Arai, Yoichi; Ichimura, Hidehiko

Working Paper

Optimal bandwidth selection for differences of nonparametric estimators with an application to the sharp regression discontinuity design

cemmap working paper, No. CWP27/13

Provided in Cooperation with:
The Institute for Fiscal Studies (IFS), London

Supplement to "Optimal Bandwidth Selection for Differences of Nonparametric Estimators with an Application to the Sharp Regression Discontinuity Design"

Yoichi Arai and Hidehiko Ichimura

A Introduction

In this supplemental material, we present omitted discussions, an algorithm to implement the proposed method for the sharp RDD and proofs for the main results.

B Uniqueness of the AFO Bandwidths for the Difference of Densities

In this section, we verify the uniqueness of the AFO bandwidths for the difference of densities.

(i) When $f^{(2)}(x_1)f^{(2)}(x_2) < 0$, the first-order conditions are given by

$$\left.\frac{\partial AMSE_{1n}(h)}{\partial h_1}\right|_{h_1=h_1^*,h_2=h_2^*} = \mu_2^2 f^{(2)}(x_1)h_1^*\left[f^{(2)}(x_1)h_1^{*2} - f^{(2)}(x_2)h_2^{*2}\right] - \frac{\nu_0}{n}\frac{f(x_1)}{h_1^{*2}} = 0,$$

$$\left.\frac{\partial AMSE_{1n}(h)}{\partial h_2}\right|_{h_1=h_1^*,h_2=h_2^*} = -\mu_2^2 f^{(2)}(x_2)h_2^*\left[f^{(2)}(x_1)h_1^{*2} - f^{(2)}(x_2)h_2^{*2}\right] - \frac{\nu_0}{n}\frac{f(x_2)}{h_2^{*2}} = 0.$$

Solving these gives the explicit forms of h_1^* and h_2^*.

To show that h_1^* and h_2^* are global minimizers, it is sufficient to show that $AMSE_{1n}(h)$ is strictly convex with respect to h_1 and h_2. For strict convexity, we must show that the Hessian matrix is positive definite; i.e. that

$$\frac{\partial^2 AMSE_{1n}(h)}{\partial h_1^2} > 0, \quad \frac{\partial^2 AMSE_{1n}(h)}{\partial h_1^2} \cdot \frac{\partial^2 AMSE_{1n}(h)}{\partial h_2^2} - \left[\frac{\partial^2 AMSE_{1n}(h)}{\partial h_1 \partial h_2}\right]^2 > 0.$$

Given that $f^{(2)}(x_1)$ and $f^{(2)}(x_2)$ have different signs, it follows that

$$\frac{\partial^2 AMSE_{1n}(h)}{\partial h_1^2} = \mu_2^2 f^{(2)}(x_1)\left[f^{(2)}(x_1)h_1^2 - f^{(2)}(x_2)h_2^2\right] + 2\left[\mu_2 f^{(2)}(x_1)h_1\right]^2 + \frac{2\nu_0 f(x_1)}{nh_1^3} > 0,$$

because $f(\cdot)$, μ_2, ν_0, n, h_1 and h_2 are all positive. We can also show that

$$\begin{aligned}
&\frac{\partial^2 AMSE_{1n}(h)}{\partial h_1^2} \cdot \frac{\partial^2 AMSE_{1n}(h)}{\partial h_2^2} - \left[\frac{\partial^2 AMSE_{1n}(h)}{\partial h_1 \partial h_2}\right]^2 \\
&= \left\{\mu_2^2 f^{(2)}(x_1)\left[f^{(2)}(x_1)h_1^2 - f^{(2)}(x_2)h_2^2\right] + 2\left[\mu_2 f^{(2)}(x_1)h_1\right]^2 + \frac{2\nu_0 f(x_1)}{nh_1^3}\right\} \\
&\quad \times \left\{-\mu_2^2 f^{(2)}(x_2)\left[f^{(2)}(x_1)h_1^2 - f^{(2)}(x_2)h_2^2\right] + 2\left[\mu_2 f^{(2)}(x_2)h_2\right]^2 + \frac{2\nu_0 f(x_2)}{nh_2^3}\right\} \\
&\quad - \left[2\mu_2^2 f^{(2)}(x_1)f^{(2)}(x_2)h_1h_2\right]^2.
\end{aligned}$$

Note that if we ignore the first and third terms in the two brackets of the first term on the right-hand side, what is left coincides with the last term on the right-hand side. However, both the first and third terms are positive as discussed earlier. Thus, the difference of the two terms are positive.

