Averaging Estimators for Kernel Regressions

Chu-An Liu*
University of Wisconsin - Madison†

cliu32@wisc.edu
First Draft: February 2010
Current Draft: August 2011

Abstract

This paper proposes a model averaging approach to reduce the mean squared error (MSE) and the weighted integrated mean squared error (WIMSE) of kernel estimators of regression functions. At each point of estimation, we construct a weighted average of the local constant and local linear estimators. The optimal local and global weights for averaging are chosen to minimize the MSE and WIMSE of the averaging estimator, respectively. We propose two data-driven approaches for bandwidth and weight selection and derive the rate of convergence of the cross-validated weights to their optimal benchmark values. Monte Carlo simulations show that the proposed estimator can achieve significant efficiency gains over the local constant and local linear estimators.

Keywords: Averaging estimator, Bandwidth selection, Cross-validation, Local constant estimator, Local linear estimator.
JEL Classification: C13, C14.

*I am deeply indebted to Bruce Hansen and Jack Porter for guidance and encouragement.
†Department of Economics, 1180 Observatory Drive, University of Wisconsin, Madison, WI 53706, USA.

1 Introduction

There is a very large literature on the estimation of unknown conditional mean function using kernel smoothers, including the Nadaraya-Watson estimator, or the local constant estimator, (Nadaraya, 1964; Watson, 1964), the Gasser and Müller estimator (Gasser and Müller, 1979), the local linear estimator (Stone, 1977; Cleveland, 1979; Fan, 1992, 1993; Fan and Gijbels, 1992), and the local polynomial estimator (Ruppert and Wand, 1994; Fan and Gijbels, 1996). Among these kernel methods, the most popular two are the local constant and local linear estimators. The local constant estimator is attractive due to its simplicity, while the local linear estimator is appealing because of good asymptotic properties. Fan (1992, 1993) shows that the local linear estimator is minimax efficient and adapts automatically to the boundary. However, a major drawback of the local linear estimator is that it performs poorly and has a unbounded conditional variance when the data is sparse.

Previous works on improvements and modifications of the local linear estimator using averaging estimators include Choi and Hall (1998) and Cheng, Peng, and Wu (2007). Choi and Hall (1998) suggest a convex combination of three local linear estimators which reduces the bias without changing the variance. The bias reduction comes from the average of two similar but opposite shifted estimators at each point of estimation. In contrast, Cheng, Peng, and Wu (2007) propose an averaging estimator to maintain the bias but reduce the variance. They construct a linear combination of the local linear estimators at three equally spaced nearby points and the nearby points are chosen to minimize the variance. To deal with the poor behavior arising from data sparsity and the unbounded conditional variance, Seifert and Gasser (1996) propose two small-sample modifications to improve the performance: local increase of bandwidth and local polynomial ridge regression. Frölich (2004) finds that the ridge estimator proposed by Seifert and Gasser (1996, 2000) is better than the local linear matching estimator on estimating average treatment effects in the finite sample simulations. Mammen and Marron (1997) provide a modification of the local constant estimator, which is a shift of the estimator based on the kernel weighted center of mass. Unlike the local linear estimator, the conditional variance of the local constant estimator is bounded for any kernel function, see Seifert and Gasser (1996). Thus, by averaging the local constant and local linear estimators with appropriate weights, we can avoid the infinite conditional variance.

In this paper, we develop a new model averaging estimator for the kernel regression. The proposed estimator, the local weighted estimator, is a weighted average of the local constant and local linear estimators at each point of estimation. The proposed averaging estimator effectively handles data sparsity by assigning the appropriate weights between the local constant and local linear estimators. Specifically, the optimal local and global weights are found by minimization of the mean squared error (MSE) and the weighted integrated mean squared error (WIMSE), which is conceptually similar to the ridge estimator proposed by Seifert and Gasser (1996, 2000). The key difference is that we allow the averaging estimator to have different bandwidths for local constant and the local linear estimators, while Seifert and Gasser imposed the equal bandwidth constraint on two kernel estimators, which is a special case in our framework. Also, Seifert and Gasser (2000)

only consider the case of the optimal local weight. In all, our approach is similar to Seifert and Gasser (1996, 2000), but we provide two data-dependent methods to select bandwidths and weights.

To choose bandwidths and weights of the local weighted estimator, we propose two approaches: the two-step cross-validation method and the joint cross-validation method. The cross-validation method is a frequently used technique to select bandwidths. The idea of the cross-validation method is to minimize the leave-one-out sum of squared error which is asymptotically equivalent to the weighted integrated mean squared error; for the local constant estimator, see Hardle and Marron (1985) and Hardle, Hall, and Marron (1988, 1992), for the local linear estimator, see Xia and Li (2002) and Li and Racine (2004), and for the kernel estimator with categorical data, see Racine and Li (2004). Our two-step cross-validation estimation is straightforward and easily implemented. We first select bandwidths of the local constant and local linear estimators, respectively. Then, for two given cross-validated bandwidths, the weights are selected in the second step. Unlike the two-step cross-validation method, bandwidths and weight are selected simultaneously in the joint cross-validation method. We show the leave-one-out cross-validation criterion is asymptotically equivalent to the WIMSE of the local weighted estimator. We also establish the rate of convergence of the two-step cross-validated weights to their optimal benchmark value.

The main contribution of this paper is twofold. First, the proposed estimator reduces the MSE and WIMSE of the kernel regression. Second, we provide two data-dependent approaches to choose bandwidths and weights and show the rate of convergence of data-dependent weights to their optimal benchmark values.

The outline of the paper is as follows. Section 2 introduces the local weighted estimator. Section 3 analyzes the asymptotic behavior of the proposed estimator. Section 4 presents the two-step cross-validation method and the joint cross-validation method and discusses the asymptotic properties of these two data-driven approaches. Section 5 provides a numerical study to compare the local constant, the local linear, and the local weighted estimators. Section 6 concludes. Proofs, figures, and tables are given in the appendix.

2 Model and Estimation

Consider a nonparametric regression model

$$y_i = m(x_i) + e_i \tag{2.1}$$

$$m(x) = \mathrm{E}(y_i|x_i = x) \tag{2.2}$$

$$\mathrm{E}(e_i^2|x_i = x) = \sigma^2(x) \tag{2.3}$$

where $(x_1, y_1), ..., (x_n, y_n)$ are pairs of independent and identically distributed (iid) random variables from a joint density $f(x, y)$ and e_i is the iid random error. The parameter of interest is the regression function $m(x)$ which is an unknown function. Unlike the classical regression analysis which assumes the regression function is linear or some specific form, here there is no assumption on the structure of the relationship between the dependent variable y_i and the regressor x_i. Several kernel estimators

have been proposed to estimate the regression function nonparametrically, including the Nadaraya-Watson/local constant estimator, the Gasser and Müller estimator, the local linear estimator, and the local polynomial estimator.

Let $k(u)$ be the kernel function and h the bandwidth. The local constant estimator is defined as

$$\hat{m}_{lc}(x) = \left(\sum_{i=1}^{n} k\left(\frac{x_i - x}{h}\right)\right)^{-1} \left(\sum_{i=1}^{n} k\left(\frac{x_i - x}{h}\right) y_i\right). \tag{2.4}$$

The local linear estimator is defined as the solution of following minimization problem:

$$\min_{\alpha,\beta} \sum_{i=1}^{n} \big(y_i - \alpha - \beta\,(x_i - x)\big)^2 k\left(\frac{x_i - x}{h}\right). \tag{2.5}$$

Simple calculation yields

$$\hat{m}_{ll}(x) = \hat{\alpha} = \left(\sum_{i=1}^{n} s_i\right)^{-1} \left(\sum_{i=1}^{n} s_i y_i\right), \tag{2.6}$$

$$s_i = k\left(\frac{x_i - x}{h}\right)\Big(T_{n,2}(x) - (x_i - x)T_{n,1}(x)\Big), \tag{2.7}$$

$$T_{n,\ell}(x) = \sum_{i=1}^{n} k\left(\frac{x_i - x}{h}\right)(x_i - x)^{\ell}, \quad \ell = 1, 2. \tag{2.8}$$

We now define the local weighted estimator. For each point x, the local weighted estimator is a weighted average of the local constant and local linear estimators. Let $\lambda \in (-\infty, \infty)$ be the weight for the local linear estimator. The averaging estimator is

$$\hat{m}_{lw}(x) = \lambda \hat{m}_{ll}(x) + (1 - \lambda)\hat{m}_{lc}(x). \tag{2.9}$$

The local weighted estimator is an affine combination of the local constant and local linear estimators. Here, the weights for the averaging estimator are allowed to take on positive and negative values. For $\lambda = 1$, the local weighted estimator is the local linear estimator. For $\lambda = 0$, the local weighted estimator is the local constant estimator. From a finite sample point of view, it may improve performance by restricting the weights to be non-negative and to sum to one. This seems like a reasonable restriction if the weights are interpreted as model probabilities.

3 Asymptotic Properties

In this section, we present the mean squared error (MSE) and the weighted integrated mean squared error (WIMSE) of the local weighted estimator. We assume the support of x is a compact set $\mathcal{X}$. Throughout the paper, we denote $\kappa_2 = \int_{-\infty}^{\infty} u^2 k(u) du$ and $\upsilon = \int_{-\infty}^{\infty} k(u)^2 du$. Let h_{lc} be the bandwidth for the local constant estimator and h_{ll} the bandwidth for the local linear estimator. Assume $h_{ll}/h_{lc} \to \gamma$ as $n \to \infty$ where $\gamma \in (0, \infty)$.

Assumption 1. The regression function $m(\cdot)$ has a bounded second derivative.

Assumption 2. The marginal density function $f(\cdot)$ of x satisfies $f(x) > 0$ and $|f(x) - f(y)| \leq c|x-y|^a$ for some $0 < a < 1$.

Assumption 3. The conditional variance function $\sigma^2(\cdot)$ is bounded and continuous.

Assumption 4. The kernel function $k(\cdot)$ is a symmetric density function with compact support.

Assumption 5. The weight function $w(\cdot)$ is a nonnegative and bounded function with compact support $\mathcal{W}$ which is contained in the interior of $\mathcal{X}$.

Assumptions 1-4 are similar to the assumptions used in Fan (1993). Assumption 5 is imposed to reduce the bias of the local weighted estimator at the boundary points. It is well known that the local linear estimator adapts automatically to the boundary, while the local constant estimator suffers from boundary effects and requires a boundary modification. Since the local weighted estimator is a weighted average of the local constant and local linear estimators, the rate of convergence of the local weighted estimator at the boundary points is slower than that for points in the interior of the support unless $\lambda = 1$. In order to avoid boundary effects, we restrict the support of $w(x)$ to be in the interior of $\mathcal{X}$.

Theorem 1. *Under Assumptions 1-4, if $h_{lc} \to 0$, $h_{ll} \to 0$, $nh_{lc} \to \infty$, and $nh_{ll} \to \infty$ as $n \to \infty$, then the MSE of the local weighted estimator is*

$$\begin{aligned} E\big(\hat{m}_{lw}(x) - m(x)\big)^2 = {} & \lambda^2 \xi_1(x, h_{ll}) + (1-\lambda)^2 \xi_2(x, h_{lc}) + 2\lambda(1-\lambda)\xi_{12}(x, h_{ll}, h_{lc}) \\ & + o\left(h_{ll}^4 + h_{lc}^4 + h_{ll}^2 h_{lc}^2 + (nh_{ll})^{-1} + (nh_{lc})^{-1}\right), \end{aligned} \tag{3.1}$$

where

$$\begin{aligned} \xi_1(x, h_{ll}) &= \kappa_2^2 \left(\frac{m''(x)}{2}\right)^2 h_{ll}^4 + \frac{\upsilon}{nh_{ll}} \frac{\sigma^2(x)}{f(x)}, \\ \xi_2(x, h_{lc}) &= \kappa_2^2 \left(\frac{m''(x)}{2} + \frac{m'(x)f'(x)}{f(x)}\right)^2 h_{lc}^4 + \frac{\upsilon}{nh_{lc}} \frac{\sigma^2(x)}{f(x)}, \\ \xi_{12}(x, h_{ll}, h_{lc}) &= \kappa_2^2 \left(\frac{m''(x)}{2}\right)\left(\frac{m''(x)}{2} + \frac{m'(x)f'(x)}{f(x)}\right) h_{ll}^2 h_{lc}^2 + \frac{\tilde{\upsilon}}{nh_{lc}} \frac{\sigma^2(x)}{f(x)}, \end{aligned}$$

and $\tilde{\upsilon} = \int_{-\infty}^{\infty} k(\gamma u)k(u)du$.

The weight λ which minimizes MSE is

$$\lambda^o(x, h_{ll}, h_{lc}) = \frac{\xi_2(x, h_{lc}) - \xi_{12}(x, h_{ll}, h_{lc})}{\xi_1(x, h_{ll}) + \xi_2(x, h_{lc}) - 2\xi_{12}(x, h_{ll}, h_{lc})}, \tag{3.2}$$

and the minimized MSE is

$$\frac{\xi_1(x, h_{ll})\xi_2(x, h_{lc}) - \xi_{12}(x, h_{ll}, h_{lc})^2}{\xi_1(x, h_{ll}) + \xi_2(x, h_{lc}) - 2\xi_{12}(x, h_{ll}, h_{lc})}. \tag{3.3}$$

Theorem 1 states that the covariance between the local constant and local linear estimators is $\tilde{\upsilon}\sigma^2(x)/(nh_{lc}f(x))$ where $\tilde{\upsilon}$ is a convolution kernel function. The value of $\tilde{\upsilon}$ depends on the ratio of two bandwidths γ. In general, the local linear estimator tends to choose the larger bandwidths than the local constant estimator. When $h_{ll} > h_{lc}$, γ is greater than 1, and $\tilde{\upsilon}$ is always smaller than υ, which also implies that the covariance matrix of the local constant and local linear estimators is positive definite. If $h_{ll} = h_{lc}$, then the covariance term degenerates into the variance term of the local constant/linear estimator. Since the local weighted estimator allows the different bandwidths for the local constant and local linear estimators, the ridge estimator proposed by Seifert and Gasser (2000) is a particular case in this more general approach.

The values of $\xi_1(x, h_{ll})$ and $\xi_2(x, h_{lc})$ in Theorem 1 represent the leading terms of MSE of the local linear and local constant estimators, respectively. As long as $\xi_1(x, h_{ll}) \neq \xi_{12}(x, h_{ll}, h_{lc})$ and $\xi_2(x, h_{lc}) \neq \xi_{12}(x, h_{ll}, h_{lc})$, the MSE given in (3.3) is strictly less than the MSE of any linear combination of the local linear and local constant estimators. Note that the weight given in (3.2) is the optimal local weight for the local weighted estimator which is a function of x, h_{lc}, and h_{ll}. Here, the weights for the local weighted estimator are not restricted to be positive. The minimized MSE given in (3.3) is also hold for the negative weights. The only restriction is the weights are required to sum to 1. Otherwise, the averaging estimator is not consistent.

To illustrate the efficiency gains from the local weighted estimator, consider Figure 1 and 2. We plot the pointwise asymptotic MSE, asymptotic squared bias, and asymptotic variance for the local constant estimator (the dotted line), the local linear estimator (the dashed line), and the local weighted estimators (the solid line) for sample size $n = 100$, respectively. The optimal local bandwidths are used for the local constant and local linear estimators while the optimal local weight (3.2) is used for the local weighted estimator.

