Probab. Theory Relat. Fields 94, 489–504 (1993)

Probability Theory and Related Fields

Hierarchies of higher order kernels

Alain Berlinet
Unité de Biométrie, ENSA.M, INRA, Montpellier II. 9, place Pierre Viala,
F-34060 Montpellier Cedex 1, France

Received October 22, 1991; in revised form May 18, 1992

Summary. Recent literature on functional estimation has shown the importance of kernels with vanishing moments although no general framework was given to build kernels of increasing order apart from some specific methods based on moment relationships. The purpose of the present paper is to develop such a framework and to show how to build higher order kernels with nice properties and to solve optimization problems about kernels. The proofs given here, unlike standard variational arguments, explain why some hierarchies of kernels do have optimality properties. Applications are given to functional estimation in a general context. In the last section special attention is paid to density estimates based on kernels of order (m, r), i.e., kernels of order r for estimation of derivatives of order m. Convergence theorems are easily derived from interpretation by means of projections in L^2 spaces.

Mathematics Subject Classification (1980): 62 G 05, 62 G 20, 49 B 34

1 Introduction

Before entering into a very general context we will introduce hierarchies of higher order kernels through the simple and understandable example of density estimation. Let $(X_i)_{i\in\mathbb{N}}$ be a sequence of real-valued independent random variables with common unknown density f. Consider the standard Parzen-Rosenblatt kernel estimate:

$$f_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right)$$

where $(h_n)_{n\in\mathbb{N}}$ is a sequence of positive real numbers tending to zero and K is a measurable function integrating to one. The bias of $f_n(x)$ is

$$Ef_n(x) - f(x) = \int [f(x - hu) - f(x)] K(u)\, du\ .$$

If the p^{th} order derivative of $f (p \geqq 2)$ exists and is bounded in a neighbourhood of x a Taylor series expansion gives

$$Ef_n(x) - f(x) = \sum_{k=1}^{p-1} h^k \frac{(-1)^k}{k!} f^{(k)}(x) \int u^k K(u)\, du + O(h^p) . \tag{1.1}$$

Formula (1.1) shows the importance of kernels with vanishing moments. K is said to be of order p if $\int uK(u)du = \int u^2 K(u)du = \ldots = \int u^{p-1} K(u)du = 0$ and $\int u^p K(u)du$ is finite and non null. For a kernel of order p the bias is $O(h^p)$. The first consequence of the theory introduced in Sect. 2 is that kernels can be grouped into hierarchies with the following property: each hierarchy is identified by a density K_0 belonging to it and contains kernels of order 2, 3, 4, . . . which are products of polynomials with K_0. Examples of hierarchies and algorithms for computing each element of a hierarchy from the "basic kernel" K_0 are presented in Sect. 3. Section 4 gives a technical result about sequences of hierarchies. Subsection 5.1 is devoted to properties of roots of higher order kernels. Let us now suppose that we want to use a kernel of order p to reduce the asymptotic bias but that we also want to minimize the asymptotic variance which is equivalent to $\frac{f(x)}{nh_n} \int K^2(u)du$ (Singh 1979). We have to choose K of order p so as to minimize the criterion $\int K^2(u)du$ (with some additional conditions that remove degenerate cases). Our description of finite order kernels provides a powerful portmanteau theory for such optimization problems:

- it suffices to solve the problem for the basic kernel K_0 in order to obtain a hierarchy in which every kernel will optimize the criterion at its own order. We recall that K_0 is a density, thus a positive function, which makes the problem easy to solve.
- our proofs explain why a kernel has optimal properties: we write the value of the criterion for this kernel as the difference between the value for a general kernel of the same order and an explicit positive functional.

The multiple kernel method which can be applied in any context of kernel estimation (e.g. probability density, spectral density, regression, hazard rate, intensity functions, . . .) is described in Sect. 6. It provides an estimate minimizing a criterion over the smoothing parameter h and the order of the kernel. This is applied to density estimates in Sects. 6 and 7.

Let us now turn to a more general and technical setting. When smoothing data by a kernel-type method two parameters have to be specified: a kernel K and a window-width h. As far as only positive kernels are concerned, it is known that their shape is not of crucial importance whenever h is chosen accurately. Unfortunately curve estimates built from positive kernels are usually highly biased and the improvement of kernel-type estimates requires kernels of order r (Schucany and Sommers 1977; Schucany 1989). When fitting data with such kernels the problems facing us will be:

- How to choose simultaneously the order of the kernel and the window-width?
- How to choose the shape of higher order kernels?

To deal with these practical questions we first address the following theoretical one:

• Is it possible to build hierarchies of kernels of increasing order associated with an "initial shape" from which they inherit their properties?

The answer is affirmative: the initial shape will be determined by a density K_0 and each kernel of the hierarchy will be of the form $K_r(x) = \mathscr{K}_r(x, 0)\, K_0(x)$ where $\mathscr{K}_r$ is the reproducing kernel of the space of polynomials of degree at most r imbedded in $L^2(K_0)$. It is equally easy to deal with kernels $K_r^{(m)}$ of order (m, r), i.e., kernels of order r for estimating derivatives of order m (as defined in Sect. 2); they can also be written as products of K_0 with polynomials and therefore inherit the properties of K_0: any choice of shape, support, regularity conditions (such as continuity, differentiability, etc.) or tail heaviness is possible. This possibility of choice is one of the main points of the present theory, in accordance with papers that try to dismiss the commonly held idea that practically, the kernel characteristics are at best secondary. In particular some asymmetric kernels are known to overcome boundary effects (Gasser and Müller 1979a, Müller 1991). Our framework provides easy ways of solving optimization problems about kernels. We give two examples: minimum variance kernels and minimum MISE kernels for which calculus of variations is not understood at a comfortable intuitive level. We show how the old results can be thought of as simple projection plus remainder in L^2 space and extend them to any order. Indeed if K_0 is optimal in a certain sense, each kernel of the hierarchy has an optimality property at its own order. Two hierarchies have already appeared in the literature: the Legendre and Gram-Charlier hierarchies were studied by Deheuvels in 1977. The latter has recently been reexamined by Wand and Schucany (1990), under the name of gaussian-based kernels; a paper by Granovsky and Müller shows that they can be interpreted as limiting cases of some optimal kernels. We extend this last property.

A natural extension of the concept of positivity to higher order kernels is the notion of minimal number of sign changes. This has been introduced by Gasser and Müller (1979a) to remove degenerate solutions in some optimization problems. Keeping the initial density K_0 unspecified we give in Sect. 5 very general properties about the number and the multiplicity of roots of our kernels. It turns out that kernels of order $(0, r)$ and $(1, r)$ defined from a non-vanishing density K_0 have only real roots of multiplicity one.