(ii) Next, we consider the case where $f^{(2)}(x_1)f^{(2)}(x_2) > 0$. With the restriction that $f^{(2)}(x_1)h_1^2 - f^{(2)}(x_2)h_2^2 = 0$, $AMSE_{2n}(h)$ can be written as

$$AMSE_{2n}(h) = \left\{\frac{\mu_4}{4!}\left[f^{(4)}(x_1) - \lambda^{**4} f^{(4)}(x_2)\right]h_1^4\right\}^2 + \frac{\nu_0}{nh_1}\left\{f(x_1) + \frac{f(x_2)}{\lambda^{**}}\right\}.$$

The first-order condition becomes

$$\left.\frac{dAMSE_{2n}(h)}{dh_1}\right|_{h_1=h_1^{**}} = \frac{1}{72}\mu_4^2\left\{f^{(4)}(x_1) - \lambda^{**4} f^{(4)}(x_2)\right\}^2 h_1^{**7} - \frac{\nu_0}{nh_1^{**2}}\left\{f(x_1) + \frac{f(x_2)}{\lambda^{**}}\right\} = 0.$$

Solving this with respect to h_1^{**} yields the expression of Definition 1. To see that

$AMSE_{2n}(h_1)$ has a unique minimum, observe that

$$\frac{d^2AMSE_{2n}(h)}{dh_1^2} = \frac{7}{56}\mu_4^2\left\{f^{(4)}(x_1) - \lambda^{**4}f^{(4)}(x_2)\right\}^2 h_1^6 + \frac{2\nu_0}{h_1^3}\left\{f(x_1) + \frac{f(x_2)}{\lambda^{**}}\right\}.$$

Both terms on the right-hand side being positive proves strict convexity. ■

C Implementation for the Sharp RDD

In this section, we provide a detailed procedure to implement the proposed method for the sharp RDD. To obtain the proposed bandwidths, we need pilot estimates of the density and its first derivative, the second and third derivatives of the regression functions, and the conditional variances at the discontinuity point. We obtain these pilot estimates in a number of steps. Before we describe them, note that the discontinuity points for all designs are at $x = 0$. When the discontinuity point is at $x = c$ rather than $x = 0$, one proceeds by replacing X_i with $X_i - c$ in the following steps.

C.1 Step 1: Obtain pilot estimates for the density $f(0)$ and its first derivative $f^{(1)}(0)$

We calculate the density of the forcing variable at the discontinuity point $f(0)$, which is estimated by using the kernel density estimator with an Epanechnikov kernel.[1] A pilot bandwidth for kernel density estimation is chosen by using the normal scale rule, given by $\hat{\sigma} \cdot (15\phi(c)/(n\phi^{(2)}(c)^2))^{1/5}$ evaluated at $c = 0$, where $\hat{\sigma}$ is the square root of the sample variance of X_i and $\phi(\cdot)$ is the normal density (see Wand and Jones, 1994 for the normal scale rules). The first derivative of the density is estimated by using the method proposed by Jones (1994). The kernel first derivative density estimator is given by $\sum_{i=1}^n L((c - X_i)/h)/(nh^2)$, where L is the kernel function proposed by Jones (1994), $L(u) = -15u(1-u^2)1_{\{|u|<1\}}/4$ and c is the evaluation point (zero in our experiments). Again, a pilot bandwidth is obtained by using the normal scale rule,

[1] IK estimated the density in a simpler manner (see Section 4.2 of IK). We used the kernel density estimator to be consistent with the estimation method used for the first derivative. Our unreported simulation experiments produced similar results for both methods.