In Figure 1, the local weighted estimator has the smallest asymptotic MSE, asymptotic squared bias, and asymptotic variance, while the local constant estimator performs worst. It can be observed that the performances of these three estimators get worse as $|x|$ increases, since both the component of bias term $m''(x)$ and the component of variance terms $f(x)^{-1}$ become larger. Comparing the row (a) with the row (b), the performances of the local constant and local linear estimators are getting close to each other as $|f'(x)/f(x)|$ becomes smaller. A similar result can be seen in Figure 2, where the local weighted estimator performs better than other two estimators. Compared with Figure 1, we have irregular behavior in Figure 2 because the covariate x follows a mixed normal distribution. In addition, the three estimators have poor performances when the data has a peak, the region between 0 and 1, in the row (b). We observe the interesting fact that the local linear estimator does not always outperform the local constant estimator in the row (a).

We now present the WIMSE and the optimal global weight of the local weighted estimator.

Theorem 2. *Under Assumptions 1-5, if* $h_{lc} \to 0$, $h_{ll} \to 0$, $nh_{lc} \to \infty$, *and* $nh_{ll} \to \infty$ *as* $n \to \infty$,

then the WIMSE of the local weighted estimator is

$$E\int_{-\infty}^{\infty}\left(\hat{m}_{lw}(x)-m(x)\right)^2 w(x)dx = \lambda^2\zeta_1(h_{ll})+(1-\lambda)^2\zeta_2(h_{lc})+2\lambda(1-\lambda)\zeta_{12}(h_{ll},h_{lc}) + o\left(h_{ll}^4+h_{lc}^4+h_{ll}^2h_{lc}^2+(nh_{ll})^{-1}+(nh_{lc})^{-1}\right),$$

where

$$\zeta_1(h_{ll}) = \kappa_2^2\int_{-\infty}^{\infty}\left(\frac{m''(x)}{2}\right)^2 w(x)dxh_{ll}^4+\frac{\upsilon}{nh_{ll}}\int_{-\infty}^{\infty}\frac{\sigma^2(x)}{f(x)}w(x)dx,$$

$$\zeta_2(h_{lc}) = \kappa_2^2\int_{-\infty}^{\infty}\left(\frac{m''(x)}{2}+\frac{m'(x)f'(x)}{f(x)}\right)^2 w(x)dxh_{lc}^4+\frac{\upsilon}{nh_{lc}}\int_{-\infty}^{\infty}\frac{\sigma^2(x)}{f(x)}w(x)dx,$$

$$\zeta_{12}(h_{ll},h_{lc}) = \kappa_2^2\int_{-\infty}^{\infty}\left(\frac{m''(x)}{2}\right)\left(\frac{m''(x)}{2}+\frac{m'(x)f'(x)}{f(x)}\right)w(x)dxh_{ll}^2h_{lc}^2+\frac{\tilde{\upsilon}}{nh_{lc}}\int_{-\infty}^{\infty}\frac{\sigma^2(x)}{f(x)}w(x)dx,$$

and $\tilde{\upsilon}$ is defined in Theorem 1.

The weight λ which minimizes WIMSE is

$$\lambda^o(h_{ll},h_{lc}) = \frac{\zeta_2(h_{lc})-\zeta_{12}(h_{ll},h_{lc})}{\zeta_1(h_{ll})+\zeta_2(h_{lc})-2\zeta_{12}(h_{ll},h_{lc})}, \tag{3.4}$$

and the minimized WIMSE is

$$\frac{\zeta_1(h_{ll})\zeta_2(h_{lc})-\zeta_{12}(h_{ll},h_{lc})^2}{\zeta_1(h_{ll})+\zeta_2(h_{lc})-2\zeta_{12}(h_{ll},h_{lc})}. \tag{3.5}$$

The values of $\zeta_1(h_{ll})$ and $\zeta_2(h_{lc})$ in Theorem 2 represent the leading terms of WIMSE of the local linear and local constant estimators, respectively. From Theorem 2, the efficiency gain from averaging two estimators is

$$\frac{\left(\zeta_1(h_{ll})-\zeta_{12}(h_{ll},h_{lc})\right)^2}{\zeta_1(h_{ll})+\zeta_2(h_{lc})-2\zeta_{12}(h_{ll},h_{lc})} \tag{3.6}$$

compared with the local linear estimator and

$$\frac{\left(\zeta_2(h_{lc})-\zeta_{12}(h_{ll},h_{lc})\right)^2}{\zeta_1(h_{ll})+\zeta_2(h_{lc})-2\zeta_{12}(h_{ll},h_{lc})} \tag{3.7}$$

compared with the local constant estimator. Since the denominator of (3.6) and (3.7) is positive, the local weighted estimator is more efficient than both of the local linear and local constant estimators if $\zeta_1(h_{ll}) \neq \zeta_{12}(h_{ll},h_{lc})$ and $\zeta_2(h_{lc}) \neq \zeta_{12}(h_{ll},h_{lc})$. Furthermore, for any given h_{lc} and h_{ll}, the WIMSE of the local weighted estimator with the optimal global weight (3.4) is strictly smaller than the WIMSE of any linear combination of the local linear and local constant estimators.

The optimal global weight (3.4) is a function of h_{lc} and h_{ll} only. Also, the optimal global weight is independent of the sample size n and the conditional variance function $\sigma^2(x)$. Because the two bandwidths, h_{lc} and h_{ll}, have the same order, the rate of convergence of the optimal global weight

is $O(1)$, which does not depend on n, h_{lc}, and h_{ll}. If $h_{lc} = h_{ll} \equiv h$, then the optimal global weight can be simplified as

$$\lambda^o = \frac{\int_{-\infty}^{\infty} \big(\varphi_2(x)^2 - \varphi_1(x)\varphi_2(x)\big) w(x) dx}{\int_{-\infty}^{\infty} \big(\varphi_1(x) - \varphi_2(x)\big)^2 w(x) dx}$$

where $\varphi_1(x) = m''(x)/2$ and $\varphi_2(x) = m'(x)f'(x)/f(x) + m''(x)/2$, and the minimized WIMSE is

$$\frac{\upsilon}{nh} \int_{-\infty}^{\infty} \frac{\sigma^2(x)}{f(x)} w(x) dx$$

which is the weighted integrated variance of the local constant/linear estimator.

4 Cross-Validation

As shown in the previous section, the AMSE/WIMSE of local weighted estimator with the optimal local/global weights achieves a lower bound on the AMSE/WISME for any weighted average of the local constant and local linear estimators. We now develop two data-driven approaches to bandwidths and weights selection for the local weighted estimator. The method we propose is to minimize the leave-one-out cross-validation criterion. We show the leave-one-out cross-validation criterion is asymptotically equivalent to the WIMSE of the local weighted estimator. We also derive the rate of convergence of the cross-validated weights to the optimal global weights.

The first approach we propose is the two-step cross-validation method. The strategy is to choose the bandwidths in the first step and then choose the weights in the second step. Define the leave-one-out local linear, the leave-one-out local constant, and the leave-one-out local weighted estimators as follows:

$$\hat{m}_{ll,-i}(x_i) = \Big(\sum_{j \neq i} s_{j,-i} \Big)^{-1} \sum_{j \neq i} s_{j,-i} y_j, \tag{4.1}$$

$$\hat{m}_{lc,-i}(x_i) = \Big(\sum_{j \neq i} k\Big(\frac{x_j - x_i}{h}\Big) \Big)^{-1} \sum_{j \neq i} k\Big(\frac{x_j - x_i}{h}\Big) y_j, \tag{4.2}$$

$$\hat{m}_{lw,-i}(x) = \lambda \hat{m}_{ll,-i}(x) + (1-\lambda)\hat{m}_{lc,-i}(x), \tag{4.3}$$

where the subscript $-i$ denotes that we exclude the $i'th$ observation, $s_{j,-i} = k((x_j - x_i)/h)(T_{n,2,-i}(x_i) - (x_j - x_i)T_{n,1,-i}(x_i))$, and $T_{n,\ell,-i}(x_i) = \sum_{j \neq i} k((x_j - x_i)/h)(x_j - x_i)^\ell$ for $\ell = 1, 2$.

The two-step cross-validated local weighted estimator can be obtained by the following procedure. Let $M(x_i)$ be a nonnegative weight function that trims out boundary observations and

$$CV(h_{ll}) = n^{-1} \sum_{i=1}^{n} \big(y_i - \hat{m}_{ll,-i}(x_i)\big)^2 M(x_i), \tag{4.4}$$

$$CV(h_{lc}) = n^{-1} \sum_{i=1}^{n} \big(y_i - \hat{m}_{lc,-i}(x_i)\big)^2 M(x_i), \tag{4.5}$$

$$CV(\lambda, h_{ll}, h_{lc}) = n^{-1} \sum_{i=1}^{n} \big(y_i - \hat{m}_{lw,-i}(x)\big)^2 M(x_i), \tag{4.6}$$

be the cross-validation criterions for the local linear, the local constant, and the local weighted estimators. First, the bandwidths are chosen to minimize the cross-validation criterions as

$$\hat{h}_{ll} = \underset{h_{ll} \in \mathcal{H}_n}{\operatorname{argmin}} CV(h_{ll}), \tag{4.7}$$

$$\hat{h}_{lc} = \underset{h_{lc} \in \mathcal{H}_n}{\operatorname{argmin}} CV(h_{lc}), \tag{4.8}$$

where $\mathcal{H}_n = \{h : 0 < h < \eta(n), nh \geq \tau(n)\}$ and $\eta(n)$ is a positive sequence that goes to zero as $n \to \infty$ and $\tau(n)$ is a positive sequence that diverges to $+\infty$ as $n \to \infty$. Here we restrict $\hat{h}_{ll}$ and $\hat{h}_{ll}$ to lie in a shrinking set. This condition is also used in Hardle and Marron (1985).

For any given the cross-validated bandwidths, $\hat{h}_{ll}$ and $\hat{h}_{lc}$, the two-step cross-validated weight of the local weighted estimator is defined as

$$\hat{\lambda}(\hat{h}_{ll}, \hat{h}_{lc}) = \underset{\lambda \in \mathcal{L}_n}{\operatorname{argmin}} CV(\lambda, \hat{h}_{ll}, \hat{h}_{lc}), \tag{4.9}$$

where $\mathcal{L}_n = \{\lambda : \lambda \in (-\infty, \infty)\}$. If we only consider the positive weights, then we replace $\mathcal{L}_n$ as $\mathcal{L}_n^* = \{\lambda : \lambda \in [0, 1]\}$. Thus, the two-step cross-validated estimate can be obtained by

$$\hat{m}_{lw}(x) = \hat{\lambda}(\hat{h}_{ll}, \hat{h}_{lc})\hat{m}_{ll}(x) + (1 - \hat{\lambda}(\hat{h}_{ll}, \hat{h}_{lc}))\hat{m}_{lc}(x). \tag{4.10}$$

In the appendix, we show that the leading term of $CV(\lambda, h_{ll}, h_{lc})$ is given by

$$CV_0(\lambda, h_{ll}, h_{lc}) = \lambda^2 \zeta_1(h_{ll}) + (1 - \lambda)^2 \zeta_2(h_{lc}) + 2\lambda(1 - \lambda)\zeta_{12}(h_{ll}, h_{lc})$$

which is equivalent to the leading term of the WIMSE of the local weighted estimator given in Theorem 2. Let h_{ll}^o and h_{lc}^o be the values that minimize $\zeta_1(h_{ll})$ and $\zeta_1(h_{ll})$, respectively. Then it is easy to show that $h_{ll}^o = c_1 n^{-1/5}$ and $h_{lc}^o = c_2 n^{-1/5}$ where c_1 and c_2 are some constants defined in the appendix. Based on these results, we show that $\hat{\lambda}(\hat{h}_{ll}, \hat{h}_{lc}) = \lambda^o(h_{ll}^o, h_{lc}^o) + o_p(1)$.

We now present the rate of convergence of $\hat{\lambda}(\hat{h}_{ll}, \hat{h}_{lc})$ to $\lambda^o(h_{ll}^o, h_{lc}^o)$.

Assumption 6. $m(x)$, $f(x)$, and $\sigma^2(x)$ are all fourth order differentiable in $\mathcal{X}$. $e_i = y_i - m(x_i)$ has finite fourth moment.

Assumption 7. The kernel function $k(u)$ is m times differentiable. Define $k^{(j)}(u)$ as the jth order derivative of $k(u)$. Then the kernel function satisfies $\int_{-\infty}^{\infty} |u^j k^{(j)}(u)| du < \infty$ for all $j = 1, ..., m$, where $m > 6$ is a positive integer.

Assumption 6 imposes some standard moment and smoothness conditions. Assumption 7 is required to show that the remainder term of a Taylor expansion is negligible.

Theorem 3. *Under Assumptions 1-5 and Assumptions 6-7, if $h_{lc} \to 0$, $h_{ll} \to 0$, $nh_{lc} \to \infty$, and $nh_{ll} \to \infty$ as $n \to \infty$, then we have*

$$\hat{\lambda}(\hat{h}_{ll}, \hat{h}_{lc}) = \lambda^o(h_{ll}^o, h_{lc}^o) + O_p(n^{-3/10}).$$

Theorem 3 shows that the two-step cross-validated weight $\hat{\lambda}(\hat{h}_{ll}, \hat{h}_{lc})$ converges to the optimal global weight $\lambda^o(h_{ll}^o, h_{lc}^o)$ at the rate $O_p(n^{-3/10})$. This is exactly of the same order as the rate of convergence of the cross-validated bandwidths. Note that the two-step cross-validated weight is a function of the bandwidths. This result also implies that the rate of convergence of the two-step cross-validated weight is determined by the rate of convergence of the cross-validated bandwidths.

The second approach we propose is the joint cross-validation method. The strategy is to choose the bandwidths and weights simultaneously. In our two-step cross-validation method, we select the cross-validated weight given the cross-validated bandwidths fixed. Intuitively, by considering one more turning parameter λ, we can achieve a smaller minimum of the cross-validation function (4.6), which also implies a lower WIMSE of the local weighted estimator. Thus, the joint cross-validated parameters are defined as

$$\{\tilde{\lambda}, \tilde{h}_{ll}, \tilde{h}_{lc}\} = \underset{\lambda, h_{ll}, h_{lc}}{\operatorname{argmin}} CV(\lambda, h_{ll}, h_{lc}). \tag{4.11}$$

Let λ^*, h_{ll}^*, and h_{lc}^* denote the values that minimize the leading term of the WIMSE of the local weighted estimator $CV_0(\lambda, h_{ll}, h_{lc})$. Unlike the optimal global weight given in Theorem 2, there is no closed-form solution for λ^*, h_{ll}^*, and h_{lc}^*, and the solutions must be found numerically. Based on the fact that the leading term of $CV(\lambda, h_{ll}, h_{lc})$ is $CV_0(\lambda, h_{ll}, h_{lc})$, we expect that $\tilde{\lambda} \to_p \lambda^*$, $\tilde{h}_{ll} \to_p h_{ll}^*$, and $\tilde{h}_{lc} \to_p h_{lc}^*$. A rigorous proof of this result is beyond the scope of this paper.

