.11) imply (3.10), and the proof is complete. □

Note that the optimum rate in Lemma 3.3 is achieved when $1 - \delta - \tau^{-1} - t = 2(1+s)t$, i.e., when $t = (1 - \delta - \tau^{-1})/(3 + 2s)$, in which case $2(1+s)t < (\frac{1}{2}) + (2+s)t$. Thus we obtain the following corollary to Lemma 3.3:

COROLLARY 3.1. *Under assumptions* A_1 *and* A_2, *we have that*

$$|n^{-2} \textstyle\sum_{i=1}^n \sum_{j=1}^n v(X_i)^{-1} v(X_j)^{-1} h^{-1} k((X_i - X_j)/h) - 1| = o_p(n^{-\xi}),$$

where $\xi = 1 - \tau^{-1} - t$, $t = [(1 - \tau^{-1})/(3+2s)]^-$, $s = 0, 1, \cdots$.

REMARK 3.1. Note that if f has all its moments finite, e.g., f has compact support, then the result of Corollary 3.1 is valid with $\xi < (1 - t)$ and $t < (3 + 2s)^{-1}$, $s = 0, 1, \cdots$

LEMMA 3.4. *Under assumptions* A_1 *and* A_2, *we have that w.p.*1:

$$|n^{-2} \textstyle\sum_{i=1}^n \sum_{j=1}^n v(X_i)^{-1} v(X_j)^{-1} h^{-1} k((X_i - X_j)/h) - 1|$$
$$= O(n^{-(1-\delta-\tau^{-1}-t)}) + O(n^{-\xi'}) + O(n^{-2(1+s)t}),$$

for $\delta > 0$ *and* $\xi' < [(\frac{1}{4}) + (2+s)t] \wedge [(\frac{3}{4}) - \tau^{-1} - (t/2)]$.

PROOF. In view of (3.5) through (3.8) it is enough to show the a.s. version of (3.9) with the present rate. To simplify the notation we define

$$W_{ij}^h \equiv v^{-1}(x_i) v^{-1}(X_j) h^{-1} k(Z_{ij}^c/h), \quad U_{ij}^h \equiv W_{ij}^h - EW_{ij}^h$$

and

$$S_n(h) \equiv \textstyle\sum_{\substack{i=1 \\ i\neq j}}^n U_{ij}^h.$$

Next, let $n^2 \le m \le (n+1)^2$ (it turns out that from the subsequences $[n^\eta]$ the optimal corresponds to $\eta = 2$). Then

$$|S_m(h_m)| \le |S_{n^2}(h_m)| + D_n, \tag{3.16}$$

where $D_n = \sup\{|S_m(h_m) - S_{n^2}(h_m)|: n^2 \le m \le (n+1)^2\}$. From (3.10) we see that

$$E\{[n^2(n^2-1)]^{-1} m^{\xi'} S_{n^2}(h_m)\}^2 = O(n^{-2-(8+4s)t+4\xi'}) + O(n^{-4+2\zeta+2t+4\xi'}),$$

so that if $-2 - (8+4s)t + 4\xi' < -1$ and $-4 + 2\zeta + 2t + 4\xi' < -1$, i.e.,

$$\xi' < [(\tfrac{1}{4}) + (2+s)t] \wedge [(\tfrac{3}{4}) - (\zeta + t)/2], \tag{3.17}$$

by the Chebyshev inequality and the Borel-Cantelli Lemma, we conclude that for $n^2 \le m \le (n+1)^2$ and ξ as in (3.17), w.p.1

