Robust kernel estimator for densities of unknown smoothness

By Yulia Kotlyarova* and Victoria Zinde-Walsh†

June 20, 2005

*Department of Economics, Dalhousie University, Halifax, Nova Scotia B3H3J5 Canada; ph. (902)494-7360; fax (902)494-6917; e-mail: yulia.kotlyarova@dal.ca

†Department of Economics, McGill University, 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; it is often assumed that the density function is at least twice differentiable. However, there are cases where non-smooth density functions may be of interest. We provide asymptotic results for kernel estimation of a continuous density for an arbitrary bandwidth/kernel pair. We also derive the limit joint distribution of kernel density estimators corresponding to different bandwidths and kernel functions. Using these results, we construct an estimator that combines several estimators for different bandwidth/kernel pairs to protect against the negative consequences of errors in assumptions about order of smoothness. The results of a Monte Carlo experiment confirm the usefulness of the combined estimator. We demonstrate that while in the standard normal case the combined estimator has a relatively higher mean squared error than the standard kernel estimator, both estimators are highly accurate. On the other hand, for a non-smooth density where the MSE gets very large, the combined estimator provides uniformly better results than the standard estimator.

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 in the literature 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 such estimators, their robustness. 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 leads to asymptotic bias and makes the estimator concentrate around the wrong value; 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 yields a consistent estimator but 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 no rely on smoothness assumptions. Because of the nonparametric rates of convergence, each bandwidth/kernel pair may provide additional information. We derive the joint limit process for such estimators (similar to the joint distribution of smoothed least median of squares estimators (Zinde-Walsh 2002) and smoothed maximum score estimators (Kotlyarova and Zinde-Walsh 2004)) that 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. The weights in the linear combination can be chosen to minimize an estimate of the mean squared error; the resulting estimator is what we call a "combined estimator". The combined estimator can protect against the negative consequences of errors in assumptions about the order of smoothness.

The results of a Monte Carlo experiment confirm the usefulness of the combined estimator in finite samples. We demonstrate that while in the standard normal case

the combined estimator has a relatively larger MSE than the standard kernel estimator, both estimators are highly accurate. On the other hand, for a multimodal smooth density and a non-smooth density where the MSE gets very large, the combined estimator provides uniformly better results than the standard estimator that incorrectly assumes smoothness. 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. Appendices A and B provide the proofs of the results in Section 3 and contain additional information on how to construct polynomial kernel functions.

2 Definitions, notation, 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 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 Liapunov'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 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). There exist other, data-driven methods of bandwidth selection such as the least squares cross validation (Rudemo 1982, Bowman 1984) and the likelihood cross validation (Duin 1976). These methods do not assume differentiability of the density function and may be asymptotically optimal under weak underlying assumptions (Hall 1983 and Stone 1984). The general consensus is that plug-in methods perform well when the density is relatively smooth and that cross-validation methods identify very well steep peaks and other irregularities of the density but tend to undersmooth in more conventional settings (Park and Turlach 1992, Loader 1999a,b). However, when the dataset is large, cross validation will take a long time to compute since the computation time is a quadratic function of the sample size. And it is precisely in the large samples where suboptimality of plug-in methods, applied to the density which is not at least twice differentiable or sufficiently well-behaved, becomes obvious.

In this paper we develop a method to circumvent the choice-of-bandwidth problem using asymptotic results in Section 3, where the assumption of the existence of the second derivative of the density function is replaced with a much weaker requirement of continuity of density.

3 Asymptotic properties of kernel estimators

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

To express the conditions under which we can state asymptotic results without necessarily requiring differentiability of the density, we define the bias of the kernel density estimator

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

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

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

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

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

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

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

The proof is given in Appendix A.

Thus for case (a) (undersmoothing) we obtain a limiting normal distribution and for (b) and (c) the estimator is asymptotically biased. 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 (no asymptotic bias and a limiting Gaussian distribution), and for some slower convergence rate of h there is oversmoothing. Existence of an optimal rate depends on convergence properties of $B(K, h, x)$ that cannot be asserted without

strengthening the assumptions.

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

Assume that $\hat{f}(K,h,x)$ represents the estimator when the function K and bandwidth $h(n)$ are utilized. Consider a number of bandwidths h: $\{h_i(n)\}_{i=1}^m$. Assume that $h_i(n)$ for $i \leq m'$ corresponds to undersmoothing (part (a) of Theorem 1) while $h_i(n)$ for i such that $m' \leq m'' < i \leq m$ corresponds to oversmoothing (part (c) of Theorem 1). If an optimal rate exists then one could have $m'' \geq m'+1$ and $h_i(n)$ 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

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

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*[1].

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

(a) If each $h_1, ..., h_{m''}$ $(m'' \leq m)$ *satisfies condition (a) or (b) of Theorem 1 then*

$$\begin{aligned} \eta_a &= (\eta(h_1, K_1)', ..., \eta(h_1, K_l)', ..., \eta(h_{m''}, K_1)', ..., \eta(h_{m''}, K_l)')' \\ &\xrightarrow{d} N(0, f(x)\Psi), \end{aligned}$$

where the $lm'' \times lm''$ *matrix* Ψ *has elements*

$$\{\Psi\}_{ij} = \begin{cases} \sqrt{q} \int K_i(w) K_j(qw) dw & \text{if } h_i/h_j = q < \infty, \\ 0 & \text{if } h_i/h_j \to 0 \text{ or } h_i/h_j \to \infty; \end{cases}$$

(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 Appendix A.