given by $\hat{\sigma} \cdot (105\phi(c/\hat{\sigma})/(n\phi^{(3)}(c/\hat{\sigma}))^{1/7}$ evaluated at $c = 0.1$.[2]

C.2 Step 2: Obtain pilot bandwidths for estimating the second and third derivatives $m_j^{(2)}(0)$ and $m_j^{(3)}(0)$ for $j = 1, 2$

We next estimate the second and third derivatives by using the third-order LPR. We obtain pilot bandwidths for the LPR based on the estimated fourth derivatives $m_1^{(4)}(0) = \lim_{x\to 0+} m^{(4)}(x)$ and $m_2^{(4)}(0) = \lim_{x\to 0-} m^{(4)}(x)$. Following IK, we use estimates that are not necessarily consistent by fitting global polynomial regressions. In doing so, we construct a matrix whose ith row is given by $[1\ X_i\ X_i^2\ X_i^3\ X_i^4]$. It turns out that the matrix has an average condition number (the ratio of the largest eigenvalue to the smallest.) of 28.80. This number suggests potential multicollinearity, which typically makes the polynomial regression estimates very unstable. Hence, we use the ridge regression proposed by Hoerl, Kennard, and Baldwin (1975). This is implemented in two steps. First, using observations for which $X_i \geq 0$, we regress Y_i on 1, X_i, X_i^2, X_i^3 and X_i^4 to obtain the standard OLS coefficients $\hat{\gamma}_1$ and the variance estimate $\hat{s}_1^2$. This yields the ridge coefficient proposed by Hoerl, Kennard, and Baldwin (1975): $r_1 = (5\hat{s}_1^2)/(\hat{\gamma}_1'\hat{\gamma}_1)$. Using the data with $X_i < 0$, we repeat the procedure to obtain the ridge coefficient, r_2. Let Y be a vector of Y_i, and let X be the matrix whose ith row is given by $[1\ X_i\ X_i^2\ X_i^3\ X_i^4]$ for observations with $X_i \geq 0$, and let I_k be the $k \times k$ identity matrix. The ridge estimator is given by $\hat{\beta}_{r1} = (X'X + r_1 I_5)^{-1}X'Y$, and $\hat{\beta}_{r2}$ is obtained in the same manner. The estimated fourth derivatives are $\hat{m}_1^{(4)}(0) = 24 \cdot \hat{\beta}_{r1}(5)$ and $\hat{m}_2^{(4)}(0) = 24 \cdot \hat{\beta}_{r2}(5)$, where $\hat{\beta}_{r1}(5)$ and $\hat{\beta}_{r2}(5)$ are the fifth elements of $\hat{\beta}_{r1}$ and $\hat{\beta}_{r2}$, respectively. The estimated conditional variance is $\sigma_{r1}^2 = \sum_{i=1}^{n_1}(Y_i - \hat{Y}_i)^2/(n_1 - 5)$, where $\hat{Y}_i$ denotes the fitted values, n_1 is the number of observations for which $X_i \geq 0$, and the summation is over i with $X_i \geq 0$. σ_{r2}^2 is obtained analogously. The plug-in bandwidths for the third-order LPR used to

[2]The normal scale rules do not work when the evaluation point is zero because the third derivative of the normal density at zero is equal to zero. Hence, we use $c = 0.1$. The following results are robust to the value of c, unless c differs greatly from zero.

estimate the second and third derivatives are calculated by

$$h_{\nu,j} = C_{\nu,3}(K)\left(\frac{\sigma_{rj}^2}{\hat{f}(0)\cdot\hat{m}_j^{(4)}(0)\cdot n_j}\right)^{1/9},$$

where $j = 1, 2$ (see Fan and Gijbels, 1996, Section 3.2.3 for information on plug-in bandwidths and the definition of $C_{\nu,3}$). We use $\nu = 2$ and $\nu = 3$ for estimating the second and third derivatives, respectively.