$$[n^2(n^2-1)]^{-1} m^{\xi'} S_{n^2}(h_m) = o(1). \tag{3.18}$$

Also,

$$\begin{aligned}
&E[S_m(h_m) - S_{n^2}(h_m)]^2\\
&\quad = E\{2 \textstyle\sum_{i=n^2+1}^{m} \sum_{j=1}^{n^2} U_{ij}(h_m) + \sum_{i=n^2+1, i\ne j}^{m} \sum_{j=n^2+1}^{m} U_{ij}(h_m)\}^2\\
&\quad = 4 \textstyle\sum_{i=n^2+1}^{m} \sum_{j=1}^{n^2} EU_{ij}^2(h_m) + 4 \sum_{i=n^2+1}^{m} \sum_{\substack{j=1\\ j\ne \ell}}^{n^2} \sum_{\ell=1}^{n^2} EU_{ij}(h_m)U_{i\ell}(h_m)\\
&\qquad + 4 \textstyle\sum_{\substack{i=n^2+1\\ i\ne \ell}}^{m} \sum_{\ell=n^2+1}^{m} \sum_{j=1}^{n^2} EU_{ij}(h_m)U_{\ell j}(h_m) + E\{\sum_{i=n^2+1}^{m} \sum_{j=n^2+1}^{m} U_{ij}(h_m)\}^2\\
&\qquad + 8 \textstyle\sum_{i=n^2+1}^{m} \sum_{j=1}^{n^2} \sum_{\ell=n^2+1}^{m} EU_{ij}(h_m)U_{i\ell}(h_m)\\
&\quad = O(n^{3+2\zeta+2t}) + \{O(n^3) + O(n^{5-(8+4s)t})\} + \{O(n^3) + O(n^{4-(8+4s)t})\}\\
&\qquad + \{O(n^{2+2\zeta+2t}) + O(n^{3-(8+4s)t})\} + \{O(n^3) + O(n^{4-(8+4s)t})\}\\
&\quad = O(n^{3+2\zeta+2t}) + O(n^{5-(8+4s)t}),
\end{aligned}$$

where the rates indicated above, were derived using (3.6), the inequality used in (3.11), and (3.12) through (3.14), and proceeding as in the proof of (3.10). Then,

$$ED_n^2 \le \textstyle\sum_{m=n^2+1}^{(n+1)^2} E[S_m(h_m) - S_{n^2}(h_m)]^2 = O(n^{4+2\zeta+2t}) + O(n^{6-(8+4s)t}),$$

so that

$$E\{[n^2(n^2-1)]^{-1} m^{\xi'} D_n\}^2 = O(n^{-4+2\zeta+2t+4\xi'}) + O(n^{-2-(8+4s)t+4\xi'}),$$

and hence, for ξ' as in (3.17), by the Chebyshev inequality and the Borel-Cantelli Lemma, we have that w.p.1

$$[n^2(n^2-1)]^{-1} m^{\xi'} D_n = o(1). \tag{3.19}$$

Then the result follows from (3.16) through (3.19). □

COROLLARY 3.2. *Under assumptions* A_1 *and* A_2, *we have that w.p.*1

$$|\, n^{-2} \textstyle\sum_{i=1}^{n} \sum_{j=1}^{n} v(X_i)^{-1} v(X_j)^{-1} h^{-1} k((X_i - X_j)/h) - 1 \,| = o(n^{-\xi'}),$$

where $\xi' = (3/4) - \tau^{-1} - (t/2)$, *with*

$$t = [(3 - 4\tau^{-1})/(10 + 8s)]^- \quad \textit{if} \quad s \le 5(1 - 4\tau^{-1})^{-1}/2$$

and $t = [(1 - 2\tau^{-1})/(5 + 2s)]^-$ *otherwise*, $s = 0, 1, \cdots$.

REMARK 3.2. Note that if f has all its moments finite, e.g., f has compact support, then the result of Corollary 3.2 is valid with $\xi' = [(3/4) - (t/2)]^-$ with $t = [3(10 + 8s)^{-1}]^-$ if $s < 5/2$ and $t = [(5 + 2s)^{-1}]^-$ if $s > 5/2$, $s = 0, 1, \cdots$.

4. On the consistency of the estimators.

In this section we establish the consistency in probability and a.s., of the MPLE's u of $f^{1/2}$ and of the density estimators $f_n = u^2$, in the $L_p(\mathbb{R})$ norms, $p = 1, 2, \infty$, and in Sobolev norms, in a unified manner. Also, we indicate the optimal rates of convergence that our proofs allow in each case. We denote the MPLE's of v, f by u and $f_n = u^2$ respectively, the normalizing parameter λ_n by λ and set $h = O(n^{-t})$, $t < ½$.

THEOREM 4.1. *Under the assumptions of Lemma* 3.3 *and* $v \in H$, *we have that*:

(i) $|(\lambda/n) - 1| = O_p(n^{-(1-\delta-\tau^{-1}-t)/2}) + O_p(n^{-\xi/2}) + O(n^{-(s+1)t})$,
(ii) $\| u - v \|_2 = O_p(n^{-(1-\delta-\tau^{-1}-t)/2}) + O_p(n^{-\xi/2}) + O(n^{-(s+1)t})$,
(iii) $\| u - v \|_\infty = O_p(n^{-(1-\delta-\tau^{-1}-2t)/2}) + O_p(n^{-(\xi-t)/2}) + O(n^{-(2s+1)t/2})$,

with $\xi = ½ + (s + 2)t$ *for convergence in probability and*

$$\xi = [¼ + (s + 2)t] \wedge [¾ - \tau^{-1} - (t/2)]$$

for a.s. convergence.