Thus, if the bandwidths approach 0 at different rates or $\int K_i(w)K_j(w)dw = 0$, the corresponding estimators $\hat{f}(K, h, x)$ 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 can be easily extended to the case of multivariate density functions. Consider the simplest estimator that uses a multivariate density function K_d as kernel and applies the same bandwidth to every coordinate $1, ..., d$ of the random vector:

$$\widehat{f}_d(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K_d \left(\frac{X_{i1} - x_1}{h}, ..., \frac{X_{id} - x_d}{h} \right).$$

The asymptotic results for a single density estimator and joint distribution of estimators can be obtained by using h such that $nh^d \to \infty$ and replacing normalization

$n^{1/2}h^{1/2}$ with d-dimensional normalization $n^{1/2}h^{d/2}$. If the variation of some components of the random vector x is greater than in the others, the use of the same bandwidths for all dimensions may be inappropriate. Pagan and Ullah (1999) suggest to linearly transform the data to have a unit covariance matrix, and then to apply a single bandwidth.

4 The combined estimator

As the results in Section 3 show, finding an optimal rate for a density estimator may be problematic. Here we use the results of Theorem 2 to construct a new combined estimator that optimally combines several standard kernel estimators with various bandwidths and smoothing functions instead of focusing on a single bandwidth/kernel combination. Although efficiency may suffer in straightforward cases when an optimal rate can be found, the Monte Carlo experiments show that the combined estimator provides remarkably robust performance over a variety of cases. Section 4.1 defines the combined estimator. Section 4.2 addresses practical issues of construction of the combined estimator. Performance in a Monte Carlo experiment is discussed in Section 5.

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 $\hat{f}(K_j, h_i, x)$.

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

$MSE(\hat{f}(\{a_{ij}\})) = \sum a_{i_1 j_1} a_{i_2 j_2} \{bias(\hat{f}(K_{j_1}, h_{i_1}, x)) bias(\hat{f}(K_{j_2}, h_{i_2}, x))$

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

The MSE of this combined estimator will not be larger than the smallest MSE of individual estimators $\hat{f}(K_j, h_i, x)$ that are included in the combination. It may be possible to improve upon the best individual estimator in the set by using its combinations with other kernel estimators. If individual estimators are uncorrelated, their combination reduces the variance. The robust method of the least squares cross validation (LSCV) is considered to be appropriate under very weak assumptions, but it produces just one bandwidth (chosen on a grid) for a prespecified kernel. Therefore, if the kernel is not appropriate or the grid of bandwidths not fine enough, the results may be suboptimal and the combined estimator may outperform the LSCV. 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 determine the weights in practice we need to estimate the biases and covariances of all $\hat{f}(K_j, h_i, x)$.

Denote estimated biases and covariances by "hats".

Then ,

$\widehat{MSE}(\hat{f}(\{a_{ij}\})) = \sum a_{i_1 j_1} a_{i_2 j_2} \{\widehat{bias}(\hat{f}(K_{j_1}, h_{i_1}, x))\widehat{bias}(\hat{f}(K_{j_2}, h_{i_2}, x))$

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

Define the combined density estimator $\widehat{f_c}$ by

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

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, $Var(K_j, h_i, x)$, use $\dfrac{\widehat{f(x)} \int K_j(w)^2 dw}{h_i n}$, where the estimate of the density, $\widehat{f(x)}$, has to be specified. Since the smallest bandwidth corresponds to asymptotically unbiased estimator, the candidates for the estimate are $\hat{f}(K_j, h_1, x)$ or a weighted average of individual estimators evaluated at h_1 using kernels $K_1, .., K_l$;

for all off-diagonal elements, covariances $Cov(\hat{f}(K_{j_1}, h_{i_1}, x), \hat{f}(K_{j_2}, h_{i_2}, x))$ can be approximated by $\sqrt{\dfrac{q_{i_1 i_2} Var(\hat{f}(K_{j_1}, h_{i_1}, x)) Var(\hat{f}(K_{j_2}, h_{i_2}, x))}{\delta_{j_1} \delta_{j_2}}} \cdot \int K_{j_1}(w) K_{j_2}(q_{i_1 i_2} w) dw$, where $q_{i_1 i_2} = h_{i_1}/h_{i_2}$, $\delta_j = \int K_j^2(w) dw$.

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):

$$
\begin{aligned}
& Cov(\hat{f}(K_{j_1}, h_{i_1}, x), \hat{f}(K_{j_2}, h_{i_2}, x)) \\
= \; & B^{-1} \sum_{s=1}^{B} \left(\hat{f}_s(K_{j_1}, h_{i_1}, x) - B^{-1} \sum_{t=1}^{B} \hat{f}_t(K_{j_1}, h_{i_1}, x) \right) \\
& \times \left(\hat{f}_s(K_{j_2}, h_{i_2}, x) - B^{-1} \sum_{t=1}^{B} \hat{f}_t(K_{j_2}, h_{i_2}, x) \right),
\end{aligned}
$$

where B is the number of bootstrap replications.

In our Monte Carlo experiment we used the first, asymptotic, method.

The 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.