Up to now the methods for building kernels used some specific arguments based on moment relationships and gave no natural link between the initial kernel and the higher order ones. This is the case for the following properties:

• if $K(x)$ is of order 2, $(3\ K(x) + xK'(x))/2$ is of order 4 (Schucany and Sommers 1977; Silverman 1986). This has been generalized by Jones (1990).

• Twicing and other methods (Stuetzle and Mittal 1979; Devroye 1989): if $K(x)$ is of order s, $2K(x) - (K*K)(x)$ is of order $2s$ and $3K(x) - 3(K*K)(x) + (K*K*K)(x)$ is of order $3s$. On the contrary, our framework makes clear the relationships between kernels of different orders in the same hierarchy.

The relevant computational questions are easy to solve: two kernels of the same hierarchy differ by a product of K_0 and a linear combination of polynomials which are orthonormal in $L^2(K_0)$ and are therefore very easy to compute. When the Fourier Transform is used, choosing K_0 in a clever way may considerably reduce computational costs.

The selection of the order of a Parzen-Rosenblatt kernel was first considered by Hall and Marron (1988) in the case of density estimation. By performing a mean

integrated squared error analysis of the problem, they investigated theoretical properties of kernels with Fourier transform $\exp(-|t|^p)$ and proposed cross-validation as a method for choosing the kernel order and the smoothing parameter. We define here a multi-stage procedure for constructing curve estimates based on increasing order kernels and leading to a data-driven choice of both the order and the window-width which applies to a wide variety of smoothing problems. In the last part we will focus on density estimates based on hierarchies of kernels for which strong consistency results are available (Berlinet 1990). The interpretation of these estimates by means of projections provides exponential upper bounds for the probability of deviation.

2 Definition of K_0-based hierarchies

A common construction of finite order kernels is obtained through piecewise polynomials (Singh 1979; Müller 1984 and Gasser et al. 1985) or Fourier transform (Devroye 1987; Hall and Marron 1988). We shall be mainly concerned here with products of polynomials and densities; it turns out that almost all reasonable kernels are of this type.

Throughout the paper $V_r (r \geqq 0)$ will denote the space of polynomials of degree at most r. Unless otherwise stated integrals will be taken with respect to the Lebesgue measure on $\mathbb{R}$.

A measurable function K is said to be a kernel of order $(m, p) \in \mathbb{N}^2$, $m \leqq p-2$, if

$$\int x^j K(x)\,dx = \begin{cases} 0 & \text{for } j \in \{0, \ldots, p-1\} \text{ and } j \neq m \\ m! & \text{for } j = m \\ C_p \neq 0 & \text{for } j = p\,. \end{cases}$$

Kernels of order (m, p) are used to estimate m^{th} derivatives of functions with a reduced bias (typically of order h^p, h being the window-width; see Sect. 7 below in the case of density estimation).

A kernel of order $(0, p)$ is simply called a kernel of order p: it integrates to one, has a finite non null moment of order p and its moments of order 1 to $(p-1)$ vanish. A new and useful characterization of kernel order is given in the following lemma by means of evaluation maps for derivatives in function space.

Lemma 1 *A function K is a kernel of order (m, p) if and only if*

$$\begin{cases} \forall P \in V_{p-1} \int P(x) K(x)\,dx = P^{(m)}(0) \\ \text{and} \qquad \int x^p K(x)\,dx = C_p \neq 0\,. \end{cases}$$

Proof. Let $P \in V_{p-1}$. Let us suppose that K is of order (m, p) and expand P in Taylor series. This gives $\int P(x) K(x)\,dx = \sum_{i=0}^{p-1} \frac{P^{(i)}(0)}{i!} \int x^i K(x)\,dx = P^{(m)}(0)$.

The converse is true since $\forall j < m \quad \dfrac{d^m(x^j)}{dx^m} = 0$

$$\frac{d^m(x^m)}{dx^m} = m!$$

$$\forall j > m \quad \frac{d^m(x^j)}{dx^m} = \frac{j!}{(j-m)!} x^{j-m}\,. \quad \square$$

In other words if K is a kernel of order (m, p) the linear form on V_{p-1}:

$$P \to \int P(x)K(x)dx$$

is nothing else than the evaluation of $P^{(m)}$ at the point zero. This suggests introducing Reproducing Kernel Hilbert Subspaces (RKHS) of L^2 spaces (Berlinet 1990) and more particularly polynomial spaces, because on such spaces the evaluation maps have nice representations in terms of orthonormal bases. We will see that a convenient general structure for the construction of hierarchies of higher order kernels can be established from RKHS theory, through using a succession of reproducing kernels applied to a "basic kernel".

Let K_0 be a density function (called the basic kernel) and let V be a RKHS of $L^2(K_0)$:

V, endowed with the scalar product $(\varphi, \psi) = \int \varphi(x)\psi(x)K_0(x)dx$ is a Hilbert space of real functions and there is a function $\mathscr{K}(x, y)$ (the reproducing kernel) such that

$$\forall x \in \mathbb{R},\ \mathscr{K}(x, .) \in V$$

$$\forall \varphi \in V,\ \forall x \in \mathbb{R},\ \int \mathscr{K}(x, u)\varphi(u)K_0(u)du = \varphi(x) \quad \text{(reproduction property)}.$$

The existence of $\mathscr{K}$ is equivalent to the continuity on V of all the evaluation forms $f \to f(x)$. If $(\varphi_i)_{i \in I \subset \mathbb{N}}$ is an orthonormal basis in V we have:

$$\forall x \in \mathbb{R},\ \mathscr{K}(x, .) = \sum_{i \in I} \varphi_i(x)\varphi_i(.) \text{ (convergence in } V \text{ and pointwise)}.$$

If K_0 has finite moments up to order $2r$, then V_r is a RKHS of $L^2(K_0)$ just like any finite dimensional subspace of functions. Let $(P_i)_{0 \leqq i \leqq r}$ be the sequence of the first $(r+1)$ orthonormal polynomials in L^2K_0 (see Sect. 3 below), let $m \in \mathbb{N}$ and let

$$\mathscr{K}_r^{(m)}(x, y) = \sum_{i=0}^{r} P_i^{(m)}(y)P_i(x).$$

Note that $\mathscr{K}_r^{(m)}(x, y) = \sum_{i=m}^{r} P_i^{(m)}(y)P_i(x)$ since P_i is of exact degree i. $\mathscr{K}_r^{(m)}(., y)$ represents (in V_r) the derivation of order m as stated in

Lemma 2 $\forall \varphi \in L^2(K_0)$, $\int \varphi(x)\mathscr{K}_r^{(m)}(x, y)K_0(x)dx = \dfrac{d^m(\Pi_r(\varphi))}{dx^m}(y)$ *where* Π_r *is the projection from* $L^2(K_0)$ *onto* V_r.