PROOF. If k is a probability density, it has characteristic exponent $q = 2$, i.e., $\lim_{t\to 0}(1 - \tilde{k}(t))/|t|^q = c_q$, a finite nonzero constant (e.g., see Tapia and Thompson, 1978). In the case that $\int z^{2s}k(z)\,dz = 0$, $s \in \mathbb{Z}_+$, then $q = 2(s + 1)$. Then, using Parseval's Theorem for the Fourier transform we have that $\| u \|_2^2 = (2\pi)^{-1} \| \tilde{u} \|_2^2$ and hence

$$h^{-2(s+1)}(\| v \|^2 - 1) = (2\pi)^{-1} \int | \tilde{v}(t) |^2 t^{2(s+1)} \tilde{k}(ht)^{-1} \frac{[1 - \tilde{k}(ht)]}{(ht)^{2(s+1)}}\, dt$$

$$\to \| v^{(s+1)} \|_2^2,$$

as $h \to 0$, by virtue of the dominated convergence Theorem and Proposition 8.44 in Breiman (1968). Hence,

$$\| v \|^2 - 1 = O(h^{2(s+1)}). \tag{4.1}$$

Then, (i) is a consequence of (4.1) and Lemmas 3.2 and 3.3 For part (ii), note that

$$\begin{aligned} \| u - v \|^2 \le {} & (n/\lambda)^2 \{ n^{-2} \textstyle\sum_{i=1}^n \sum_{j=1}^n v^{-1}(X_i) v^{-1}(X_j) h^{-1} k((X_i - X_j)/h) - 1\} \\ & + [(n/\lambda) - 1]^2 + \| v \|^2 - 1; \end{aligned} \tag{4.2}$$

the result then follows from (3.3), (4.1), part (i) and Lemma 3.3.

Next, note that $\kappa^*(x, y) = h^{-1}k((x - y)/h)$ is the kernel of the RKHS H and $\| \kappa_x^* \|^2 = k(0)h^{-1}$, where $\kappa_x^*(\cdot) = \kappa^*(\cdot, x)$. Then,

$$| u(x) - v(x) | = | \langle \kappa_x^*, u - v \rangle \le h^{-1/2} \| k \|_\infty \| u - v \|,$$

giving part (iii) as a consequence of (4.1), (4.2), part (i) and Lemma 3.3. The a.s. convergence results follow similarly using Lemma 3.4 instead of Lemma 3.3. □

COROLLARY 4.1. *Under the assumption of Theorem* 4.1, *we have that* f_n *converges to* f *in probability and a.s., in the* $L_1(\mathbb{R})$ *and* $L_2(\mathbb{R})$ *norms with the rate in part* (i) *and in the* $\sup_{\mathbb{R}}$-*norm with the rate of part* (ii) *of Theorem* 4.1.

PROOF. These results are consequences of Theorem 4.1 and the inequalities:

$$\|f_n - f\|_1 \le 2\ \|u - v\|_2,$$

$$\|f_n - f\|_2 \le (2\ \|v\|_\infty + \|u - v\|_\infty)\ \|u - v\|_2,$$

$$\|f_n - f\|_\infty \le (2\ \|v\|_\infty + \|u - v\|_\infty)\ \|u - v\|_2,$$

respectively. □

Note that the optimal rates of convergence for the results of Theorem 4.1 and Corollary 4.1, can be deduced from Corollaries 3.1 and 3.2. In the following proposition we summarize the maximum rates which our proofs allow, and in order to simplify the notation we assume that f has all its moments finite, e.g., f has compact support.