In our Monte Carlo study, we will use the fact that the estimators with the smallest bandwidth (undersmoothing) are asymptotically unbiased. To find individual biases, we can subtract the average of estimators with the smallest bandwidth from actual

estimators:

$$Bias\hat{f}(K_j, h_i, x) = \hat{f}(K_j, h_1, x) - \overline{\hat{f}(K., h_1, x)}.$$

Other possible estimators of bias could be based on twicing kernels (Newey, Hsieh, and Robins 2004, Kauerman, Mueller, and Carroll 1998) or on Hall's (1992) interpretation of the expected value of $\hat{f}(K_j, h_i, x)$. Hall (1992) observes that the usual bootstrap estimates the expected value of a smooth functional of the empirical distribution function, while $E\hat{f}(K_j, h_i, x)$ is a smooth functional of the estimated density. Therefore, its expected value can be estimated by

$$E_{\hat{f}}(\hat{f}^{bootstrapped}) = \int K(w)\widehat{f}(x - hw)dw$$

and the estimate of the bias will be

$$\widehat{Bias}\hat{f}(K_j, h_i, x) = \int K(w)\widehat{f}(K_j, h_i, x - h_i w)dw - \hat{f}(K_j, h_i, x).$$

4.2.2 Procedure for computing the combined estimator

To determine the set of bandwiths we start with the largest bandwidth in the set that is Silverman's (1986) rule-of-thumb bandwidth. It is optimal when the underlying density is normal. Several studies (Park and Turlach 1992, Loader 1999a) indicate that this bandwidth is usually larger than other methods. If this bandwidth belongs to a truly optimal function/bandwidth combination then as the sample size increases it should yield the fastest convergence rate. Otherwise, it will correspond to oversmoothing. Other bandwidths represent various degrees of undersmoothing and are determined as $2^{i-m}h_m$, for $i = 1, ..., m-1$. If we work with several different kernels, it is desirable to adjust their scale in such a way as for them to have the same rule of thumb bandwidth.

This can be done by considering a transformation $K_\delta(w) = \delta^{-1}K(w/\delta)$, and estimating for both $K(w)$ and $K_\delta(w)$ their rule-of-thumb bandwidths.

We recommend using smoothing functions of order two and above. In theory, one can utilize even lower order kernels since if the density is not differentiable there is nothing to be gained from using kernels of order 2; on the other hand, if the density is sufficiently smooth, a second order kernel would provide an advantage. 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 densities. There may be some advantage in using orthogonal kernels since then the corresponding covariance matrix is zero and they may provide complementary information. To construct several orthogonal polynomial kernels of a given order, we will follow the procedure described in Appendix B.

The entire procedure for a combined estimator includes the following steps: (i) compute the rule-of-thumb bandwidth and other $m-1$ bandwidths; (ii) find the density estimators for all smoothing functions and bandwidths; (iii) estimate the biases and the covariance matrix; and (iv) find the optimal weights for the linear combination and compute (5).

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; thus, the density estimator evaluated at the rule-of-thumb bandwidth should be the optimal choice. The properties of kernel density estimators for this case are well established and it is important to see

how the combined estimator will perform. The combined estimator is not expected to outperform the standard kernel density estimator. The question is, how much worse it will fare. The extra noise in the combined estimator relative to the optimal one is introduced by estimators of biases and the covariance matrix.

In the second model we consider the normal mean mixture $f_2(x) = 0.5\phi(x) + 3\phi(10(x-0.8)) + 2\phi(10(x-1.2))$ analyzed by Hardle et al (1998). This density is also infinitely differentiable; however, it is trimodal and much more wiggly than the standard normal density. Theoretically, its rate of convergence can be made very close to the square root of n and is determined in practice, as well 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 but should produce an estimator converging at the rate $n^{-2/5}$.

The third model contains a non-smooth density that satisfies the Lipschitz condition everywhere except $x = -2$, where it is discontinuous. The rule of thumb bandwidth will converge to zero too slowly, while the combined estimator is expected to perform well everywhere outside of a small neighbourhood of -2, where the density does not satisfy Assumption 1b.

$$f_3(x) = \begin{cases} 5.25 - 5x \text{ if } x \in [0.95, 1.05], \\ 0.5 \text{ if } x \in [0, 0.95), \\ 0.5 + 5x \text{ if } x \in [-0.1, 0), \\ -\frac{1}{38} - \frac{10}{38}x \text{ if } x \in [-2; -0.1), \\ 0 \text{ otherwise.} \end{cases}$$

The sample sizes considered in the experiments are $n = 1000, 2000$, and 4000. 2000 replications per model were performed. The combined estimators are constructed using three bandwidths: h_{opt} = rule of thumb, $h_{opt}/2$ and $h_{opt}/4$.

The results for two different sets of kernels are reported. In both studies we estimate MSEs at 121 points between -3 and 3, and compute simulated MISEs (integrated mean

squared errors).

In the first study we use just one kernel of order two, the standard normal density $K2$. The largest bandwidth is estimated according to the "better rule of thumb" $0.9An^{-1/5}$, where $A = \min(sd, R/1.34)$, R is the interquartile range, and sd is the standard deviation. The results are obtained for (i) the standard kernel density estimator with the "better rule of thumb" bandwidth h_{opt}; (ii) the least squares cross validation method; and (iii) the combined estimator. For cross validation we performed a grid search over 75 bandwidths, starting with $h_{opt}/25$ and the increment $h_{opt}/25$.