Proof. Let $Q(x) = \sum_{i=0}^{r} \alpha_i P_i(x)$ be any polynomial of degree at most r:

$$\int Q(x)\mathscr{K}_r^{(m)}(x, y)K_0(x)dx = \sum_{i=0}^{r} \alpha_i P_i^{(m)}(y) = Q^{(m)}(y).$$

Now, let $\varphi \in L^2(K_0)$ and $\Pi_r(\varphi)$ be the projection of φ onto V_r. As $\mathscr{K}_r^{(m)}(\cdot, y)$ lies in V_r

$$\int \varphi(x)\mathscr{K}_r^{(m)}(x, y)K_0(x)dx = \int \Pi_r(\varphi)(x)\mathscr{K}_r^{(m)}(x, y)K_0(x)dx = \frac{d^m(\Pi_r(\varphi))}{dx^m}(y). \quad \square$$

Theorems 1 and 2 show that the product $\mathscr{K}_r^{(m)}(., 0)K_0(.)$ is precisely the form under which any reasonable kernel of order $(m, r+1)$ can be written. A real

function g is said to have a change of sign at a point z if there is $\eta > 0$ such that $g(x)$ does not vanish and keeps a fixed sign on $]z-\eta, z[$ (almost everywhere) $g(x)$ does not vanish and keeps the opposite sign on $]z, z+\eta[$ (a.e.).

Theorem 1 *Let K be an integrable function with a finite number $N \geqq 1$ of sign changes at distinct (ordered) points $z_1, z_2, \ldots, z_N$ at which it is differentiable. If K keeps* (a.e.) *a fixed sign on the intervals $]-\infty, z_1[,]z_1, z_2[, \ldots,]z_N, +\infty[$ then there is a constant A and a density K_0 such that*

$$\forall x \in \mathbb{R}, \quad K(x) = AK_0(x) \prod_{i=1}^{N} (x - z_i) .$$

Proof. Let ε be $+1$ or -1 so that $\varepsilon K(x) \prod_{i=1}^{N} (x - z_i)$ be a non-negative function (a.e.) and let H be the function defined as follows:

$$H(x) = \varepsilon K(x) \prod_{i=1}^{N} (x - z_i)^{-1} \quad \text{if } x \notin \{z_1, z_2, \ldots, z_N\}$$

$$H(z_j) = \varepsilon K'(z_j) \prod_{\substack{1 \leqq i \leqq N \\ i \neq j}} (z_j - z_i)^{-1} \quad \text{for } j = 1, \ldots, N .$$

H is non-negative (a.e.) and has a finite moment of order N. Thus $K_0 = H/\int H$ is a density and $\forall x \in \mathbb{R}$, $K(x) = \varepsilon K_0(x)\left(\prod_{i=1}^{N} (x - z_i)\right) \int H$. □

Theorem 2 *Let P be a polynomial of degree at most r, K_0 be a density with finite moments up to order $(2r+1)$ and $\mathscr{K}_r$ be the reproducing kernel of V_r in $L^2(K_0)$. Then $P(x)K_0(x)$ is a kernel of order $(m, r+1)$ if and only if*

$$\begin{cases} \forall x \in \mathbb{R}, P(x) = \mathscr{K}_r^{(m)}(x, 0) \\ \int x^{r+1} P(x) K_0(x) dx = C_{r+1} \neq 0 . \end{cases}$$

Proof. Writing a polynomial $R(x)$ of V_r in the basis $1, x, x^2, \ldots, x^r$ and applying Lemmas 1 and 2 one gets:

$$\int R(x) P(x) K_0(x) dx = R^{(m)}(0) = \int R(x) \mathscr{K}_r^{(m)}(x, 0) K_0(x) dx .$$

Thus $[P(.) - \mathscr{K}_r^{(m)}(., 0)]$ is orthogonal to V_r and the necessary condition follows. The converse is obvious. □

Theorem 2 suggests the following definition:
Let K_0 be a density and $(P_i)_{i \in I \subset \mathbb{N}}$ be the sequence of orthonormal polynomials in $L^2(K_0)$. The hierarchy of kernels associated with K_0 is the family of kernels:

$$\mathscr{K}_r^{(m)}(x, 0) K_0(x) = \sum_{i=m}^{r} P_i^{(m)}(0) P_i(x) K_0(x), \quad (r, m) \in I^2, \quad r \geqq m .$$

Note that V_r is embedded in $L^2(K_0)$ and $\mathscr{K}_r^{(m)}(., 0)$ is well defined if and only if K_0 has finite moments up to order $2r$. The set I may be reduced to $\{0\}$, as it is the case when K_0 is the Cauchy density. I is always equal to an interval of $\mathbb{N}$ with lower bound equal to zero.

Each kernel $\mathscr{K}_r^{(m)}(x, 0) K_0(x)$ with finite and non null moment of order $(r+1)$ is a kernel of order $(m, r+1)$.

We actually obtain a hierarchy of sets of kernels, the initial set being the set of densities $\left(\frac{1}{h}K_0\left(\frac{\cdot}{h}\right)\right)$, $h>0$ and a rescaling of the initial kernel does not affect this hierarchy, as stated in the next theorem.

Theorem 3 *$K(.)$ is a kernel of order (m, p) with p^{th} moment equal to C_p if and only if for any $h>0$, $\frac{1}{h^{m+1}}K\left(\frac{\cdot}{h}\right)$ is a kernel of order (m, p) whose p^{th} moment is equal to $h^{p-m}C_p$. Let $(K_r^{(m)}(.))$ be the hierarchy of kernels associated with $K_0(.)$. Then, the hierarchy associated with $\frac{1}{h}K_0\left(\frac{\cdot}{h}\right)$ is the family of kernels $\left(\frac{1}{h^{m+1}}K_r^{(m)}\left(\frac{\cdot}{h}\right)\right)$.*

Proof. The first assertion follows from the following equality:

$$\forall j\in\{0,\ldots,p\}\int x^j\frac{1}{h^{m+1}}K\left(\frac{x}{h}\right)dx=h^{j-m}\int x^jK(x)dx\ ;$$

the second one from the fact that, for any polynomial P of degree at most r, we have:

$$\int P(x)\frac{1}{h^{m+1}}\mathscr{K}_r^{(m)}\left(\frac{x}{h},0\right)K_0\left(\frac{x}{h}\right)dx=\frac{1}{h^m}\int P(hu)\mathscr{K}_r^{(m)}(u,0)K_0(u)du=P^{(m)}(0)\ .\quad\square$$

Each kernel to be used to smooth data is determined by K_0, h, p and m. To choose the shape (for instance following optimality arguments) and the smoothing parameter one chooses a suitably rescaled version of K_0. To choose (m, p) one moves along the hierarchy. The order of these operations has no importance. Bearing this in mind, we will continue to speak simply of kernel hierarchies.