PROPOSITION 4.1. *Under the assumptions of Theorem* 4.1 *and* $\tau = +\infty$, *we have*:

(i) $\|f_n - f\|_1 = o_p(n^{-\rho})$
(ii) $\|f_n - f\|_2 = o_p(n^{-\rho})$
(iii) $\|f_n - f\|_\infty = o_p(n^{-\rho+(t/2)})$,

where $\rho = (1/2) - (4s + 6)^{-1}$, *with* $t = [(2s + 3)^{-1}]^-$ *for convergence in probability, and for a.s. convergence* $\rho = (3/8) - 3(40 + 32s)^{-1}$ *with* $t = [3/(10 + 8s)]^-$ *if* $s \le 2$ *and* $\rho = (3/8) - (20 + 8s)^{-1}$ *with* $t = [(5 + 2s)^{-1}]^-$ *if* $s \ge 3$.

REMARK 4.1. Note that $\|u - v\|_2 \le 2$ and $\|f_n - f\|_1 \le 2$ and hence by the dominated convergence Theorem, part (ii) of Theorem 4.1 and part (i) of Proposition 4.1, we obtain the convergence of u in IMSE and of f_n in integrated mean absolute deviation, i.e.,

$$E\|u - v\|_2^2 = o(n^{-2\rho}), \quad E\|f_n - f\|_1 = o(n^{-\rho}),$$

where $\rho = (1/2) - (4s + 6)^{-1}$, with $t = [(2s + 3)^{-1}]^-$.

In the case that the norm of H is defined through a measure μ_h with density

$$m(ht) = 1 + \textstyle\sum_{j=s+1}^{\ell} (ht)^{2j}, \quad s = 0,\ 1,\ \cdots,\ \ell - 1, \tag{4.3}$$

the corresponding MPLE's u of $v = f^{1/2}$ are equivalent to those treated in Good and Gaskins (1971) and DeMontricher et al. (1975). Note the penalty defined through (4.3) leads to a kernel k symmetric around zero with

$$\tilde{k}(t) = 1 - (m(t) - 1)m(t)^{-1}$$

and hence k has s even moments equal to zero. Since $\tilde{k} > 0$, the consistency results of this section also apply to these estimates and in fact the rate in part (iii) of Theorem 4.1 can be improved slightly. Also, we show next that derivates of the estimators f_n as well as u corresponding to (4.3), converge in the $L_2(\mathbb{R})$ norm. To simplify the statement of the theorem below, we take $\tau = \infty$ as in Proposition 4.1, although the proof is based on Theorem 4.1.

THEOREM 4.2. *For μ_h defined through* (4.3) *under the assumptions of Proposition* 4.1, *we have that in probability and a.s.*:

$$\| u^{(j)} - v^{(j)} \|_2 = O_p(n^{-\rho(1-j(s+1)^{-1})}),$$

where ρ is as in Proposition 4.1 *and $j \leq s$.*

PROOF. From (4.2) we have that

$$\begin{aligned} & h^{2(s+1)} \| u^{(s+1)} - v^{(s+1)} \|_2^2 \\ & \quad \leq \| u - v \|^2 \\ (4.4) \qquad & \quad \leq (n/\lambda)^2 \{ n^{-2} \textstyle\sum_{i=1}^n \sum_{j=1}^n v^{-1}(x_i) v^{-1}(X_j) h^{-1} k((X_i - X_j)/h) -1 \} \\ & \qquad + [(n/\lambda) -1]^2 + \textstyle\sum_{j=s+1}^{\ell} h^{2j} \| v^{(j)} \|_2^2. \end{aligned}$$

Also, as in Lemma 3.2, we have that

$$\begin{aligned} | (\lambda/n)^{1/2} - 1 |^2 \leq & \{ n^{-2} \textstyle\sum_{i=1}^n \sum_{j=1}^n v^{-1}(X_i) v^{-1}(X_j) h^{-1} k((X_i - X_j)/h) - 1 \} \\ & + \textstyle\sum_{j=s+1}^{\ell} h^{2j} \| v^{(j)} \|_2^2. \end{aligned}$$

Hence, as in Theorem 4.1, we have that

$$(4.5) \quad [(n/\lambda) - 1]^2, \| u - v \|^2 = O_p(n^{-(1-\delta-\tau^{-1}-t)}) + O_p(n^{-\xi}) + O(n^{-2(s+1)}),$$

with the ξ of Theorem 4.1. Then, from (4.4) and (4.5) we conclude that

$$(4.6) \qquad \| u^{(s+1)} - v^{(s+1)} \|_2 = O_p(1).$$

Next, note that $\| g^{(j)} \|_2^2 \leq \| g^{(j-1)} \|_2 \| g^{(j+1)} \|_2$, $g \in H$, so that

$$\| u^{(j)} - v^{(j)} \|_2 \leq \| u - v \|_2^{1-j(s+1)^{-1}} \| u^{(s+1)} - v^{(s+1)} \|_2^{j(s+1)^{-1}}$$

which, along with (4.5), and (4.6) and Lemmas 3.3 and 3.4, give the result. The rate has been obtained as in Proposition 4.1. □