The algorithm for the least squares cross validation is discussed in Silverman (1986). Since the LSCV estimator is optimal under very weak conditions, it should have quite small MSE and MISE (integrated mean squared error) in each case. The major problem with the LSCV estimator is its long computational time. On the computer with processor AMD Athlon64 3000+, when the sample size is 4000 and the underlying density is the standard normal, it takes 3 min per replication to calculate both the combined estimator and the LSCV estimator at 121 points while only 0.007 min to compute the combined estimator alone (although with the combined estimator we reestimate the coefficients at each point, and the LSCV bandwidth is determined just once); the results for 8000 observations are 12 minutes versus 0.014 min.

In the second study the combined estimator is based on two orthogonal kernels of order 3, defined on [-1,1]:

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

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

These kernels are asymmetric and may be more appropriate for modelling irregular densities. But if the density function is regular and is more than three times differentiable, asymptotic 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 some fourth-order

kernel. We also analyze how sensitive the combined estimator is to the choice of kernels when each kernel in the combination is far from the optimal. The rule of thumb corresponds to $2.85sd \times n^{-1/7}$. The MSE (Fig. 1, 2) and MISE (Table 1) are provided for both individual kernel estimators at the rule-of-thumb bandwidth and for the combined estimator.

The combined estimators are constructed as described in Section 4.2.

5.2 Summary of the results

Standard normal density.

When the true data-generating process is the standard normal density, simple kernel density estimators perform uniformly better than the combined ones. This is not surprising since the rule of thumb yields the optimal kernel when the density is normal. The combined estimator based on three simple kernel estimators with the same Gaussian kernel and different bandwidths is noticeably more erratic and amplifies those small deviations from the true density that we observe in the rule-of-thumb and cross-validated estimates. Still, the combined estimator does not significantly distort the shape of the density. It is interesting that the 3rd order kernels do not perform as well as the second order kernel, the lack of symmetry being a more important factor than the potential bias reduction due to the higher kernel order. On average, the MSE of the combined estimator is 2.5 - 3 times larger than the MSE of the standard estimator. Both MSEs, however, are very small in absolute terms. The cross-validated estimator performs slightly worse than the standard kernel estimator of order 2 but is better than the combined estimators. The combined estimator based on the two asymmetric functions is somewhat less accurate than the combined estimator on the basis of the symmetric kernel.

Mixture of normal densities (Fig. 1).

The rule-of-thumb bandwidth strongly oversmoothes the mixture of normal densities, so that instead of two narrow and high peaks on the right we observe one peak which is wide and low. In Fig. 1 we see that the standard estimator has a very definite oscillating pattern of the MSE, and this oscillation becomes even more pronounced in the case of the two asymmetric kernels. The peaks correspond to the points of local extrema of the density function. The smallest values of the MSE of standard estimators are achieved on the segments of the density function that can be well approximated by a straight line. The combined estimators model very well the right half of the density but are somewhat wiggly on the left, where the density is smooth and flat. Both combined estimators have a very stable and low MSE everywhere, with maximum values more than 10 times lower than MSE of standard estimators. The cross-validated estimator behaves similarly to the combined estimators. With this irregular but infinitely differentiable density, the combined estimator from the second study constructed from asymmetric higher-order kernels achieves higher precision than the combined estimator on the basis of the symmetric function. Individual asymmetric estimators do not detect both peaks in the mixture of normal densities, while the combined estimator does it very well.

Non-smooth density (Fig. 2).

The case of a non-smooth density demonstrates that the rule-of-thumb estimator does not model well sharp features of the density. The LSCV and combined estimator oscillate a lot but can identify abrupt changes in the pattern. It is worth noting that steep increase and decrease of the density at $x = 0$ and $x = 1$ are modelled very well whereas the jump at $x = -2$ is oversmoothed (since the density is not continuous at this point, the asymptotic results from Section 3 are not applicable). In Fig. 2 (the non-smooth density) the combined estimators clearly dominate the standard estimators in precision. For both types of estimators, the points where the density is not smooth $(x = -2, -0.1, 0, 0.95, 1.05)$ cause substantial increases in the MSE but non-smoothness

affects standard estimators over larger intervals. At $x = -2$ the density is discontinuous, therefore the combined estimators are not expected to perform well either. The cross-validated estimator has more uniform MSE than the combined estimators but their MISEs are very close.

The values of MISE in Table 1 confirm that in the absence of information about the smoothness of the density the combined estimators provide more reliable results than the standard kernel estimators. The combined estimators do not outperform the least squares cross validation method, which is shown to be optimal for bounded densities (Stone 1984), however the combined estimators can be computed much faster for large sample sizes since their computational time is of order n while for the cross validation it is $O(n^2)$.

When using a combined estimator, we may lose some efficiency in cases of smooth symmetric densities. Since such densities are usually estimated very precisely, the difference in MSEs of standard and combined estimators is not large in absolute terms. At the same time, when the density is not smooth or well behaved, standard estimators can be seriously biased and the improvement offered by combined estimators is very significant.

Acknowledgements

The support of the Social Sciences and Humanities Research Council of Canada (SSHRC), the FONDS QUEBECOIS DE LA RECHERCHE SUR LA SOCIETE ET LA CULTURE (FRQSC) is gratefully acknowledged.

Appendix A: Proofs of theorems

Proof of Theorem 1.

From definition (4) we have that

$n^{1/2}h^{1/2}B(K,h,x) = n^{1/2}h^{1/2}(E\hat{f}(x) - f(x))$

$= n^{1/2}h^{1/2}\left[\hat{f}(x) - f(x)\right] - n^{1/2}h^{1/2}\left[\hat{f}(x) - E\hat{f}(x)\right].$

In condition (a) of the Theorem 1 the left-hand side is $o(1)$, thus using (3) proves (a).

