Improving bias in kernel density estimation

Mynbaev, Kairat and Nadarajah, Saralees and Withers, Christopher and Aipenova, Aziza

Kazakh-British Technical University,, School of Mathematics, University of Manchester, Applied Mathematics Group, Industrial Research Limited,, Kazakh National University,

2014

Improving bias in kernel density estimation

Kairat T. Mynbaev[a,1,*], Saralees Nadarajah[b], Christopher S. Withers[c], Aziza S. Aipenova[d]

[a]Kazakh-British Technical University, Almaty 050000, Kazakhstan
[b]School of Mathematics, University of Manchester, Manchester M13 9PL, UK
[c]Applied Mathematics Group, Industrial Research Limited, Lower Hutt, New Zealand
[d]Kazakh National University, Almaty, Kazakhstan

Abstract

For order q kernel density estimators we show that the constant b_q in $bias = b_q h^q + o(h^q)$ can be made arbitrarily small, while keeping the variance bounded. A data-based selection of b_q is presented and Monte Carlo simulations illustrate the advantages of the method.

Keywords: density estimation, bias, higher order kernel

2010 MSC: 62G07, 62G10, 62G20

1. Introduction

Let f denote a density, K an integrable function on $\mathbb{R}$ such that $\int K dt = 1$ and let $X_1, ..., X_n$ be i.i.d. random variables with density f. Consider the kernel estimator of $f(x)$

$$f_h(x, K) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{h} K\left(\frac{x - X_j}{h}\right), \ h > 0. \tag{1}$$

Denote $\alpha_i(K) = \int x^i K(x) dx$ the ith moment of K and let K be a kernel of order q, that is $\alpha_j(K) = 0$, $j = 1, ..., q-1$, $\alpha_q(K) \neq 0$. It is well-known that the

[*] Corresponding author

Email addresses: kairat_mynbayev@yahoo.com (Kairat T. Mynbaev), Saralees.nadarajah@manchester.ac.uk (Saralees Nadarajah), kit.withers@gmail.com (Christopher S. Withers), a.aipenova@mail.ru (Aziza S. Aipenova)

[1] Research supported in part by the Ministry of Education and Science of the Republic of Kazakhstan. The Ministry had no involvement in study design, collection, analysis and interpretation of data; in the writing of the report; and in the decision to submit the article for publication.

bias is proportional to $\alpha_q(K)h^q$ if f is q-smooth in some sense [Devroye, 1987, Scott, 1992, Silverman, 1986, Wand & Jones, 1995].

The usual approach is to stick to some K and be content with the resulting $\alpha_q(K)$. The purpose of this paper is to show that it pays to reduce $\alpha_q(K)$ by choosing a suitable K. Despite the bias being proportional to $\alpha_q(K)h^q$, the benefits of the suggested approach are not obvious because as the qth moment is made smaller, the variance of the estimator may go up. Our construction of K allows us to control the variance. Our results imply that among all kernels of order q with uniformly bounded variances there is no kernel with the least nonzero $|\alpha_q(K)|$. The issue of selecting the kernel order does not arise in the approach suggested in [Mynbaev and Martins Filho, 2010].

In case of L_1 convergence the main idea can be illustrated using the corresponding bias notion from Devroye [1987]. Let bias be defined as $\int |f \star K_h - f| dt$ where $K_h(x) = K(x/h)/h$. If K is of order q, f has $q-1$ absolutely continuous derivatives and an integrable derivative $f^{(q)}$, then by [Devroye, 1987, Theorem 7.2]

$$q! \int |f \star K_h - f| dt / \left(h^q \int |f^{(q)}| dt \right) \to \alpha_q(K), \ h \downarrow 0.$$

Here $\alpha_q(K)$ can be made as small as desired using our Theorem 2.

We call a free-lunch effect the fact that $\alpha_q(K)$ can be made as small as desired, without increasing the density smoothness or the kernel order. Of course, in finite samples bias cannot be eliminated completely. Put it differently, for very small $\alpha_q(K)$ sample variance starts to dominate the effect of small bias.

For simplicity, in our main results in Section 2 we consider only classical smoothness characteristics. The simulation results in Section 3 compare our kernel performance with that of three well-known kernel families. The overall conclusion is that a better estimation performance is not necessarily a consequence of some optimization criterion and can be achieved by directly targeting the bias of the estimator. All proofs are contained in Section 4.

2. Main results

Multiplication by polynomials [Deheuvels, 1977, Wand & Schucany, 1990] is one of several ways to construct higher-order kernels. [Withers and Nadarajah, 2013] have explored the procedure of transforming a kernel K into a higher-order kernel $T_{\boldsymbol{a}}K$ via multiplication of K by a polynomial of order q, $(T_{\boldsymbol{a}}K)(t) = \left(\sum_{i=0}^{q} a_i t^i\right) K(t)$, with a suitably chosen vector of coefficients $\boldsymbol{a} = (a_0, ..., a_q)' \in \mathbb{R}^{q+1}$. Unlike several authors who chose the polynomial subject to some optimization criterion (see [Berlinet, 1993, Fan & Hu, 1992, Gasser & Muller, 1979, Lejeune and Sarda, 1992, Wand & Schucany, 1990]) Withers & Nadarajah with their definition of the polynomial directly targeted moments of the resulting kernel. In their Theorem 2.1, they defined a polynomial transformation in such a way that the moments of the new kernel numbered 1 through $q-1$ are zero. They did not notice that the qth moment can be targeted in the same way and can be made as small as desired and that the variance of the resulting estimator retains the order $1/(nh)$ as the qth moment is manipulated. This is what we do here. Besides, we show that not only variance but all the higher-order terms in h in the Taylor decomposition of the bias and variance can be controlled not to increase.

We do this under two sets of assumptions. The first set is that the density is infinitely differentiable and all moments of K exist and the second is that the density has a finite number of derivatives and the kernel and its square possess a finite number of moments. We give complete proofs for the first set, because part of the argument is new and it can be extended to justify some formal infinite decompositions from [Withers and Nadarajah, 2013]. The proof for the second set goes more along traditional lines (except for controlling higher-order terms) and is therefore omitted.

