Robust kernel estimator for densities of unknown smoothness

By Yulia Kotlyarova* and Victoria Zinde-Walsh†

April 30, 2007

*Department of Economics, Dalhousie University, Halifax, Nova Scotia B3H3J5 Canada; ph. (902)494-8824; fax (902)494-6917; e-mail: yulia.kotlyarova@dal.ca

†Department of Economics, McGill University and CIREQ, 855 Sherbrooke Street West, Montreal, Quebec H3A2T7 Canada; ph. (514)398-4834; fax (514)398-4938; e-mail: victoria.zinde-walsh@mcgill.ca

Results on nonparametric kernel estimators of density differ according to the assumed degree of density smoothness. A kernel/bandwidth pair that was optimal for a twice differentiable function may not be suitable when the density is piecewise linear. If there is uncertainty about the degree of smoothness, an inappropriate choice may lead to under- or oversmoothing. To examine various possible outcomes we provide asymptotic results on kernel estimation of a continuous density for an arbitrary bandwidth/kernel pair and derive the limit joint distribution of kernel density estimators corresponding to different bandwidths and kernel functions. Using these results, we propose a combined estimator constructed as an optimal linear combination of several estimators with different bandwidth/kernel pairs. Its theoretical properties (Kotlyarova and Zinde-Walsh 2006) are such that it automatically attains the best possible rate without a priori knowledge of the degree of smoothness. Our Monte Carlo results confirm the advantages of the combined estimator of density.

Keywords: Kernel density estimation; Bandwidth selection; Combined estimator

2000 Mathematics Subject Classifications: 62G07; 62G20; 62G35

1 Introduction

Investigation of the asymptotic and finite-sample behaviour of kernel density estimators focused largely on the search for appropriate values of the bandwidth, assuming that the underlying model was sufficiently smooth. While it enabled researchers to obtain very precise expressions for the optimal bandwidth, it undermined the primary characteristic feature of non-parametric estimators, their robustness, by restricting density to belong to a class of smooth functions. If second order or higher order derivatives of the

density exist, a bandwidth that ensures an optimal convergence rate can be found for a kernel of sufficiently high order. If, however, there is no certainty that the smoothness assumptions hold, under- and especially oversmoothing are likely. Oversmoothing produces heavily-biased estimators; it occurs when the bandwidth is too large and too many irrelevant observations are used to determine the density at a particular point, which leads to elimination of peaks and troughs. Undersmoothing increases the mean squared error (MSE) as the estimate becomes very volatile. If there are no grounds on which to assume smoothness of the density, the chosen rate for the bandwidth may be in error and the estimator will suffer from the problems associated with under- or oversmoothing.

In this paper we consider the asymptotic properties of kernel estimators for a continuous (but not necessarily differentiable) density based on different bandwidth/kernel pairs and investigate ways of improving efficiency that do not rely on smoothness assumptions. Because of the nonparametric rates of convergence, each bandwidth/kernel pair may provide additional information. Similarly to the joint distribution of smoothed least median of squares estimators (Zinde-Walsh 2002) and smoothed maximum score estimators (Kotlyarova and Zinde-Walsh 2004), we derive the joint limit distribution for kernel density estimators. The result demonstrates that some estimators of density at a point may be asymptotically independent, thus a linear combination of several such estimators may improve the accuracy relative to each individual estimator.

Kotlyarova and Zinde-Walsh [KZW] (2006) showed how a linear combination of semi-parametric or non-parametric estimators can protect against negative consequences of errors in assumptions about the order of smoothness and possible oversmoothing by automatically attaining the best rate that would have been possible had we known a priori the optimal bandwidth/kernel pair in the set. The weights in the linear combination are selected to minimize an estimate of the mean squared error; the resulting estimator is what we call a "combined estimator". We demonstrate here that the conditions un-

der which the asymptotic advantages of the combined estimator hold are satisfied for our kernel density estimators. Mostly, the use of combined estimators in the literature is restricted to convex combinations (see, e.g. Fan and Ullah (1999) where a convex combination of a parametric and a non-parametric regression estimators is employed to protect against misspecification of regression functions). Here we do not impose such restrictions allowing for trade-off between biases in addition to variance reduction.

The results of a Monte Carlo experiment confirm the usefulness of the combined estimator in finite samples. We demonstrate that combined estimators perform as well or even better than the best individual cross-validated estimator. This is an important result since the best individual kernel is not the same for the three models considered in our study. For the standard normal and the mixture of normal densities, the most accurate individual estimator is based on the fourth-order kernel and is significantly better than the estimator with the second-order kernel. As for the non-smooth, piecewise linear density, the second-order kernel yields more precise results than the fourth-order kernel. Thus, the use of the combined estimators protects against losses in accuracy caused by the absence of information about the properties of the density. Moreover, the combined estimator is less sensitive to the choice of smoothing functions.

The paper is organized as follows. Section 2 contains the definitions, assumptions and known results for the kernel density estimator. Section 3 provides asymptotic results under weak (only continuity, no smoothness) assumptions for the kernel density estimator, as well as for the joint limit process for several estimators. The new combined estimator is defined in Section 4, where we also discuss how to compute it (selection of bandwidths, smoothing kernels, estimation of the MSE of a linear combination). Performance of combined estimators is evaluated in a Monte Carlo experiment in Section 5. Appendix provides the proof of Theorem 2 in Section 3.