Similarly, if condition (b) holds then

$n^{1/2}h^{1/2}\left[\hat{f}(x) - f(x)\right] - n^{1/2}h^{1/2}\left[\hat{f}(x) - E\hat{f}(x)\right] - B(K) = o(1),$

and (b) follows from (3).

For (c) we get from (3) that

$(nh)^{\frac{1}{2}}\left(\hat{f}(x) - E\hat{f}(x)\right) = O_p(1)$ and therefore

$n^{1/2}h^{1/2}\left[\hat{f}(x) - f(x)\right] - n^{1/2}h^{1/2}B(K,h,x) = O_p(1).$

Since $\left(n^{1/2}h^{1/2}\left|B(K,h,x)\right|\right)^{-1} = o(1)$, we can show that

$\left(n^{1/2}h^{1/2}\left|B(K,h,x)\right|\right)^{-1}\left[n^{1/2}h^{1/2}\left[\hat{f}(x) - f(x)\right] - n^{1/2}h^{1/2}B(K,h,x)\right] = o_p(1)$

and (c) obtains.■

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

$E(\eta(h_{i_1},K_{j_1})\eta(h_{i_2},K_{j_2}))$

$$= n\left(h_{i_1}h_{i_2}\right)^{\frac{1}{2}} E\left[\left(\frac{1}{nh_{i_1}}\sum K_{j_1}(\frac{X_s-x}{h_{i_1}}) - f(x)\right)\left(\frac{1}{nh_{i2}}\sum K_{j_2}(\frac{X_t-x}{h_{i_2}}) - f(x)\right)\right]$$
$$= n\left(h_{i_1}h_{i_2}\right)^{\frac{1}{2}} \frac{1}{n^2 h_{i_1}h_{i_2}}\sum EK_{j_1}(\frac{X_i-x}{h_{i_1}})K_{j_2}(\frac{X_i-x}{h_{i_2}})$$
$$+n\left(h_{i_1}h_{i_2}\right)^{\frac{1}{2}} [\frac{1}{n^2 h_{i_1}h_{i_2}}\sum_{l\neq m} EK_{j_1}(\frac{X_l-x}{h_{i_1}})EK_{j_2}(\frac{X_m-x}{h_{i_2}})$$
$$-f(x)\frac{1}{nh_{i_1}}\sum EK_{j_1}(\frac{X_l-x}{h_{i_1}}) - f(x)\frac{1}{nh_{i_2}}\sum EK_{j_2}(\frac{X_l-x}{h_{i_2}}) + f(x)^2]$$
$$= n\left(h_{i_1}h_{i_2}\right)^{\frac{1}{2}} \frac{1}{n^2 h_{i_1}h_{i_2}}\sum EK_{j_1}(\frac{X_i-x}{h_{i_1}})K_{j_2}(\frac{X_i-x}{h_{i_2}})$$
$$+n\left(h_{i_1}h_{i_2}\right)^{\frac{1}{2}} \left[\left(\frac{1}{nh_{i_1}}\sum EK_{j_1}(\frac{X_s-x}{h_{i_1}}) - f(x)\right)\left(\frac{1}{nh_{i2}}\sum EK_{j_2}(\frac{X_t-x}{h_{i_2}}) - f(x)\right)\right]$$
$$-n^{-1}\left(h_{i_1}h_{i_2}\right)^{-\frac{1}{2}}\sum EK_{j_1}(\frac{X_i-x}{h_{i_1}})EK_{j_2}(\frac{X_i-x}{h_{i_2}})$$
$$= n\left(h_{i_1}h_{i_2}\right)^{\frac{1}{2}} \frac{1}{n^2 h_{i_1}h_{i_2}}\sum EK_{j_1}(\frac{X_i-x}{h_{i_1}})K_{j_2}(\frac{X_i-x}{h_{i_2}}) + o(1)$$

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

For the first term, introduce a new variable $q = h_{i_1}/h_{i_2}$ and compute the expectation as $h_{i_1} \to 0$:

$$\frac{q^{\frac{1}{2}}}{nh_{i_1}}\sum EK_{j_1}(\frac{X_i-x}{h_{i_1}})K_{j_2}(\frac{q(X_i-x)}{h_{i_1}}) = \frac{q^{\frac{1}{2}}}{h_{i_1}}\int K_{j_1}(\frac{w-x}{h_{i_1}})K_{j_2}(\frac{q(w-x)}{h_{i_1}})f(w)dw$$
$$\overset{z=\frac{w-x}{h_{i_1}}}{=} q^{\frac{1}{2}}\int K_{j_1}(z)K_{j_2}(qz)f(x+h_{i_1}z)dz = q^{\frac{1}{2}}f(x)\int K_{j_1}(z)K_{j_2}(qz)dz + o(1).$$

Thus,

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

If $q \to \infty$ or $q \to 0$, $q^{\frac{1}{2}}f(x)\int K_{j_1}(z)K_{j_2}(qz)dz \to 0$ under Assumption 2.