5 Simulations

In this section we investigate the finite-sample behavior of the nonparametric regressors. We use the same design as Seifert and Gasser (2000) and Cheng, Peng, and Wu (2007). Three regression functions are considered:

1. peak: $m(x_i) = 2 - 5x_i + 5\exp(-400(x_i - 0.5)^2)$,
2. bimodal: $m(x_i) = 0.3\exp(-16(x_i - 0.25)^2) + 0.7\exp(-64(x_i - 0.75)^2)$,
3. sine: $m(x_i) = \sin(5\pi x_i)$.

In our experiment design, the covariate x follows normal distribution $N(0, 1)$, uniform distribution $U(0, 1)$, or mixed normal distribution $0.5N(-1, 1)+0.5N(1.75, 0.25)$. The random error e is normal distributed $N(0, \rho\sigma)$, where $\rho = 0.5$, 1, or 2 and $\sigma = \sqrt{0.5}$, 0.1, and 0.5 for the peak, bimodal, and sine regression functions. respectively.

We compare the performance of the following estimators: Cross-validated local constant estimator (labeled LC); Cross-validated local linear estimator (labeled LL); Two-step cross-validated local weighted estimator with unrestricted weights (labeled LW2SU); Two-step cross-validated local weighted estimator with positive weights (labeled LW2SP); Joint cross-validated local weighted estimator with unrestricted weights (labeled LWJU); Joint cross-validated local weighted estimator with positive weights (labeled LWJP); Least squares estimator (labeled LS). The bandwidths for the local constant and local linear estimators range over $\{0.008 \cdot 1.1^k, k = 1, \ldots, 60\}$. To save

computation time, we search the unrestricted weights between -3 and 3. For the joint cross-validation method, we use two-step cross-validated bandwidths and weights as the initial points. A second-order Gaussian Kernel is used in the simulations. We evaluate the finite sample performances via the mean and median of mean squared estimation error (MSEE) based on 1000 Monte Carlo replications. The mean squared estimation error is computed as MSEE $= n^{-1}\sum_{i=1}^{n}(m(x_i) - \hat{m}(x_i))^2$ where $m(x_i)$ is the regression function from true DGP and $\hat{m}(x_i)$ is the cross-validated kernel estimator or least squares estimator. Finally, the sample size is $n = (50, 100, 200, 400)$.

Table 1 through 12 summarize the results. In order to save space, we only report the results for $\rho = 1$. Other values of ρ have similar results. From Table 1 through 12, it is obvious that the least squares estimator has a poor performance compared to the kernel estimators in all simulations. Table 1 through 3 present the results for DGPs with $x \sim N(0,1)$ and $e \sim N(0,1)$. Table 1 illustrate the results when the regression function is linear with peak which is a case of data sparsity, see Seifert and Gasser (1996). When the data is sparse, the LL estimator performs worse than the LC estimator as we expect. Both LW2S and LWD have better performances than LL and LC. Examining Table 1 more closely, we observe that choosing the bandwidths and weights simultaneously and the restriction on the positive weights can improve the efficiency of the local weighted estimator. Table 2 presents the results when the regression function is bimodal. All kernel estimators have similar performance except for the LL estimator which has a larger mean and median MSEE. Table 3 presents the results when the regression function is a sine function. It is clear that the LC estimator does better than the LL estimator in the small sample sizes. Both LW2S and LWD outperform LL and LC.

Table 4 through 6 present the results for DGPs with $x \sim U(0,1)$ and $e \sim N(0,1)$. Asymptotically, when the covariate x is uniformly distributed, the performances of the LL, LC, and LW estimator are the same. However, Table 4 shows that the LL estimator has better performance than the LC estimator except for $n = 50$. The LW estimator achieves a lower mean and median of MSEE than the LL estimator. Among the LW estimators, LW2SP and LWJP show good performance over LW2SU and LWJU when the data is sparse. From table 6, we observe that when the regression function is bimodal and $x \sim U(0,1)$, LL and LW2SP do better than other kernel estimators. Table 7 through 9 show the results for DGPs with $x \sim 0.5N(-1,1) + 0.5N(1.75, 0.25)$ and $e \sim N(0,1)$. Overall, we have similar results as shown in Table 1-3 when x follows a standard normal distribution. Table 10 through 12 present the results for DGPs with $x \sim N(0,1)$ and $e \sim N(0,\sigma_i)$ where $\sigma_i = x_i^2$. When the error term is heteroskedastic, all estimators have larger mean and median MSEE. From Table 10 we observe that the LW estimator performs better than the LL and LC estimators. Among the LW estimators, LWJU and LWJP do better than LW2SU and LW2SP. Furthermore, LWJP performs slightly better than LWJU. Table 12 shows that the LW estimator does significantly better than the LL estimator. The LC estimator is slightly better than the LW estimator except for $n = 50$.

Next we consider the behavior of the cross-validated weights. The densities of the two-step cross-validated weights and joint cross-validated weights are displayed in Figures 3-6 for $n = 100$. In each figure, the three panels display regression functions. The solid line represents the two-step

cross-validated local weighted estimator with unrestricted weights, the dashed line represents the two-step cross-validated local weighted estimator with positive weights, the dotted lines represents the joint cross-validated local weighted estimator with unrestricted weights, and the dash-dotted lines represents the joint cross-validated local weighted estimator with positive weights.

In each panel, the unrestricted weights of the two-step cross-validation method and the joint cross-validation method are concentrated between 0 and 1 which implies that the restriction on the positive weights can improve the efficiency of the local weighted estimator. This also explains why the local weighted estimator with positive weights slightly outperforms the local weighted estimator with unrestricted weights in our simulation results. Figures 3-6 also show that the densities of the two-step cross-validated weights and the joint cross-validated are quite similar except for the case that x is uniformly distributed. Compared the densities of weights across the different regression functions, the local weighted estimator tends to choose larger weights on the local constant estimator for DGP_2, see Figure 3, 5, and 6.

6 Conclusion

This paper proposes an averaging kernel estimator for the nonparametric regression. The proposed estimator is a weighted average of the local constant and local linear estimators at each point of estimation. The goal in model averaging is to reduce the MSE and WIMSE of kernel estimators of regression functions. Two data-driven methods of bandwidths and weights selection are proposed, and we derive the rate of convergence of the cross-validated bandwidths and weights. Simulations show that the averaging estimator performs substantially better than the local constant and local linear estimators, especially when the data is sparse. One possible direction for future research is to apply the model averaging approach to the nonparametric conditional distribution estimation. Among the nonparametric conditional distribution estimators, the local constant estimator has many attractive properties including that it is non-negative and monotonicity preserving, while the local linear estimator has the advantages that it achieves minimax efficiency and adapts automatically to the boundary. Hence, it seems reasonable to consider averaging these two kernel estimators and applying the proposed MSE and WIMSE reduction approach.

Appendix

A Proofs of Theorems

Proof of Theorem 1. Suppose Assumptions 1 to 4 hold. Observe that

$$\begin{aligned}
\mathrm{E}(\hat{m}_{lw}(x)-m(x))^2 &= \lambda^2\mathrm{E}\big(\hat{m}_{ll}(x)-m(x)\big)^2+(1-\lambda)^2\mathrm{E}\big(\hat{m}_{lc}(x)-m(x)\big)^2 \\
&\quad +2\lambda(1-\lambda)\mathrm{E}\big((\hat{m}_{ll}(x)-m(x))(\hat{m}_{lc}(x)-m(x))\big) \\
&= \lambda^2\mathrm{E}\big(\hat{m}_{ll}(x)-m(x)\big)^2+(1-\lambda)^2\mathrm{E}\big(\hat{m}_{lc}(x)-m(x)\big)^2 \\
&\quad +2\lambda(1-\lambda)\Big(\mathrm{E}(\hat{m}_{ll}(x)-m(x))\mathrm{E}(\hat{m}_{lc}(x)-m(x)) \\
&\qquad +\mathrm{E}\big((\hat{m}_{ll}(x)-\mathrm{E}\hat{m}_{ll}(x))(\hat{m}_{lc}(x)-\mathrm{E}\hat{m}_{lc}(x))\big)\Big). \qquad \text{(A.1)}
\end{aligned}$$

Following Fan (1992, 1993), we have

$$\mathrm{E}(\hat{m}_{lc}(x)-m(x)) = \kappa_2\left(\frac{m''(x)}{2}+\frac{m'(x)f'(x)}{f(x)}\right)h_{lc}^2+o\left(h_{lc}^2\right) \qquad \text{(A.2)}$$

$$Var(\hat{m}_{lc}(x)) = \frac{\upsilon}{nh_{lc}}\frac{\sigma^2(x)}{f(x)}+o\left(\frac{1}{nh_{lc}}\right) \qquad \text{(A.3)}$$

$$\mathrm{E}(\hat{m}_{ll}(x)-m(x)) = \kappa_2\left(\frac{m''(x)}{2}\right)h_{ll}^2+o\left(h_{ll}^2\right) \qquad \text{(A.4)}$$

$$Var(\hat{m}_{ll}(x)) = \frac{\upsilon}{nh_{ll}}\frac{\sigma^2(x)}{f(x)}+o\left(\frac{1}{nh_{ll}}\right). \qquad \text{(A.5)}$$

By Lemma 1, we have

$$Cov\big(\hat{g}_{lc}(x),\hat{g}_{ll}(x)\big) = \frac{\tilde{\upsilon}}{nh_{lc}}\frac{\sigma^2(x)}{f(x)}+o\left(h_{lc}^2h_{ll}^2+\frac{1}{nh_{lc}}\right) \qquad \text{(A.6)}$$

where $\tilde{\upsilon}=\int_{-\infty}^{\infty}k(\gamma u)k(u)du$. Substituting (A.2)-(A.6) into (A.1), we have (3.1). The optimal local weight (3.2) and the minimized MSE (3.3) are found by minimizing (3.1) with respect to λ. ■

Proof of Theorem 2. The proof follows that of Theorem 1. ■

Proof of Theorem 3. Our proof is similar to that of Theorem 2.1 in Li and Racine (2004). In the following proof, we use the short hand notation $M_i=M(x_i)$, $f_i=f(x_i)$, $\hat{f}_i=\hat{f}(x_i)=(nh)^{-1}\sum_{j\neq i}k((x_j-x_i)/h)$, $K_{lc,ij}=h_{lc}^{-1}k((x_j-x_i)/h_{lc})$, and $K_{ll,ij}=h_{ll}^{-1}k((x_j-x_i)/h_{ll})$.

Define

$$CV(h_{ll},h_{lc}) = n^{-1}\sum_{i=1}^{n}\left(y_i-\hat{m}_{ll,-i}(x_i)\right)\left(y_i-\hat{m}_{lc,-i}(x_i)\right)M_i.$$

Then we can write the equation (4.6) as

$$CV(\lambda,h_{ll},h_{lc}) = \lambda^2CV(h_{ll})+(1-\lambda)^2CV(h_{lc})+2\lambda(1-\lambda)CV(h_{ll},h_{lc}). \qquad \text{(A.7)}$$

We first consider the first term of (A.7). By a standard Taylor series expansion of $m(x_j)$ at x_i, we have $m(x_j) = m(x_i) + (x_j - x_i)m'(x_i) + R_{ij}$, where $R_{ij} = m(x_j) - m(x_i) - (x_j - x_i)m'(x_i)$. Then we can write the equation (2.1) as

$$y_j = m(x_i) + (x_j - x_i)m'(x_i) + R_{ij} + e_j. \tag{A.8}$$

Substituting (A.8) into (4.1), we have

$$\begin{aligned}\hat{m}_{ll,-i}(x_i) &= m(x_i) + n^{-1}\sum_{j\neq i} K_{ll,ij}(R_{ij} + e_j)/\hat{f}_i + o_p(h_{ll}^2) \\ &= m(x_i) + r_{n,-i} + o_p(h_{ll}^2)\end{aligned}$$

where $r_{n,-i} = n^{-1}\sum_{j\neq i} K_{ll,ij}(R_{ij} + e_j)/\hat{f}_i$. Thus, the leave-one-out cross-validation criterion can be written as

$$\begin{aligned}CV(h_{ll}) &= n^{-1}\sum_{i=1}^{n}\left(y_i - \hat{m}_{ll,-i}(x_i)\right)^2 M_i \approx n^{-1}\sum_{i=1}^{n}\left(y_i - m(x_i) - r_{n,-i}\right)^2 M_i \\ &= n^{-1}\sum_{i=1}^{n} r_{n,-i}^2 M_i - 2n^{-1}\sum_{i=1}^{n} e_i r_{n,-i} M_i + n^{-1}\sum_{i=1}^{n} e_i^2 M_i.\end{aligned}$$

Define $CV_1(h_{ll}) = n^{-1}\sum_i r_{n,-i}^2 M_i - 2n^{-1}\sum_i e_i r_{n,-i} M_i$. Note that minimizing $CV(h_{ll})$ over h_{ll} is equivalent to minimizing $CV_1(h_{ll})$ since $n^{-1}\sum_{i=1}^{n} e_i^2 M_i$ is not related to h_{ll}. By some algebra, we have

$$\begin{aligned}CV_1(h_{ll}) = {} & n^{-3}\sum_i\sum_{j\neq i}\sum_{\ell\neq i} K_{ll,ij}K_{ll,i\ell}(R_{ij}R_{i\ell} + e_j e_\ell + 2R_{ij}e_\ell)M_i/\hat{f}_i^2 \\ & - 2n^{-2}\sum_i\sum_{j\neq i} K_{ll,ij}(e_i(R_{ij} + e_j))M_i/\hat{f}_i.\end{aligned} \tag{A.9}$$

Note that $\hat{f}_i - f_i = o_p(1)$ and

$$\frac{1}{\hat{f}_i} = \frac{1}{f_i} + \frac{(f_i - \hat{f}_i)}{f_i^2} + \frac{(f_i - \hat{f}_i)^2}{f_i^2\hat{f}_i}. \tag{A.10}$$

By replacing the random denominator $\hat{f}_i$ by f_i in (A.9), we have $CV_2(h_{ll}) = U_{11} + U_{12} + 2U_{13}$, where

$$\begin{aligned}U_{11} &= n^{-3}\sum_i\sum_{j\neq i}\sum_{\ell\neq i} K_{ll,ij}K_{ll,i\ell}R_{ij}R_{i\ell}M_i/f_i^2 \\ U_{12} &= n^{-3}\sum_i\sum_{j\neq i}\sum_{\ell\neq i} K_{ll,ij}K_{ll,i\ell}e_j e_\ell M_i/f_i^2 - 2n^{-2}\sum_i\sum_{j\neq i} K_{ll,ij}e_i e_j M_i/f_i \\ U_{13} &= n^{-3}\sum_i\sum_{j\neq i}\sum_{\ell\neq i} K_{ll,ij}K_{ll,i\ell}R_{ij}e_\ell M_i/f_i^2 - n^{-2}\sum_i\sum_{j\neq i} K_{ll,ij}R_{ij}e_i M_i/f_i.\end{aligned}$$

