ISSN 1440-771X

MONASH University

Australia

Department of Econometrics and Business Statistics

http://www.buseco.monash.edu.au/depts/ebs/pubs/wpapers/

Semiparametric Localized Bandwidth Selection For Kernel Density Estimation

Tingting Cheng, Jiti Gao and Xibin Zhang

November 2014

Working Paper 27/14
(revised 14/14)

Semiparametric Localized Bandwidth Selection for Kernel Density Estimation

Tingting Cheng* AND Jiti Gao AND Xibin Zhang

Abstract

Since conventional cross–validation bandwidth selection methods don't work for the case where the data considered are dependent time series, alternative bandwidth selection methods are needed. In recent years, Bayesian based global bandwidth selection methods have been proposed. Our experience shows that the use of a global bandwidth is however less suitable than using a localized bandwidth in kernel density estimation in the case where the data are dependent time series as discussed in an empirical application of this paper. Nonetheless, a difficult issue is how we can consistently estimate a localized bandwidth. In this paper, we propose a semiparametric estimation method, for which we establish a completely new asymptotic theory for the proposed semiparametric localized bandwidth estimator. Applications of the new bandwidth estimator to the kernel density estimation of Eurodollar deposit rate and the S&P 500 daily return demonstrate the effectiveness and competitiveness of the proposed semiparametric localized bandwidth.

KEYWORDS: Hyperparameter estimation; likelihood function; localized bandwidth.

*Corresponding author. Department of Econometrics and Business Statistics, Monash University, Caulfield East, Victoria 3145, Australia. Telephone: +61 3 99034532. Fax: +61 3 99032007. Email: chengtingting1986@gmail.com.

1 Introduction

Kernel density estimation is an important tool for exploring the distributional properties of a random variable in an unknown population (Silverman, 1986). Such kernel estimation techniques have been widely used in many application studies (see for example, Aït-Sahalia, 1996; Bithell, 1990; Seaman and Powell, 1996; Elgammal, Duraiswami, Harwood, and Davis, 2002). It is known that the performance of a kernel density estimator is mainly determined by its bandwidth. This paper aims to present a semiparametric estimation method for localized bandwidth selection.

There exists a large body of literature on bandwidth selection for kernel density estimation. Jones, Marron, and Sheather (1996) presented surveys on bandwidth selection methods, including the rule–of–thumb, the least squares cross–validation, the biased cross–validation, the plug–in method and a smoothed bootstrapping method. Fan and Yao (2003) discussed bandwidth selection methods for nonparametric density estimation in Section 5.4 of Chapter 5, including the normal reference rule, the cross–validation and the plug–in method. An overview on kernel density estimation and relevant bandwidth selectors can also be found in Sheather (2004). Recently, Heidenreich, Schindler, and Sperlich (2013) compared the existing bandwidth methods on a set of designs and suggested a mixture of the simple plug–in and cross–validation methods. There also exist some investigations on using Bayesian sampling approaches to bandwidth estimation (see Zhang, King, and Hyndman, 2006, among others). All these methods mainly aim to choose a global bandwidth, with which a kernel estimator is likely to simultaneously under– and over–smooth the underlying density function in different areas on its support (Sain and Scott, 1996). The recent development on kernel density estimation with localized bandwidth suggests that a small bandwidth be assigned to the observations in the high–density region and a large bandwidth be assigned to those in the low–density region. The localized kernel density estimation method attempts to solve this problem by allowing for different bandwidths in different regions on the support of the underlying density (see for example, Brewer, 2000; Gangopadhyay and Cheung, 2002; Kulasekera and Padgett, 2006).

In this paper, we construct a localized bandwidth through the conditional density of the bandwidth parameter given a data set. The resulting bandwidth estimate is a function of the density point, which is the value where the density estimator is calculated. The resulting density estimator is

$$\widehat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h(x)}K\left(\frac{X_i - x}{h(x)}\right).$$

Gangopadhyay and Cheung (2002) and de Lima and Atuncar (2011) considered this type of local-

ized bandwidth and derived a closed–form bandwidth estimate, where the prior of bandwidth belongs to a conjugate family of the likelihood.[†]