Then consider for $\lambda : \lambda'\lambda = 1$ variables $z_{in} = \lambda'\Sigma^{-1/2}\eta_i$, 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_{in}^2 exists (see Pagan and Ullah (1999, p. 40)) and so the Lyapunov condition is satisfied. By Lyapunov's central limit theorem we have $z_{in} \xrightarrow{d} N(0,1)$. 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 (b) of Theorem 1. For (c) the covariances are zero because the estimators have different convergence rates.■

Appendix B: Polynomial kernels

The smoothing functions are selected to be polynomials that satisfy these assumptions:

(a) The smoothing function K is a continuously differentiable function with support in $[-1, 1]$;

(b) $\int K(w)dw = 1$;

(c) K is a kernel function of order s: $\int w^i K(w)dw = 0$ if $0 < i < s$, $s \geq 2$.

Consider an n-degree polynomial, $\sum_{i=0}^{n} a_i x^i$.

If the following restrictions are imposed on the coefficients of the polynomial:

1. $K(-1) = K(1) = 0 : \sum_{i=0}^{n} a_i(-1)^i = 0; \quad \sum_{i=0}^{n} a_i = 0;$
2. $K'(-1) = K'(1) = 0 : \sum_{i=0}^{n} i a_i(-1)^{i-1} = 0; \ \sum_{i=0}^{n} i a_i = 0;$
3. $\int K(w)dw = 1 : \sum_{i=0}^{n} \frac{a_i}{i+1}(1 - (-1)^{i+1}) = 1;$
4. $\int wK(w)dw = 0 : \sum_{i=0}^{n} \frac{a_i}{i+2}(1 - (-1)^{i+2}) = 0,$

then the quartic second-order kernel can be obtained:

$K = \frac{15}{16}\left(1 - x^2\right)^2.$

To construct orthogonal kernels, add the requirement

5. $\int K_i(x)K_j(x)dx = 0$, for $i \neq j$.

It may be helpful to work with pairs of asymmetric kernels such as

6. $K_i(x) = K_j(-x)$.

Finally, for the 3rd-order kernels

7. $\int w^2 K(w)dw = 0$.

We construct two kernels of third order, $K3a$ and $K3b$, that satisfy conditions 1-7:

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

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

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 C. R. Loader (1996). Smoothing by local regression: principles and methods, in *Statistical theory and computational aspects of smoothing,* W. Hardle and M. G. Schimek (eds.), 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] Gerard, P. and W. Schucany (1999). Local bandwidth selection for kernel estimation of population densities with line transect sampling, *Biometrics* 55, 769-773.

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

[6] Hall, P. (1992). *The bootstrap and Edgeworth expansion*, New York: Springer-Verlag.

[7] Hardle, W., G. Kerkyacharian, D. Picard, and A. Tsybakov (1998). *Wavelets, approximation, and statistical applications*, New York: Springer-Verlag.

[8] Kauerman G., Mueller M., and R. Carroll (1998) The efficiency of bias-corrected estimators for nonparametric kernel estimation based on local estimating equations. Working paper, Texas A&M University.

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

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

[11] Loader, C. R. (1999b). *Local regression and likelihood*, New York: Springer.

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

[13] Newey, W. , F. Hsieh, and J. Robins (2004). Twicing kernels and a small bias property of semiparametric estimators, *Econometrica* 72, 947-962.

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

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

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

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

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

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

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

[21] Schuster, E. F. and G. G. Gregory (1981). On the inconsistency of maximum likelihood nonparametric density estimators, in *Computer Science and Statistics:*

Proceedings of the 13th Symposium on the Interface, W. F. Eddy (ed.), Berlin: Springer, 295-298.

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

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

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

[25] Zinde-Walsh, V. (2002). Asymptotic Theory for some High Breakdown Point Estimators, *Econometric Theory* 18, 1172-1196.

Table 1. Integrated mean squared errors of kernel density estimators

	kernel	$K2$	$K2$	$K2$	$K3a$	$K3b$	$K3ab$
density	method	*standard*	*LSCV*	*comb*2	*standard*	*standard*	*comb*33
normal	$n = 1000$	0.00112	0.00137	0.00307	0.00267	0.00269	0.00388
	$n = 2000$	0.00066	0.00077	0.00178	0.00151	0.00150	0.00217
	$n = 4000$	0.00039	0.00044	0.00104	0.00085	0.00085	0.00121
mixture	$n = 1000$	0.0721	0.0087	0.0083	0.1483	0.1493	0.0065
	$n = 2000$	0.0617	0.0050	0.0047	0.1340	0.1342	0.0036
	$n = 4000$	0.0507	0.0029	0.0028	0.1237	0.1237	0.0020
non-	$n = 1000$	0.0362	0.0117	0.0113	0.0816	0.0901	0.0118
smooth	$n = 2000$	0.0306	0.0084	0.0084	0.0693	0.0776	0.0089
	$n = 4000$	0.0259	0.0064	0.0066	0.0603	0.0686	0.0071