2 Definitions, assumptions, known results

Consider a univariate random variable X and the corresponding density function $f()$. We are interested in estimating the value of the density function at x.

Assumption 1.

(a) (X_i), $i = 1, ..., n$, is a random sample of X;

(b) the density function $f(x)$ exists and is continuous at x.

To estimate the density we utilize kernel functions but do not restrict kernels to be symmetric or nonnegative density functions; as will be clear later, this may give us some extra flexibility.

Assumption 2.

(a) The kernel smoothing function K is a continuous real-valued function;

(b) $\int K(z)dz = 1$;

(c) (Parzen 1962) $\int |K(z)|dz < \infty$, $|z||K(z)| \to 0$ as $|z| \to \infty$, $\sup |K(z)| < \infty$;

$\int K(z)^{2+\delta}dz < \infty$ for some $\delta > 0$.

Assumption 3.

(a) The bandwidth parameter $h_n \to 0$ as $n \to \infty$;

(b) $h_n n \to \infty$ as $n \to \infty$.

The kernel density estimator (Rosenblatt 1956, Parzen 1962) is defined as

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

Assumptions 1-3 are sufficient to prove that the kernel density estimator is MSE-consistent and has a normal limiting distribution (Parzen (1962) applies Lyapunov's

central limit theorem for triangular arrays to prove normality):

$$E(\hat{f}(x) - f(x))^2 \to 0 \text{ as } n \to \infty, \tag{2}$$

$$(nh_n)^{\frac{1}{2}} \left(\hat{f}(x) - E\hat{f}(x) \right) \xrightarrow{d} N \left(0, f(x) \int K^2(z)dz \right). \tag{3}$$

Assumption 3a ensures that the estimator is asymptotically unbiased; Assumptions 3b and 2c guarantee that the variance of the estimator will tend to zero.

If the existence of continuous second order derivatives of the density function is assumed then the sharp rate of bandwidth $h_n = cn^{-\frac{1}{5}}$ will be optimal for a second-order kernel and the convergence rate of the density estimator is $n^{-2/5}$ (see Pagan and Ullah (1999) for discussion). If higher order derivatives of density exist, further improvements in efficiency can be obtained by using a higher order kernel[1] to reduce the bias (Cleveland and Loader 1996, Marron and Wand 1992).

The assumption of continuity of the second derivative of the density function can not be easily verified although it is routinely made when determining the optimal bandwidth using Silverman's (1986) "rule of thumb", or plug-in methods by Park and Marron (1990) and Sheather and Jones (1991). However, the bandwidth selection methods that are based on this assumption may behave very poorly when it is violated (Loader 1999a,b). The data-driven methods of bandwidth selection such as the least squares cross validation (Rudemo 1982, Bowman 1984) and the likelihood cross validation (Duin 1976) do not assume differentiability of the density function and may be asymptotically optimal under weak underlying assumptions (Hall 1983, Stone 1984).[2]

Several Monte Carlo experiments (Park and Turlach 1992, Loader 1999a,b) show that plug-in methods perform well when the density is relatively smooth (and prefer-

[1] Higher order kernel may produce negative values for the density at some points.

[2] When the data set is large, cross validation will take a long time to compute since the computation time is a quadratic function of the sample size; in such situations one could use the recently proposed modified versions such as in Lambert et al (1999).

ably unimodal) but often oversmooth more irregular densities. Cross-validation methods identify very well steep peaks and other irregularities of the density but tend to undersmooth in more conventional settings.

The kernel order also plays an important role, with higher order kernel delivering improved performance in highly smooth cases and lower order being better suited when the density is not sufficiently smooth. Without a priori knowledge of density smoothness there is no clear indication of how to select the order of kernel and the corresponding bandwidth.

In this paper we consider the combined estimator designed to circumvent the choice of bandwidth and kernel problem. In the next section we develop the asymptotic results that prove that the assumptions in KZW (2006) are satisfied and thus the combined estimator will automatically deliver the optimal rate.

3 Asymptotic properties of kernel estimators

3.1 Distribution of a single univariate density estimator when density is continuous

We define the bias of the kernel density estimator

$$B(h_n, K, x) = E(\hat{f}(x) - f(x)) = \int K(z)\,[f(x + zh_n) - f(x)]\,dz. \tag{4}$$

Under Assumption 3a, $B(h_n, K, x)$ converges to 0. Under more stringent differentiability assumptions, a sharp rate for $B(h_n, K, x)$ could be determined but we do not make such assumptions. To simplify notation, the subscript n will be omitted in h_n.

The following theorem is a corollary of Parzen's results (2) and (3).

Theorem 1. *Under Assumptions 1 - 3, if h is such that as $n \to \infty$*

(a) $n^{1/2}h^{1/2}B(h,K,x) \to 0$

then $n^{1/2}h^{1/2}(\hat{f}(x) - f(x)) \xrightarrow{d} N(0, f(x)\int K^2(w)dw)$;

(b) $n^{1/2}h^{1/2}B(h,K,x) \to B(K)$, *where* $0 < |B(K)| < \infty$,

then $n^{1/2}h^{1/2}(\hat{f}(x) - f(x)) \xrightarrow{d} N(B(K), f(x)\int K^2(w)dw)$;

(c) $n^{1/2}h^{1/2}\,|B(h,K,x)| \to \infty$

then $|B(h,K,x)|^{-1}\left[\hat{f}(x) - f(x) - B(h,K,x)\right] = o_p(1)$.