C.3 Step 3: Estimation of the second and third derivatives $m_j^{(2)}(0)$ and $m_j^{(3)}(0)$ as well as the conditional variances $\hat{\sigma}_j^2(0)$ for $j = 1, 2$

We estimate the second and third derivatives at the threshold by using the third-order LPR with the pilot bandwidths obtained in Step 2. Following IK, we use the uniform kernel, which yields constant values of $C_{2,3} = 5.2088$ and $C_{3,3} = 4.8227$. To estimate $\hat{m}_1^{(2)}(0)$, we construct a vector $Y_a = (Y_1, \ldots, Y_{n_a})'$ and an $n_a \times 4$ matrix, X_a, whose ith row is given by $[1\ X_i\ X_i^2\ X_i^3]$ for observations with $0 \leq X_i \leq h_{2,3}$, where n_a is the number of observations with $0 \leq X_i \leq h_{2,3}$. The estimated second derivative is given by $\hat{m}_1^{(2)}(0) = 2 \cdot \hat{\beta}_{2,1}(3)$, where $\hat{\beta}_{2,1}(3)$ is the third element of $\hat{\beta}_{2,1}$ and $\hat{\beta}_{2,1} = ({X_a}'X_a)^{-1}X_aY_a$. We estimate $\hat{m}_2^{(2)}(0)$ in the same manner. Replacing $h_{2,3}$ with $h_{3,3}$ leads to an estimated third derivative of $\hat{m}_1^{(3)}(0) = 6 \cdot \hat{\beta}_{3,1}(4)$, where $\hat{\beta}_{3,1}(4)$ is the fourth element of $\hat{\beta}_{3,1}$, $\hat{\beta}_{3,1} = ({X_b}'X_b)^{-1}X_bY_b$, $Y_b = (Y_1, \ldots, Y_{n_b})'$, X_b is an $n_b \times 4$ matrix whose ith row is given by $[1\ X_i\ X_i^2\ X_i^3]$ for observations with $0 \leq X_i \leq h_{3,3}$, and n_b is the number of observations with $0 \leq X_i \leq h_{3,3}$. The conditional variance at the threshold $\sigma_1^2(0)$ is calculated as $\hat{\sigma}_1(0) = \sum_{i=1}^{n_2}(Y_i - \hat{Y}_i)^2/(n-4)$, where $\hat{Y}_i$ denotes the fitted values from the regression used to estimate the second derivative.[3] $\hat{\beta}_{2,2}$ and $\hat{\beta}_{3,2}$ can be obtained analogously.

[3]One can use the fitted values from the regression used to estimate the third derivatives, having replaced n_a with n_b. However, because these values produce simulation results that are almost identical to those produced by the fitted values described in the main text, we present the latter.

C.4 Step 4

The final step is to plug the pilot estimates into the MMSE given by (10) and to use numerical minimization over the compact region to obtain $\hat{h}_1$ and $\hat{h}_2$. Unlike $AMSE_{1n}(h)$ and $AMSE_{2n}(h)$ subject to the restriction given in Definition 3, the MMSE is not necessarily strictly convex, particularly when the sign of the product is positive. In conducting numerical optimization, it is important to try optimization with several initial values, so as to avoid finding only a local minimum. $(\hat{h}_1^E, \hat{h}_2^E)$ and $(\hat{h}_1^R, \hat{h}_2^R)$ can be computed using the MMSE given by (11) and (13), respectively.

D Proof of Theorem 3

Recall that the objective function is:

$$\widehat{MMSE}_n(h) = \left\{ \frac{b_1}{2} \left[\hat{m}_1^{(2)}(x) h_1^2 - \hat{m}_2^{(2)}(x) h_2^2 \right] \right\}^2 + \left[\hat{b}_{2,1}(x) h_1^3 - \hat{b}_{2,2}(x) h_2^3 \right]^2 + \frac{\nu}{n\hat{f}(x)} \left\{ \frac{\hat{\sigma}_1^2(x)}{h_1} + \frac{\hat{\sigma}_2^2(x)}{h_2} \right\}.$$