Let $\beta_j(K) = \int_{\mathbb{R}} |K(t)t^j| dt$ denote the jth absolute moment of K. The estimator of $f^{(l)}(x)$ is obtained by differentiating both sides of (1) l times.

Theorem 1. *Suppose that f is infinitely differentiable and K has a continuous derivative of order l. Further assume that K and $K^{(l)}$ have absolute moments* *of all orders,*

$$\limsup_{j\to\infty}\left|\frac{f^{(j)}(x)}{j!}\max\left\{\beta_{j+1}(K),\beta_{j+1}(K^{(l)})\right\}\right|^{1/j}=0, \tag{2}$$

$$\left\|K^{(l)}\right\|_{C(\mathbb{R})}=\sup_{t\in\mathbb{R}}\left|K^{(l)}(t)\right|<\infty. \tag{3}$$

Then

$$Ef_h^{(l)}(x,K)=\sum_{i=0}^{\infty}\frac{f^{(i+l)}(x)}{i!}(-h)^i\alpha_i(K), \tag{4}$$

$$var\left(f_h^{(l)}(x,K)\right)=\frac{1}{nh^{2l+1}}\left\{\sum_{i=0}^{\infty}\frac{f^{(i)}(x)\alpha_i(M)}{i!}(-h)^i-h\left[h^lEf_h^{(l)}(x,K)\right]^2\right\} \tag{5}$$

where $M=\left[K^{(l)}\right]^2$ and the series converge for all $h\in\mathbb{R}$. Consequently, if K is a kernel of order q, then

$$\begin{aligned} Ef_h^{(l)}(x,K)-f^{(l)}(x) &= \frac{f^{(q+l)}(x)}{q!}(-h)^q\alpha_q(K)+O(h^{q+1}), && (6)\\ var\left(f_h^{(l)}(x,K)\right) &= \frac{1}{nh^{2l+1}}\left\{f(x)\int_{\mathbb{R}}M(t)dt+O(h)\right\}. && (7)\end{aligned}$$

Further, for the ISE convergence of the estimator of the lth derivative of f the asymptotically optimal bandwidth is given by

$$h_{opt}=\left\{\frac{(2l+1)\alpha_0((K^{(l)})^2)}{2qn\alpha_0(f^{(q+l)})^2}\left[\frac{q!}{\alpha_q(K)}\right]^2\right\}^{1/(2q+2l+1)}. \tag{8}$$

With the function K we can associate symmetric matrices

$$\mathbf{A}_q(K)=\begin{pmatrix}\alpha_0(K) & \alpha_1(K) & ... & \alpha_q(K)\\ \alpha_1(K) & \alpha_2(K) & ... & \alpha_{q+1}(K)\\ ... & ... & ... & ...\\ \alpha_q(K) & \alpha_{q+1}(K) & ... & \alpha_{2q}(K)\end{pmatrix},\ \mathbf{B}_q=\mathbf{A}_q(K^2).$$

In the next theorem we prove the free-lunch effect, for simplicity limiting ourselves to estimation of $f(x)$.

Theorem 2. *Suppose that f is infinitely differentiable, K is continuous and has absolute moments of all orders, $\|K\|_{C(\mathbb{R})} < \infty$,*

$$\limsup_{j\to\infty}\left|\frac{f^{(j)}(x)}{j!}\beta_{q+j+1}(K)\right|^{1/j} = 0, \tag{9}$$

$$\det \mathbf{A}_q(K) \neq 0. \tag{10}$$

Let a vector $\boldsymbol{b}\in\mathbb{R}^{q+1}$ have components $b_0 = 1$, $b_1 = ... = b_{q-1} = 0$, $b_q > 0$ and set $\boldsymbol{a} = \mathbf{A}_q(K)^{-1}\boldsymbol{b}$. Then

$$Ef_h(x, T_{\boldsymbol{a}}K) - f(x) = \frac{f^{(q)}(x)}{q!}(-h)^q b_q + O(h^{q+1}), \tag{11}$$

$$var\left(f_h(x, T_{\boldsymbol{a}}K)\right) = \frac{1}{nh}\left\{f(x)\boldsymbol{b}'\mathbf{C}_q\boldsymbol{b} + O(h)\right\} \tag{12}$$

where $\mathbf{C}_q = \mathbf{A}_q(K)^{-1}\mathbf{B}_q\mathbf{A}_q(K)^{-1}$ and $\boldsymbol{b}'\mathbf{C}_q\boldsymbol{b} > 0$. The terms of higher order in h in (11) and (12) retain their magnitude as $b_q \to 0$.

Remark 1. Taking $0 < m < q$, $b_0 = ... = b_{m-1} = b_{m+1} = ... = b_{q-1} = 0$, $b_m = 1$, $b_q \neq 0$ we obtain an (m, q)-kernel, see the related definitions and theory in [Berlinet & Thomas-Agnan, 2004].

Corollary 1. *Denote the elements of $\mathbf{A}_q(K)^{-1}$ by $\mathbf{A}_q^{ij}$, $i, j = 0, ..., q$, $\boldsymbol{c} = (1, 0, ..., 0)' \in \mathbb{R}^{q+1}$ and $\boldsymbol{d} = (0, ..., 0, b_q)' \in \mathbb{R}^{q+1}$. Then $\boldsymbol{b} = \boldsymbol{c} + \boldsymbol{d}$. As $b_q \to 0$, one has $(T_{\boldsymbol{a}}K)(t) \to \left(\sum_{i=0}^{q}\mathbf{A}_q^{i,0}t^i\right)K(t)$, $\boldsymbol{b}'\mathbf{C}_q\boldsymbol{b} = \boldsymbol{c}'\mathbf{C}_q\boldsymbol{c} + O(b_q) \to (\mathbf{C}_q)_{11} = \sum_{i,j}\mathbf{A}_q^{i,0}(\mathbf{B}_q)_{ij}\mathbf{A}_q^{j,0}$. It follows that in (11) b_q can be made as small as desired, while (12) retains its magnitude as we do this.*

Remark 2. In the course of the proof of Theorem 2 we show that $\mathbf{B}_q$ is positive definite. The argument can be adapted to show that (10) holds if K is nonnegative.