Thus, for case (a) (undersmoothing) we obtain a limiting normal distribution, a limit normal with a non-zero mean for (b), and in case (c) the bias dominates. Without making assumptions about the degree of smoothness of density all that is known is that for some rate of $h \to 0$ there is undersmoothing and an unbiased limiting Gaussian distribution, and for some slower convergence rate of h there is oversmoothing. Existence of an optimal rate depends on convergence properties of $B(h,K,x)$ that cannot be asserted without strengthening the assumptions.

3.2 The joint limit process for univariate density estimators of continuous densities

Assume that $\hat{f}(h,K,x)$ represents the estimator when the function K and bandwidth h are utilized. Consider a number of bandwidths h: $\{h_i\}_{i=1}^m$. Assume that h_i for $i \le m'$ corresponds to undersmoothing (part (a) of Theorem 1) while h_i for i such that $m' \le m'' < i \le m$ corresponds to oversmoothing (part (c) of Theorem 1). If an optimal rate exists then one could have $m'' \ge m' + 1$ and h_i for $i = m' + 1, ..., m''$ corresponding to the optimal rate. For example, for an s times continuously differentiable density and using some s-order kernel, the optimal bandwidth is $O(n^{-\frac{1}{2s+1}})$ (see, e.g., Pagan and Ullah (1999), p. 30).

We combine each h_i with each smoothing function K_j from some set of functions that satisfy Assumption 2, $j = 1, ..., l$. Define

$$\eta(h_i, K_j) = \begin{cases} n^{1/2} h_i^{1/2} (\hat{f}(h_i, K_j, x) - f(x)) \text{ for } i = 1, ..., m', \\ n^{1/2} h_i^{1/2} (\hat{f}(h_i, K_j, x) - f(x) - B(h_i, K_j, x)) \\ \text{for } i = m' + 1, ..., m'', \\ |B(h_i, K_j, x)|^{-1} \left[\hat{f}(h_i, K_j, x) - f(x) - B(h_i, K_j, x) \right] \\ \text{for } i = m'' + 1, ..., m. \end{cases}$$

Theorem 2. *Suppose that Assumptions 1-3 hold for each bandwidth* $h_i, 1 \leq i \leq m$, *and for each kernel* K_j, $1 \leq j \leq l$, *and that the functions* $\{K_j\}_{j=1}^{l}$ *form a linearly independent set*[3].

(a) If each $h_1, ..., h_{m''}$ $(m'' \leq m)$ *satisfies condition (a) or (b) of Theorem 1 then* $\eta_a \equiv (\eta(h_1, K_1)', ..., \eta(h_1, K_l)', ..., \eta(h_{m''}, K_1)', ..., \eta(h_{m''}, K_l)')' \xrightarrow{d} N(0, f(x)\Psi)$, *where the covariance between* $\eta(h_{i_1}, K_{j_1})$ *and* $\eta(h_{i_2}, K_{j_2})$ *is determined by the following element of the* $lm'' \times lm''$ *matrix* Ψ:

$$\Psi_{(i_1-1)l+j_1,(i_2-1)l+j_2} = \begin{cases} \sqrt{q} \int K_{j_1}(w) K_{j_2}(qw) dw \quad \text{if} \quad h_{i_1}/h_{i_2} \to q < \infty, \\ 0 \quad \text{if} \quad h_{i_1}/h_{i_2} \to 0 \quad \text{or} \quad h_{i_1}/h_{i_2} \to \infty; \end{cases}$$

(b) If each $h_{m''+1}, ..., h_m$ $(m'' \leq m)$ *satisfies condition (c) of Theorem 1 then* $(\eta(h_{m''+1}, K_1)', ..., \eta(h_{m''+1}, K_l)', ..., \eta(h_m, K_1)', ..., \eta(h_m, K_l)')' \xrightarrow{p} 0$;

(c) $Cov(\eta(h_{i_1}, K_{j_1}), \eta(h_{i_2}, K_{j_2})) \to 0$ *for* $1 \leq i_1 \leq m''$ *and* $m'' + 1 \leq i_2 \leq m$, *and any* j_1, j_2.

The proof is provided in the Appendix. Theorems 1 and 2 can be easily extended to the case of multivariate density functions at the cost of more notational complexity.

Thus, if the bandwidths approach 0 at different rates or $\int K_{j_1}(w) K_{j_2}(w) dw = 0$, the corresponding estimators $\hat{f}(h_{i_1}, K_{j_1}, x)$ and $\hat{f}(h_{i_2}, K_{j_2}, x)$ are asymptotically independent. This is a consequence of the fact that only a small fraction of observations have any effect on the estimator, therefore reweighting observations with different kernel functions can produce estimators with independent limit processes.

Theorems 1 and 2 of this section correspond to Assumptions 1 and 2 in KZW(2006)

[3]If some linear combination of smoothing kernels K_j is zero then the joint distribution at each bandwidth is degenerate.

where it is shown that once these assumptions hold there will be gains from combining kernel density estimators.