However, to the best of our knowledge, there is no asymptotic theory for the above–derived bandwidth estimation method in the current literature. This paper aims to fill this gap and presents an asymptotic theory for the bandwidth estimation method. We propose a semiparametric localized bandwidth estimation method and then establish an asymptotically normal distribution with root–n rate of convergence. Härdle, Hall, and Marron (1988) showed that $n^{3/10}(\widehat{h}_{\text{cv}} - h_0)$ converges in distribution to a normal random variable for the case where $h_0 \sim n^{-1/5}$ and $\widehat{h}_{\text{cv}}$ was chosen based on the conventional cross–validation method. In this work, we considerably improve this by showing that the proposed semiparametric localized bandwidth can achieve a much faster rate of convergence at an order of $\sqrt{n}$. Meanwhile, we will demonstrate that the semiparametric localized bandwidth estimation method has satisfactory finite–sample properties and better out–of–sample performance than those based its natural competitors when estimating the density of Eurodollar daily deposit rate, and the density of S&P 500 daily return.

In Bayesian inference, the prior of a parameter is often chosen from a family of densities characterized by other parameter(s), known as hyperparameter(s). Although there are several prior choices of a parameter, in the simulation studies and the empirical applications, we consider the case where the prior of bandwidth is an inverse Gamma density with two hyperparameters. We find that these two hyperparameters play an important role in the performance of the resulting bandwidth. It is well known that expressing honest prior information can be difficult. From a Bayesian point of view, when there is an uncertainty on hyperparameters, a solution would be to put a "hyperprior" on a hyperparameter. This solution might have a basis because there exists a certain level of "robustness" with respect to the specification of parameters in hyperpriors. However, there is no guarantee for this hierarchical procedure to achieve robustness universally, and the choice of a hyperprior is very subjective.

Casella (2001) investigated hyperparameter estimation based on the EM algorithm and the Markov chain Monte Carlo (MCMC) simulation. Atchadé (2011) developed an adaptive Monte Carlo strategy for sampling from posterior in empirical Bayes analysis. However, these methods are not applicable to bandwidth estimation discussed in this paper, mainly because the

[†]In these papers, as well as in our approach to bandwidth selection, bandwidth is treated as a parameter, whose posterior is employed to derive a posterior estimate of bandwidth. Therefore, it should be stressed here that this approach is different from the Bayesian nonparametric approach, which has been extensively investigated in the literature (see for example, Lo, 1984; Ghosh and Ramamoorthi, 2003; Hjort, Holmes, Müller, and Walker, 2010).

information available from the original data may not be sufficient for consistently estimating the hyperparameters concerned. We propose using a likelihood approach to hyperparameter estimation, where a pseudo random sample is generated from the marginal likelihood, a function of hyperparameters. A likelihood function is constructed based on this pseudo sample and is then maximized to derive hyperparameter estimates. This likelihood approach is semiparametric because the density for constructing the likelihood is approximated by its kernel estimator. It is expected that the resulting hyperparameter estimates are more appropriate than those subjectively chosen.

The main contributions of this paper are summarized as follows.

(i) We develop an asymptotic theory for our proposed semiparametric localized bandwidth selection method.

(ii) We present a likelihood approach to hyperparameter estimation and show that it works well in both the Monte Carlo simulation studies and the empirical studies.

(iii) We conduct simulation studies to examine the finite–sample performance of our proposed semiparametric localized bandwidth selection, as well as the performance of the resulting kernel density estimators.

(iv) We apply the proposed semiparametric localized bandwidth selection method to the kernel density estimation of Eurodollar daily deposit rate and the S&P 500 daily return.

The rest of this paper is organized as follows. Section 2 briefly describes the construction of a semiparametric localized bandwidth selection method. In Section 3, we investigate the asymptotic properties of the proposed semiparametric localized bandwidth selection method. Section 4 presents Monte Carlo simulation studies to evaluate the performance of the semiparametric localized bandwidth selection method. In Sections 5 and 6, two empirical examples are presented to illustrate the application of the proposed semiparametric localized bandwidth selection method. Section 7 concludes the paper. Appendix A discusses a consistent estimation method for the hyperparameters involved in the prior distribution. The proofs of the main theorems are given in Appendix B.