COROLLARY 4.2 *Under the assumptions of Theorem* 4.2, *we have that in probability and a.s.*:

$$\| f_n' - f' \|_2 = O_p(n^{-\rho(1-(s+1)^{-1})}),$$

where ρ is an in Proposition 4.1.

PROOF. Note that

$$\| f_n' - f' \|_2 \leq (\| v' \|_2^{1/2} + \| u - v \|_\infty) \| u' - v' \|_2 + \| v' \|_\infty \| u - v \|_2,$$

and the result follows from Theorems 4.1 and 4.2. □

Similar results can be derived analogously for higher order derivatives.

The strong consistency in the Hellinger distance and in the L_1-norm of the subclass of MPLE's based on kernels (1.5) can be shown (with reduced rates) without the assumption that $v \in H$. For completeness we sketch the ideas here;

for details see Klonias (1982a). Setting

$$v_*(x) = \int h^{-1}k\left(\frac{x-y}{h}\right)v(y)\,dy,$$

we have that

$$\| v_* \|^2 = \int |\tilde{v}(t)|^2 \tilde{k}(ht)\,dt \le \| k \|_1,$$

i.e., $v_* \in H$, and hence, as in Lemma 5.1 in Klonias (1982b), with v_* in place of v, we have

$$\int \hat{f}_n \log f \le \int \log u^2 dF_n \le \int f \log f + \int \log u^2 d(F_n - F), \tag{4.7}$$

$$0 < 1 - (\lambda/n) \le \int f \log f - \int \hat{f}_n \log f + \int \log u_n^2 d(F_n - F), \tag{4.8}$$

where $\hat{f}_n$ is the kernel estimate based on the R.K. of H. Also, under the assumption that $E|X|^\tau < +\infty$, $\tau > 4$, we have that for $d < 2^{-1} - \gamma(t + 2\tau^{-1})$, as $n \to \infty$,

$$\int \log u d(F_n - F), \quad \int (\hat{f}_n - f)\log f = O(n^{-d}) \text{ a.s.}, \tag{4.9}$$

where the last convergence is a consequence of the a.s. convengence of

$$\| (\hat{f}_n - f)/f^{1/2} \|_2;$$

see Bickel and Rosenblatt (1973), page 1073.

Then, from (4.7), (4.8) and (4.9) we have that

$$\lambda/n - 1, \quad \int \log u^2 dF_n - \int \log f dF = O(n^{-d}),$$

and hence (see Lemma 5.3 in Klonias, 1982b),

$$\| u - v \|_2, \quad \| f_n - f \|_1 = O(n^{-d/2}).$$

5. Examples and data based choice of the smoothing parameter.

In this section we present a number of numerical examples as well as a data based approach for the choice of the smoothing parameter h, concentrating on the univarate case.

In all figures the underlying density is represented by a dotted line. For the identification of the other curves, see the legends of the figures. For the purpose of uniformity, all MPLE's, except the one in Figure 2C, were constructed with kernel $k = \phi - \phi''$, where ϕ denotes the standard normal density (see the Introduction for a rationale), although a number of more complicated kernels performed slightly better in the estimation of bimodal curves. The kernel estimates Figures 1B, 2B, presented for comparison, are based on $k = \phi$. We generated two normal samples of size 100 each, using the IMSL random number generator GGNML with DSEED's 255866175 and 1949292845 respectively, and used the first sample for the estimation of the normal, Figures 1, 3 and the combined sample for the bimodal density, Figures 2, 4. In all cases, except Figures 3C and 4C where h is data based, we selected the curves giving the best visual fit.