Since $T_{\boldsymbol{a}}K$, the optimal bandwidth and the minimized value of the asymptotic ISE all depend on the number b_q in (11), application of the optimal bandwidth (8) is not straightforward. We find it more convenient to discuss the choice of b_q in the simulations section.

In the next theorem we give conditions sufficient for the free-lunch effect when f is not infinitely differentiable and K does not possess moments of all orders.

Kernel family	q=2	q=4	q=6	q=8	q=10	q=12
Epanechnikov	0.2000	-0.0476	0.0117	-0.0029	0.0007	-0.0002
Gram-Charlier	1	-3	15	-105	945	-10395

Table 1: Moments of two types of kernels

Theorem 3. *Suppose that* (10) *holds, f is $(q+1)$-times continuously differentiable, $\|f'\|_{C(\mathbb{R})} + \left\|f^{(q+1)}\right\|_{C(\mathbb{R})} < \infty$ and $\beta_{2q+1}(K) + \beta_{2q+1}(K^2) < \infty$. Then* (11) *and* (12) *are true. For the ISE convergence the optimal bandwidth is given by* (8) *where $l = 0$.*

3. Monte Carlo simulations

3.1. Description of kernel families and target densities

We focus on the category of kernels obtained from second-order kernels by multiplying by polynomials, because our estimator is in this category. This type of kernel construction is also known to be computationally efficient. For the purpose of comparison with our kernels, we select two classes of kernels. One is based on the Gaussian kernel and the other extends Epanechnikov's approach. We take the two families from [Berlinet & Thomas-Agnan, 2004].

Epanechnikov-type kernels are given in [Berlinet & Thomas-Agnan, 2004, p.142]. The entry for the 8th order should look like this: $(11025 - 132300x^2 + 436590x^4 - 540540x^6 + 225225x^8)/4096$. Outside the segment $[-1, 1]$ the kernels are zero, inside the segment they are defined by the formulas in that table.

The Gram-Charlier kernels are taken from [Berlinet & Thomas-Agnan, 2004, p.140]. The entry for the 8th order has also been corrected to $(105 - 105x^2 + 21x^4 - x^6)/48\phi(x)$. The corrections are based on equations from [Berlinet & Thomas-Agnan, 2004, p.162] implemented in Mathematica. Here $\phi(x)$ is the Gaussian density. All these kernels have even orders, and we also use only even orders. The moments of the kernels of two types are given in Table 1.

The target densities, that is the densities to be estimated, are those proposed in [Marron & Wand, 1992]. They are normal mixtures defined as follows:

1. Gaussian ($f_1(x) \equiv N(0,1)$),
2. Bimodal ($f_2(x) \equiv N(-1, 4/9)/2 + N(1, 4/9)/2$),
3. Separated-Bimodal ($f_3(x) \equiv N(-1.5, 1/4)/2 + N(1.5, 1/4)/2$) and
4. Trimodal ($f_4(x) \equiv (9N(-6/5, 9/25) + 9N(6/5, 9/25) + 2N(0, 1/16))/20$).

They are listed in the order of increasing curvature, the Trimodal being the most difficult to estimate.

3.2. Bandwidth choice

Equations (6) and (7) imply that $ISE = \int (var + bias^2)dx$ asymptotically is $\phi(h)$ where

$$\phi(h) = c_1 h^{2q} + c_2 h^{-(2l+1)},\ c_1 = \left(\frac{\alpha_q(K)}{q!}\right)^2 \int \left(f^{(q+l)}\right)^2 dx,\ c_2 = \frac{1}{n}\int \left(K^{(l)}\right)^2 dx. \tag{13}$$

Minimizing ϕ we obtain

$$h_{opt} = \left(\frac{(2l+1)c_2}{2qc_1}\right)^{1/(2q+2l+1)} \tag{14}$$

from which the usual expression for the optimal bandwidth (8) obtains.

In what follows we consider only estimation of densities ($l = 0$). In this case the minimized value of ϕ is

$$\phi(h_{opt}) = c_1^{1/(2q+1)} c_2^{2q/(2q+1)} (2q+1)(2q)^{-2q/(2q+1)}. \tag{15}$$

For conventional kernels the constants c_1, c_2 are given by (13) with $l = 0$ and in case of $T_{\boldsymbol{a}}K$ we have functions of b_q

$$c_1(b_q) \ = \left(\frac{b_q}{q!}\right)^2 \int \left(f^{(q)}\right)^2 dx,\ c_2(b_q) = \frac{1}{n}\boldsymbol{b}'\mathbf{C}_q\boldsymbol{b}. \tag{16}$$

Plugging (16) in (14) we obtain definitions of $h_{opt}(b_q)$. Substituting $h_{opt}(b_q)$ in (15) we obtain $\phi(h_{opt}(b_q))$.

Obviously, (15) tends to zero as $b_q \to 0$. However, setting $b_q = 0$ would not eliminate bias completely. There is a general fact that for kernel estimators bias can be zero only in case of special densities and kernels [Devroye, 1987, p.113]. In our situation, we illustrate this fact in Figure 1, which depicts the behavior of average bias and MSE as functions of b_q. Both increase as $b_q \to 0$. (Note: in Figure 1, average bias is the average over iterations of integrals $\int (f(x) - \hat{f}(x, b_q))dx$ for each value of b_q; $f(x)$, $\hat{f}(x, b_q)$ are a density and its estimator, resp.). When $b_q \to 0$, the "optimal" bandwidth (14) tends to infinity. The estimator becomes oversmoothed, thus the behavior of average bias and MSE observed in Figure 1.

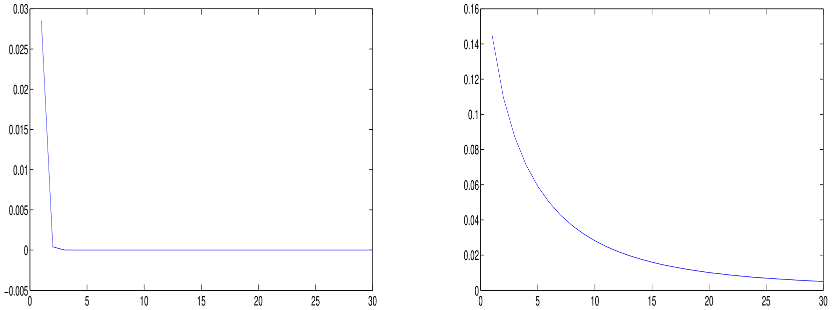

Figure 1: Left pane: average bias. Right pane: Mean Squared Error. In both panes the b_q values are the values on the x axis times 2×10^{-4}. The numbers of observations and repetitions are 1000; the density is Gaussian, and the transformed kernel $T_{\boldsymbol{a}}K$ is of order 2 based on the Epanechnikov family.