To begin with, we show that $\hat{h}_1$ and $\hat{h}_2$ satisfy Assumption 3. If we choose a sequence of h_1 and h_2 to satisfy Assumption 3, then $\widehat{MMSE}_n(h)$ converges to 0. Assume to the contrary that either one or both of $\hat{h}_1$ and $\hat{h}_2$ do not satisfy Assumption 3. Since $m_2^{(2)}(x)^3 b_{2,1}(x)^2 \neq m_1^{(2)}(x)^3 b_{2,2}(x)^2$ by assumption, $\hat{m}_2^{(2)}(x)^3 \hat{b}_{2,1}(x)^2 \neq \hat{m}_1^{(2)}(x)^3 \hat{b}_{2,2}(x)^2$ with probability approaching 1. Without loss of generality, we assume this as well. Then at least one of the first-order bias term, the second-order bias term and the variance term of $\widehat{MMSE}_n(\hat{h})$ does not converge to zero in probability. Then $\widehat{MMSE}_n(\hat{h}) > \widehat{MMSE}_n(h)$ holds for some n. This contradicts the definition of $\hat{h}$. Hence $\hat{h}$ satisfies Assumption 3.

We first consider the case in which $m_1^{(2)}(x) m_2^{(2)}(x) < 0$. In this case, with probability approaching 1, $\hat{m}_1^{(2)}(x) \hat{m}_2^{(2)}(x) < 0$, so that we assume this without loss of generality. When this holds, note that the leading terms are the first term and the last term of $\widehat{MMSE}_n(\hat{h})$ since $\hat{h}_1$ and $\hat{h}_2$ satisfy Assumption 3. Define the plug-in version of $AMSE_{1n}(h)$ provided in Definition 3 by

$$\widehat{AMSE}_{1n}(h) = \left\{ \frac{b_1}{2} \left[\hat{m}_1^{(2)}(x) h_1^2 - \hat{m}_2^{(2)}(x) h_2^2 \right] \right\}^2 + \frac{\nu}{n \hat{f}(x)} \left\{ \frac{\hat{\sigma}_1^2(x)}{h_1} + \frac{\hat{\sigma}_2^2(x)}{h_2} \right\}.$$

Let the minimizer of $\widehat{AMSE}_{1n}(h)$ by $\tilde{h}_1$ and $\tilde{h}_2$. Also define

$$\hat{\theta}_1 = \left\{ \frac{v \hat{\sigma}_1^2(x)}{\hat{b}_1^2 \hat{f}(x) \hat{m}_1^{(2)}(x) \left[\hat{m}_1^{(2)}(x) - \hat{\lambda}_1^2 \hat{m}_2^{(2)}(x) \right]} \right\}^{1/5} \quad \text{and} \quad \hat{\lambda}_1 = \left\{ - \frac{\hat{\sigma}_2^2(x) \hat{m}_1^{(2)}(x)}{\hat{\sigma}_1^2(x) \hat{m}_2^{(2)}(x)} \right\}^{1/3}.$$

A calculation yields $\tilde{h}_1 = \hat{\theta}_1 n^{-1/5} \equiv \tilde{C}_1 n^{-1/5}$ and $\tilde{h}_2 = \hat{\theta}_1 \hat{\lambda}_1 n^{-1/5} \equiv \tilde{C}_2 n^{-1/5}$. With this choice, $\widehat{AMSE}_{1n}(\tilde{h})$ and hence $\widehat{MMSE}_n(\tilde{h})$ converges at the rate of $n^{-4/5}$. Note that if $\hat{h}_1$ or $\hat{h}_2$ converges at the rate slower than $n^{-1/5}$, then the bias term converges at the rate slower than $n^{-4/5}$. If $\hat{h}_1$ or $\hat{h}_2$ converges at the rate faster than $n^{-1/5}$, then the variance term converges at the rate slower than $n^{-4/5}$. Thus the minimizer of $\widehat{MMSE}_n(h)$, $\hat{h}_1$ and $\hat{h}_2$ converges to 0 at rate $n^{-1/5}$.