By the same argument, it follows that $CV_2(h_{lc}) = U_{21} + U_{22} + 2U_{23}$, where

$$
\begin{aligned}
U_{21} &= n^{-3} \sum_i \sum_{j \neq i} \sum_{\ell \neq i} K_{lc,ij} K_{lc,i\ell} Q_{ij} Q_{i\ell} M_i / f_i^2, \\
U_{22} &= n^{-3} \sum_i \sum_{j \neq i} \sum_{\ell \neq i} K_{lc,ij} K_{lc,i\ell} e_j e_\ell M_i / f_i^2 - 2n^{-2} \sum_i \sum_{j \neq i} K_{lc,ij} e_i e_j M_i / f_i, \\
U_{23} &= n^{-3} \sum_i \sum_{j \neq i} \sum_{\ell \neq i} K_{lc,ij} K_{lc,i\ell} Q_{ij} e_\ell M_i / f_i^2 - n^{-2} \sum_i \sum_{j \neq i} K_{lc,ij} Q_{ij} e_i M_i / f_i.
\end{aligned}
$$

and $Q_{ij} = m(x_j) - m(x_i)$. Also, we have $CV_2(h_{ll}, h_{lc}) = U_{31} + U_{32} + U_{33}$, where

$$
\begin{aligned}
U_{31} &= n^{-3} \sum_i \sum_{j \neq i} \sum_{\ell \neq i} K_{ll,ij} K_{lc,i\ell} R_{ij} Q_{i\ell} M_i / f_i^2, \\
U_{32} &= n^{-3} \sum_i \sum_{j \neq i} \sum_{\ell \neq i} K_{ll,ij} K_{lc,i\ell} e_j e_\ell M_i / f_i^2 \\
&\quad - n^{-2} \sum_i \sum_{j \neq i} K_{ll,ij} e_i e_j M_i / f_i - n^{-2} \sum_i \sum_{j \neq i} K_{lc,ij} e_i e_j M_i / f_i, \\
U_{33} &= n^{-3} \sum_i \sum_{j \neq i} \sum_{\ell \neq i} K_{ll,ij} K_{lc,i\ell} (R_{ij} e_\ell + Q_{i\ell} e_j) M_i / f_i^2 \\
&\quad - n^{-2} \sum_i \sum_{j \neq i} K_{ll,ij} R_{ij} e_i M_i / f_i - n^{-2} \sum_i \sum_{j \neq i} K_{lc,ij} Q_{ij} e_i M_i / f_i.
\end{aligned}
$$

By Lemma 5, 6, and 7, we have

$$
\begin{aligned}
U_{11} &= \kappa_2^2 \int_{-\infty}^{\infty} \left(\frac{m''(x)}{2} \right)^2 M(x) f(x) dx h_{ll}^4 + O(h_{ll}^6 + n^{-1/2} h_{ll}^4 + n^{-1} h_{ll}^{-1/2}) \\
U_{12} &= \frac{\upsilon}{n h_{ll}} \int_{-\infty}^{\infty} \sigma^2(x) M(x) dx + O(n^{-1} h_{ll}^{-1/2} + n^{-3/2} h_{ll}^{-1}) \\
U_{13} &= O_p(n^{-1/2} h_{ll}^2).
\end{aligned}
$$

If $M(x)f(x) = w(x)$, then we have

$$
\begin{aligned}
CV_2(h_{ll}) &= \kappa_2^2 \int_{-\infty}^{\infty} \left(\frac{m''(x)}{2} \right)^2 w(x) dx h_{ll}^4 + \frac{\upsilon}{n h_{ll}} \int_{-\infty}^{\infty} \frac{\sigma^2(x)}{f(x)} w(x) dx \\
&\quad + O(h_{ll}^6 + n^{-1/2} h_{ll}^2 + n^{-1} h_{ll}^{-1/2} + n^{-3/2} h_{ll}^{-1}) \\
&= \zeta_1(h_{ll}) + (s.o.).
\end{aligned} \tag{A.11}
$$

where $(s.o.)$ denotes the small order term. Similarly, by Lemma 8, 9, and 10, we have $CV_2(h_{lc}) = \zeta_2(h_{lc}) + (s.o.)$ and by Lemma 2, 3, and 4, we have $CV_2(h_{ll}, h_{lc}) = \zeta_{12}(h_{ll}, h_{lc}) + (s.o.)$. Therefore,

$$
\begin{aligned}
CV(\lambda, h_{ll}, h_{lc}) &= \lambda^2 \zeta_1(h_{ll}) + (1-\lambda)^2 \zeta_2(h_{lc}) + 2\lambda(1-\lambda) \zeta_{12}(h_{ll}, h_{lc}) + n^{-1} \sum_i e_i^2 M_i + (s.o.) \\
&= CV_0(\lambda, h_{ll}, h_{lc}) + \tilde{\sigma}^2 + (s.o.),
\end{aligned}
$$

where $CV_0(\lambda, h_{ll}, h_{lc}) = \lambda^2 \zeta_1(h_{ll}) + (1-\lambda)^2 \zeta_2(h_{lc}) + 2\lambda(1-\lambda) \zeta_{12}(h_{ll}, h_{lc})$ and $\tilde{\sigma}^2 = n^{-1} \sum_i e_i^2 M_i$.

Recall that the optimal bandwidth h_{ll}^o is the value that minimizes $\zeta_1(h_{ll})$. Then, we have $h_{ll}^o = c_1 n^{-1/5}$ where $c_1 = \big(\upsilon \int_{-\infty}^{\infty} \sigma^2(x)w(x)/f(x)dx\big)^{1/5}\big(\kappa_2^2 \int_{-\infty}^{\infty} m''(x)^2 w(x)dx\big)^{-1/5}$. Note that $\hat{h}_{ll}$ is the value that minimizes $CV(h_{ll})$ and $CV(h_{ll}) = \zeta_1(h_{ll}) + (s.o.)+$ terms not related to h_{ll}. Thus, we have $\hat{h}_{ll} = h_{ll}^o + o_p(h_{ll}^o)$. Similarly, we have $h_{lc}^o = c_2 n^{-1/5}$ where $c_2 = \big(\upsilon \int_{-\infty}^{\infty} \sigma^2(x)w(x)/f(x)dx\big)^{1/5} \cdot \big(\kappa_2^2 \int_{-\infty}^{\infty}(2m'(x)f'(x)/f(x) + m''(x))^2 w(x)dx\big)^{-1/5}$ and $\hat{h}_{lc} = h_{lc}^o + o_p(h_{lc}^o)$. Let $\lambda^o(h_{ll}, h_{lc})$ defined in (3.4) be the value that minimizes $CV_0(\lambda, h_{ll}, h_{lc})$ for given h_{ll} and h_{lc}. Let $\hat{\lambda}(h_{ll}, h_{lc})$ be the value that minimizes $CV(\lambda, h_{ll}, h_{lc})$ for given h_{ll} and h_{lc}. Noting that $CV(\lambda, h_{ll}, h_{lc}) = CV_0(\lambda, h_{ll}, h_{lc}) + (s.o.)+$ terms not related to λ, we have $\hat{\lambda}(h_{ll}, h_{lc}) = \lambda^o(h_{ll}, h_{lc}) + o_p(\lambda^o(h_{ll}, h_{lc}))$. Because $\lambda^o(h_{ll}^o, h_{lc}^o) = O(1)$, $\hat{h}_{ll} = h_{ll}^o + o_p(h_{ll}^o)$, and $\hat{h}_{lc} = h_{lc}^o + o_p(h_{lc}^o)$, we have $\hat{\lambda}(\hat{h}_{ll}, \hat{h}_{lc}) = \lambda^o(h_{ll}^o, h_{lc}^o) + o_p(1)$.

We now derive the rate of convergence of $\hat{\lambda}(\hat{h}_{ll}, \hat{h}_{lc}) - \lambda^o(h_{ll}^o, h_{lc}^o)$. By Lemmas 2-10 and the fact that $h_{ll}^o = c_1 n^{-1/5}$ and $h_{lc}^o = c_1 n^{-1/5}$, we have

$$CV(\lambda, h_{ll}, h_{lc}) = CV_0(\lambda, h_{ll}, h_{lc}) + O\Big(\lambda^2 n^{-1} h_{ll}^{-1/2} + \lambda^2 n^{-1} h_{lc}^{-1/2} + \lambda^2 n^{-1} h_{ll}^{-1/4} h_{lc}^{-1/4}\Big) + \tilde{\sigma}^2 + (s.o.).$$

Minimizing $CV(\lambda, h_{ll}, h_{lc})$ over λ, it follows that

$$\hat{\lambda}(h_{ll}, h_{lc}) = \lambda^o(h_{ll}, h_{lc}) + \lambda^o(h_{ll}, h_{lc}) O_p\left(n^{-1} h_{ll}^{-1/2} + n^{-1} h_{lc}^{-1/2} + n^{-1} h_{ll}^{-1/4} h_{lc}^{-1/4}\right) + (s.o.).$$

Hardle, Hall, and Marron (1988) show that $\hat{h}_{lc} = h_{lc}^o + O_p(n^{-3/10})$. Li and Racine (2004) show that $\hat{h}_{ll} = h_{ll}^o + O_p(n^{-3/10})$. Therefore, we can show that $\lambda^o(\hat{h}_{ll}, \hat{h}_{lc}) = \lambda^o(h_{ll}^o, h_{lc}^o) + O_p(n^{-3/10})$. Then, based on the facts that $\lambda^o(h_{ll}^o, h_{lc}^o) = O(1)$ and $\lambda^o(\hat{h}_{ll}, \hat{h}_{lc}) = O(1)$, we have

$$\begin{aligned}\hat{\lambda}(\hat{h}_{ll}, \hat{h}_{lc}) &= \lambda^o(h_{ll}^o, h_{lc}^o) + \Big(\lambda^o(\hat{h}_{ll}, \hat{h}_{lc}) - \lambda^o(h_{ll}^o, h_{lc}^o)\Big) + O_p(n^{-9/10}) \\ &= \lambda^o(h_{ll}^o, h_{lc}^o) + O_p(n^{-3/10}).\end{aligned}$$

This completes the proof. ■

B Lemmas

Lemma 1. *Under Assumptions 1-4, if $h_{lc} \to 0$, $h_{ll} \to 0$, $nh_{lc} \to \infty$, and $nh_{ll} \to \infty$ as $n \to \infty$, then the covariance between the local constant and local linear estimators is given by*

$$Cov\big(\hat{m}_{lc}(x), \hat{m}_{ll}(x)\big) = \frac{1}{nh_{lc}} \frac{\sigma^2(x)}{f(x)} \int_{-\infty}^{\infty} k(\gamma u)k(u)du + o\left(h_{lc}^2 h_{ll}^2 + \frac{1}{nh_{lc}}\right).$$

Proof. In the following proof, we use the short hand notation $k_i = k\left((x_i - x)/h\right)$. Note that $y_i = m(x_i) + e_i = m(x) + (m(x_i) - m(x)) + e_i$. Then, we can rewrite equation (2.4) and equation (2.6) as

$$\begin{aligned}\hat{m}_{lc}(x) &= m(x) + \frac{\sum_{i=1}^n k_i(m(x_i) - m(x))}{\sum_{i=1}^n k_i} + \frac{\sum_{i=1}^n k_i e_i}{\sum_{i=1}^n k_i}, \\ \hat{m}_{ll}(x) &= m(x) + \frac{\sum_{i=1}^n s_i(m(x_i) - m(x))}{\sum_{i=1}^n s_i} + \frac{\sum_{i=1}^n s_i e_i}{\sum_{i=1}^n s_i}.\end{aligned}$$

Then, using mean and variance decomposition, the covariance between the local constant estimator and the local linear estimator can be expanded into the following four terms,

$$Cov(\hat{m}_{lc}(x), \hat{m}_{ll}(x)) = \Omega_1 + \Omega_2 + \Omega_3 + \Omega_4$$

where

$$\begin{aligned}
\Omega_1 &= Cov\left(\frac{\sum_{i=1}^n k_i(m(x_i)-m(x))}{\sum_{i=1}^n k_i}, \frac{\sum_{i=1}^n s_i(m(x_i)-m(x))}{\sum_{i=1}^n s_i}\right), \\
\Omega_2 &= Cov\left(\frac{\sum_{i=1}^n k_i e_i}{\sum_{i=1}^n k_i}, \frac{\sum_{i=1}^n s_i e_i}{\sum_{i=1}^n s_i}\right), \\
\Omega_3 &= Cov\left(\frac{\sum_{i=1}^n k_i(m(x_i)-m(x))}{\sum_{i=1}^n k_i}, \frac{\sum_{i=1}^n s_i e_i}{\sum_{i=1}^n s_i}\right), \\
\Omega_4 &= Cov\left(\frac{\sum_{i=1}^n k_i e_i}{\sum_{i=1}^n k_i}, \frac{\sum_{i=1}^n s_i(m(x_i)-m(x))}{\sum_{i=1}^n s_i}\right).
\end{aligned}$$

First, consider the third term Ω_3. Using mean and variance decomposition and taking conditional expectations on x_i, we have

$$\begin{aligned}
\Omega_3 &= \mathrm{E}\left(\frac{\sum_{i=1}^n k_i(m(x_i)-m(x))\sum_{i=1}^n s_i \mathrm{E}(e_i|x_i)}{\sum_{i=1}^n k_i \sum_{i=1}^n s_i}\right) \\
&\quad - \mathrm{E}\left(\frac{\sum_{i=1}^n k_i(m(x_i)-m(x))}{\sum_{i=1}^n k_i}\right)\mathrm{E}\left(\frac{\sum_{i=1}^n s_i \mathrm{E}(e_i|x_i)}{\sum_{i=1}^n s_i}\right) \\
&= 0. \qquad \text{(B.1)}
\end{aligned}$$

where the last equality follows from $\mathrm{E}(e_i|x_i) = 0$ and the law of iterated expectations. By the same argument, we have $\Omega_4 = 0$.

Next, consider the first term Ω_1. Following Fan (1993), denote $X_n = O_r(a_n)$ if $\mathrm{E}|X_n|^r = O(a_n^r)$ and $X_n = o_r(a_n)$ if $\mathrm{E}|X_n|^r = o(a_n^r)$. By using the method of the kernel density estimate, it follows that

$$\left(\sum_{i=1}^n k_i + n^{-2}\right)^{-1} = (nh_{lc}f(x))^{-1} + o_2\left((nh_{lc})^{-1}\right), \qquad \text{(B.2)}$$

$$\left(\sum_{i=1}^n s_i + n^{-2}\right)^{-1} = \left(n^2 h_{ll}^4 \kappa_2 f^2(x)\right)^{-1} + o_4\left((n^2 h_{ll}^4)^{-1}\right), \qquad \text{(B.3)}$$

$$\sum_{i=1}^n k_i(m(x_i)-m(x)) = nh_{lc}^3 \kappa_2 f(x)\left(\frac{m'(x)f'(x)}{f(x)} + \frac{m''(x)}{2}\right) + o_2(nh_{lc}^3), \qquad \text{(B.4)}$$

$$\sum_{i=1}^n s_i(m(x_i)-m(x)) = n^2 h_{ll}^6 \kappa_2 f(x) S_n + o_4(n^2 h_{ll}^6), \qquad \text{(B.5)}$$

where $\kappa_2 = \int_{-\infty}^{\infty} u^2 k(u)du$ and $S_n = h_{ll}^{-3}\mathrm{E}(m(x_i)-m(x)-m'(x)(x_i-x))k\left((x_i-x)/h_{ll}\right)$. Here we use $\sum_{i=1}^n s_i + n^{-2}$ instead of $\sum_{i=1}^n s_i$ to avoid zero in the denominator, see Fan (1993) for detail.