3 Computational aspects

Only straightforward methods of numerical analysis are needed to calculate these kernels and the associated curve estimates. The orthonormal polynomials can be computed by means of the following relationships:

$$P_n(x)=\frac{Q_n(x)}{\|Q_n\|},\ \|Q_n\|=\left(\int Q_n^2(x)K_0(x)dx\right)^{1/2},\quad\forall n\in\mathbb{N}\ ;$$

$$Q_0(x)=1;\ Q_1(x)=x-\int xK_0(x)dx;\ Q_n(x)=(x-\alpha_n)Q_{n-1}(x)-\beta_n Q_{n-2}(x),\ n\geqq 2$$

$$\text{with}\quad\alpha_n=\frac{\int xQ_{n-1}^2(x)K_0(x)dx}{\int Q_{n-1}^2(x)K_0(x)dx}\quad\text{and}\quad\beta_n=\frac{\int Q_{n-1}^2(x)K_0(x)dx}{\int Q_{n-2}^2(x)K_0(x)dx}\ .$$

The associated kernel $K_r^{(m)}$ of order $(m, r+1)$ is given by:

$$K_r^{(m)}(x)=\sum_{i=m}^{r}P_i(x)P_i^{(m)}(0)K_0(x)\ .$$

When K_0 is symmetric, we have

$$Q_0(x)=1;\quad Q_1(x)=x;\quad Q_n(x)=xQ_{n-1}(x)-\beta_nQ_{n-2}(x),\qquad n\geqq 2$$

and $\forall n \in \mathbb{N}$, Q_{2n} is even and Q_{2n+1} is odd. Therefore, in that case, the condition $\int x^{r+1} K_r^{(m)}(x)\,dx = C_{r+1} \neq 0$ can be satisfied only if $(r+m)$ is odd; this last condition entails that $P_r^{(m)}(0) = 0$ and $\mathscr{K}_r^{(m)}(x, 0) = \mathscr{K}_{r-1}^{(m)}(x, 0)$.

The reproducing kernel can be computed either iteratively or by means of the Christoffel-Darboux formulas, when the Q_i's are known explicitly:

$$\forall x \neq y, \quad \mathscr{K}_r(x, y) = \sum_{i=0}^{r} P_i(x) P_i(y) = \frac{1}{\|Q_r\|^2}\left(\frac{Q_{r+1}(x)Q_r(y) - Q_{r+1}(y)Q_r(x)}{x - y}\right)$$

$$\forall x, \quad \mathscr{K}_r(x, x) = \sum_{i=0}^{r} [P_i(x)]^2 = \frac{1}{\|Q_r\|^2}\,[Q'_{r+1}(x)Q_r(x) - Q_{r+1}(x)Q'_r(x)]\,.$$

3.1 Determinantal expressions

To give an explicit formula for $K_r^{(m)}$, we introduce some notation: for $k \geqq 1$ and any sequence $\mu = (\mu_i)$ of real numbers, let us denote by M_k^q the Hankel matrix of order k built from $\mu_q, \mu_{q+1}, \ldots, \mu_{q+2k-2}$, and by H_k^q its determinant:

$$M_k^q = \begin{pmatrix} \mu_q & \mu_{q+1} & \cdots & \mu_{q+k-1} \\ \mu_{q+1} & & & \\ \vdots & & & \\ \mu_{q+k-1} & \cdots & & \mu_{q+2k-2} \end{pmatrix}, \quad H_k^q = \det(M_k^q)\,.$$

Finally, let $H_{k,m}^q(x)$, $x \in \mathbb{R}$, $m \in \{1, \ldots, k\}$ be the determinant of the matrix obtained from M_k^q by replacing the m^{th} line by $1, x, x^2, \ldots, x^{k-1}$. We will suppose that all the principal minors of M_{r+1}^0 are different from zero.

Theorem 4 *Let $\mu = (\mu_i)_{0 \leqq i \leqq 2s}$ be the sequence of $(2s+1)$ first moments of K_0.*

Then $\forall x \in \mathbb{R}, \forall k \in \mathbb{N}$, $Q_k(x) = H_{k+1,k+1}^0(x)/H_k^0$

$$\forall k \in \mathbb{N}, \|Q_k\| = (H_{k+1}^0/H_k^0)^{1/2}$$

$$\forall k \in \mathbb{N}, \beta_k = H_k^0 H_{k-2}^0/(H_{k-1}^0)^2$$

$$\forall k \in \mathbb{N}, P_k(x) = H_{k+1,k+1}^0(x)(H_k^0 H_{k+1}^0)^{-1/2}$$

$$\forall x \in \mathbb{R}, K_r^{(m)}(x) = m!\, H_{r+1,m+1}^0(x) K_0(x)/H_{r+1}^0\,.$$

Proof. The first four equalities are well known and easy to check (Brézinski 1980). Now, writing $K_r^{(m)}(x)$ in the basis $1, x, x^2, \ldots, x^r$ and applying the definition of a kernel of order (m, p) yields a linear system in the coefficients of $K_r^{(m)}(x)$ with matrix M_k^0.

Straightforward algebra gives the result. □

Remark. The determinantal form of $\mathscr{K}_r^{(m)}(x)$ can be used either in practical computations with small values of r, or in theoretical considerations, for instance to show that the kernels derived in Gasser et al. (1985) are those of Legendre and Epanechnikov hierarchies (we give a direct proof in Sect. 5 below).

3.2 Examples

Any choice of K_0, with finite moments up to order $2r$ ($r \geqq 1$), provides a sequence of kernels $K_r^{(m)}(x) = \mathscr{K}_r^{(m)}(x, 0)\, K_0(x)$. This choice, possibly made from the observations, has to be further investigated, especially when information is available on the support of f. As we shall see in part 5, optimal densities (in a sense to be defined) give rise to optimal hierarchies.

(a) $K_0(x) = \frac{1}{2}\mathbb{1}_{[-1,1]}(x)$ leads to piecewise polynomial kernels: Legendre kernels.
(b) $K_0(x) = \frac{3}{4}(1-x^2)_+$ is the basic kernel of the Epanechnikov hierarchy.

(a) and (b) are particular cases of the Jacobi hierarchies obtained with $K_0(x) = A(1-x)_+^{\alpha}(1+x)_+^{\beta}$.

(c) $K_0(x) = (2^{1+1/2k}\Gamma(1+1/2k))^{-1} \exp\left(-\frac{x^{2k}}{2}\right)$ gives rise to Gram–Charlier kernels when $k = 1$. The derivatives of the orthonormal polynomials are in this case linear combinations of a bounded number of polynomials of the same system.