Thus we can write $\hat{h}_1 = \hat{C}_1 n^{-1/5} + o_p(n^{-1/5})$ and $\hat{h}_2 = \hat{C}_2 n^{-1/5} + o_p(n^{-1/5})$ for some $O_P(1)$ sequences $\hat{C}_1$ and $\hat{C}_2$ that are bounded away from 0 and ∞ as $n \to \infty$. Using this expression,

$$\begin{aligned} \widehat{MMSE}_n(\hat{h}) &= n^{-4/5} \left\{ \frac{b_1}{2} \left[\hat{m}_1^{(2)}(x) \hat{C}_1^2 - \hat{m}_2^{(2)}(x) \hat{C}_2^2 \right] \right\}^2 \\ &\quad + \frac{\nu}{n^{4/5} \hat{f}(x)} \left\{ \frac{\hat{\sigma}_1^2(x)}{\hat{C}_1} + \frac{\hat{\sigma}_2^2(x)}{\hat{C}_2} \right\} + o_p(n^{-4/5}). \end{aligned}$$

Note that

$$\begin{aligned} \widehat{MMSE}_n(\tilde{h}) &= n^{-4/5} \left\{ \frac{b_1}{2} \left[\hat{m}_1^{(2)}(x) \tilde{C}_1^2 - \hat{m}_2^{(2)}(x) \tilde{C}_2^2 \right] \right\}^2 \\ &\quad + \frac{\nu}{n^{4/5} \hat{f}(x)} \left\{ \frac{\hat{\sigma}_1^2(x)}{\tilde{C}_1} + \frac{\hat{\sigma}_2^2(x)}{\tilde{C}_2} \right\} + O_P(n^{-8/5}). \end{aligned}$$

Since $\hat{h}$ is the optimizer, $\widehat{MMSE}_n(\hat{h})/\widehat{MMSE}_n(\tilde{h}) \leq 1$. Thus

$$\frac{\left\{\frac{b_1}{2}\left[\hat{m}_1^{(2)}(x)\hat{C}_1^2 - \hat{m}_2^{(2)}(x)\hat{C}_2^2\right]\right\}^2 + \frac{\nu}{\hat{f}(x)}\left\{\frac{\hat{\sigma}_1^2(x)}{\hat{C}_1} + \frac{\hat{\sigma}_2^2(x)}{\hat{C}_2}\right\} + o_p(1)}{\left\{\frac{b_1}{2}\left[\hat{m}_1^{(2)}(x)\tilde{C}_1^2 - \hat{m}_2^{(2)}(x)\tilde{C}_2^2\right]\right\}^2 + \frac{\nu}{\hat{f}(x)}\left\{\frac{\hat{\sigma}_1^2(x)}{\tilde{C}_1} + \frac{\hat{\sigma}_2^2(x)}{\tilde{C}_2}\right\} + O_P(n^{-4/5})} \leq 1.$$

Note that the denominator converges to

$$\left\{\frac{b_1}{2}\left[m_1^{(2)}(x)C_1^{*2} - m_2^{(2)}(x)C_2^{*2}\right]\right\}^2 + \frac{\nu}{f(x)}\left\{\frac{\sigma_1^2(x)}{C_1^*} + \frac{\sigma_2^2(x)}{C_2^*}\right\},$$

where C_1^* and C_2^* are the unique optimizers of

$$\left\{\frac{b_1}{2}\left[m_1^{(2)}(x)C_1^2 - m_2^{(2)}(x)C_2^2\right]\right\}^2 + \frac{\nu}{f(x)}\left\{\frac{\sigma_1^2(x)}{C_1} + \frac{\sigma_2^2(x)}{C_2}\right\},$$

with respect to C_1 and C_2. This implies that $\hat{C}_1$ and $\hat{C}_2$ also converge to the same respective limit C_1^* and C_2^* because the inequality will be violated otherwise.

Next we consider the case with $m_1^{(2)}(x)m_2^{(2)}(x) > 0$. In this case, with probability approaching 1, $\hat{m}_1^{(2)}(x)\hat{m}_2^{(2)}(x) > 0$, so that we assume this without loss of generality.