The choice of b_q should reflect the trade-off between the free-lunch effect and estimator variance in finite samples. In case of conventional kernels, this trade-off is incorporated in the optimal bandwidth, and the bandwidth choice ends there. Here we discuss two approaches we tried in our simulations: I) in one b_q is proportional to $\alpha_q(K)$ with some scaling coefficient m, that is, $b_q = m\alpha_q(K)$ and II) the other is based on comparison of minimized values of ISE (this was the suggestion of one of the reviewers). After a lot of experimenting we found that in fact Approach I works for $q = 2$, $q = 4$ giving $m = 0.25$, while Approach II is better for $q \geq 6$ giving $m = 0.4$. The following is the summary of our experiments.

Approach I. Comparison of (6) and (11) shows that it makes sense to select

$b_q = m\alpha_q(K_q)$ with some multiplier m. With this idea in mind, we looked at empirical ISE for conventional kernels. It turned out that for small sample sizes (around 100) the theoretical optimal bandwidth was not so optimal. The best bandwidth was about $0.4h_{opt}$. For large sample sizes (around 1000) the empirical ISE was flat in a large interval around the theoretical optimal h. That large interval contained the number $0.4h_{opt}$. Thus, $0.4h_{opt}$ was at least as good as h_{opt} in our simulations for all sample sizes and all conventional kernels. By analogy we set $m = 0.4$ for $T_{\boldsymbol{a}}K$. This choice turned out to be robust with respect to the choice of the estimated density. Unfortunately, estimation results with $T_{\boldsymbol{a}}K$ were strictly better than with conventional kernels only for kernel orders $q = 6, 8, 10, 12$. In cases $q = 2,\ q = 4$ the transformed kernel with $m = 0.4$ was about as good as the conventional ones, and to find a better multiplier we turned to the second approach.

Approach II. Here we explore the idea to choose b_q satisfying

$$\phi(h_{opt}(b_q)) \leq \phi(h_{opt}), \tag{17}$$

see the definitions in the beginning of Section 3.2. Plugging the numbers from (16) and (15), resp., into (17) and canceling out common factors (they depend only on n, q and $f^{(q)}$) we obtain an equivalent condition

$$b_q \left(\boldsymbol{b}'\mathbf{C}_q\boldsymbol{b}\right)^q \leq |\alpha_q(K_q)| \left(\alpha_0(K_q^2)\right)^q. \tag{18}$$

The notation K_q is used to emphasize that K depends on q. Luckily, this condition does not involve the density to be estimated. The left side is a polynomial of b_q of degree $2q + 1$. By Corollary 1 this polynomial is of order $O(b_q)$ in the neighborhood of zero and (18) always holds for all small b_q. However, selecting b_q very small or zero is not an option because of the estimator oversmoothing problem illustrated in Figure 1.

Denote $\max b_q$ the largest positive b_q satisfying (18). Here we consider only $q = 2,\ q = 4$. We tried to see if setting b_q to $\max b_q$ would work. For Gram-Charlier kernels the values of $\max b_q$ were 1 ($q = 2$) and 0.7612 ($q = 4$). For Epanechnikov kernels the respective values of $\max b_q$ were 0.1162 and 0.0067.

Of these numbers, only 0.7612 worked well. Note that for the Gram-Charlier kernel of order 4 one has $\alpha_q(K) = 3$, and the number 0.7612 is approximately $0.25\alpha_q(K)$. This suggests the choice $m = 0.25$. Surprisingly, the multiplier $m = 0.25$ worked well for all kernels considered in this paragraph (Gram-Charlier and Epanechnikov of orders $q = 2,\ q = 4$), which ended our multiplier selection.

Just as a side remark, we explain why the second approach could not be used for all q.

1) The values of $\max b_q$ behave too irregularly to be useful.

2) Another difficulty is that the polynomial $b_q\left(\boldsymbol{b}'\mathbf{C}_q\boldsymbol{b}\right)^q$ may not be strictly monotone in the interval between zero and the upper bound. For instance, for Gram-Charlier kernels of orders 2, 6, 10 the derivative of the said polynomial has positive (sometimes double) roots, while the remaining kernels (of orders 4, 8, 12) are monotone, see Figure 2. Picking an arbitrary b_q between zero and $\max b_q$ may not provide the right balance between bias and variance. In simulations we found excellent choices for all kernel orders but could not formulate a general rule for selecting on "optimal" b_q.

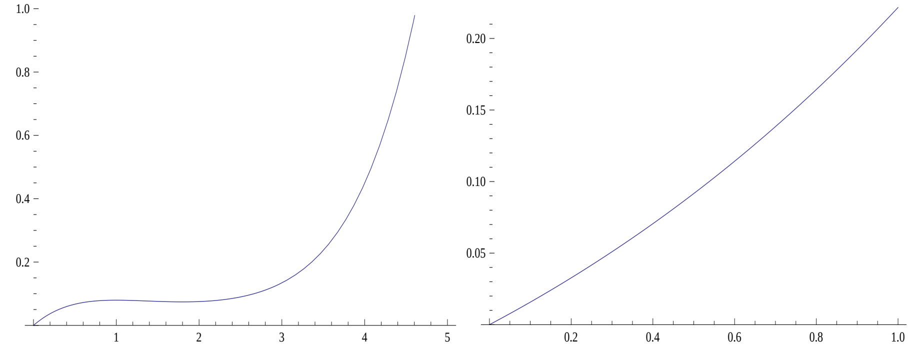

Figure 2: Graphs of h_{opt} as a function of b_q. Left pane: $q = 2$. Right pane: $q = 4$

Summarizing, the "best" values of b_q are $b_q = m\alpha_q(K_q)$ where the multiplier m is 0.25 for $q = 2,\ q = 4$ and $m = 0.4$ for the remaining orders. The results reported in the next section are based on these values. After having made these choices we tried them on the kernels proposed by [Fan & Hu, 1992]. They worked even better in that case (the statistics are not reported to preserve space). Fan and Hu kernels are similar to Gram-Charlier in the sense that both families are based on the Gaussian density $\phi(x)$. To show how different they are, we are giving the expression for the Fan and Hu kernel of order $q = 8$: $\phi(x)(40320 - 282240x^2 + 352800x^4 - 147840x^6 + 26145x^8 - 2121x^{10} + 77x^{12} - x^{14})/5040$ (the corresponding Gram-Charlier kernel is obtained by multiplying the polynomial given in Section 3.1 by $\phi(x)$).