(d) $K_0(x) = \left(\frac{2}{\beta}\Gamma\left(\frac{\alpha+1}{\beta}\right)\right)^{-1} |x|^{\alpha} \exp(-|x|^{\beta})$ gives rise to Laguerre kernels when $\alpha = 0$ and $\beta = 1$.

(e) $K_0(x) = A \exp\left(-\frac{\alpha}{4}(x-\beta)^4 - \frac{\gamma}{2}(x-\beta)^2\right)$ ($\alpha \geqq 0$; $\gamma > 0$ if $\alpha = 0$).

This family of distributions is characterized by the same property as in (c) with a number of polynomials less than or equal to two.

Some of these kernels have been discussed in detail in the literature (Deheuvels 1977a). Numerous results concerning orthogonal polynomials with weights, such as those given above and many others, can be found in Freud (1973); Nevai (1973a, b, 1979); Brézinski (1980).

4 Sequences of hierarchies

Now study how different hierarchies of kernels and different families of densities approximate each other. Let K_0 and $K_{0,\ell}$, $\ell \in \mathbb{N}$, be densities associated with families of orthonormal polynomials $(P_i)_{i\in I}$ and $(P_{i,\ell})_{i\in I}$. From Theorem 4 it is clear that the convergence, as ℓ tends to infinity, of the moments of $K_{0,\ell}$ to the corresponding moments of K_0 entails the convergence of the coefficients of $P_{i,\ell}$ to the coefficients of P_i and therefore each element of the K_0-hierarchy can appear as a limiting case of the $K_{0,\ell}$-hierarchies. From the Lebesgue dominated convergence theorem it follows that the condition of convergence of the moments is fulfilled provided that the functions $|K_{0,\ell}(x)|$, $\ell \in \mathbb{N}$, are bounded by a function with corresponding finite moments and $K_{0,\ell}$ tends to K_0 almost surely. As an example, Theorem 5 below shows that a number of hierarchies with unbounded support can appear as limiting cases of hierarchies with compact support.

Theorem 5 *Let $(K_{r,\ell}^{(m)})$ be the hierarchy of kernels associated with the density:*

$$K_{0,\ell}(x) = A_\ell |x|^{\alpha}\left(1 - \frac{\varphi(x)}{\ell}\right)^{\ell} \mathbb{1}_{\varphi(x) \leqq \ell}$$

where φ is a positive function such that $\exp(-\varphi(x))$ *has finite moments of any order. Then*

$$\forall x \in \mathbb{R}, \lim_{\ell \to +\infty} K_{r,\ell}^{(m)}(x) = K_r^{(m)}(x)$$

where $(K_r^{(m)})$ *is the hierarchy associated with the density* $K_0(x) = A/x^{\alpha} \exp(-\varphi(x))$.

Proof. The key idea is that for any positive function φ we have:

$$\forall x \in \mathbb{R}, \forall \ell \geqq 1, 0 \leqq \exp(-\varphi(x)) - \left(1 - \frac{\varphi(x)}{\ell}\right)^{\ell} \mathbb{1}_{\varphi(x) \leqq \ell} \leqq \frac{x_\ell}{\ell} \exp(-x_\ell)$$

where (x_ℓ) lies in $]1, 2[$ and satisfies $\lim_{\ell \to +\infty} x_\ell = 2$. Therefore, if φ is such that $\exp(-\varphi(x))$ has moments of any order, the conclusion follows from the Lebesgue theorem. □

Application of Theorem 5 to Example (c) above and its extension to Example (d) are straightforward. A particular case is the Gauss hierarchy with initial kernel $(2\pi)^{-1/2} \exp\left(\frac{-x^2}{2}\right)$ which is the limit, as ℓ tends to infinity, of the hierarchies associated with the densities $A_\ell \left(1 - \frac{x^2}{\ell}\right)^{\ell} \mathbb{1}_{x^2 \leqq \ell}$. Indeed, Theorem 5 makes a wide family of analytical kernels appear as limiting cases of compact support kernels with attractive properties (Granovsky and Müller 1991).

5 Optimality properties of higher order kernels

5.1 Roots of higher order kernels

A natural extension of the concept of positivity to higher order kernels is the concept of minimal number of sign changes. This has been introduced by Gasser and Müller (1979a) to remove degenerate solutions in some optimization problems. They have proved that kernels of Examples (a) and (b) have a minimal number of sign changes ($(p-2)$ for a kernel of order p). Mimicking their proof, such results can be extended to all commonly used hierarchies, once K_0 has been specified. The polynomials $\mathscr{K}_{p-1}^{(m)}(x, 0)$ do have orthogonality properties, but with respect to non necessarily positive definite functionals and the classical properties of roots of orthogonal polynomials cannot be carried over. Letting K_0 unspecified we give hereafter very general properties about the number and the multiplicity of roots of our kernels. Theorems 6 and 7 are technical. Their corollary states that kernels of order $(0, r)$ and $(1, r)$ defined from a non-vanishing density K_0 have only real roots of multiplicity one.

Theorem 6 *Let K_0 be a density of probability, let $r \geqq 2$, $m \in [0, r-1]$ and $(P_i)_{0 \leqq i \leqq r}$ be the sequence of the first $(r+1)$ orthonormal polynomials in $L^2(K_0)$. The polynomial $\mathscr{K}_r^{(m)}(x, 0) = \sum_{i=m}^{r} P_i^{(m)}(0) P_i(x)$ (of degree $d \in [1, r]$) has at least one real root of odd multiplicity.*

Proof. As K_0 is a density of probability, the equalities

$$\int \mathscr{K}_r^{(m)}(x, 0) K_0(x)\, dx = 0 \qquad (m > 0)$$

and

$$\int x^2 \mathscr{K}_r^{(0)}(x, 0) K_0(x)\, dx = x^2|_{x=0} = 0$$

show that $\mathscr{K}_r^{(m)}(x, 0)$ has at least one real root where it changes sign. □

Theorem 7 *Let r_i be the multiplicity of each real root z_i of $\mathscr{K}_r^{(m)}(x, 0)$ and let q_0 be the sum of the numbers $[r_i/2]$ (brackets denote the integer part).*

$$\textit{Then} \quad \begin{cases} \textit{either } m \textit{ is even, } 2m < r \textit{ and } 2q_0 = d + m + 1 - r \\ \textit{or } \quad 2q_0 < \min(d + 1 - m,\ d + m + 1 - r)\,. \end{cases}$$