For the estimation of the smoothing parameter h, note that for the MPLE's we have a data based parameter λ such that $\lambda/n \to 1$ a.s., as $n \to \infty$ and for $k \geq 0$, $\lambda < n$. Also, the roughness of the MPLE, as measured by the Sobolev norm $\| \cdot \|$, is given by $\| u \|^2 = \langle u, u \rangle = n/\lambda \to 1$ a.s. as $n \to \infty$. Based on these considerations, we propose estimating h by the maximizer of $\lambda = \lambda(h)$, subject to $\lambda \leq n$. In Figures 3C and 4C the estimates were constructed with the data based h obtained by the approach described above. For sample sizes $n = 100$, the CPU time on a VAX 11/780 needed for the construction of a data based MPLE is of the order of one minute and for $n = 200$ of the order of 90 seconds. For a given h the order of the CPU times becomes 10 and 20 seconds respectively.

6. A nonparametric regression estimator. For simplicity of notation let us consider a bivariate random variable $\mathbf{Z} = (X, Y)$. We construct the regression function estimator $\hat{m}(x) = \int y f_n(x, y)\, dy / \int f_n(x, y)\, dy$, where $f_n(x, y) = u^2(x, y)$ and u is the solution to the optimization problem (2.1) over the RKHS $H(\mathbb{R}^2)$ with kernel $\kappa_\mu(x - x', y - y')$. An interesting feature of this estimator is that if we take κ_μ to be of the form $(h_1 h_2)^{-1} k_1((x - x')/h_1) k_2((y - y')/h_2)$ the regression estimator, after some algebraic manipulation, is given by:

$$(6.1) \qquad \hat{m}(x) = \{\textstyle\sum_{i=1}^n \sum_{j=1}^n w_{ij}(Y_i + Y_j)/2\}/(\sum_{i=1}^n \sum_{j=1}^n w_{ij}),$$

where

$$w_{ij} \equiv [u(X_i, Y_i)u(X_j, Y_j)]^{-1} k_1\left(\frac{x - X_i}{h_1}\right) k_1\left(\frac{x - X_j}{h_1}\right) (k_2 * k_2) \left(\frac{Y_i - Y_j}{h_2}\right).$$

Note that the weights w_{ij} being functions of the density estimate at the knots are functions of the whole sample and not of $\mathbf{Z}_i$, $\mathbf{Z}_j$ alone.

Furthermore note that if we let $h_2 \to 0$, (6.1) reduces to the classical nonparametric regression estimator based on a kernel estimate of the density; see Watson (1964). To obtain this, note that for $i \neq j$ we have

$$(6.2) \qquad \lim(k_2 * k_2)((Y_i - Y_j)/h_2) = 0 \quad \text{as} \quad h_2 \to 0.$$

Also, note that dividing both the numerator and denominator in (6.1) by h_2, we can express the weights w_{ij} in terms of $q_i(h_2) \equiv (\lambda h_2)^{1/2} u(X_i, Y_i)$, $i = 1, \cdots, n$, rather than $u(X_i, Y_i)$. But as in (2.2), we have that

$$(6.3) \qquad q_j(h_2) = \textstyle\sum_{i=1}^n q_i(h_2)^{-1} h_1^{-1} k_1\left(\dfrac{X_j - X_i}{h_1}\right) k_2\left(\dfrac{Y_j - Y_i}{h_2}\right), \quad j = 1, \cdots, n.$$

Then, (6.2), (6.3) and the continuity of the functions involved, imply that

$$q_j(0) = q_j(0)^{-1} h_1^{-1} k_1(0) k_2(0),$$

so that

$$(6.4) \qquad q_j(0) = (h_1^{-1} k_1(0) k_2(0))^{1/2}, \quad j = 1, \cdots, n.$$

Hence, from (6.2) and (6.4) letting $h_2 \to 0$, we obtain as a limiting case of (6.1),