4 The combined estimator

In this section we define the combined estimator for density and discuss some of the specifics of constructing it.

4.1 Definition of the combined estimator

Suppose that bandwidths $h_1 < h_2 < ... < h_m$ correspond to various convergence rates, where h_1 corresponds to undersmoothing and h_m to oversmoothing; the optimal rate may or may not exist. For a set of smoothing functions $K_1, ..., K_l$, Theorem 2 indicates the structure of the joint limit distribution of $\hat{f}(h_i, K_j, x)$.

Construct a linear combination $\hat{f}(a) = \sum_{i=1}^{m}\sum_{j=1}^{l} a_{ij}\hat{f}(h_i, K_j, x)$, $\sum_{i,j} a_{ij} = 1$. Assume that the biases, variances and covariances for all $\hat{f}(h_i, K_j, x)$ are known. Then one could find weights $\{a_{ij}\}$ that minimize the mean squared error $MSE(\hat{f}(a))$ and provide an optimal estimator:

$$MSE(\hat{f}(a)) = \sum_{i_1,j_1,i_2,j_2} a_{i_1 j_1} a_{i_2 j_2}\{B(h_{i_1}, K_{j_1}, x)B(h_{i_2}, K_{j_2}, x)$$
$$+Cov(\hat{f}(h_{i_1}, K_{j_1}, x), \hat{f}(h_{i_2}, K_{j_2}, x))\}.$$

In KZW (2006) the limit weights are derived and it is shown that the convergence rate for the optimal combination is at least as fast as that for the best individual estimator. In fact, since the weights are not necessarily non-negative there is a possibility of trading off the biases of individual estimators. It should be emphasized that the proposed combined estimator is local and the weights change from point to point, allowing for additional flexibility in fitting the data.

To implement this approach we need to estimate the biases and covariances of all $\hat{f}(h_i, K_j, x)$.

Denote estimated biases and covariances by "hats".

Then, for a univariate density,

$\widehat{MSE}(\hat{f}(a)) = \sum_{i_1,j_1,i_2,j_2} a_{i_1 j_1} a_{i_2 j_2} \{\widehat{B}(h_{i_1}, K_{j_1}, x)\widehat{B}(h_{i_2}, K_{j_2}, x)$

$+\widehat{Cov}(\hat{f}(h_{i_1}, K_{j_1}, x), \hat{f}(h_{i_2}, K_{j_2}, x))\}.$

Define the combined density estimator $\widehat{f_c}$ as a linear combination $\widehat{f_c} = \hat{f}(\widehat{a})$, where

$$\widehat{a} = \arg\min \widehat{MSE}(\hat{f}(a)), \quad \sum_{i,j} a_{ij} = 1. \tag{5}$$

In KZW (2006) it is demonstrated that as long as Assumptions 1 and 2 hold and the covariances and biases are consistently estimated the combined estimator performs similarly to the (infeasible) optimal combination.

4.2 Construction of the combined estimator

4.2.1 Estimation of variances and biases

Consistent estimators for biases and covariances can be obtained by various procedures; we require that these estimators do not rely on information about density smoothness.

Consider first the covariance matrix. For large sample sizes, one can rely on the joint asymptotic distribution (Theorem 2). For the diagonal elements, $Var(\hat{f}(h_i, K_j, x))$, use $\dfrac{\widehat{f(x)} \int K_j(w)^2 dw}{h_i n}$. The estimate of the density, $\widehat{f(x)}$, has to be specified. Since the smallest bandwidth corresponds to an estimator with the smallest bias, the candidates for the estimate are $\hat{f}(h_1, K_j, x)$ or a weighted average of individual estimators evaluated at h_1 using kernels $K_1, .., K_l$. For all off-diagonal elements, covariances $Cov(\hat{f}(h_{i_1}, K_{j_1}, x), \hat{f}(h_{i_2}, K_{j_2}, x))$ can be approximated by $\dfrac{\widehat{f(x)}}{n h_{i_2}} \int K_{j_1}(w) K_{j_2}(\frac{h_{i_1}}{h_{i_2}} w) dw$.

For small sample sizes, it would be more appropriate to apply the bootstrap (see Hall (1992) for a discussion of the bootstrap for nonparametric estimators) for both variances and covariances:

$$\widehat{Cov}(\hat{f}(h_{i_1}, K_{j_1}, x), \hat{f}(h_{i_2}, K_{j_2}, x))$$
$$= (M-1)^{-1} \sum_{s=1}^{M} \left(\hat{f}_s(h_{i_1}, K_{j_1}, x) - M^{-1} \sum_{t=1}^{M} \hat{f}_t(h_{i_1}, K_{j_1}, x) \right)$$
$$\times \left(\hat{f}_s(h_{i_2}, K_{j_2}, x) - M^{-1} \sum_{t=1}^{M} \hat{f}_t(h_{i_2}, K_{j_2}, x) \right),$$

where M is the number of bootstrap replications, while $\hat{f}_s$ and $\hat{f}_t$ denote kernel density estimates based on the sth and tth bootstrapped samples, respectively.