Proof. $\mathscr{K}_r^{(m)}(x, 0) = (u(x)v(x)$ where $u(x) = \prod_i (x - z_i)^{2[r_i/2]}$ and $v(x)$ are polynomials of degrees $2q_0$ and $(d - 2q_0)$ respectively. We have

$$\forall q \in \mathbb{N},\ \int x^{2q} v(x) \mathscr{K}_r^{(m)}(x, 0) K_0(x)\, dx = \int x^{2q} u(x) [v(x)]^2 K_0(x)\, dx > 0\,.$$

The first integral would vanish if we had $2q + d - 2q_0 \leqq r$ and ($m \leqq 2q - 1$ or $m > 2q + d - 2q_0$). Therefore no integer number $q \geqq 0$ satisfies

$$m + 1 \leqq 2q \leqq r + 2q_0 - d \text{ or } 2q < \min(r + 2q_0 - d + 1,\ m + 2q_0 - d)\,.$$

The first condition is equivalent to

$$\begin{cases} (m \text{ is even and } m + 1 = r + 2q_0 - d) \\ \text{or} \\ (r + 2q_0 - d < m + 1) \end{cases}$$

while the second one is equivalent to

$$\begin{cases} (r + 2q_0 - d + 1 \leqq 0) \\ \text{or} \\ (m + 2q_0 - d \leqq 0)\,. \end{cases}$$

Since $r \geqq d$ and $q_0 \geqq 0$ the condition $(r + 2q_0 - d + 1 \leqq 0)$ cannot be satisfied. The conclusion follows. □

Corollary. *If $m \in \{0, 1\}$, $\mathscr{K}_r^{(m)}(x, 0)$ has only real roots of multiplicity one.*

Proof. If $m \in \{0, 1\}$, $2q_0 \leqq d + 1 - r \leqq 1$ thus $q_0 = 0$. □

Remark. Kernels of order $(0, r)$ and $(1, r)$ may have roots with multiplicity higher than one if K_0 has such roots or if $\mathscr{K}_r^{(m)}(x, 0)$ and $K_0(x)$ have roots in common. An example of order (0, 3) with a root of order two has been presented by Mammitzsch (1989).

5.2 Two optimal hierarchies

Our description of finite order kernels turns out to be a powerful tool in the search for asymptotically optimal kernels. It enables production of very short proofs and confirmation of a conjecture claimed by Gasser et al. (1985).

The functionals to be minimized are the same in almost all nonparametric estimation problems (cumulative distribution function, density, regression, spectral density, hazard function, . . . , and derivatives) and lead to two important families of kernels: minimum variance and minimum MISE hierarchies.

5.2.1 Minimum variance hierarchy. Minimum variance kernels of order $(m, r+1)$ on $[-1, 1]$ are solutions to the following variational problem:

$$\text{(P1)} \quad \begin{cases} W(K) = \int_{-1}^{1} K^2(x)\,dx \text{ is minimized} \\ \text{subject to } \forall P \in V_r \int_{-1}^{1} P(x)K(x)\,dx = P^{(m)}(0)\,. \end{cases}$$

They are known to be uniquely defined polynomials of degree $(r-1)$ with $(r-1)$ real roots in $[-1, 1]$, symmetric for m even and antisymmetric for m odd. Explicit formulas have been derived for their coefficients in Gasser et al. (1985) as mentioned above. We show that the minimum variance family of order $(m, r+1)$ kernels is identical to the hierarchy associated with the density $K_0(x) = \frac{1}{2}\mathbb{1}_{[-1,1]}(x)$ which is the minimum variance kernel of order (0, 2).

Theorem 8 *The solution to problem* (P1) *is given by*:

$$K_r^{(m)}(x) = \sum_{i=m}^{r} P_i^{(m)}(0) P_i(x) \mathbb{1}_{[-1,1]}(x)$$

where the P_i's are the orthonormal polynomials in $L^2(\mathbb{1}_{[-1,1]})$, i.e. the Legendre polynomials.

Proof. Let $\mathscr{K}_r^{(m)}(x, 0) = \sum_{i=m}^{r} \sqrt{2} P_i^{(m)}(0) \sqrt{2} P_i(x)$ and $K_0(x) = \frac{1}{2}\mathbb{1}_{[-1,1]}(x)$. Then, by Theorem 2, $K_r^{(m)}(x) = \mathscr{K}_r^{(m)}(x, 0)\, K_0(x)$ is a kernel of order $(m, r+1)$. Let K be an other polynomial kernel on $[-1, 1]$ of order $(m, r+1)$. K has necessarily a degree d greater than r and has the same first $(r+1)$ coordinates as $\mathscr{K}_r^{(m)}(x, 0)$.

Thus
$$K(x) = \left(\mathscr{K}_r^{(m)}(x, 0) + \sum_{i=r+1}^{d} \alpha_i P_i(x) \right) K_0(x)$$

and
$$W(K) = W(K_r^{(m)}(x)) + W\left(\sum_{i=r+1}^{d} \alpha_i P_i(x) K_0(x) \right).$$

This shows that $K_r^{(m)}(x)$ is the unique solution to problem (P1). □

5.2.2 Minimum MISE *hierarchy.* Gasser et al. introduced polynomial kernels for which they proved optimality up to order 5 and conjectured the same property for any order. This conjecture can now be proved using the unifying variational principle introduced in Granovsky and Müller (1991). We give here a general very simple proof. The minimum MISE family of order (m, p) kernels is identical to the hierarchy associated with the Epanechnikov density: $(3/4)\,(1-x^2)_+$ which is the minimum MISE kernel of order (0, 2).

Minimum MISE kernels of order $(m, r+1)$ on $[-1, 1]$ are solutions to the following variational problem ($(r+m)$ is supposed to be odd):

(P2)

$$\begin{cases} T(K) = \left(\int_{-1}^{1} K^2(x)\,dx \right)^{r+1-m} \left| \int_{-1}^{1} x^{r+1} K(x)\,dx \right|^{2m+1} \text{ is minimized} \\ \text{subject to } \forall P \in V_r \int_{-1}^{1} P(x)K(x)\,dx = P^{(m)}(0)\,. \end{cases}$$

Theorem 9 *The polynomial solution to problem* (P2) *vanishing at the ends of* $[-1, 1]$ *is given by*:

$$K_r^{(m)}(x) = \sum_{i=m}^{r} P_i^{(m)}(0)\, P_i(x)(3/4)(1-x^2)_+$$

where the P_i*'s are the orthonormal polynomials in* $L^2(K_0)$ *with* $K_0(x) = (3/4)(1-x^2)_+$.