Finally, we compared all four families of kernels: our kernels are the best, if the multiplier is chosen as indicated, Epanechnikov is the second best, followed by Gram-Charlier, which is followed by Fan and Hu. However, our simulations do not guarantee that our multiplier choice will deliver improvement over any other kernel family.

3.3. Estimation results

Let us say we want to estimate the trimodal density. With the chosen sample size, we estimate it twice: once using the conventional kernel and then using its rival $T_{\boldsymbol{a}}K$ (of the same order and based on the kernel from the same family; the multiplier value is either 0.25 or 0.4). This is repeated 1000 times and the Mean Squared Error (MSE) for the transformed kernel is divided by the MSE for the conventional kernel, to see the percentage gain/loss. The results are reported in Table 2.

It is evident that the relative performance of the proposed kernels improves as the sample size and kernel order grow. The improvement ranges from 5% to 30% in the lower right corner of each subtable (where $n = 1000$ and $q = 12$). The improvement over Gram-Charlier kernels is much larger than over Epanechnikov ones. The overall conclusion is that the proposed method delivers better estimation, at least for the densities considered.

Family	n	$q = 2$	$q = 4$	$q = 6$	$q = 8$	$q = 10$	$q = 12$
Gaussian density							
Epanechnikov	$n = 100$	0.883	0.944	0.932	0.882	0.971	0.933
	$n = 500$	0.917	0.956	0.881	0.869	0.966	0.914
	$n = 1000$	0.927	0.965	0.950	0.862	0.953	0.951
Gram-Charlier	$n = 100$	0.979	0.891	0.961	0.690	0.798	0.614
	$n = 500$	0.897	0.868	0.998	0.819	0.807	0.634
	$n = 1000$	0.858	0.904	0.955	0.895	0.846	0.635
Bimodal density							
Epanechnikov	$n = 100$	0.983	0.964	0.938	0.834	0.927	0.942
	$n = 500$	0.974	0.941	0.945	0.887	0.970	0.932
	$n = 1000$	0.939	0.951	0.966	0.855	0.905	0.942
Gram-Charlier	$n = 100$	0.960	0.926	0.957	0.675	0.783	0.627
	$n = 500$	0.991	0.901	0.945	0.802	0.780	0.618
	$n = 1000$	0.920	0.899	0.947	0.887	0.829	0.647
Separated bimodal density							
Epanechnikov	$n = 100$	0.921	0.979	0.980	0.866	0.956	0.993
	$n = 500$	0.923	0.9730	0.970	0.831	0.960	0.910
	$n = 1000$	0.905	0.987	0.938	0.851	0.951	0.962
Gram-Charlier	$n = 100$	0.928	0.925	0.993	0.651	0.818	0.625
	$n = 500$	0.847	0.953	0.974	0.729	0.809	0.668
	$n = 1000$	0.907	0.906	0.966	0.776	0.803	0.670
Trimodal density							
Epanechnikov	$n = 100$	0.945	0.937	0.958	0.822	1.006	1.028
	$n = 500$	0.944	0.974	0.946	0.878	0.935	0.948
	$n = 1000$	0.968	0.960	0.917	0.888	0.965	0.967
Gram-Charlier	$n = 100$	0.886	0.879	0.937	0.678	0.869	0.661
	$n = 500$	0.923	0.893	0.974	0.659	0.828	0.667
	$n = 1000$	0.959	0.880	0.952	0.675	0.815	0.666

Table 2: MSE ratios for estimation with sample sizes $n = 100, 500, 1000$ for two kernel families and four densities; the number of iterations is 1000 everywhere

4. Proofs

By $c_1, c_2, ...$ we denote various positive constants whose precise value does not matter.

Lemma 1. *Under condition* (2)

$$\limsup_{j\to\infty} \left| f^{(j)}(x)/j! \right|^{1/j} = 0, \tag{19}$$

$$f^{(k)}(x+h) = \sum_{i=0}^{\infty} \frac{f^{(i+k)}(x)}{i!} h^i, \ k = 0, 1, 2, ... \tag{20}$$

where all the series converge for any $h \in \mathbb{R}$.

Proof. We start with a simple generalization of inequality (1.4.7) from [Lukacs, 1970]. Let $1 \le i < j < \infty$. By Hölder's inequality with $p = j/i$, $1/q + 1/p = 1$ we have

$$\begin{aligned} \int \left| K(t) t^i \right| dt &\le \left(\int |K(t)|\, dt \right)^{1/q} \left(\int |K(t)|\, |t|^{ip}\, dt \right)^{1/p} \\ &\le \left(\int |K(t)|\, dt \right)^{(j-i)/j} \left(\int \left| K(t) t^j \right| dt \right)^{i/j} \end{aligned}$$

or $[\beta_i(K)]^{1/i} \le [\beta_0(K)]^{1/i-1/j} [\beta_j(K)]^{1/j}$. For i, j in the range under consideration one has $0 < 1/i - 1/j < 1$, so the above inequality implies

$$[\beta_i(K)]^{1/i} \le c_K [\beta_j(K)]^{1/j} \text{ for all } 1 \le i < j < \infty \tag{21}$$

where $c_K = \max\{1, \beta_0(K)\}$. This bound yields $1 \le c_1 \beta_j(K)^{1/j}$. Using also (2) we see that (19) is true. By the Cauchy-Hadamard theorem then the series $f(x+h) = \sum_{i=0}^{\infty} f^{(i)}(x) h^i / i!$ converges for any $h \in \mathbb{R}$. By the properties of power series all the series (20) converge. □

Lemma 2. *If* $\beta_i(K) + \beta_i(K^{(l)}) < \infty$, *then for* $j = 0, 1, ..., l-1$ *one has* $\sup_{s\in\mathbb{R}} \left| K^{(j)}(s) s^i \right| \le c_1 \left[\beta_i(K) + \beta_i(K^{(l)}) \right]$.