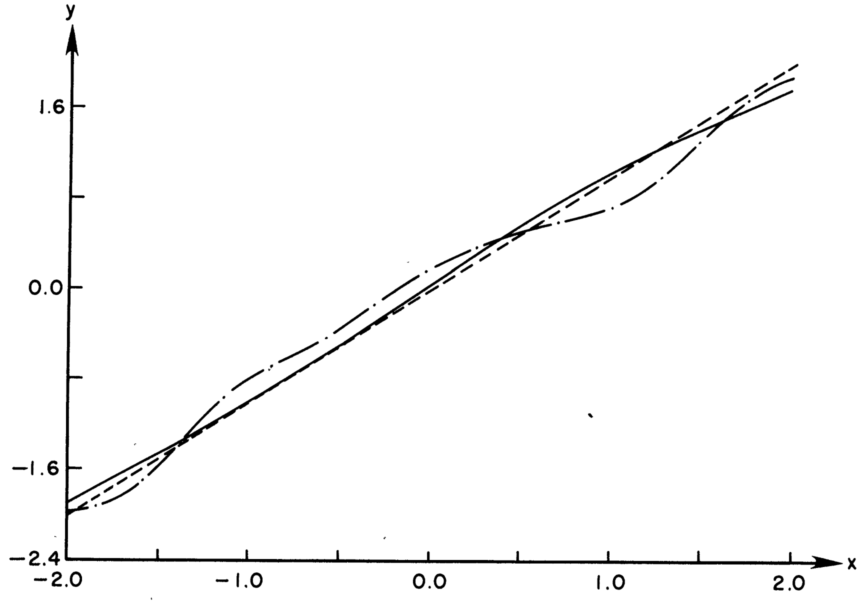

FIG. 5. *Regression* $Y = x + \varepsilon$, A: --- *equation* $y = x$. B: −·− *kernel estimator with* $h = .65$. C: —— MPLE *with* $h_1 = 1.5$, $h_2 = 50$.

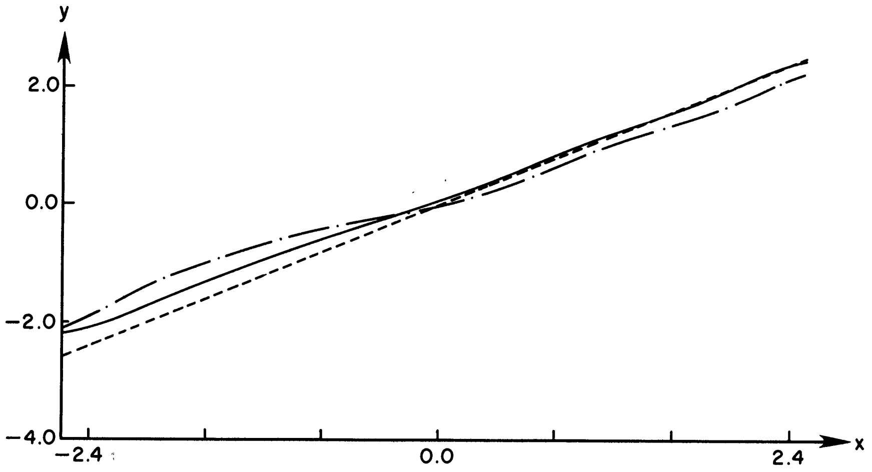

FIG. 6. *Regression* $Y = X + \varepsilon$, A: --- *equation* $y = x$. B: −·− *kernel estimator with* $h = 1$. C: —— MPLE *with* $h_1 = h_2 = 1$.

the kernel estimate of the regression function given by

$$(6.5)\qquad \hat{m}_0(x) = \{\textstyle\sum_{i=1}^{n} k_1^2((x - X_i)/h_1)Y_i\}/\{\sum_{i=1}^{n} k_1^2((x - X_i)/h_1)\}.$$

Note that the numerical evaluation of $m(x)$ is straightforward after the density estimator f_n has been computed and for $n = 100$ the numerical effort is of the order of 50 seconds CPU time on a VAX 11/780. As an example we applied (6.1) to the model $Y = x + \varepsilon$, where the x's were deterministic, uniformly spaced on $[-2, 2]$ and for errors we used the first normal data of Section 5. We used $k_1 = k_2 = \phi - \phi''$. We thought this example interesting because the uniform density is not a member of $W^{2,1}$ over $\mathbb{R}$ and certainly does not belong to the RKHS H corresponding to k_1. The estimates (6.1) and (6.5) appear in Figure 5. The setting for Figure 6 is the same as above, but now X is random. For the x-sample we used the second normal data of Section 5.

Acknowledgements. I would like to thank the referees for a number of key references.

REFERENCES

ANDERSON, J. A. and BLAIR, V. (1982). Penalized likelihood estimation in logistic regression and discrimination. *Biometrika* **69** 123–136.

BARTOSZYNSKI, R., BROWN, B. W., MCBRIDE, C. M. and THOMPSON, J. R. (1981). Some nonparametric techniques for estimating the intensity function of a cancer related nonstationary Poisson process. *Ann. Statist.* **9** 1050–1060.