By the mean and variance decomposition and using (B.2), (B.3), (B.4), and (B.5), we have

$$\begin{aligned}\Omega_1 &= \mathrm{E}\left(\frac{\sum_{i=1}^n k_i(m(x_i)-m(x))\sum_{i=1}^n s_i(m(x_i)-m(x))}{(\sum_{i=1}^n k_i+n^{-2})(\sum_{i=1}^n s_i+n^{-2})}\right)\\ &\quad - \mathrm{E}\left(\frac{\sum_{i=1}^n k_i(m(x_i)-m(x))}{\sum_{i=1}^n k_i+n^{-2}}\right)\mathrm{E}\left(\frac{\sum_{i=1}^n s_i(m(x_i)-m(x))}{\sum_{i=1}^n s_i+n^{-2}}\right)\\ &= h_{lc}^2 h_{ll}^2\kappa_2\frac{S_n}{f(x)}\left(\frac{m'(x)f'(x)}{f(x)}+\frac{m''(x)}{2}\right)+o(h_{lc}^2h_{ll}^2)\\ &\quad -\left(h_{lc}^2\kappa_2\left(\frac{m'(x)f'(x)}{f(x)}+\frac{m''(x)}{2}\right)+o(h_{lc}^2)\right)\left(h_{ll}^2\frac{S_n}{f(x)}+o(h_{ll}^2)\right)\\ &= o(h_{lc}^2h_{ll}^2) \end{aligned} \tag{B.6}$$

Now, consider the second term Ω_2. After some calculation and the law of iterated expectations, the second term can be simplified as

$$\begin{aligned}\Omega_2 &= \mathrm{E}\left(\frac{(\sum_{i=1}^n k_ie_i)(\sum_{i=1}^n s_ie_i)}{(\sum_{i=1}^n k_i+n^{-2})(\sum_{i=1}^n s_i+n^{-2})}\right)-\mathrm{E}\left(\frac{\sum_{i=1}^n k_ie_i}{\sum_{i=1}^n k_i+n^{-2}}\right)\mathrm{E}\left(\frac{\sum_{i=1}^n s_ie_i}{\sum_{i=1}^n s_i+n^{-2}}\right)\\ &= \mathrm{E}\left(\frac{\sum_{i=1}^n k_is_ie_i^2}{(\sum_{i=1}^n k_i+n^{-2})(\sum_{i=1}^n s_i+n^{-2})}+\frac{\sum_{i\neq j}^n k_is_je_ie_j}{(\sum_{i=1}^n k_i+n^{-2})(\sum_{i=1}^n s_i+n^{-2})}\right)\\ &= \mathrm{E}\left(\frac{\sum_{i=1}^n k_is_i\sigma^2(x_i)}{(\sum_{i=1}^n k_i+n^{-2})(\sum_{i=1}^n s_i+n^{-2})}\right)\end{aligned}$$

where $\sigma^2(x_i)=\mathrm{E}(e_i^2|x_i)$ is the conditional variance function. Recall that

$$\sum_{i=1}^n k_is_i\sigma^2(x_i)=\sum_{i=1}^n k\left(\frac{x_i-x}{h_{lc}}\right)k\left(\frac{x_i-x}{h_{ll}}\right)\Big(T_{n,2}(x)-(x_i-x)T_{n,1}(x)\Big)\sigma^2(x_i), \tag{B.7}$$

$$T_{n,\ell}(x)=\sum_{i=1}^n k\left(\frac{x_i-x}{h_{ll}}\right)(x_i-x)^\ell, \quad \ell=1,2. \tag{B.8}$$

Suppose $h_{ll}/h_{lc}\to\gamma$ as $n\to\infty$ where $-\infty<\gamma<\infty$. By the method of the kernel density estimate, we have

$$\sum_{i=1}^n k\left(\frac{x_i-x}{h_{ll}}\right)(x_i-x)^\ell = nh_{ll}^{\ell+1}f(x)\int_{-\infty}^{\infty}u^\ell k(u)du+o_4(nh_{ll}^{\ell+1}), \tag{B.9}$$

$$\sum_{i=1}^n k\left(\frac{x_i-x}{h_{lc}}\right)k\left(\frac{x_i-x}{h_{ll}}\right) = nh_{ll}f(x)\int_{-\infty}^{\infty}k(\gamma u)k(u)du+o_2\left(nh_{ll}\right). \tag{B.10}$$

Substituting (B.9) and (B.10) into (B.7), we have

$$\sum_{i=1}^n k_is_i\sigma^2(x_i)=n^2h_{ll}^4\kappa_2f^2(x)\sigma^2(x)\int_{-\infty}^{\infty}k(\gamma u)k(u)du+o_4(n^2h_{ll}^4) \tag{B.11}$$

Thus, by (B.2), (B.3), and (B.11), it follows that

$$\Omega_2=\frac{1}{nh_{lc}}\frac{\sigma^2(x)}{f(x)}\int_{-\infty}^{\infty}k(\gamma u)k(u)du+o\left((nh_{lc})^{-1}\right). \tag{B.12}$$

The lemma follows from (B.1), (B.6) and (B.12). ■

Lemma 2. *Under Assumptions 1-4 and Assumptions 6-7, if* $h_{lc} \to 0$, $h_{ll} \to 0$, $nh_{lc} \to \infty$, *and* $nh_{ll} \to \infty$ *as* $n \to \infty$, *then* $U_{31} = \kappa_2^2 \int_{-\infty}^{\infty} (m''(x)/2)(m''(x)/2 + m'(x)f'(x)/f(x))M(x)f(x)dx h_{ll}^2 h_{lc}^2 + O(h_{ll}^3 h_{lc}^3 + n^{-1/2} h_{ll}^2 h_{lc}^2 + n^{-1} h_{ll}^{-1/4} h_{lc}^{-1/4})$.

Proof. Note that $U_{31} = n^{-3} \sum_i \sum_{j \neq i} \sum_{\ell \neq i} K_{ll,ij} K_{lc,i\ell} R_{ij} Q_{i\ell} M_i / f_i^2 = U_{31a} + U_{31b}$ where

$$U_{31a} = n^{-3} \sum \sum \sum_{i \neq j \neq \ell} K_{ll,ij} K_{lc,i\ell} R_{ij} Q_{i\ell} M_i / f_i^2$$
$$U_{31b} = n^{-3} \sum_i \sum_{j \neq i} K_{ll,ij} K_{lc,ij} R_{ij} Q_{ij} M_i / f_i^2.$$

We first consider the first term U_{31a}. Define $U_{31a} = n^{-3} \sum \sum \sum_{i \neq j \neq \ell} H_{31a}(x_i, x_j, x_\ell)$ where $H_{31a}(x_i, x_j, x_\ell) = (1/3)(K_{ll,ij} K_{lc,i\ell} R_{ij} Q_{i\ell} M_i / f_i^2 + K_{ll,ji} K_{lc,j\ell} R_{ji} Q_{j\ell} M_j / f_j^2 + K_{ll,\ell j} K_{lc,\ell i} R_{lj} Q_{\ell i} M_\ell / f_\ell^2)$ is a symmetrized version of $K_{ll,ij} K_{lc,i\ell} R_{ij} Q_{i\ell} M_i / f_i^2$. Note that

$$\begin{aligned} &\mathrm{E}[H_{31a}(x_i, x_j, x_\ell)|x_i] \sim \mathrm{E}[K_{ll,ij} K_{lc,i\ell} R_{ij} Q_{i\ell} M_i / f_i^2 | x_i] \\ &\quad = \mathrm{E}[K_{ll,ij} R_{ij} | x_i] \mathrm{E}[K_{lc,i\ell} Q_{i\ell} | x_i] M_i / f_i^2 = \mathrm{E}[K_{ll,ij} R_{ij} / f_i | x_i] \mathrm{E}[K_{lc,i\ell} Q_{i\ell} / f_i | x_i] M_i. \end{aligned} \tag{B.13}$$

By Assumption 6 and the method of the kernel density estimate we have, uniformly in i,

$$\begin{aligned} \mathrm{E}[K_{ll,ij} R_{ij} / f_i | x_i] &= \mathrm{E}[K_{ll,ij}(m(x_j) - m(x_i) - (x_j - x_i)m'(x_i))/f_i | x_i] \\ &= \kappa_2 (m''(x)/2) h_{ll}^2 + O(h_{ll}^3), \end{aligned} \tag{B.14}$$

$$\begin{aligned} \mathrm{E}[K_{lc,i\ell} Q_{i\ell} / f_i | x_i] &= \mathrm{E}[K_{lc,i\ell}(m(x_\ell) - m(x_i))/f_i | x_i] \\ &= \kappa_2 (m''(x)/2 + m'(x)f'(x)/f(x)) h_{lc}^2 + O(h_{lc}^3). \end{aligned} \tag{B.15}$$

Plugging (B.14) and (B.15) into (B.13), we have

$$\mathrm{E}[H_{31a}(x_i, x_j, x_\ell)|x_i] = \kappa_2^2 (m''(x)/2)(m''(x)/2 + m'(x)f'(x)/f(x))M(x) h_{ll}^2 h_{lc}^2 + O(h_{ll}^3 h_{lc}^3). \tag{B.16}$$

Also by the law of iterated expectations, we have

$$\begin{aligned} &\mathrm{E}[H_{31a}(x_i, x_j, x_\ell)] = \mathrm{E}[\mathrm{E}[H_{31a}(x_i, x_j, x_\ell)|x_i]] \\ &\quad = \mathrm{E}\big[\kappa_2^2 (m''(x)/2)(m''(x)/2 + m'(x)f'(x)/f(x))M(x) h_{ll}^2 h_{lc}^2\big] + O(h_{ll}^3 h_{lc}^3) \\ &\quad = \kappa_2^2 \int_{-\infty}^{\infty} (m''(x)/2)(m''(x)/2 + m'(x)f'(x)/f(x))M(x)f(x)dx h_{ll}^2 h_{lc}^2 + O(h_{ll}^3 h_{lc}^3). \end{aligned} \tag{B.17}$$

By the U-statistic H-decomposition and using (B.14), (B.15), (B.16), and (B.17), we have

$$\begin{aligned} U_{31a} &= \mathrm{E}[H_{31a}(x_i, x_j, x_\ell)] + 3n^{-1} \sum_i \big(\mathrm{E}[H_{31a}(x_i, x_j, x_\ell|x_i)] - \mathrm{E}[H_{31a}(x_i, x_j, x_\ell)]\big) + (s.o.) \\ &= \mathrm{E}[H_{31a}(x_i, x_j, x_\ell)] + n^{-1/2} O(h_{ll}^2 h_{lc}^2) + n^{-1} O(h_{ll}^{-1/4} h_{lc}^{-1/4}) \\ &= \kappa_2^2 \int_{-\infty}^{\infty} (m''(x)/2)(m''(x)/2 + m'(x)f'(x)/f(x))M(x)f(x)dx h_{ll}^2 h_{lc}^2 \\ &\quad + O(h_{ll}^3 h_{lc}^3 + n^{-1/2} h_{ll}^2 h_{lc}^2 + n^{-1} h_{ll}^{-1/4} h_{lc}^{-1/4}). \end{aligned} \tag{B.18}$$

Note that the last term ($s.o.$) is the small order term because the last term in the decomposition is a degenerate U-statistic, see Li and Racine (2004) for detail.

Next, we consider U_{31b}. Define $H_{31b}(x_i, x_j) = (1/2)K_{ll,ij}K_{lc,ij}R_{ij}Q_{ij}(M_i/f_i^2 + M_j/f_j^2)$ as a symmetrized version of $K_{ll,ij}K_{lc,ij}R_{ij}Q_{ij}M_i/f_i^2$. Then $U_{31b} = n^{-1}(n^{-2}\sum\sum_{j\neq i} H_{31b}(x_i, x_j))$. By Assumption 6, (B.13), (B.14) and (B.15), we have, uniformly in i,

$$\begin{aligned}\mathrm{E}[H_{31b}(x_i, x_j)|x_i] &= \mathrm{E}[K_{ll,ij}K_{lc,ij}R_{ij}Q_{ij}M_i/f_i^2|x_i] = O(h_{ll}^2 h_{lc}^2)\\ \mathrm{E}[H_{31b}(x_i, x_j)] &= \mathrm{E}[\mathrm{E}[H_{31b}(x_i, x_j)|x_i]] = O(h_{ll}^2 h_{lc}^2)\end{aligned}$$

Thus, by the U-statistic H-decomposition, we have

$$\begin{aligned}U_{31b} &= n^{-1}\Big(\mathrm{E}[H_{31b}(x_i, x_j)] + 2n^{-1}\sum_i \big(\mathrm{E}[H_{31b}(x_i, x_j)|x_i] - \mathrm{E}[H_{31b}(x_i, x_j)]\big) + (s.o.)\Big)\\ &= O(n^{-1}h_{ll}^2 h_{lc}^2). \qquad \text{(B.19)}\end{aligned}$$

The Lemma follows from (B.18) and (B.19). ■

Lemma 3. *Under Assumptions 1-4 and Assumptions 6-7, if $h_{lc} \to 0$, $h_{ll} \to 0$, $nh_{lc} \to \infty$, and $nh_{ll} \to \infty$ as $n \to \infty$, then $U_{32} = \tilde{v}/(nh_{lc})\int_{-\infty}^{\infty}\sigma^2(x)M(x)dx + O(n^{-1}h_{ll}^{-1/4}h_{lc}^{-1/4} + n^{-3/2}h_{lc}^{-1})$.*

Proof. Note that $U_{32} = n^{-3}\sum_i\sum_{j\neq i}\sum_{\ell\neq i}K_{ll,ij}K_{lc,i\ell}e_j e_\ell M_i/f_i^2 - n^{-2}\sum_i\sum_{j\neq i}K_{ll,ij}e_i e_j M_i/f_i - n^{-2}\sum_i\sum_{j\neq i}K_{lc,ij}e_i e_j M_i/f_i = U_{32a} + U_{32b} - U_{32c} - U_{32d}$ where

$$\begin{aligned}U_{32a} &= n^{-3}\sum_i\sum_{j\neq i}K_{ll,ij}K_{lc,ij}e_j^2 M_i/f_i^2\\ U_{32b} &= n^{-3}\sum\sum\sum_{i\neq j\neq \ell}K_{ll,ij}K_{lc,i\ell}e_j e_\ell M_i/f_i^2\\ U_{32c} &= n^{-2}\sum_i\sum_{j\neq i}K_{ll,ij}e_i e_j M_i/f_i\\ U_{32d} &= n^{-2}\sum_i\sum_{j\neq i}K_{lc,ij}e_i e_j M_i/f_i.\end{aligned}$$

We first consider U_{32a}. Define $H_{32a}(z_i, x_j) = (1/2)K_{ll,ij}K_{lc,ij}(e_i^2 M_i/f_i^2 + e_j^2 M_j/f_j^2)$ as a symmetrized version of $K_{ll,ij}K_{lc,ij}e_j^2 M_i/f_i^2$. Then $U_{32a} = n^{-1}[n^{-2}\sum\sum_{j\neq i}H_{32a}(x_i, x_j)]$. By Assumption 6 and the method of the kernel density estimate we have,

$$\begin{aligned}\mathrm{E}[H_{32a}(x_i, x_j)|x_i] &= (1/2)\big((e_i^2/f_i^2)\mathrm{E}[K_{ll,ij}K_{lc,ij}|x_i]M_i + \mathrm{E}[K_{ll,ij}K_{lc,ij}e_j^2 M_j/f_j^2|x_i]\big)\\ &= (1/2)(e_i^2/f_i^2)\mathrm{E}[K_{ll,ij}K_{lc,ij}|x_i]M_i + (1/2)\mathrm{E}[\sigma^2(x_j)K_{ll,ij}K_{lc,ij}M_j/f_j^2|x_i]\\ &= (1/2)(\tilde{v}/h_{lc})\big(e_i^2 + \sigma^2(x_i)\big)M_i/f_i + O(h_{ll}), \qquad \text{(B.20)}\end{aligned}$$

and

$$\mathrm{E}[H_{32a}(x_i, x_j)] = \mathrm{E}[\mathrm{E}[H_{32a}(x_i, x_j)|x_i]] = \tilde{v}/h_{lc}\int_{-\infty}^{\infty}\sigma^2(x)M(x)dx + O(h_{ll}). \qquad \text{(B.21)}$$

Then by the U-statistic H-decomposition and using (B.20) and (B.21), we have

$$\begin{aligned} U_{32a} &= n^{-1}\Big(\mathrm{E}[H_{32a}(x_i,x_j)] + 2n^{-1}\sum_i \big(\mathrm{E}[H_{32b}(x_i,x_j)|x_i] - \mathrm{E}[H_{32a}(x_i,x_j)]\big) + (s.o.)\Big) \\ &= n^{-1}\mathrm{E}[H_{32a}(x_i,x_j)] + n^{-1/2}O((nh_{lc})^{-1}) \\ &= \tilde{v}/(nh_{lc})\int_{-\infty}^{\infty}\sigma^2(x)M(x)dx + O(n^{-1}h_{ll} + n^{-3/2}h_{lc}^{-1}) \end{aligned} \tag{B.22}$$

where $\tilde{v} = \int_{-\infty}^{\infty} k(\gamma u)k(u)du$ and $n^{-1/2}O((nh_{lc})^{-1})$ comes from U-statistic H-decomposition.