Proof. Let $s > 0$. It is well-known that the Sobolev space $W_1^l[0,1]$ is embedded in $C^j[0,1]$ for $j = 0, 1, ..., l-1$, that is, with some constant c_2 independent of K

one has $\left\|K^{(j)}\right\|_{C[0,1]} \leq c_2 \int_0^1 \left[|K(t)| + |K^{(l)}(t)|\right] dt$. Applying this bound to the segment $[s, s+1]$ and using the fact that $|t/s|^i \geq 1$ for $t \in [s, s+1]$ we obtain

$$
\begin{aligned}
\left|K^{(j)}(s)\right| &\leq c_2 \int_s^{s+1} \left[|K(t)| + |K^{(l)}(t)|\right] dt \\
&\leq \frac{c_2}{|s|^i} \int_s^{s+1} \left[|K(t)| + |K^{(l)}(t)|\right] |t|^i \, dt \leq \frac{c_2}{|s|^i} \left[\beta_i(K) + \beta_i(K^{(l)})\right].
\end{aligned}
$$

The case $s < 0$ is treated similarly. This proves the lemma. □

Lemma 3. *If condition* (2) *holds, we have the representation* $\int_{\mathbb{R}} K(-s) f^{(l)}(x + sh) ds = \sum_{i=0}^{\infty} f^{(i+l)}(x) \alpha_i(K)(-h)^i / i!$.

Proof. Substituting $f^{(l)}(x+sh)$ from (20) and changing the order of integration and summation produces

$$
\begin{aligned}
\int_{\mathbb{R}} K(-s) f^{(l)}(x + sh) ds &= \int_{\mathbb{R}} K(-s) \sum_{i=0}^{\infty} \frac{f^{(i+l)}(x)}{i!} (sh)^i ds \\
&= \sum_{i=0}^{\infty} \frac{f^{(i+l)}(x)}{i!} \int_{\mathbb{R}} K(-s)(-s)^i ds (-h)^i \\
&= \sum_{i=0}^{\infty} \frac{f^{(i+l)}(x)}{i!} \alpha_i(K)(-h)^i. \qquad (22)
\end{aligned}
$$

The main problem is to prove that here the series can be integrated term-wise. Consider

$$
\left|\frac{f^{(i+l)}(x)}{i!} \beta_i(K)\right|^{1/i} = \left|\frac{f^{(i+l)}(x)}{(i+l)!} \frac{(i+l)!}{i!} \beta_i(K)\right|^{1/i}.
$$

Here $\left|\frac{(i+l)!}{i!}\right|^{\frac{1}{i}} = |(i+l)...(i+1)|^{\frac{1}{i}} \leq c_1$. By (21) $\beta_i(K)^{\frac{1}{i}} \leq c_2 \beta_{i+l+1}(K)^{1/(i+l+1)}$. Hence,

$$
\begin{aligned}
\left|\frac{f^{(i+l)}(x)}{i!} \beta_i(K)\right|^{1/i} &\leq c_3 \left|\frac{f^{(i+l)}(x)}{(i+l)!}\right|^{1/i} \beta_{i+l+1}(K)^{1/(i+l+1)} \\
&= c_3 \left[\left|\frac{f^{(i+l)}(x)}{(i+l)!}\right|^{\frac{i+l}{i}}\right]^{1/(i+l)} \beta_{i+l+1}(K)^{1/(i+l+1)}. (23)
\end{aligned}
$$

From Lemma 1 we know that $\left|f^{(j)}(x)/j!\right|^{1/j} < 1$ for all large j. Since $(i+l)/i > (i+l)/(i+l+1)$, we have for all large i $\left|\frac{f^{(i+l)}(x)}{(i+l)!}\right|^{\frac{i+l}{i}} \leq \left|\frac{f^{(i+l)}(x)}{(i+l)!}\right|^{\frac{i+l}{i+l+1}}$ which together with (23) and (2) implies

$$\left|\frac{f^{(i+l)}(x)}{i!}\beta_i(K)\right|^{1/i} \leq c_3 \left|\frac{f^{(i+l)}(x)}{(i+l)!}\beta_{i+l+1}(K)\right|^{1/(i+l+1)} \to 0,\ i \to \infty.$$

By the Cauchy-Hadamard theorem therefore the series $\sum |f^{(i+l)}(x)|\beta_i(K)h^i/i!$ converges for any h. This means that $\int_{\mathbb{R}} \sum_{i=0}^{\infty} |f^{(i+l)}(x)| \left|K(-s)s^i\right| /i!ds\, |h|^i < \infty,\ h \in \mathbb{R}$. Attaching a unit mass to each $i = 0, 1, ...,$ by the Fubini theorem we see that in (22) the order of integration and summation can be changed:

$$\int_{\mathbb{R}} \sum_{i=0}^{\infty} \frac{f^{(i+l)}(x)}{i!} K(-s)s^i h^i ds = \sum_{i=0}^{\infty} \frac{f^{(i+l)}(x)}{i!} \int_{\mathbb{R}} K(-s)s^i dsh^i.$$

□

Proof of Theorem 1. **Step 1**. To justify integration by parts below, we start with the proof that

$$\lim_{s\to\infty} K^{(j)}(-s)f^{(l-1-j)}(x+sh) = \lim_{s\to-\infty} K^{(j)}(-s)f^{(l-1-j)}(x+sh) = 0 \quad (24)$$

for any $h > 0$ and $j = 0, ..., l-1,\ l \geq 1$ (if $l = 0$, no integration by parts is needed). Consider the case $s \to \infty$ (the case $s \to -\infty$ is similar). By Lemmas 1 (take $k = l-1-j$) and 2 (select $i+1$ in place of i)

$$\begin{aligned}\left|K^{(j)}(-s)f^{(l-1-j)}(x+sh)\right| &= \left|\sum_{i=0}^{\infty} \frac{f^{(i+l-1-j)}(x)}{i!} K^{(j)}(-s)s^i h^i\right| \\ &\leq \frac{c_1}{s} \sum_{i=0}^{\infty} \left|\frac{f^{(i+l-1-j)}(x)}{i!}\right| h^i \left[\beta_{i+1}(K) + \beta_{i+1}(K^{(l)})\right]. \quad (25)\end{aligned}$$

The series on the right converges for all $h > 0$. As an example, we show this just for the part of the series that contains $K^{(l)}$ using (2).