List of Figures

Fig. 1

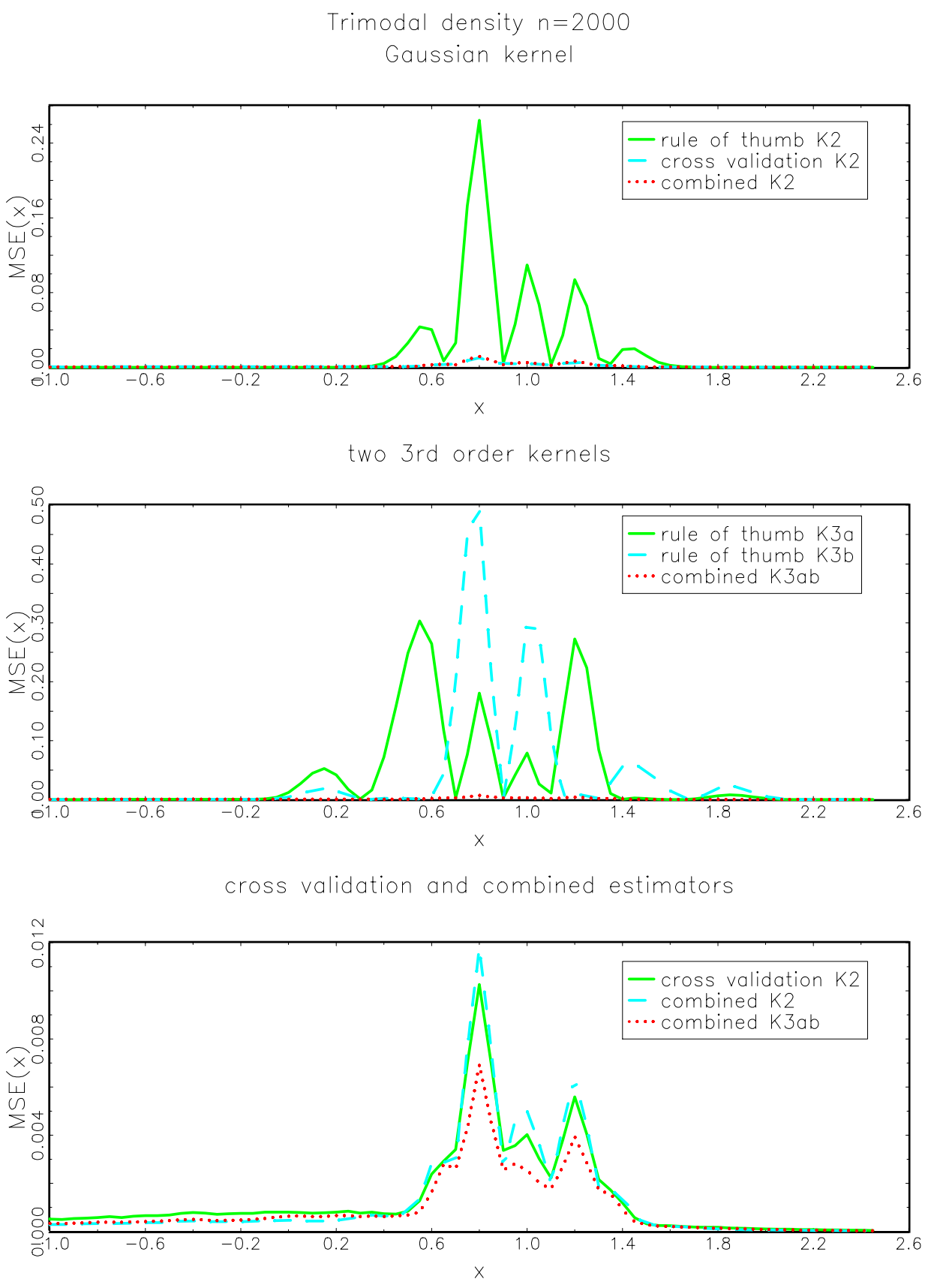
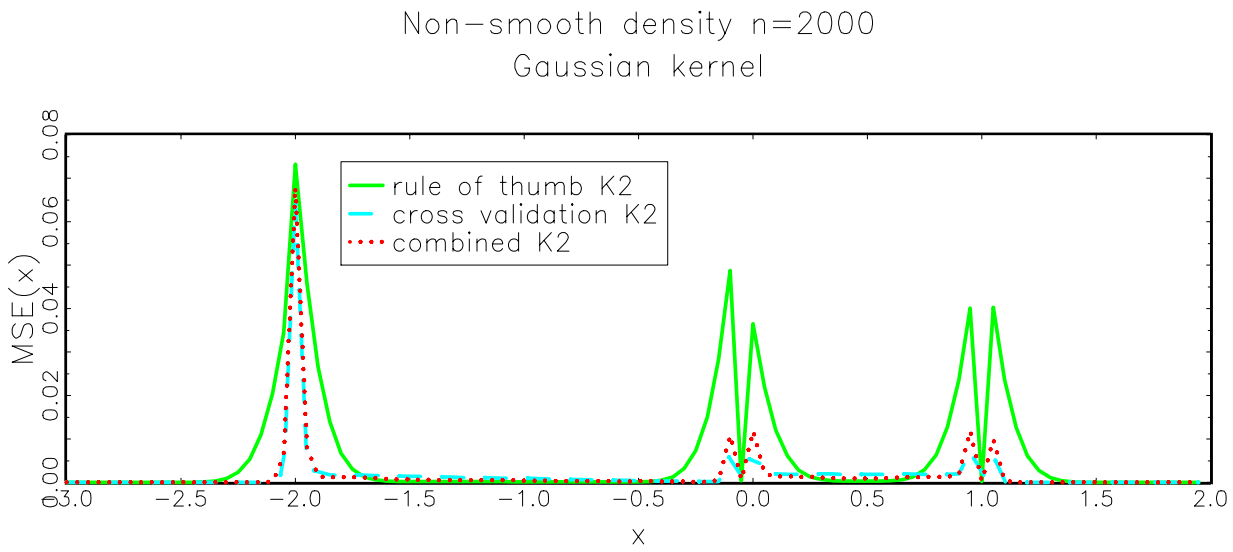
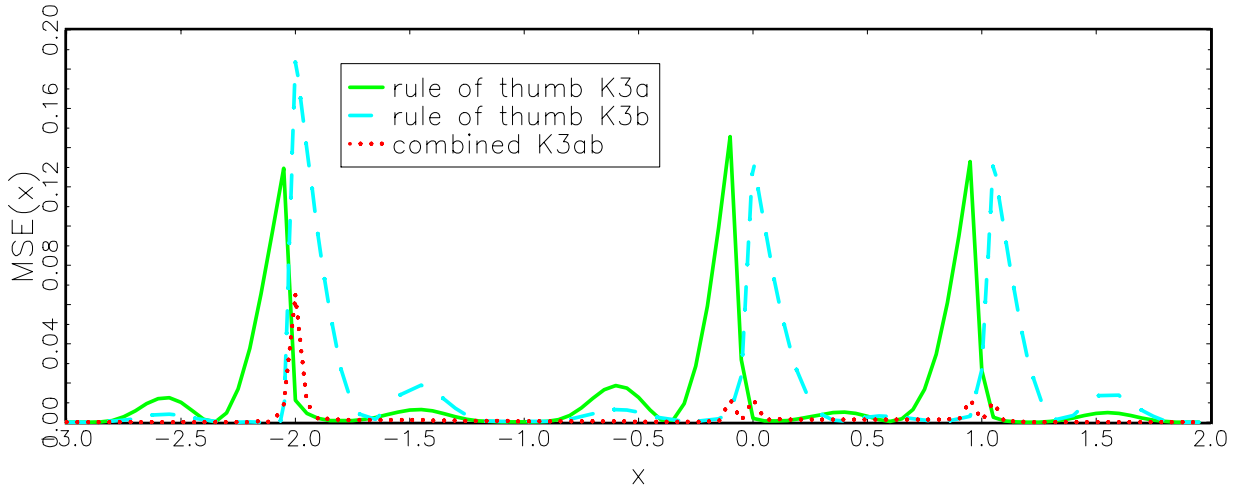
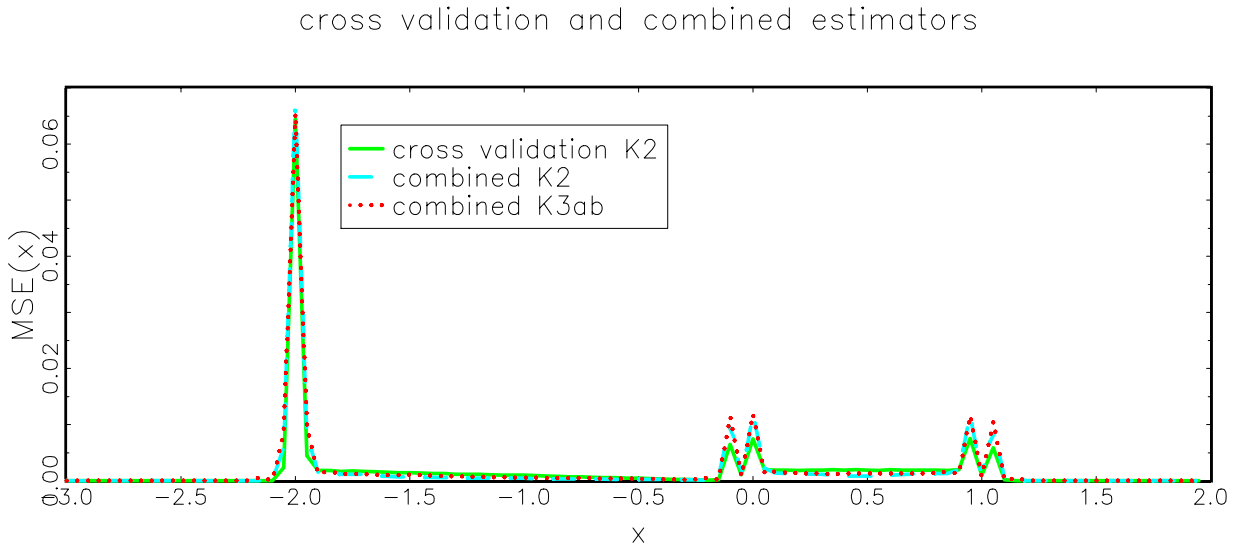

Trimodal density n=2000
Gaussian kernel
rule of thumb K2
cross validation K2
combined K2
MSE(x)
0.00
0.08
0.16
0.24
−1.0
−0.6
−0.2
0.2
0.6
1.0
1.4
1.8
2.2
2.6
X
two 3rd order kernels
rule of thumb K3a
rule of thumb K3b
combined K3ab
MSE(x)
0.00
0.10
0.20
0.30
0.40
0.50
X
cross validation and combined estimators
cross validation K2
combined K2
combined K3ab
MSE(x)
0.000
0.004
0.008
0.012
X

Fig. 2


Non−smooth density n=2000
Gaussian kernel
rule of thumb K2
cross validation K2
combined K2
MSE(x)
0.00
0.02
0.04
0.06
0.08
−3.0
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
X


two 3rd order kernels
rule of thumb K3a
rule of thumb K3b
combined K3ab
MSE(x)
0.00
0.04
0.08
0.12
0.16
0.20
−3.0
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
X


cross validation and combined estimators
cross validation K2
combined K2
combined K3ab
MSE(x)
0.00
0.02
0.04
0.06
−3.0
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
X