Next, we consider the second term U_{32b}. Define $U_{32b} = n^{-3}\sum\sum\sum_{i\neq j\neq \ell} H_{32b}(x_i,x_j,x_\ell)$ where $H_{32b}(x_i,x_j,x_\ell) = (1/3)(K_{ll,ij}K_{lc,i\ell}e_je_\ell M_i/f_i^2 + K_{ll,ji}K_{lc,j\ell}e_\ell e_i M_j/f_j^2 + K_{ll,\ell j}K_{lc,i\ell}e_ie_jM_\ell/f_\ell^2)$ is a symmetrized version of $K_{ll,ij}K_{lc,i\ell}e_je_\ell M_i/f_i^2$. Since $\mathrm{E}[e_\ell|x_j] = 0$, we have

$$\mathrm{E}[H_{32b}(x_i,x_j,x_\ell)|x_j] = 0.$$

Then, U_{32b} is a second-order degenerate U-statistics with

$$\mathrm{E}[H_{32b}(x_i,x_j,x_\ell)|x_i,x_j] = (1/3)e_ie_jM_\ell\mathrm{E}[K_{ll,\ell j}K_{lc,\ell i}/f_\ell^2|x_i,x_j].$$

By Assumption 6 and the method of the kernel density estimate we have,

$$\mathrm{E}[K_{ll,\ell j}K_{lc,\ell i}/f_\ell^2|x_i,x_j] = \tilde{v}/(h_{lc}f(x)) + O(h_{ll}h_{lc}).$$

By the U-statistic H-decomposition, we have

$$\begin{aligned} U_{32b} &= 3\Big(n^{-2}\sum_i\sum_{j\neq i}\mathrm{E}[H_{32b}(x_i,x_j,x_\ell)|x_i,x_j] + (s.o.)\Big) \\ &= n^{-2}\sum_i\sum_{j\neq i} e_ie_jM_\ell\mathrm{E}[K_{ll,\ell j}K_{lc,\ell i}/f_\ell^2|x_i,x_j] + (s.o.) \\ &= n^{-2}(\tilde{v}/h_{lc})\sum_i\sum_{j\neq i} e_ie_jM_i/f_i + (s.o.) = n^{-1}h_{ll}^{-1/4}h_{lc}^{-1/4}\mathcal{Z}_{32b,i} + (s.o.) \\ &= O(n^{-1}h_{ll}^{-1/4}h_{lc}^{-1/4}), \end{aligned} \tag{B.23}$$

where $\mathcal{Z}_{32b,i} = (nh_{ll}^{1/4}h_{lc}^{1/4})(n^{-2}(\tilde{v}/h_{lc})\sum_i\sum_{j\neq i}e_ie_j/f_i)$ is a mean zero $O_p(1)$ random variable. Similarly, we can show that

$$U_{32c} = n^{-2}\sum_i\sum_{j\neq i}K_{ll,ij}e_ie_jM_i/f_i = n^{-1}h_{ll}^{-1/4}h_{lc}^{-1/4}\mathcal{Z}_{32c,i} + (s.o.), \tag{B.24}$$

$$U_{32d} = n^{-2}\sum_i\sum_{j\neq i}K_{lc,ij}e_ie_jM_i/f_i = n^{-1}h_{ll}^{-1/4}h_{lc}^{-1/4}\mathcal{Z}_{32d,i} + (s.o.), \tag{B.25}$$

where $\mathcal{Z}_{32c,i}$ and $\mathcal{Z}_{32d,i}$ are mean zero $O_p(1)$ random variables.

The Lemma follows from (B.22), (B.23), (B.24), and (B.25). ■

Lemma 4. *Under Assumptions 1-4 and Assumptions 6-7, if $h_{lc}\to 0$, $h_{ll}\to 0$, $nh_{lc}\to\infty$, and $nh_{ll}\to\infty$ as $n\to\infty$, then $U_{33} = O_p(n^{-1/2}h_{ll}h_{lc} + n^{-1/2}h_{ll}^2 + n^{-1/2}h_{lc}^2)$.*

Proof. $U_{33} = n^{-3}\sum_i \sum_{j\neq i}\sum_{\ell\neq i} K_{ll,ij}K_{lc,i\ell}(R_{ij}e_\ell + Q_{i\ell}e_j)M_i/f_i^2 - n^{-2}\sum_i\sum_{j\neq i} K_{ll,ij}R_{ij}e_iM_i/f_i - n^{-2}\sum_i\sum_{j\neq i}K_{lc,ij}Q_{ij}e_iM_i/f_i = U_{33a} + U_{33b} + U_{33c} + U_{33d} - U_{33e} - U_{33f}$ where

$$
\begin{aligned}
U_{33a} &= n^{-3}\sum_i\sum_{j\neq i} K_{ll,ij}K_{lc,ij}R_{ij}e_jM_i/f_i^2 \\
U_{33b} &= n^{-3}\sum\sum\sum_{i\neq j\neq \ell} K_{ll,ij}K_{lc,i\ell}R_{ij}e_\ell M_i/f_i^2 \\
U_{33c} &= n^{-3}\sum_i\sum_{j\neq i} K_{ll,ij}K_{lc,ij}Q_{ij}e_jM_i/f_i^2 \\
U_{33d} &= n^{-3}\sum\sum\sum_{i\neq j\neq \ell} K_{ll,ij}K_{lc,i\ell}Q_{i\ell}e_jM_i/f_i^2 \\
U_{33e} &= n^{-2}\sum_i\sum_{j\neq i} K_{ll,ij}R_{ij}e_iM_i/f_i \\
U_{33f} &= n^{-2}\sum_i\sum_{j\neq i} K_{lc,ij}Q_{ij}e_iM_i/f_i.
\end{aligned}
$$

We first consider U_{33e}. Define $H_{33e}(x_i,x_j) = (1/2)K_{ll,ij}[R_{ij}e_iM_i/f_i + R_{ji}e_jM_j/f_j]$ as a symmetrized version of $K_{ll,ij}R_{ij}e_iM_i/f_i$. Then $U_{33e} = n^{-2}\sum_i\sum_{j\neq i}H_{33e}(x_i,x_j)$. By Assumption 6 and (B.14) we have

$$
\begin{aligned}
\mathrm{E}[H_{33e}(x_i,x_j)|x_i] &= e_iM_i\mathrm{E}[K_{ll,ij}R_{ij}/f_i|x_i] \\
&= e_iM_i\kappa_2(m''(x)/2)h_{ll}^2 + O_p(h_{ll}^4) = h_{ll}^2\mathcal{Z}_{33e,i}M_i + (s.o.)
\end{aligned} \tag{B.26}
$$

where $\mathcal{Z}_{33e,i} = e_i\kappa_2(m''(x)/2)$. Also by the law of iterated expectations, we have $\mathrm{E}[H_{33e}(x_i,x_j)] = 0$. Then by the U-statistic H-decomposition and B.26, we have

$$
\begin{aligned}
U_{33e} &= 2n^{-1}\sum_i \mathrm{E}[H_{33e}(x_i,x_j)|x_i] + (s.o.) \\
&= 2n^{-1}h_{ll}^2\sum_i \mathcal{Z}_{33e,i}M_i + (s.o.) = O_p(n^{-1/2}h_{ll}^2)
\end{aligned} \tag{B.27}
$$

where the last equality follows from the fact that $n^{-1/2}\mathcal{Z}_{33e,i}$ is a zero mean $O_p(1)$ random variable. By the same argument, it can be shown that $U_{33f} = O_p(n^{-1/2}h_{lc}^2)$.

Next, we consider U_{33a}. Note that $U_{33a} = (nh_{lc})^{-1}O_p(U_{33e})$. Then we have

$$
U_{33a} = O_p((nh_{lc})^{-1}n^{-1/2}h_{ll}^2) = o_p(n^{-1/2}h_{ll}^2). \tag{B.28}
$$

Similarly, we have $U_{33c} = o_p(n^{-1/2}h_{lc}^2)$.

Finally, we consider the U_{33b} term. Define $U_{33b} = n^{-3}\sum\sum\sum_{i\neq j\neq\ell}H_{33b}(x_i,x_j,x_\ell)$ where $H_{33b}(x_i,x_j,x_\ell) = (1/3)[K_{ll,ij}K_{lc,i\ell}R_{ij}e_\ell M_i/f_i^2 + K_{ll,j\ell}K_{lc,ji}R_{jl}e_iM_j/f_j^2 + K_{ll,\ell i}K_{lc,\ell j}R_{li}e_jM_\ell/f_\ell^2]$ is a symmetrized version of $K_{ll,ij}K_{lc,i\ell}R_{ij}e_\ell M_i/f_i^2$. By Assumption 6 and (B.14) we have

$$
\mathrm{E}[H_{33b}(x_i,x_j,x_\ell)|x_i] = e_iM_i\kappa_2(m''(x)/2) + O(h_{ll}h_{lc}) = M_i\mathcal{Z}_{33e,i} + (s.o.).
$$

Also by the law of iterated expectations we have $\mathrm{E}[H_{33b}(x_i, x_j, x_\ell)] = 0$. Then by the U-statistic H-decomposition, we have

$$\begin{aligned} U_{33b} &= 3n^{-1} \sum_i \mathrm{E}[H_{33b}(x_i, x_j, x_\ell)|x_i] + (s.o.) \\ &= 3n^{-1/2} h_{ll} h_{lc} (n^{-1/2} \sum_i \mathcal{Z}_{33e,i} M_i) + (s.o.) = O_p(n^{-1/2} h_{ll} h_{lc}). \end{aligned} \tag{B.29}$$

By the same argument, we have $U_{33d} = O_p(n^{-1/2} h_{ll} h_{lc})$.

The Lemma follows from (B.27), (B.28), and (B.29). ■

Lemma 5. *Under Assumptions 1-4 and Assumptions 6-7, if $h_{ll} \to 0$ and $nh_{ll} \to \infty$ as $n \to \infty$, then $U_{11} = \kappa_2^2 \int_{-\infty}^{\infty} (m''(x)/2)^2 M(x) f(x) dx h_{ll}^4 + O(h_{ll}^6 + n^{-1/2} h_{ll}^4 + n^{-1} h_{ll}^{-1/2})$.*

Proof. The argument is similar to the proof of Lemma 2. ■

Lemma 6. *Under Assumptions 1-4 and Assumptions 6-7, if $h_{ll} \to 0$ and $nh_{ll} \to \infty$ as $n \to \infty$, then $U_{12} = \upsilon/(nh_{ll}) \int_{-\infty}^{\infty} \sigma^2(x) M(x) dx + O(n^{-1} h_{ll}^{-1/2} + n^{-3/2} h_{ll}^{-1})$.*

Proof. The argument is similar to the proof of Lemma 3. ■

Lemma 7. *Under Assumptions 1-4 and Assumptions 6-7, if $h_{ll} \to 0$ and $nh_{ll} \to \infty$ as $n \to \infty$, then $U_{13} = O_p(n^{-1/2} h_{ll}^2)$.*

Proof. The argument is similar to the proof of Lemma 4. ■

Lemma 8. *Under Assumptions 1-4 and Assumptions 6-7, if $h_{lc} \to 0$ and $nh_{lc} \to \infty$ as $n \to \infty$, then $U_{21} = \kappa_2^2 \int_{-\infty}^{\infty} (m''(x)/2 + m'(x) f'(x)/f(x))^2 M(x) f(x) dx h_{lc}^4 + O(h_{lc}^6 + n^{-1/2} h_{lc}^4 + n^{-1} h_{lc}^{-1/2})$.*

Proof. The argument is similar to the proof of Lemma 2. ■

Lemma 9. *Under Assumptions 1-4 and Assumptions 6-7, if $h_{lc} \to 0$ and $nh_{lc} \to \infty$ as $n \to \infty$, then $U_{22} = \upsilon/(nh_{lc}) \int_{-\infty}^{\infty} \sigma^2(x) M(x) dx + O(n^{-1} h_{lc}^{-1/2} + n^{-3/2} h_{lc}^{-1})$.*

Proof. The argument is similar to the proof of Lemma 3. ■

Lemma 10. *Under Assumptions 1-4 and Assumptions 6-7, if $h_{lc} \to 0$ and $nh_{lc} \to \infty$ as $n \to \infty$, then $U_{23} = O_p(n^{-1/2} h_{lc}^2)$.*

Proof. The argument is similar to the proof of Lemma 4. ■

C Figures

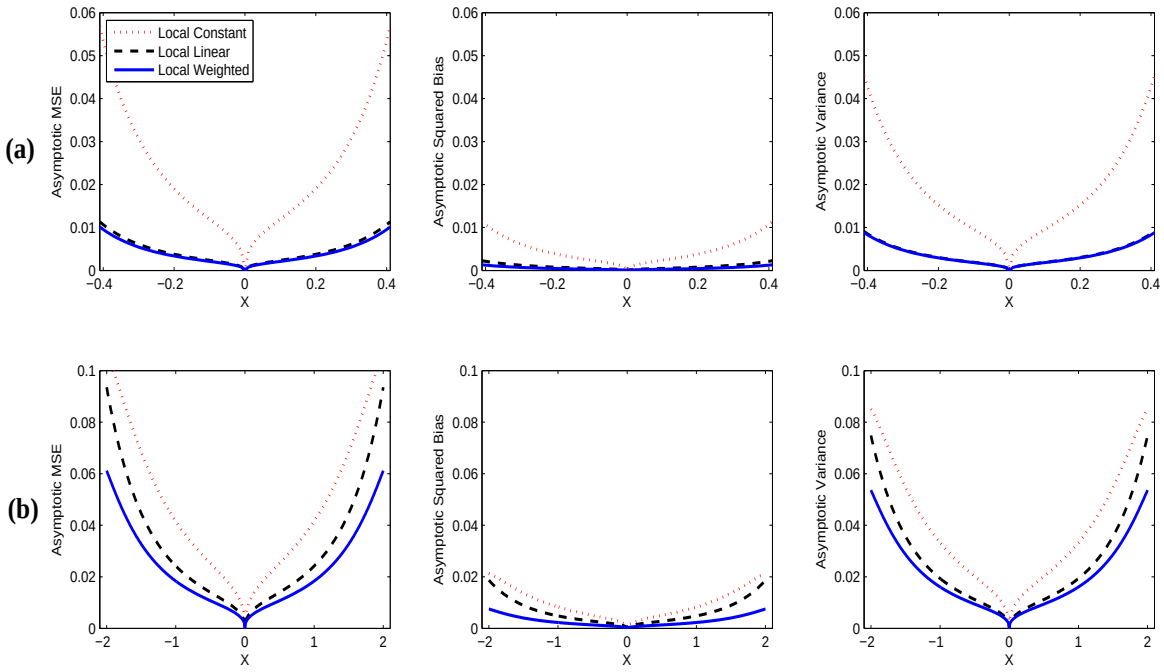

Figure 1: The pointwise asymptotic MSE, asymptotic squared bias, and asymptotic variance curve with sample size $n = 100$. From left to right are asymptotic MSE, asymptotic squared bias, and asymptotic variance against the covariate x for the regression model $y_i = \sin(0.75x_i) + e_i$ with $x \sim N(0, 0.25^2)$ and $e \sim N(0, 1)$ in row (a), and $x \sim N(0, 1)$ and $e \sim N(0, 1)$ in row (b). The solid line represents the local weighted estimator, the dotted line represents the local constant estimator, and the dashed line represents the local linear estimator.