Case $j = l-1$. In this case (2) is directly applicable.

Case $j < l-1$. Denote $m = i+l-1-j \geq i+1$. By (21) $\beta_{i+1}(K^{(l)})^{1/(i+1)} \leq$

$c_1\beta_{m+1}(K^{(l)})^{1/(m+1)}$. Using this bound we obtain

$$\begin{aligned}
&\left|\frac{f^{(m)}(x)}{i!}\beta_{i+1}(K^{(l)})\right|^{1/i} = \left|\frac{f^{(m)}(x)}{m!}\frac{m!}{i!}\beta_{i+1}(K^{(l)})\right|^{1/i} \\
&\le c_2\left|\frac{f^{(m)}(x)}{m!}\left[\beta_{m+1}(K^{(l)})\right]^{\frac{i+1}{m+1}}\right|^{1/i} = c_2\left[\left|\frac{f^{(m)}(x)}{m!}\right|^{\frac{m+1}{i+1}}\beta_{m+1}(K^{(l)})\right]^{\frac{i+1}{m+1}\frac{1}{i}} \\
&\le c_2\left[\left|\frac{f^{(m)}(x)}{m!}\right|\beta_{m+1}(K^{(l)})\right]^{\frac{1}{m}\frac{m(i+1)}{i(m+1)}} \to 0,\ i\to\infty.
\end{aligned}$$

In the last transition we used the facts that $|f^{(m)}(x)|/m! < 1$ for all large i and $(m+1)/(i+1) > 1$ (as in the proof of Lemma 3).

Thus, by the Cauchy-Hadamard theorem (25) converges and (24) obtains upon letting $s\to\infty$.

Step 2. (24) allows us to integrate l times by parts:

$$\begin{aligned}
&Ef_h^{(l)}(x,K) = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{h^{l+1}}\int_{\mathbb{R}}K^{(l)}\left(\frac{x-t}{h}\right)f(t)dt \\
&= \frac{1}{h^{l+1}}\int_{\mathbb{R}}K^{(l)}\left(\frac{x-t}{h}\right)f(t)dt = \frac{1}{h^{l}}\int_{\mathbb{R}}K^{(l)}\left(-s\right)f(x+sh)ds \\
&= -\frac{1}{h^{l}}K^{(l-1)}\left(-s\right)f(x+sh)\left.\right|_{s=-\infty}^{s=\infty} + \frac{1}{h^{l-1}}\int_{\mathbb{R}}K^{(l-1)}\left(-s\right)f'(x+sh)ds \\
&= ... = -\frac{1}{h}K\left(-s\right)f^{(l-1)}(x+sh)\left.\right|_{s=-\infty}^{s=\infty} + \int_{\mathbb{R}}K(-s)f^{(l)}(x+sh)ds. \quad (26)
\end{aligned}$$

(4) follows from this equation and Lemma 3.

Step 3. Next we evaluate the variance. Since X_j are i.i.d. we have

$$\begin{aligned}
&var\left(f_h^{(l)}(x,K)\right) = \frac{1}{nh^{2l+2}}var\left(K^{(l)}\left(\frac{x-X_1}{h}\right)\right) \\
&= \frac{1}{nh^{2l+2}}\left\{E\left[K^{(l)}\left(\frac{x-X_1}{h}\right)\right]^2 - \left[EK^{(l)}\left(\frac{x-X_1}{h}\right)\right]^2\right\} \\
&= \frac{1}{nh^{2l+2}}\left\{\int_{\mathbb{R}}M\left(\frac{x-t}{h}\right)f(t)dt - \left[\int_{\mathbb{R}}K^{(l)}\left(\frac{x-t}{h}\right)f(t)dt\right]^2\right\}. \quad (27)
\end{aligned}$$

From (26) and (4)

$$\int_{\mathbb{R}}K^{(l)}\left(\frac{x-t}{h}\right)f(t)dt = h^{l+1}Ef_h^{(l)}(x,K) = h^{l+1}\sum_{i=0}^{\infty}\frac{f^{(i+l)}(x)}{i!}(-h)^i\alpha_i(K). \quad (28)$$

Conditions (3) and (2) imply an analog of (2) for the kernel M :

$$\left|\frac{f^{(j)}(x)}{j!}\beta_{j+1}(M)\right|^{1/j} \leq \left|\frac{f^{(j)}(x)}{j!}\left\|K^{(l)}\right\|_{C(\mathbb{R})}\beta_{j+1}(K^{(l)})\right|^{1/j} \to 0,$$

so Lemma 3 applies to M in place of K (with $l = 0$). Hence,

$$\frac{1}{h}\int_{\mathbb{R}} M\left(\frac{x-t}{h}\right)f(t)dt = \int_{\mathbb{R}} M(-s)f(x+sh)ds = \sum_{i=0}^{\infty}\frac{f^{(i)}(x)}{i!}\alpha_i(M)(-h)^i. \tag{29}$$

(27), (28) and (29) lead to (5). Equations (6), (7) follow from (4), (5).