Proof. Obviously, $K_r^{(m)}$ satisfies the condition. The functional T is invariant under scale transformations $K(.) \to \frac{1}{h^{m+1}} K\left(\frac{\cdot}{h}\right)$. Therefore we have to compare $W(K_r^{(m)})$ with $W(RK_0)$ where R is a polynomial such that

$$\begin{cases} \int_{-1}^{1} x^{r+1} R(x) K_0(x)\, dx = \int_{-1}^{1} x^{r+1} K_r^{(m)}(x)\, dx \\ \forall P \in V_r \int_{-1}^{1} P(x) R(x) K_0(x)\, dx = P^{(m)}(0)\,. \end{cases}$$

It turns out that $(R - \mathscr{K}_r^{(m)}(x, 0))$ is orthogonal to V_{r+1} in $L^2(K_0)$. Now,

$$W(RK_0) = \int (RK_0 - K_r^{(m)})^2 + W(K_r^{(m)}) + 2\int K_r^{(m)}(x)(R(x) - \mathscr{K}_r^{(m)}(x, 0)) K_0(x)\, dx\,.$$

As K_0 is symmetric, $K_r^{(m)}$ is of degree $(r+1)$ at most. Thus $W(RK_0) = \int (RK_0 - K_r^{(m)})^2 + W(K_r^{(m)})$ and the conclusion follows. □

Granovsky and Müller (1989) proved that $K_r^{(m)}$ minimizes the same criterion over the set of square integrable kernels of order (m, p) with a fixed number $(p-2)$ of sign changes on $\mathbb{R}$.

6 The multiple kernel method

Let us suppose that a function f (e.g. a probability density function, a spectral density function, a regression function, an intensity function, . . .) has to be estimated from a sample of points and that a criterion C has been chosen to judge the accuracy of any kernel estimate f_n: C is a score function of the sample estimating some measure of deviation between f_n and the true unknown function f. Once the sample is given, C is a function of the rescaled kernel $\frac{1}{h^{m+1}} K_r^{(m)}\left(\frac{\cdot}{h}\right)$. The initial kernel K_0 is chosen regarding the asymptotic behaviour of C.

As an example one can think of the problem of density estimation from a sample $X_1, \ldots, X_n$ of independent random variables with common density f. If the criterion is the MISE (Mean Integrated Squared Error) $= E(\int (f_n(x) - f(x))^2\, dx)$ where $f_n(x)$ is the standard Parzen–Rosenblatt kernel estimate $\frac{1}{nh} \sum_{j=1}^{n} \mathscr{K}_r\left(\frac{x - X_j}{h}, 0\right) K_0\left(\frac{x - X_j}{h}\right)$ built from the sample, a natural choice for K_0 is the Epanechnikov optimal kernel, or a nearly optimal kernel (under suitable assumptions on f, see Epanechnikov 1969). A natural choice for C is the L_2 cross-validation criterion: $\int f_n^2(x)\, dx - \frac{2}{n} \sum_{i=1}^{n} f_{n,i}(X_i)$ where $f_{n,i}$ is the kernel estimate

based on the $(n-1)$ observations different from X_i. For relevant discussion and references, see Berlinet and Devroye (1989). Once K_0 has been chosen one can compute for any order r the value h_r of the smoothing parameter optimizing (at least over a grid G) $C\left(\frac{1}{h^{m+1}} K_r^{(m)}\left(\frac{\cdot}{h}\right)\right)$. Let C_r be the value of C at the optimal h_r. Then, the optimal order $\hat{r}$ in a bounded interval $[0, R]$ is defined so as to optimize C_r over $[0, R]$ and the corresponding rescaled kernel $\frac{1}{h_{\hat{r}}} K_{\hat{r}}^{(m)}\left(\frac{\cdot}{h_{\hat{r}}}\right)$ is used to build f_n.

The multiple kernel method can also be used with estimates $f_{n,r}$ and $f_{n,s}$ of different orders r and s to provide best smoothing parameters h_r and h_s at these orders as proposed by Devroye (1989): h_r and h_s are chosen so as to minimize for instance the L^1 distance between $f_{n,r}$ and $f_{n,s}$.

7 The estimation procedure for the density and its derivatives

As in Sect. 6 above let $X_1, \ldots, X_n$ be independent random variables with common unknown density f and cumulative distribution function F. We give in this section some specific properties of estimates of f, F and of derivatives of f based on higher order kernels. These estimates can be interpreted by means of projections in L^2 spaces. Let $f_n(x) = \frac{1}{nh} \sum_{j=1}^{n} K_r\left(\frac{x - X_j}{h}\right)$ be the standard kernel estimate of f built from the kernel $K_r(x) = \mathscr{K}_r(x, 0)\, K_0(x)$. Let μ_n be the measure with density f_n and $\bar{\mu}_n$ be the empirical measure associated with the sample. Theorem 10 shows that estimating the measure $\mu(A)$ of a Borel set A with a kernel like K_r and smoothing parameter h is nothing else than deriving the best L^2-approximation with weight K_0 of the function $\bar{\mu}_n(A - h.)$ by a polynomial Π_A of degree at most r and taking $\Pi_A(0)$ as an estimate of $\mu(A)$:

Theorem 10 *For any Borel set A, we have $\mu_n(A) = \Pi_A(0)$ where*

$$\Pi_A = \arg \min_{\pi \in V_r} \int (\pi(u) - \bar{\mu}_n(A - hu))^2 K_0(u)\, du \,.$$

Proof.

$$\mu_n(A) = \int \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_A(X_j + hv) \mathscr{K}_r(v, 0) K_0(v)\, dv \,. \tag{7.1}$$

The integral in (7.1) is the value at 0 of the projection of $\frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_A(X_j + h.)$ on the subspace V_r i.e. the solution of the following problem: find π in V_r minimizing the norm of $(\pi(.) - \bar{\mu}_n(A - h.))$ and evaluate π at the point 0. The conclusion follows. □

Now let us see how to handle the deviation

$$(f^{(m)}(x) - f_n^{(m)}(x)) = (f^{(m)}(x) - Ef_n^{(m)}(x)) + (Ef_n^{(m)}(x) - f_n^{(m)}(x))$$

between the m^{th} derivative of f and its standard kernel estimate. Let us suppose, as it is usually the case, that the function $d(.) = f(x - h.)$ belongs to $L^2(K_0)$. Theorem 11 gives the relationship between the expectation of $f_n^{(m)}(x)$ and the function d and provides an exponential upper bound for the probability of deviation: $\Pr(|f_n^{(m)}(x) - Ef_n^{(m)}(x)| \geqq \varepsilon)$.