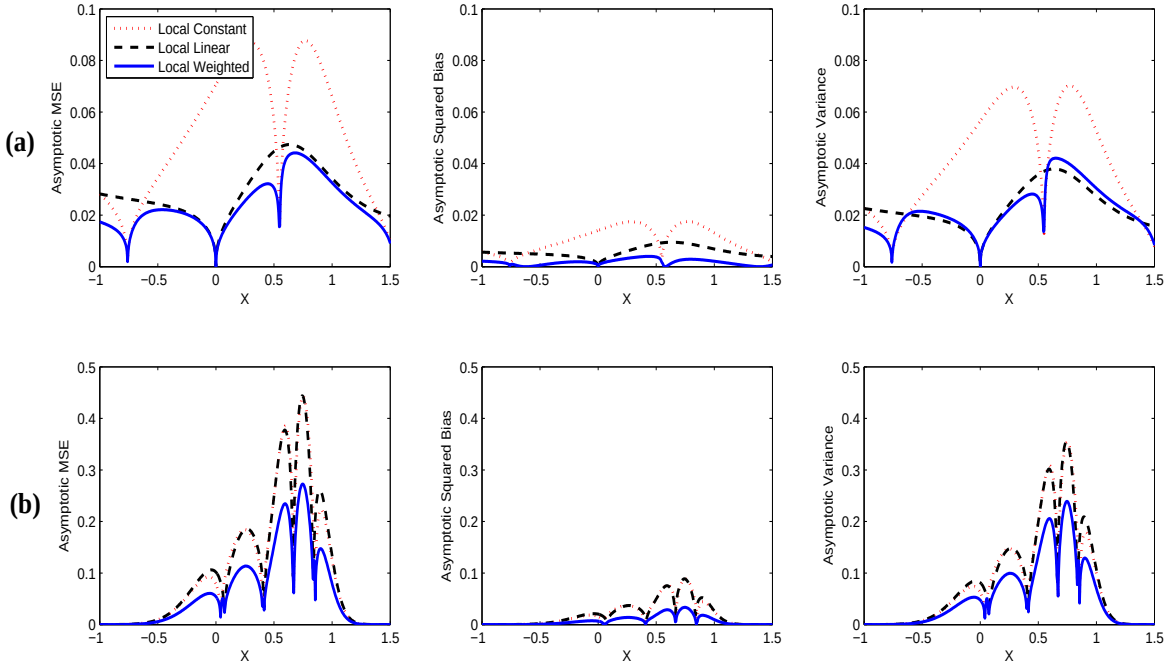

Figure 2: The pointwise asymptotic MSE, asymptotic squared bias, and asymptotic variance curve with sample size $n = 100$. From left to right are asymptotic MSE, asymptotic squared bias, and asymptotic variance against the covariate x. In row (a), the regression is sine function, $y_i = \sin(0.75x_i) + e_i$ with $x \sim 0.5N(-1, 1) + 0.5N(1.75, 0.25)$ and $e \sim N(0, 1)$, and in row (b) the regression is bimodal function, $y_i = 0.3\exp(-16(x_i - 0.25)^2) + 0.7\exp(-64(x_i - 0.75)^2) + e_i$, with $x \sim 0.5N(-1, 1) + 0.5N(1.75, 0.25)$ and $e \sim N(0, 1)$. The solid line represents the local weighted estimator, the dotted line represents the local constant estimator, and the dashed line represents the local linear estimator.

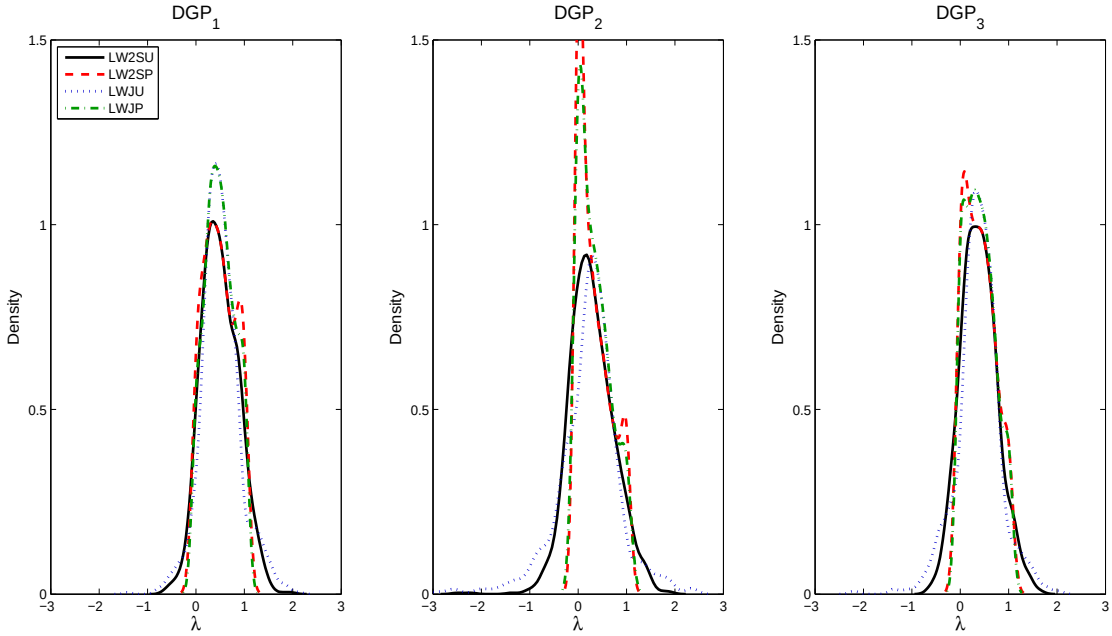

Figure 3: Densities of the cross-validated weights with $x \sim N(0,1)$, $e \sim N(0,1)$, and $n = 100$.

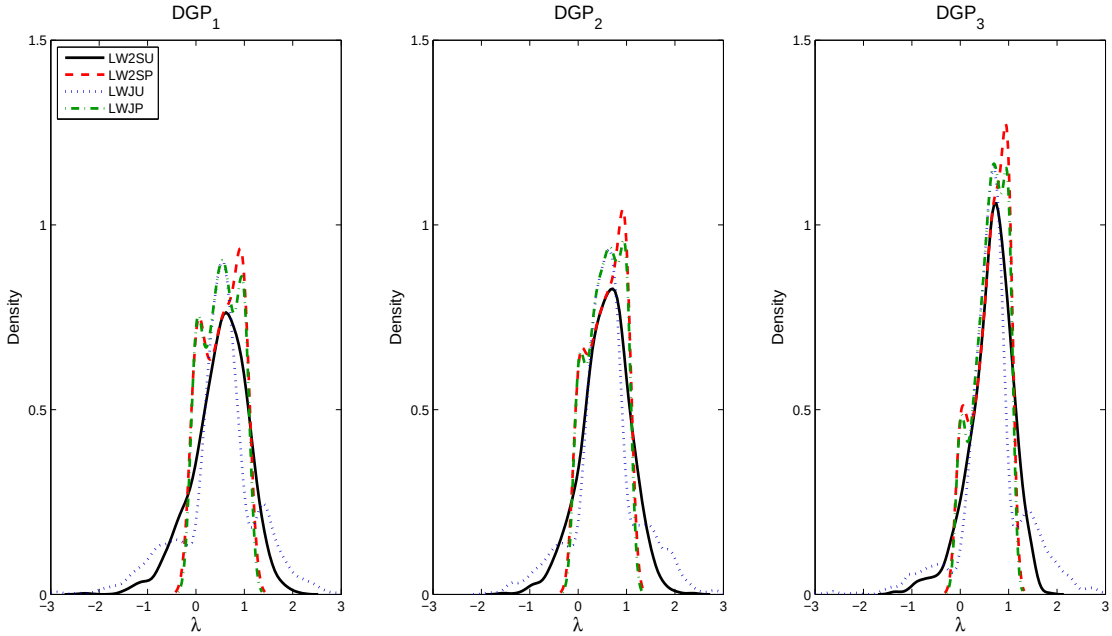

Figure 4: Densities of the cross-validated weights with $x \sim U(0,1)$, $e \sim N(0,1)$, and $n = 100$.

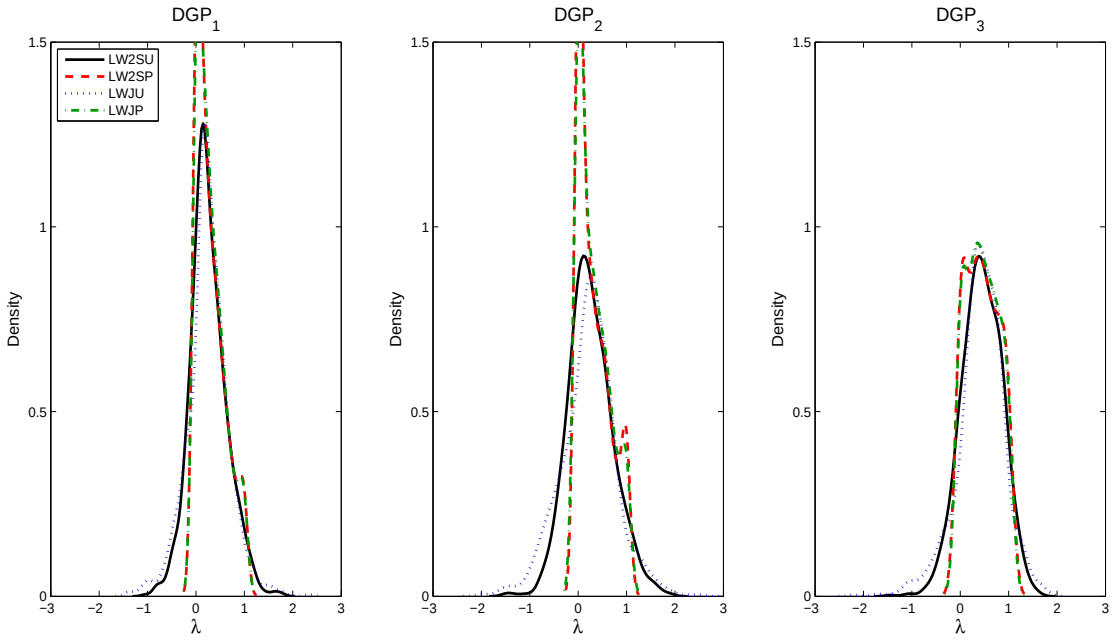

Figure 5: Densities of the cross-validated weights with $x \sim 0.5N(-1,1)+0.5N(1.75,0.25)$, $e \sim N(0,1)$, and $n = 100$.

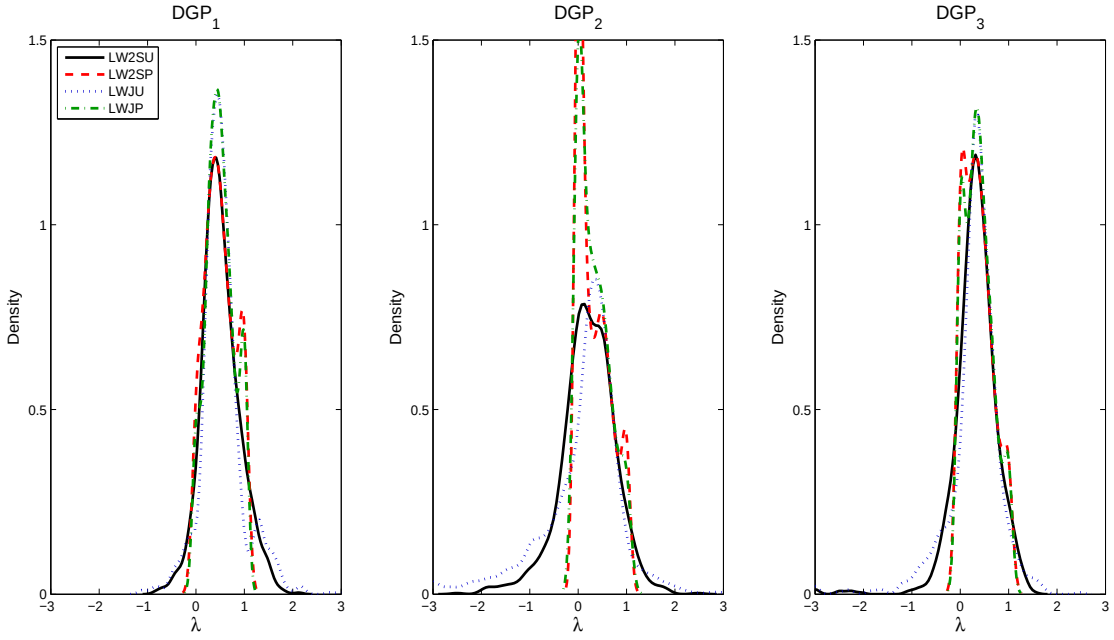

Figure 6: Densities of the cross-validated weights with $x \sim N(0,1)$, $e \sim N(0,\sigma_i)$, $\sigma_i = x_i^2$, and $n = 100$.

D Tables

Table 1: MSEE Results for DGP_1

	N	LC	LL	$LW2SU$	$LW2SP$	$LWJU$	$LWJP$	LS
mean	50	0.2992	0.3245	0.2838	0.2800	0.2798	0.2699	0.5193
	100	0.2031	0.2383	0.1986	0.1977	0.1927	0.1906	0.5135
	200	0.1359	0.1500	0.1352	0.1347	0.1343	0.1333	0.5209
	400	0.0872	0.0919	0.0871	0.0868	0.0876	0.0871	0.5214
median	50	0.2871	0.3243	0.2792	0.2768	0.2659	0.2645	0.4850
	100	0.1972	0.2278	0.1954	0.1952	0.1891	0.1884	0.4920
	200	0.1356	0.1473	0.1351	0.1348	0.1332	0.1329	0.5089
	400	0.0863	0.0903	0.0858	0.0857	0.0865	0.0858	0.5135

* DGP: $m(x) = 2 - 5x + 5\exp(-400(x - 0.5)^2)$, $x \sim N(0, 1)$, $e \sim N(0, 1)$.