Step 4. The optimal bandwidth has been derived in Section 3.2. □

Proof of Theorem 2. To apply Theorem 1, we check that the kernel $T_{\boldsymbol{a}}K$ satisfies its conditions with $l = 0$. Using $\beta_{j+1}(T_{\boldsymbol{a}}K) \leq \sum_{i=0}^{q}|a_i|\,\beta_{i+j+1}(K)$ and $\left(\sum|b_i|^p\right)^{1/p} \leq \sum|b_i|$ we obtain

$$\begin{aligned}
&\left|\frac{f^{(j)}(x)}{j!}\beta_{j+1}(T_{\boldsymbol{a}}K)\right|^{1/j} \leq \left|\frac{f^{(j)}(x)}{j!}\sum_{i=0}^{q}|a_i|\,\beta_{i+j+1}(K)\right|^{1/j} \\
&= \left|\sum_{i=0}^{q}\left[\left|\frac{f_j(x)}{j!}a_i\right|\beta_{i+j+1}(K)\right]^{j/j}\right|^{1/j} \leq \sum_{i=0}^{q}\left[\left|\frac{f_j(x)}{j!}a_i\right|\beta_{i+j+1}(K)\right]^{1/j} \\
&\leq c_1\max_{i=0,\dots,q}\left|\frac{f_j(x)}{j!}\beta_{i+j+1}(K)\right|^{1/j} \leq c_2\max_{i=0,\dots,q}\left|\frac{f^{(j)}(x)}{j!}\left[\beta_{q+j+1}(K)\right]^{\frac{i+j+1}{q+j+1}}\right|^{1/j} \\
&= c_2\max_{i=0,\dots,q}\left[\left|\frac{f^{(j)}(x)}{j!}\right|^{\frac{q+j+1}{i+j+1}}\beta_{q+j+1}(K)\right]^{\frac{i+j+1}{q+j+1}\frac{1}{j}} \\
&\leq c_2\max_{i=0,\dots,q}\left[\frac{f^{(j)}(x)}{j!}\beta_{q+j+1}(K)\right]^{\frac{i+j+1}{q+j+1}\frac{1}{j}} \to 0,\ j\to\infty.
\end{aligned}$$

Here we have used (21), (9), (19) and the fact that $(q+j+1)/(i+j+1) \geq 1$ for $i = 0, ..., q$. Thus, $T_{\boldsymbol{a}}K$ satisfies (2), Theorem 1 is applicable and, in particular, all the series involved converge.

The definitions of $\boldsymbol{a}$, $T_{\boldsymbol{a}}K$, $\mathbf{B}_q$ and $\mathbf{C}_q$ imply

$$\begin{aligned}
\mathbf{A}_q(K)\boldsymbol{a} &= \begin{pmatrix} \sum a_i\alpha_i(K) \\ \dots \\ \sum a_i\alpha_{i+q}(K) \end{pmatrix} = \begin{pmatrix} \alpha_0(T_{\boldsymbol{a}}K) \\ \dots \\ \alpha_q(T_{\boldsymbol{a}}K) \end{pmatrix} = \boldsymbol{b}, \\
\alpha_0\left((T_{\boldsymbol{a}}K)^2\right) &= \int (T_{\boldsymbol{a}}K)^2(s)ds = \sum_{i,j=0}^{q} a_i a_j \int K^2(t)t^{i+j}dt \\
&= \sum_{i,j=0}^{q} a_i a_j \alpha_{i+j}(K^2) = \boldsymbol{a}'\mathbf{B}_q\boldsymbol{a} = \boldsymbol{b}'\mathbf{C}_q\boldsymbol{b}.
\end{aligned}$$

These equations, (6) and (7) give (11) and (12).

The system $\phi_j(t) = K(t)t^j$, $j = 0, ..., q$, is linearly independent because the equation $\sum c_i\phi_i(t) = 0$ almost everywhere would imply $\sum c_i t^i = 0$ on the set $\{t : K(t) \neq 0\}$ of positive measure. $\mathbf{B}_q$ is the Gram matrix of this system:

$$\mathbf{B}_q = \begin{pmatrix} \int \phi_0^2(t)dt & \dots & \int \phi_0(t)\phi_q(t)dt \\ \dots & \dots & \dots \\ \int \phi_q(t)\phi_0(t)dt & \dots & \int \phi_q^2(t)dt \end{pmatrix}.$$

Linear independence of ϕ_j implies positive definiteness of $\mathbf{B}_q$ and $\boldsymbol{b}'\mathbf{C}_q\boldsymbol{b} > 0$, see [Gantmacher, 1959]. The final remark about the terms of higher order in h is warranted by Theorem 1. The proof is complete. □

Acknowledgements

The authors thank two anonymous referees for their thoughtful suggestions.

References

Berlinet, A., 1993. Hierarchies of higher order kernels. Probab. Theory Related Fields. 94, 489–504.

Berlinet, A., Thomas-Agnan, Ch., 2004. Reproducing Kernel Hilbert spaces in Probability and Statistics. Kluwer Academic Publishers, Boston, MA.

Deheuvels, P., 1977. Estimation non parametrique de la densite par histogrammes generalises. Rev. Statist. Appl. 25, 5–42.

Devroye, L., 1987. A Course in Density Estimation. Progress in Probability and Statistics, 14. Birkhäuser Boston Inc., Boston, MA.

Fan, J., Hu, T. Ch., 1992. Bias correction and higher order kernel functions. Statist. Probab. Lett. 13, 235–243.

Gantmacher, F. R., 1959. The Theory of Matrices. Vols. 1, 2. Chelsea Publishing Co, New York.

Gasser, Th., Muller, H. G., 1979. Kernel estimation of regression functions, in: Gasser, Th., Rosenblatt, M. (Eds.), Smoothing Techniques for Curve Estimation. Springer, Berlin, pp. 23–68.

Lejeune, M., Sarda, P., 1992. Smooth estimators of distribution and density functions. Comput. Statist. Data Anal. 14, 457–471.

Lukacs, E., 1970. Characteristic Functions, second ed. Griffin.

Marron, J. S., Wand, M. P., 1992. Exact mean integrated squared error. Ann. Stat. 20, 712–736.

Mynbaev, K. T., Martins Filho, C., 2010. Bias reduction in kernel density estimation via Lipschitz conditions. J. Nonparametr. Statist. 22, 219–235.

Scott, D. W., 1992. Multivariate Density Estimation: Theory, Practice, and Visualization. John Wiley & Sons.

Silverman, B. W., 1986. Density Estimation for Statistics and Data Analysis. Chapman & Hall, London.

Wand, M. P.; Jones, M. C., 1995. Kernel Smoothing. Monographs on Statistics and Applied Probability, 60. Chapman and Hall, London.

Wand, M. P., Schucany, W. R., 1990. Gaussian-based kernels. Canad. J. Statist. 18, 197–204.

Withers, C. S., Nadarajah, S., 2013. Density estimates of low bias. Metrika, 76, 357–379.