BICKEL, P. J. and ROSENBLATT, M. (1973). On some global measures of the deviations of density function estimates. *Ann. Statist.* **1** 1071–1095.

BREIMAN, L. (1968). *Probability.* Addison-Wesley, New York.

BREIMAN, L., MEISEL, W. and PURCELL, E. (1977). Variable kernel estimates of probability densities. *Technometrics* **19** 135–144.

CHUNG, K. L. (1974). *A Course in Probability Theory.* Academic Press, New York.

COX, D. D. (1983). A penalty method for nonparametric estimation of the logarithmic derivative of a density function. Tech. Report No. 704, Dept. of Statist., Univ. of Wisconsin.

DEMONTRICHER, G. F., TAPIA, R. A. and THOMPSON, J. R. (1975). Nonparametric maximum likelihood estimation of probability densities by penalty function methods. *Ann. Statist.* **3** 1329–1348.

FARRELL, R. H. (1972). On the best obtainable asymptotic rates of convergence in estimation of a density function at a point. *Ann. Math. Statist.* **43** 170–180.

GOOD, I. J. and GASKINS, R. A. (1971). Nonparametric roughness penalties for probability densities. *Biometrika* **58** 255–277.

GOOD, I. J. and GASKINS. R. A. (1980). Density estimation and bumphunting by the penalized likelihood methods exemplified by scattering and meteorite data. (Invited paper). *J. Amer. Statist. Assoc.* **75** 42–73.

HALL, P. (1983). On near neighbour estimates of a multivariate density. *J. Multivariate Anal.* **13** 24–39.

IBRAGIMOV, I. A. (1956). On the composition of unimodal distributions. *Theory Probab. Appl.* **1** 255–260.

KLONIAS, V. K. (1982a). On a class of maximum penalized likelihood estimators of the probability density function. Tech. Report No. 364, Department of Mathematical Sciences, The Johns Hopkins University.

KLONIAS, V. K. (1982b). Consistency of two nonparametric maximum penalized likelihood estimators of the probability density function. *Ann. Statist.* **10** 811–824.

KLONIAS, V. K. and NASH, S. G. (1983a). On the computation of a class of maximum penalized likelihood estimators of the probability density function. Computer Science and Statistics: Fifteenth Annual Symposium on the Interface 310–314. Houston, Texas, March 1983.

KLONIAS, V. K. and NASH, S. G. (1983b). On the numerical evaluation of a class of nonparametric density and regression estimators. Tech. Report No. 376, Department of Mathematical Sciences, The Johns Hopkins University.

LOÈVE, M. (1977). *Probability Theory.* Springer-Verlag, Berlin. Graduate Texts in Mathematics 45.

LUKACS, E. (1970). *Characteristic Functions.* Griffin, London.

NASH, S. G. (1982). Preconditioning of truncated-Newton methods. Tech. Report No. 371, Department of Mathematical Sciences, The Johns Hopkins University.

PARZEN, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.* **33** 1065–1076.

REINSCH, C. (1967). Smoothing by spline functions. *Numer. Math.* **24** 383–393.

RICE, J. and ROSENBLATT, M. (1983). Smoothing splines: regression, derivatives and deconvolution. *Ann. Statist.* **11** 141–156.

RAO, C. R. (1973). *Linear Statistical Inference and its Applications.* Series in Probability and Mathematical Statistics, Wiley, New York.

SCOTT, D. W., TAPIA, R. A. and THOMPSON, J. R. (1980). Nonparametric probability density and estimation by discrete maximum penalized-likelihood criteria. *Ann. Statist.* **8** 820–832.

SILVERMAN, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statist.* **10** 795–810.

STONE, CHARLES J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8** 1348–1360.

TAPIA, R. A. and THOMPSON, J. R. (1978). *Nonparametric Probàbility Density Estimation.* The Johns Hopkins University Press, Baltimore and London.

WAHBA, G. (1975). Smoothing noisy data with spline functions. *Numer. Math.* **24** 309–317.

WAHBA, G. (1976). Histosplines with knots which are order statistics. *J. Royal Statist. Soc. B* **38** 140–151.

WATSON, G. S. (1964). Smooth regression analysis. *Sankhya A* **26** 359–372.

YOSIDA, K. (1970). *Functional Analysis.* Springer-Verlag, Berlin.

MATHEMATICAL SCIENCES
THE JOHNS HOPKINS UNIVERSITY
BALTIMORE, MARYLAND 21218