In our Monte Carlo experiment we used the first, asymptotic, method, with $\widehat{f(x)} = l^{-1} \sum_{j=1}^{l} \hat{f}(h_1, K_j, x)$.

Estimation of the bias is more complicated. Without assumptions regarding smoothness of the density function, we do not know the precise convergence rate of the bias. Existing methods of bias correction and approximation (e.g., Schucany and Sommers 1977, Gerard and Schucany 1999) are based on the assumption that the density is several times differentiable. The standard bootstrap procedure is not applicable either: due to the linear structure of kernel density estimators the expected value of the bootstrapped bias is zero (Hall 1992).

In our Monte Carlo study, we will use the fact that the estimators with the smallest bandwidth (undersmoothing) have biases that converge to zero the fastest. To find individual biases, we subtract the average of estimators with the smallest bandwidth from actual estimators[4]: $\widehat{B}(h_i, K_j, x) = \hat{f}(h_i, K_j, x) - l^{-1} \sum_{p=1}^{l} \hat{f}(h_1, K_p, x)$.

4.2.2 Procedure for computing the combined estimator

To determine a set of bandwidths we start with the bandwidth obtained using least squares cross validation (LSCV). The robust method of the LSCV is considered to be appropriate under very weak assumptions (see Li and Racine (2007) for discussion). Several studies (Park and Turlach 1992, Loader 1999a) indicate that this bandwidth

[4]According to Hall (1992, p. 207), another approach to bias estimation is to use an estimate $\widehat{f}$ of the density f in the definition of the bias: $\int K(w)\widehat{f}(x - hw)dw - \hat{f}(x)$. However, the question of choosing an appropriate $\hat{f}(x)$ for this bias estimator remains open.

is usually smaller than the bandwidths obtained by other methods and corresponds to undersmoothing. We experimented with bandwidths smaller than LSCV and our results confirmed generally poorer performance of such bandwidths both in individual and combined estimators. Thus we compute other bandwidths as constant multiples of the LSCV bandwidth determined as $d^{i-1}h_{LSCV}$, where $d > 1$, for $i = 2, ..., m$. After experimenting with a wider range of m we selected $d = 1.5$ and $m = 3$.

Recall that the estimators at the lowest bandwidth are used to compute the biases; this may cause a problem of possible underestimation of the MSE of density estimators with the lowest bandwidth because, when the combined estimator is based on a single kernel, the estimate of the bias for the lowest bandwidth is zero by construction. By experimentation we found that this is indeed the case, therefore we do not include the lowest bandwidth estimators in the combination.

Symmetric kernels are appropriate when dealing with smooth densities, while asymmetric functions may pick up some irregularities of the density that will be discarded by symmetric smoothing functions. There may be some advantage in using mutually orthogonal kernels since they produce asymptotically uncorrelated estimators that provide complementary information.

The entire procedure for a combined estimator includes the following steps: (i) compute the LSCV bandwidth and the other $m - 1$ bandwidths for each kernel; (ii) find the density estimators for all smoothing functions and bandwidths; (iii) estimate the biases and the covariance matrix by the methods described in section 4.2.1; (iv) find the optimal weights for the linear combination of estimators for all kernels and all bandwidths, excluding the lowest, by solving (5) and obtain the combined estimator by using these weights.

5 Performance of the combined estimator

5.1 The DGP and combined estimator of density

We consider three different density functions.

For the first model we use the standard normal distribution: $f_1(x) = \phi(x)$. Its density is infinitely differentiable and very smooth; the density estimator evaluated at the rule-of-thumb bandwidth is the optimal choice.

In the second model we consider the mixture of three normal densities

$f_2(x) = 0.5\phi(x) + 3\phi(10(x - 0.8)) + 2\phi(10(x - 1.2))$

analyzed by Härdle et al (1998). This density is also infinitely differentiable; however, it is trimodal and much more wiggly than the standard normal density. Theoretically, the rate of convergence of the estimator is determined, as for the standard normal distribution, by the order of the smoothing function. The rule of thumb, designed for bell-shaped symmetric functions, will not be optimal in this case, though it should produce an estimator converging at the rate $n^{-2/5}$ for the second order kernel. The more meaningful comparison here is between the estimator with LSCV bandwidth and the combined estimators.

The third model contains a piecewise linear density that satisfies the Lipschitz condition everywhere.

$$f_3(x) = \begin{cases} 5.25 - 5x \text{ if } x \in [0.95, 1.05], \\ 0.5 \text{ if } x \in [0, 0.95), \\ 0.5 + 5x \text{ if } x \in [-0.1, 0), \\ -0.0475 - 0.475x \text{ if } x \in [-1.1; -0.1), \\ 0.9975 + 0.475x \text{ if } x \in [-2.1; -1.1), \\ 0 \text{ otherwise.} \end{cases}$$

The rule-of-thumb bandwidth will converge to zero too slowly. The LSCV bandwidth should perform well but the choice of kernel may affect its performance. The

question is whether the combined estimator could offer an advantage.

The sample sizes considered in the experiments are $n = 500$, 1000, and 2000; 1000 replications per model were performed.

We report results for combined estimators constructed using three bandwidths: $h_{\min}$ = least squares cross validation, $1.5h_{\min}$ and $2.25h_{\min}$.