Table 2: MSEE Results for DGP_2

	N	LC	LL	$LW2SU$	$LW2SP$	$LWJU$	$LWJP$	LS
mean	50	0.0045	0.0054	0.0045	0.0044	0.0045	0.0044	0.0227
	100	0.0029	0.0033	0.0029	0.0029	0.0030	0.0029	0.0232
	200	0.0018	0.0019	0.0018	0.0018	0.0018	0.0018	0.0232
	400	0.0011	0.0012	0.0011	0.0011	0.0011	0.0011	0.0233
median	50	0.0043	0.0051	0.0043	0.0042	0.0043	0.0042	0.0225
	100	0.0029	0.0031	0.0029	0.0029	0.0029	0.0028	0.0228
	200	0.0018	0.0019	0.0018	0.0018	0.0018	0.0018	0.0231
	400	0.0011	0.0011	0.0011	0.0011	0.0011	0.0011	0.0232

* DGP: $m(x) = 0.3\exp(-16(x - 0.25)^2) + 0.7\exp(-64(x - 0.75)^2)$, $x \sim N(0, 1)$, $e \sim N(0, 1)$.

Table 3: MSEE Results for DGP_3

	N	LC	LL	$LW2SU$	$LW2SP$	$LWJU$	$LWJP$	LS
mean	50	0.2256	0.2802	0.2151	0.2113	0.2206	0.2094	0.4901
	100	0.0970	0.1224	0.0960	0.0955	0.0969	0.0955	0.4943
	200	0.0621	0.0637	0.0602	0.0599	0.0606	0.0602	0.4974
	400	0.0384	0.0377	0.0372	0.0370	0.0373	0.0372	0.4998
median	50	0.1523	0.2070	0.1550	0.1550	0.1540	0.1533	0.4904
	100	0.0946	0.1037	0.0944	0.0935	0.0946	0.0936	0.4951
	200	0.0612	0.0614	0.0595	0.0593	0.0600	0.0596	0.4982
	400	0.0380	0.0370	0.0365	0.0365	0.0367	0.0366	0.4989

* DGP: $m(x) = \sin(5\pi x)$, $x \sim N(0, 1)$, $e \sim N(0, 1)$.

Table 4: MSEE Results for DGP_1

	N	LC	LL	$LW2SU$	$LW2SP$	$LWJU$	$LWJP$	LS
mean	50	0.1746	0.1807	0.1712	0.1682	0.1741	0.1682	1.3487
	100	0.1050	0.1044	0.1032	0.1018	0.1047	0.1020	1.3552
	200	0.0605	0.0589	0.0589	0.0583	0.0594	0.0585	1.3621
	400	0.0352	0.0341	0.0340	0.0338	0.0341	0.0339	1.3571
median	50	0.1704	0.1731	0.1648	0.1621	0.1689	0.1619	1.3222
	100	0.1015	0.1020	0.1004	0.0987	0.1014	0.0989	1.3378
	200	0.0591	0.0574	0.0574	0.0570	0.0579	0.0570	1.3534
	400	0.0338	0.0333	0.0331	0.0330	0.0329	0.0329	1.3562

* DGP: $m(x) = 2 - 5x + 5\exp(-400(x-0.5)^2)$, $x \sim U(0,1)$, $e \sim N(0,1)$.

Table 5: MSEE Results for DGP_2

	N	LC	LL	$LW2SU$	$LW2SP$	$LWJU$	$LWJP$	LS
mean	50	0.0024	0.0023	0.0024	0.0023	0.0024	0.0023	0.0286
	100	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014	0.0290
	200	0.0008	0.0007	0.0008	0.0007	0.0008	0.0008	0.0290
	400	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0292
median	50	0.0023	0.0022	0.0022	0.0021	0.0022	0.0022	0.0285
	100	0.0013	0.0013	0.0013	0.0013	0.0013	0.0013	0.0290
	200	0.0007	0.0007	0.0007	0.0007	0.0008	0.0007	0.0289
	400	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0292

* DGP: $m(x) = 0.3\exp(-16(x-0.25)^2)+0.7\exp(-64(x-0.75)^2)$, $x \sim U(0,1)$, $e \sim N(0,1)$.

Table 6: MSEE Results for DGP_3

	N	LC	LL	$LW2SU$	$LW2SP$	$LWJU$	$LWJP$	LS
mean	50	0.0647	0.0619	0.0637	0.0618	0.0645	0.0623	0.4770
	100	0.0385	0.0360	0.0369	0.0363	0.0378	0.0365	0.4798
	200	0.0218	0.0197	0.0201	0.0199	0.0207	0.0201	0.4821
	400	0.0125	0.0110	0.0113	0.0112	0.0115	0.0113	0.4836
median	50	0.0595	0.0582	0.0595	0.0581	0.0603	0.0587	0.4769
	100	0.0365	0.0343	0.0348	0.0345	0.0359	0.0348	0.4816
	200	0.0207	0.0188	0.0192	0.0191	0.0198	0.0194	0.4829
	400	0.0122	0.0106	0.0109	0.0108	0.0112	0.0109	0.4839

* DGP: $m(x) = \sin(5\pi x)$, $x \sim U(0,1)$, $e \sim N(0,1)$.

Table 7: MSEE Results for DGP_1

	N	LC	LL	$LW2SU$	$LW2SP$	$LWJU$	$LWJP$	LS
mean	50	0.2780	0.3481	0.2766	0.2755	0.2746	0.2718	0.9796
	100	0.1773	0.2208	0.1748	0.1742	0.1740	0.1731	0.9939
	200	0.1126	0.1330	0.1120	0.1116	0.1123	0.1119	0.9981
	400	0.0698	0.0759	0.0696	0.0694	0.0698	0.0697	0.9804
median	50	0.2649	0.3145	0.2620	0.2610	0.2604	0.2577	0.9547
	100	0.1713	0.2006	0.1714	0.1710	0.1705	0.1702	0.9700
	200	0.1111	0.1237	0.1104	0.1102	0.1108	0.1108	0.9834
	400	0.0690	0.0740	0.0684	0.0684	0.0688	0.0688	0.9786

* DGP: $m(x) = 2 - 5x + 5\exp(-400(x-0.5)^2)$, $x \sim 0.5N(-1,1) + 0.5N(1.75, 0.25)$, $e \sim N(0,1)$.

Table 8: MSEE Results for DGP_2

	N	LC	LL	$LW2SU$	$LW2SP$	$LWJU$	$LWJP$	LS
mean	50	0.0037	0.0045	0.0037	0.0037	0.0037	0.0037	0.0315
	100	0.0024	0.0027	0.0024	0.0023	0.0024	0.0023	0.0316
	200	0.0015	0.0016	0.0014	0.0014	0.0014	0.0014	0.0319
	400	0.0009	0.0009	0.0009	0.0009	0.0009	0.0009	0.0324
median	50	0.0035	0.0040	0.0035	0.0035	0.0036	0.0035	0.0312
	100	0.0023	0.0025	0.0023	0.0023	0.0023	0.0023	0.0313
	200	0.0014	0.0015	0.0014	0.0014	0.0014	0.0014	0.0319
	400	0.0009	0.0009	0.0009	0.0009	0.0009	0.0009	0.0324

* DGP: $m(x) = 0.3\exp(-16(x-0.25)^2) + 0.7\exp(-64(x-0.75)^2)$, $x \sim 0.5N(-1,1) + 0.5N(1.75, 0.25)$, $e \sim N(0,1)$.

Table 9: MSEE Results for DGP_3

	N	LC	LL	$LW2SU$	$LW2SP$	$LWJU$	$LWJP$	LS
mean	50	0.1173	0.1590	0.1142	0.1132	0.1151	0.1128	0.4893
	100	0.0701	0.0808	0.0693	0.0686	0.0695	0.0686	0.4947
	200	0.0432	0.0460	0.0424	0.0420	0.0425	0.0423	0.4972
	400	0.0265	0.0261	0.0258	0.0256	0.0258	0.0258	0.4988
median	50	0.1045	0.1174	0.1052	0.1049	0.1067	0.1060	0.4889
	100	0.0687	0.0728	0.0672	0.0665	0.0672	0.0664	0.4949
	200	0.0424	0.0432	0.0417	0.0413	0.0417	0.0414	0.4971
	400	0.0260	0.0254	0.0253	0.0251	0.0253	0.0254	0.4989

* DGP: $m(x) = \sin(5\pi x)$, $x \sim 0.5N(-1,1) + 0.5N(1.75, 0.25)$, $e \sim N(0,1)$.

Table 10: MSEE Results for DGP_1

	N	LC	LL	$LW2SU$	$LW2SP$	$LWJU$	$LWJP$	LS
mean	50	0.3964	0.4286	0.3688	0.3624	0.3616	0.3457	0.5512
	100	0.3163	0.3656	0.3046	0.3021	0.2882	0.2821	0.5200
	200	0.2379	0.2751	0.2400	0.2391	0.2319	0.2290	0.5231
	400	0.1727	0.1920	0.1767	0.1758	0.1772	0.1745	0.5173
median	50	0.3697	0.4180	0.3589	0.3552	0.3464	0.3398	0.5002
	100	0.2962	0.3487	0.2894	0.2873	0.2683	0.2667	0.4946
	200	0.2285	0.2661	0.2321	0.2307	0.2187	0.2169	0.5088
	400	0.1653	0.1855	0.1696	0.1692	0.1702	0.1677	0.5107

* DGP: $m(x) = 2 - 5x + 5\exp(-400(x-0.5)^2)$, $x \sim N(0,1)$, $e \sim N(0,\sigma_i)$, $\sigma_i = x_i^2$.

Table 11: MSEE Results for DGP_2

	N	LC	LL	$LW2SU$	$LW2SP$	$LWJU$	$LWJP$	LS
mean	50	0.0067	0.0082	0.0068	0.0067	0.0068	0.0067	0.0233
	100	0.0051	0.0061	0.0052	0.0052	0.0052	0.0051	0.0229
	200	0.0036	0.0041	0.0037	0.0037	0.0038	0.0037	0.0233
	400	0.0025	0.0028	0.0026	0.0026	0.0027	0.0026	0.0234
median	50	0.0062	0.0076	0.0062	0.0062	0.0062	0.0061	0.0225
	100	0.0047	0.0057	0.0048	0.0048	0.0048	0.0048	0.0227
	200	0.0034	0.0039	0.0035	0.0035	0.0036	0.0035	0.0233
	400	0.0024	0.0026	0.0025	0.0025	0.0025	0.0025	0.0233

* DGP: $m(x) = 0.3\exp(-16(x-0.25)^2) + 0.7\exp(-64(x-0.75)^2)$, $x \sim N(0,1)$, $e \sim N(0,\sigma_i)$, $\sigma_i = x_i^2$.

Table 12: MSEE Results for DGP_3

	N	LC	LL	$LW2SU$	$LW2SP$	$LWJU$	$LWJP$	LS
mean	50	0.2788	0.3070	0.2523	0.2472	0.2574	0.2435	0.4999
	100	0.1591	0.2136	0.1611	0.1600	0.1612	0.1580	0.5000
	200	0.1126	0.1271	0.1152	0.1142	0.1162	0.1143	0.5001
	400	0.0796	0.0855	0.0816	0.0810	0.0823	0.0814	0.4995
median	50	0.2240	0.2636	0.1992	0.1990	0.1953	0.1927	0.4997
	100	0.1435	0.1785	0.1471	0.1467	0.1465	0.1451	0.5014
	200	0.1066	0.1213	0.1096	0.1089	0.1111	0.1090	0.4996
	400	0.0761	0.0822	0.0784	0.0780	0.0793	0.0779	0.5002

* DGP: $m(x) = \sin(5\pi x)$, $x \sim N(0,1)$, $e \sim N(0,\sigma_i)$, $\sigma_i = x_i^2$.

References

Cheng, M.-Y., L. Peng, and J.-S. Wu (2007): "Reducing Variance in Univariate Smoothing," *The Annals of Statistics*, 35(2), 522–542.

Choi, E., and P. Hall (1998): "On Bias Reduction in Local Linear Smoothing," *Biometrika*, 85(2), 333–345.

Cleveland, W. (1979): "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, 74, 829–836.

Fan, J. (1992): "Design-adaptive Nonparametric Regression," *Journal of the American Statistical Association*, 87, 998–1004.

——— (1993): "Local Linear Regression Smoothers and Their Minimax Efficiencies," *The Annals of Statistics*, 21(1), 196–216.

Fan, J., and I. Gijbels (1992): "Variable Bandwidth and Local Linear Regression Smoothers," *The Annals of Statistics*, 20, 2008–2036.

——— (1996): *Local Polynomial Modelling and Its Applications*, vol. 66. Chapman & Hall/CRC.

Frölich, M. (2004): "Finite-Sample Properties of Propensity-Score Matching and Weighting Estimators," *Review of Economics and Statistics*, 86(1), 77–90.

Gasser, T., and H. Müller (1979): "Kernel Estimation of Regression Functions," in *Smoothing Techniques for curve Estimation*, ed. by T. Gasser, and M. Rosenblatt, Lecture Notes in Mathematics 757, pp. 23–68. Springer.

Hardle, W., P. Hall, and J. Marron (1988): "How Far Are Automatically Chosen Regression Smoothing Parameters From Their Optimum?," *Journal of the American Statistical Association*, 83, 86–95.

Hardle, W., P. Hall, and J. Marron (1992): "Regression Smoothing Parameters That Are Not Far From Their Optimum," *Journal of the American Statistical Association*, 87, 227–233.

Hardle, W., and J. Marron (1985): "Optimal Bandwidth Selection in Nonparametric Regression Function Estimation," *The Annals of Statistics*, 13(4), 1465–1481.

Li, Q., and J. Racine (2004): "Cross-Validated Local Linear Nonparametric Regression," *Statistica Sinica*, 14(2), 485–512.

Mammen, E., and J. Marron (1997): "Mass Recentred Kernel Smoothers," *Biometrika*, 84(4), 765–777.

Nadaraya, E. (1964): "On Estimating Regression," *Theory of Probability and Its Applications*, 9(1), 141–142.

Racine, J., and Q. Li (2004): "Nonparametric Estimation of Regression Functions with Both Categorical and Continuous Data," *Journal of Econometrics*, 119(1), 99–130.

Ruppert, D., and M. Wand (1994): "Multivariate Locally Weighted Least Squares Regression," *The Annals of Statistics*, 22, 1346–1370.

Seifert, B., and T. Gasser (1996): "Finite-Sample Variance of Local Polynomials: Analysis and Solutions," *Journal of the American Statistical Association*, 91, 267–275.

——— (2000): "Data Adaptive Ridging in Local Polynomial Regression," *Journal of Computational and Graphical Statistics*, 9, 338–360.

Stone, C. (1977): "Consistent Nonparametric Regression," *The Annals of Statistics*, 5(4), 595–620.

Watson, G. (1964): "Smooth Regression Analysis," *Sankhyā: The Indian Journal of Statistics, Series A*, 26, 359–372.

Xia, Y., and W. Li (2002): "Asymptotic Behavior of Bandwidth Selected by the Cross-Validation Method for Local Polynomial Fitting," *Journal of Multivariate Analysis*, 83(2), 265–287.