When these conditions hold, define

$$\hat{\theta}_2 = \left\{\frac{v\left[\hat{\sigma}_1^2(x) + \hat{\sigma}_2^2(x)/\hat{\lambda}_2\right]}{6\hat{f}(x)\left[\hat{b}_{2,1}(x) - \hat{\lambda}_2^3\hat{b}_{2,2}(x)\right]^2}\right\}^{1/7} \quad \text{and} \quad \hat{\lambda}_2 = \left\{\frac{\hat{m}_1^{(2)}(x)}{\hat{m}_2^{(2)}(x)}\right\}^{1/2}.$$

and let $h_2 = \hat{\lambda}_2 h_1$. This sets the first-order bias term of $\widehat{MMSE}_n(h)$ equal to 0. Define the plug-in version of $AMSE_{2n}(h)$ by

$$\widehat{AMSE}_{2n}(h) = \left\{\hat{b}_{2,1}(x)h_1^3 - \hat{b}_{2,2}(x)h_2^3\right\}^2 + \frac{v}{n\hat{f}(x)}\left\{\frac{\hat{\sigma}_1^2(x)}{h_1} + \frac{\hat{\sigma}_2^2(x)}{h_2}\right\}$$

Choosing h_1 to minimize $\widehat{AMSE}_{2n}(h)$, we define $\tilde{h}_1 = \hat{\theta}_2 n^{-1/7} \equiv \tilde{C}_1 n^{-1/7}$ and $\tilde{h}_2 =$

$\hat{\lambda}_2\tilde{h}_1 \equiv \tilde{C}_2 n^{-1/7}$. Then $\widehat{MMSE}_n(\tilde{h})$ can be written as

$$\widehat{MMSE}_n(\tilde{h}) = n^{-6/7}\left\{\hat{b}_{2,1}(x)\tilde{C}_1^3 - \hat{b}_{2,2}(x)\tilde{C}_2^3\right\}^2 + n^{-6/7}\frac{\nu}{\hat{f}(x)}\left\{\frac{\hat{\sigma}_1^2(x)}{\tilde{C}_1} + \frac{\hat{\sigma}_2^2(x)}{\tilde{C}_2}\right\}.$$

In order to match this rate of convergence, both $\hat{h}_1$ and $\hat{h}_2$ need to converge at the rate slower than or equal to $n^{-1/7}$ because the variance term needs to converge at the rate $n^{-6/7}$ or faster. In order for the first-order bias term to match this rate,

$$\hat{m}_1^{(2)}(x)\hat{h}_1^2 - \hat{m}_2^{(2)}(x)\hat{h}_2^2 \equiv B_{1n} = n^{-3/7}b_{1n},$$

where $b_{1n} = O_P(1)$ so that under the assumption that $m_2^{(2)}(x) \neq 0$, with probability approaching 1, $\hat{m}_2^{(2)}(x)$ is bounded away from 0 so that assuming this without loss of generality, we have $\hat{h}_2^2 = \hat{\lambda}_2^2\hat{h}_1^2 - B_{1n}/\hat{m}_2^{(2)}(x)$. Substituting this expression to the second term and the third term, we have

$$\begin{aligned}\widehat{MMSE}_n(\hat{h}) = \left\{\frac{b_1}{2}B_{1n}\right\}^2 + \left\{\hat{b}_{2,1}(x)\hat{h}_1^3 - \hat{b}_{2,2}(x)\{\hat{\lambda}_2^2\hat{h}_1^2 - B_{1n}/\hat{m}_2^{(2)}(x)\}^{3/2}\right\}^2 \\ + \frac{\nu}{n\hat{f}(x)}\left\{\frac{\hat{\sigma}_1^2(x)}{\hat{h}_1} + \frac{\hat{\sigma}_2^2(x)}{\{\hat{\lambda}_2^2\hat{h}_1^2 - B_{1n}/\hat{m}_2^{(2)}(x)\}^{1/2}}\right\}.\end{aligned}$$