We utilize the following four kernel functions defined on [-1,1]:

(a) symmetric second-order kernel $K2 = \frac{15}{16}\left(1-x^2\right)^2 I(|x| \leq 1)$;

(b) symmetric fourth-order kernel $K4 = \frac{105}{64}\left(1-x^2\right)^2 (1-3x^2) I(|x| \leq 1)$;

and two orthogonal asymmetric kernels of order three:

(c) $K3a(x) = \frac{105}{64}\left(1-3x^2\right)\left(1+\sqrt{23}x\right)\left(1-x^2\right)^2 I(|x| \leq 1)$ and

(d) $K3b(x) = \frac{105}{64}\left(1-3x^2\right)\left(1-\sqrt{23}x\right)\left(1-x^2\right)^2 I(|x| \leq 1)$.

The two asymmetric kernels may be more appropriate for modelling irregular densities. If the density function is symmetric and more than three times differentiable, theoretical biases for the two functions are opposite in sign and equal in absolute value, and a simple average of these two estimators may produce variance reduction by a factor of 2 and a bias reduction equivalent to using a fourth-order kernel.

We report estimates of integrated MSE for the cross-validated density estimators based on K2 and K4 and the combined estimators based on (a) the second-order function K2; (b) K4 only; (c) both K2 and K4, and (d) both K3a and K3b.

The algorithm for the least squares cross validation follows Silverman (1986). For the combined estimator that uses both $K2$ and $K4$, we construct separate sets of bandwidths for the elements with $K2$ and the elements with $K4$, since the cross-validated bandwidths for these two functions are quite different. In the case of the combined estimator with third-order functions, the lowest bandwidth corresponds to the cross-validated bandwidth for the symmetric function $0.5K3a + 0.5K3b = K4$.

For the standard normal density, MSE's are estimated at 141 points between -3.5 and 3.5. For the mixture of normal densities, we consider 121 points between -3 and 3,

and for the piecewise-linear density 161 points between -2.5 and 1.5 are evaluated.

The combined estimators are constructed as described in Section 4.2.

5.2 Summary of the results

The purpose of this Monte Carlo experiment is to compare the performances of cross-validated and combined estimators. We demonstrate that the cross-validated estimates are quite sensitive to the choice of kernel function and that the combined estimators offer a more reliable way of ensuring robustness and accuracy in density estimation (by the MISE criterion).

In Table 1 we present estimated MISEs for all the models. Table 2 contains relative MISEs: for each combination of density and sample size we report the outcomes normalized by the MISE of the best cross-validated estimator. Thus, for the standard normal and the mixture of normal densities the denominator is the MISE of the cross-validated estimator with the fourth-order kernel, while for the non-smooth case the MISE estimates are normalized by the cross-validated K2.

When the true data-generating process is the standard normal density, all reported estimators are very precise in absolute terms. The combined estimators are uniformly better than the cross-validated K2 estimator, while the combined K4 estimator has about 15% lower MISE than the cross-validated K4 estimator.

For the mixture of normal densities, the second-order cross-validated estimator has 10-20% larger MISE than the corresponding fourth-order estimator. The combined K2 estimator is slightly better than the cross-validated K2, while the remaining combined estimators are as good or better in terms of MISE than the cross-validated K4. In fact, as the sample size increases, the combined K4 estimator is gaining in precision relative to the cross-validated K4. The orthogonal pair of third-order kernels also shows a strong performance.

In the case of the non-smooth density, the second-order cross-validated estimator outperforms the 4th-order kernel, though the difference is decreasing with n. The results show that all combined estimators are no worse than the 2nd-order cross-validated estimator and that their relative performance improves with sample size. At $n = 2000$, the MISE of the combined K2 estimator is 14% less than that of the cross-validated K2 estimator (10% less for the combined K4 and 6-7% less for the other two combined estimators).

The table is representative of a more extensive Monte Carlo experiment with a variety of bandwidth combinations as well as different kernel combinations. On the basis of these experiments we conclude the following:

1. No individual estimator offers a consistently good performance for all cases: the 4th order kernel with CV bandwidth is better in the normal and mixed case, but the 2d order kernel is better in the non-smooth case.

2. The combined estimator always improves on the "incorrect" choice of an individual estimator: if, for example, we use the second-order kernel in the mixed case (where the errors could be substantial), any of the combined estimators reduces the error.

3. The combined estimator often improves on the "best" individual estimator: practically always in the non-smooth case and for all combinations except the one using only the second-order kernel in the mixed case.

In the absence of knowledge of smoothness, the combined-estimator approach offers a robust way of obtaining the estimates as good or better than the best (unknown) individual cross-validated estimator. We recommend using the combined estimators based on higher-order kernels.

Acknowledgements

The support of the Social Sciences and Humanities Research Council of Canada (SSHRC), the *Fonds québécois de la recherche sur la société et la culture* (FQRSC) is gratefully acknowledged.

Appendix: Proof of Theorem 2

To prove Theorem 2 we need to consider covariances between the $\eta(h, K)$.

For (a), consider first estimators that satisfy condition (a) of Theorem 1. Recall from Theorem 1 (a) that $E\eta(h_i, K_j) \to 0$, therefore the covariance matrix is determined by the value of $E(\eta(h_{i_1}, K_{j_1})\eta(h_{i_2}, K_{j_2}))$.