Theorem 11 *Let* $f_n^{(m)}(x) = \frac{1}{nh^{m+1}} \sum_{j=1}^{n} \mathscr{K}_r^{(m)}\left(\frac{x - X_j}{h}, 0\right) K_0\left(\frac{x - X_j}{h}\right)$ *be the standard kernel estimate of the* m^{th} *derivative of* f. *Suppose that the function* $d(.) = f(x - h.)$ *belongs to* $L^2(K_0)$ *then the expectation of* $f_n^{(m)}(x)$ *is the value at* 0 *of the* m^{th} *derivative of the polynomial* P_h *such that* $P_h(h.)$ *is the projection of* d *on* V_r. *If moreover* $|K_r^{(m)}|$ *is bounded by the constant* $M(m, r)$ *we have*:

$$\forall \varepsilon > 0, \Pr(|f_n^{(m)}(x) - P_h^{(m)}(0)| \geqq \varepsilon) \leqq 2\exp\left\{\frac{-\varepsilon^2 n h^{2(m+1)}}{2M^2(m, r)}\right\}.$$

Proof. $Ef_n^{(m)}(x) = \left(\frac{1}{h^{m+1}} K_r^{(m)}\left(\frac{\cdot}{h}\right) * f\right)(x) = \frac{1}{h^m} \int f(x - hv) \mathscr{K}_r^{(m)}(v, 0) K_0(v)\, dv$

$$Ef_n^{(m)}(x) = \frac{1}{h^m} \frac{d^m(P_h(hv))}{dv^m}\bigg|_{v=0} = P_h^{(m)}(0).$$

The inequality is a consequence of Lemma 1.2 in (Mc Diarmid 1989). □

We have a similar result for $F_n(x)$ when the function $F(x - h.)$ belongs to $L^2(K_0)$. Now, once K_0 is specified deterministic approximation theorems in $L^2(K_0)$ give the behaviour of $(f^{(m)}(x) - Ef_n^{(m)}(x))$. Thus weak or strong (using Borel–Cantelli lemma) convergence theorems can be easily derived for $f_n^{(m)}(x)$. Strong consistency results covering a wide class of density estimates were given in (Berlinet 1990). They can be applied in the framework of this paper to hierarchies of density estimates.

Acknowledgements. I wish to thank Professor J.S. Marron and two referees for helpful comments about the presentation of this paper.

References

Berlinet, A.: Reproducing kernels and finite order kernels. In: Roussas, G. (ed.) Nonparametric functional estimation and related topics, pp. 3–18. London New York: Kluwer 1990

Berlinet, A., Devroye, L.: Estimation d'une densité: un point sur la méthode du noyau. Stat. Anal. Données **14** (n° 1), 1–32 (1989)

Brézinski, C.: Padé-type approximation and general orthogonal polynomials. Basel: Birkhäuser 1980

Deheuvels, P.: Estimation non-paramétrique de la densité par histogrammes généralisés. Rev. Stat. Appl. **25**, 5–42 (1977)

Devroye, L.: The double kernel method in density estimation. Ann. Inst. Henri Poincaré **25** (n° 4), 553–580 (1989)

Epanechnikov, V.A.: Nonparametric estimation of a multidimensional probability density. Theory Probab. Appl. **14**, 153–158 (1969)

Freud, G.: On polynomial approximation with the weight $\exp(-x^{2k}/2)$. Acta Math. Acad. Sci. Hung. **24**, 363–371 (1973)

Gasser, T., Müller, H.G.: Kernel estimation of regression function. In: Gasser, T., Rosenblatt, M. (eds.) Smoothing techniques for curve estimation. (Lect. Notes Math., vol. 757, pp. 23–68) Berlin Heidelberg New York: Springer 1979a

Gasser, T., Müller, H.G.: Optimal convergence properties of kernel estimates of derivatives of a density function. In: Gasser, T., Rosenblatt, M. (eds.) Smoothing techniques for curve estimation. (Lect. Notes Math., vol. 757, pp. 144–154) Berlin Heidelberg New York: Springer 1979b

Gasser, T., Müller, H.G., Mammitzsch, V.: Kernels for nonparametric curve estimation. J. R. Stat. Soc., Ser. B **47**, 238–252 (1985)

Granovsky, B.L., Müller, H.G.: On the optimality of a class of polynomial kernel functions. Stat. Decis. **7**, 301–312 (1989)

Granovsky, B.L., Müller, H.G.: Optimizing kernel methods: a unifying variational principle. Int. Stat. Rev. **59** (3), 373–388 (1991)

Hall, P., Marron, J.S.: Choice of kernel order in density estimation. Ann. Stat. **16**, 161–173 (1988)

Jones, M.C.: Changing kernels' orders. (Preprint 1990)

Mc Diarmid, C.: On the method of bounded differences. In: Surveys in combinatorics. (Lond. Math. Soc. Lect. Notes Ser., vol. 141, pp. 148–188) Cambridge: Cambridge University Press 1989

Mammitzsch, V.: A note on kernels of order v, k. In: Mande, P., Huškovich, M. (eds.) Proceedings of the Fourth Prague Symposium on Asymptotic Statistics, pp. 411–412. Prague: Charles University 1989

Müller, H.G.: Smooth optimum kernel estimators of densities, regression curves and modes. Ann. Stat. **12**, 766–774 (1984)

Müller, H.G.: Weighted local regression and kernel methods for nonparametric curve fitting. J. Am. Stat. Assoc. **82**, 231–238 (1987)

Müller, H.G.: On the construction of boundary kernels. University of California at Davis (Preprint 1991)

Nevai, P.: Some properties of orthogonal polynomials corresponding to the weight $(1 + x^{2k})^{\alpha} \exp(-x^{2k})$ and their application in approximation theory. Sov. Math. Dokl. **14**, 1116–1119 (1973a)

Nevai, P.: Orthogonal polynomials on the real line associated with the weight $|x|^{\alpha} \exp(-|x|^{\beta})$, I. Acta Math. Acad. Sci. Hung. **24**, 335–342 (1973b)

Nevai, P.: Orthogonal polynomials. Mem. Am. Math. Soc. **219** (1979)

Schucany, W.R.: On nonparametric regression with higher-order kernels. J. Stat. Plann. Inference **23**, 145–151 (1989)

Schucany, W.R.: Sommers, J.P.: Improvement of kernel type density estimators. J. Am. Stat. Assoc. **72**, 420–423 (1977)

Silverman, B.W.: Density estimation for statistics and data analysis. London: Chapman and Hall 1986

Singh, R.S.: Mean squared errors of estimates of a density and its derivatives. Biometrika **66**(1), 177–180 (1979)

Stuetzle, W., Mittal, Y.: Some comments on the asymptotic behavior of robust smoothers. In: Gasser, T., Rosenblatt, M. (eds.) Smoothing techniques for curve estimation. (Lect. Notes Math., vol. 757, pp. 191–195) Berlin Heidelberg New York: Springer 1979

Wand, M., Schucany, W.R.: Gaussian-based kernels. Can. J. Stat. **18**, 197–204 (1990)