Suppose $\hat{h}_1$ is of order slower than $n^{-1/7}$. Then because $\hat{m}_2^{(2)}(x)^3\hat{b}_{2,1}(x)^2 \neq \hat{m}_1^{(2)}(x)^3\hat{b}_{2,2}(x)^2$ and this holds even in the limit, the second-order bias term is of order slower than $n^{-6/7}$. If $\hat{h}_1$ converges to 0 faster than $n^{-1/7}$, then the variance term converges at the rate slower than $n^{-6/7}$. Therefore we can write $\hat{h}_1 = \hat{C}_1 n^{-1/7} + o_p(n^{-1/7})$ for some $O_P(1)$ sequence $\hat{C}_1$ that is bounded away from 0 and ∞ as $n \to \infty$ and as before $\hat{h}_2^2 = \hat{\lambda}_2^2\hat{h}_1^2 - B_{1n}/\hat{m}_2^{(2)}(x)$. Using this expression, we can write

$$\begin{aligned}\widehat{MMSE}_n(\hat{h}) &= n^{-6/7}\left\{\frac{b_1}{2}b_{1n}\right\}^2 \\ &+ n^{-6/7}\left\{\left[\hat{b}_{2,1}(x)\hat{C}_1^3 + o_p(1) - \hat{b}_{2,2}(x)\{\hat{\lambda}_2^2\hat{C}_1^2 + o_p(1) - n^{-1/7}b_{1n}/\hat{m}_2^{(2)}(x)\}^{3/2}\right]\right\}^2 \\ &+ n^{-6/7}\frac{\nu}{\hat{f}(x)}\left\{\frac{\hat{\sigma}_1^2(x)}{\hat{C}_1 + o_p(1)} + \frac{\hat{\sigma}_2^2(x)}{\{\hat{\lambda}_2^2\hat{C}_1^2 + o_p(1) - n^{-1/7}b_{1n}/\hat{m}_2^{(2)}(x)\}^{1/2}}\right\}.\end{aligned}$$

Thus b_{1n} converges in probability to 0. Otherwise the first-order bias term remains and that contradicts the definition of $\hat{h}_1$.

Since $\hat{h}$ is the optimizer, $\widehat{MMSE}_n(\hat{h})/\widehat{MMSE}_n(\tilde{h}) \leq 1$. Thus

$$\frac{o_p(1) + \left\{\left[\hat{b}_{2,1}(x)\hat{C}_1^3 - \hat{b}_{2,2}(x)\{\hat{\lambda}_2^2\hat{C}_1^2 + o_p(1)\}^{3/2}\right]\right\}^2 + \frac{\nu}{\hat{f}(x)}\left\{\frac{\hat{\sigma}_1^2(x)}{\hat{C}_1 + o_p(1)} + \frac{\hat{\sigma}_2^2(x)}{\{\hat{\lambda}_2^2\hat{C}_1^2 + o_p(1)\}^{1/2}}\right\}}{\left\{\hat{b}_{2,1}(x)\tilde{C}_1^3 - \hat{b}_{2,2}(x)\tilde{C}_2^3\right\}^2 + \frac{\nu}{\hat{f}(x)}\left\{\frac{\hat{\sigma}_1^2(x)}{\tilde{C}_1} + \frac{\hat{\sigma}_2^2(x)}{\tilde{C}_2}\right\}} \leq 1.$$

If $\hat{C}_1 - \tilde{C}_1$does not converge to 0 in probability, then the ratio is not less than 1 at some point. hence $\hat{C}_1 - \tilde{C}_1 = o_p(1)$. Therefore $\hat{h}_2/\tilde{h}_2$ converges in probability to 1 as well.

The result above also shows that $\widehat{MMSE}_n(\hat{h})/MSE_n(h^*)$ converges to 1 in probability in both cases. ■

References

Fan, J., and I. Gijbels (1996): *Local polynomial modeling and its applications.* Chapman & Hall.

Hoerl, A. E., R. W. Kennard, and K. F. Baldwin (1975): "Ridge regression: some simulations," *Communications in Statistics, Theory and Methods*, 4, 105–123.

Jones, M. C. (1994): "On kernel density derivative estimation," *Communications in Statistics, Theory and Methods*, 23, 2133–2139.

Wand, M. P., and M. C. Jones (1994): *Kernel Smoothing.* Chapman & Hall.