Since x_s is independent of x_t as long as $s \neq t$, their functions $K_{j_1}(\frac{X_s - x}{h_{i_1}})$ and $K_{j_2}(\frac{X_t - x}{h_{i_2}})$ are also independent. We have that

$$
\begin{aligned}
&E(\eta(h_{i_1}, K_{j_1})\eta(h_{i_2}, K_{j_2})) \\
&= n\,(h_{i_1}h_{i_2})^{\frac{1}{2}}\, E\left[\left(\tfrac{1}{nh_{i_1}}\textstyle\sum K_{j_1}(\tfrac{X_s - x}{h_{i_1}}) - f(x)\right)\left(\tfrac{1}{nh_{i2}}\textstyle\sum K_{j_2}(\tfrac{X_t - x}{h_{i_2}}) - f(x)\right)\right] \\
&= n\,(h_{i_1}h_{i_2})^{\frac{1}{2}}\, \tfrac{1}{n^2 h_{i_1}h_{i_2}}\textstyle\sum EK_{j_1}(\tfrac{X_i - x}{h_{i_1}})K_{j_2}(\tfrac{X_i - x}{h_{i_2}}) \\
&+ n\,(h_{i_1}h_{i_2})^{\frac{1}{2}}\left[\left(\tfrac{1}{nh_{i_1}}\textstyle\sum EK_{j_1}(\tfrac{X_s - x}{h_{i_1}}) - f(x)\right)\left(\tfrac{1}{nh_{i2}}\textstyle\sum EK_{j_2}(\tfrac{X_t - x}{h_{i_2}}) - f(x)\right)\right] \\
&- n^{-1}\,(h_{i_1}h_{i_2})^{-\frac{1}{2}}\textstyle\sum EK_{j_1}(\tfrac{X_i - x}{h_{i_1}})EK_{j_2}(\tfrac{X_i - x}{h_{i_2}}) \\
&= n\,(h_{i_1}h_{i_2})^{\frac{1}{2}}\, \tfrac{1}{n^2 h_{i_1}h_{i_2}}\textstyle\sum EK_{j_1}(\tfrac{X_i - x}{h_{i_1}})K_{j_2}(\tfrac{X_i - x}{h_{i_2}}) + o(1)
\end{aligned}
$$

The last equality follows from condition (a): $n^{1/2}h^{1/2}B(h, K, x) \to 0$ for all K and h, the relationship $\frac{1}{h}EK(\frac{X_s - x}{h}) - f(x) = B(h, K, x)$ and Assumption 3b.

Suppose without loss of generality that $q_{12} = h_{i_1}/h_{i_2} \to q < \infty$ (if $q_{12} \to \infty$ consider instead $q_{21} = q_{12}^{-1}$). For the first term

$\frac{q_{12}^{1/2}}{nh_{i_1}}\sum EK_{j_1}(\frac{X_i - x}{h_{i_1}})K_{j_2}(\frac{q_{12}(X_i - x)}{h_{i_1}}) = \frac{q_{12}^{1/2}}{h_{i_1}}\int K_{j_1}(\frac{w - x}{h_{i_1}})K_{j_2}(\frac{q_{12}(w - x)}{h_{i_1}})f(w)dw$; by substituting $z = \frac{w - x}{h_{i_1}}$,

$q_{12}^{\frac{1}{2}}\int K_{j_1}(z)K_{j_2}(q_{12}z)f(x + h_{i_1}z)dz$.

Thus, as $n \to \infty, h_{i_1} \to 0, q_{12} \to q$ and by continuity of f,

$E(\eta(h_{i_1}, K_{j_1})\eta(h_{i_2}, K_{j_2})) \to q^{\frac{1}{2}} f(x) \int K_{j_1}(z) K_{j_2}(qz) dz$.

If $q \to 0$, the limit is zero.

Then consider for a vector $\lambda : \lambda'\lambda = 1$ variables $z_n = \lambda' \Sigma^{-1/2} \eta_\alpha$, where $\Sigma = Var(\eta_a)$. Using Assumption 2(c) that $\int K(z)^{2+\delta} dz < \infty$ for some $\delta > 0$, it can be shown that some higher moment of z_n^2 exists (see Pagan and Ullah (1999), p. 40) and so the Lyapunov condition is satisfied. By Lyapunov's central limit theorem, we have $z_n \xrightarrow{d} N(0,1)$. Part (a) follows by Cramer-Wold theorem.

Part (a) for bandwidths corresponding to condition (b) of Theorem 1 is obtained similarly by noting that it implies $0 < h_{i_1}/h_{i_2} = q < \infty$ when $m' < i_{1,} i_2 \leq m''$.

Part (b) follows from (c) of Theorem 1. For part (c) the covariances are zero because the estimators have different convergence rates.■

References

[1] Bowman, A., 1984, An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, **71**, 353-360.

[2] Cleveland, W.S. and Loader, C.R., 1996, Smoothing by local regression: principles and methods, In: W. Hardle and M.G. Schimek (Eds.) *Statistical Theory and Computational Aspects of Smoothing* (Heidelberg: Physica), 10-49.

[3] Duin, R.P.W., 1976, On the choice of smoothing parameter for Parzen estimators of probability density functions. *IEEE Transactions on Computers*, C-25, 1175-1179.

[4] Fan, Y. and Ullah, A., 1999, Asymptotic normality of a combined regression estimator. *Journal of Multivariate Analysis*, **71**, 191-240.

[5] Gerard, P. and Schucany, W., 1999, Local bandwidth selection for kernel estimation of population densities with line transect sampling. *Biometrics*, **55**, 769-773.

[6] Hall, P., 1983, Large sample optimality of least squares cross-validation in density estimation. *Annals of Statistics*, **11**, 1156-1174.

[7] Hall, P., 1992, *The Bootstrap and Edgeworth Expansion* (New York: Springer-Verlag).

[8] Härdle, W., Kerkyacharian, G., Picard, D. and Tsybakov, A., 1998, *Wavelets, Approximation, and Statistical Applications* (New York: Springer-Verlag).

[9] Kotlyarova, Y. and Zinde-Walsh, V., 2004, Improving the efficiency of the smoothed maximum score estimator. Working paper, McGill University.

[10] Kotlyarova, Y. and Zinde-Walsh, V., 2006, Non- and semi-parametric estimation in models with unknown smoothness. *Economics Letters*, **93**, 379-386.

[11] Lambert, C.G., Harrington, S.E., Harvey, C.R. and Glodjo, A., 1999, Efficient on-line nonparametric kernel density estimation. *Algorithmica*, **25**, 37-57.

[12] Li, Q. and Racine, J.S., 2007, *Nonparametric Econometrics: Theory and Practice* (Princeton University Press).

[13] Loader, C.R., 1999a, Bandwidth selection: classical or plug-in? *The Annals of Statistics*, **27**, 415-438.

[14] Loader, C.R., 1999b, *Local Regression and Likelihood* (New York: Springer).

[15] Marron, J.S. and Wand, M.P., 1992, Exact mean integrated squared error. *Annals of Statistics*, **20**, 712-736.

[16] Pagan, A. and Ullah, A., 1999, *Nonparametric Econometrics* (Cambridge University Press).

[17] Park, B.U. and Marron, J.S., 1990, Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association*, **85**, 66-72.

[18] Park, B.U. and Turlach, B.A., 1992, Practical performance of several data driven bandwidth selectors. *Computational Statistics*, **7**, 251-270.

[19] Parzen, E., 1962, On estimation of a probability density and mode. *Annals of Mathematical Statistics*, **33**, 1065-1076.

[20] Rosenblatt, M., 1956, Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, **27**, 832-837.

[21] Rudemo, M., 1982, Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, **9**, 65-78.

[22] Schucany, W. and Sommers, J., 1977, Improvement of kernel type density estimators. *Journal of the American Statistical Association*, **72**, 420-423.

[23] Schuster, E.F. and Gregory, G.G., 1981, On the inconsistency of maximum likelihood nonparametric density estimators. In: W. F. Eddy (Ed.) *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface* (Berlin: Springer), 295-298.

[24] Sheather, S.J. and Jones, M.C., 1991, A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society*, Series B, **53**, 683-690.

[25] Silverman, B.W., 1986, *Density Estimation for Statistics and Data Analysis* (London: Chapman and Hall).

[26] Stone, C.J., 1984, An asymptotically optimal window selection rule for kernel density estimates. *Annals of Statistics*, **12**, 1285-1297.

[27] Zinde-Walsh, V., 2002, Asymptotic theory for some high breakdown point estimators. *Econometric Theory*, **18**, 1172-1196.

Table 1. Estimated MISE

Density	sample size	Estimated MISE					
		K2,CV	K4,CV	comb2	comb4	comb24	comb33
Normal	500	0.00240	0.00194	0.00221	0.00162	0.00189	0.00204
	1000	0.00132	0.00105	0.00125	0.00089	0.00105	0.00114
	2000	0.00079	0.00059	0.00076	0.00051	0.00061	0.00067
Mixed Normal	500	0.0151	0.0138	0.0148	0.0136	0.0134	0.0132
	1000	0.0086	0.0075	0.0083	0.0072	0.0074	0.0071
	2000	0.0050	0.0042	0.0048	0.0038	0.0042	0.0039
Non-smooth	500	0.0116	0.0125	0.0107	0.0116	0.0115	0.0114
	1000	0.0073	0.0078	0.0065	0.0070	0.0071	0.0070
	2000	0.0044	0.0046	0.0037	0.0039	0.0041	0.0041

Table 2. Relative MISE

Density	sample size	Relative MISE					
		K2,CV	K4,CV	comb2	comb4	comb24	comb33
Normal	500	1.24	1.00	1.14	0.83	0.97	1.05
	1000	1.25	1.00	1.19	0.84	1.00	1.09
	2000	1.35	1.00	1.28	0.86	1.04	1.13
Mixed Normal	500	1.10	1.00	1.08	0.99	0.97	0.96
	1000	1.14	1.00	1.10	0.95	0.98	0.95
	2000	1.20	1.00	1.15	0.92	1.01	0.94
Non-smooth	500	1.00	1.09	0.93	1.01	0.99	0.99
	1000	1.00	1.07	0.90	0.96	0.97	0.97
	2000	1.00	1.05	0.86	0.90	0.94	0.93