2 Semiparametric localized bandwidth

Let X denote a univariate random variable with a density $f(x)$, which is approximated by

$$f(x|h) = f * K_h(x) = \int f(y)K_h(y-x)dy = \mathbb{E}_1\left[K_h(X-x)\right], \tag{1}$$

where $K_h(\cdot) = K(\cdot/h)/h$, and $K(\cdot)$ is the kernel function, and $\mathbb{E}_1[\cdot]$ denotes the expectation under the conditional distribution of X given h. In the rest of this paper, we use $\mathbb{E}_2[\cdot]$ to denote the expectation under the marginal distribution of h and $\mathbb{E}[\cdot]$ denote the full expectation under the joint distribution of (X,h). In other words, $\mathbb{E}[\cdot] = \mathbb{E}_2(\mathbb{E}_1[\cdot])$.

In fact, $f(x|h)$ can be considered as the density of $X+\xi$, where ξ is a random variable with mean 0 and density $K_h(\cdot)$. When h is small, the difference between $f(x|h)$ and $f(x)$ is practically negligible. Let $\pi(h)$ denote the prior density of h. We define the posterior of h given $X=x$ as $\pi(h|x) = \frac{f(x|h)\pi(h)}{\int f(x|h)\pi(h)dh}$. A Bayes estimate of h is the posterior mean and is given by

$$h_0(x) = \int h\pi(h|x)dh = \frac{\int hf(x|h)\pi(h)dh}{\int f(x|h)\pi(h)dh} \equiv \frac{q(x)}{p(x)}, \tag{2}$$

where $q(x) = \int hf(x|h)\pi(h)dh$ and $p(x) = \int f(x|h)\pi(h)dh$. Since $\pi(h)$ is usually unknown in practice, we consider the case where $\pi(h)$ belongs to a parametric family indexed by a vector of unknown parameters, that is, $\pi(h) \in \left\{\pi(h|\theta_0) : \int \pi(h|\theta_0)dh = 1, \theta_0 \in \Theta\right\}$, in which θ_0 is the true value of a vector of hyperparameters. In this situation, the prior of h is denoted as $\pi(h|\theta)$, and the Bayes estimate of h given by (2) is denoted as $h_0(x|\theta) = q(x|\theta)/p(x|\theta)$, where $q(x|\theta) = \int hf(x|h)\pi(h|\theta)dh$ and $p(x|\theta) = \int f(x|h)\pi(h|\theta)dh$.

The resulting kernel density estimate is $\widehat{f}_n(x) = \frac{1}{n}\sum_{i=1}^n \frac{1}{h_0(x|\theta)} K\left(\frac{X_i - x}{h_0(x|\theta)}\right)$. In the rest of this paper, we will focus on the issue of how to estimate $h_0(x)$ semiparametrically. As mentioned in Remark 4 that follows Theorem 4 listed in Section 3 below, we also discuss the case where $h_0(x) = a_n \cdot b_0(x)$ and the corresponding estimation of $b_0(x)$, where $a_n \to 0$ and $b_0(x)$ is being treated as an unknown function of x.

Any inference or computation based on $\pi(h|x)$ cannot be directly conducted because $f(x|h)$ is unknown. When a random sample, denoted as $\{X_i : i = 1,2,\cdots,n\}$, is observed from X, we can approximate $f(x|h)$ by its sample mean:

$$\widehat{f}(x|h) = \frac{1}{n}\sum_{i=1}^n K_h(X_i - x) = \frac{1}{nh}\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right), \tag{3}$$

which is known as the kernel estimator of $f(x)$. Therefore, we can estimate $\pi(h|x)$ by

$$\widehat{\pi}(h|x) = \frac{\widehat{f}(x|h)\pi(h|\theta)}{\int \widehat{f}(x|h)\pi(h|\theta)dh}.$$

Thus, for a given θ, we estimate $h_0(x|\theta)$ by

$$h_n(x|\theta) = \frac{\int h\widehat{f}(x|h)\pi(h|\theta)dh}{\int \widehat{f}(x|h)\pi(h|\theta)dh} \equiv \frac{q_n(x|\theta)}{p_n(x|\theta)}, \tag{4}$$

which we call the sample–based estimator of $h_0(x|\theta)$, where $q_n(x|\theta) = \int h\widehat{f}(x|h)\pi(h|\theta)dh$ and $p_n(x|\theta) = \int \widehat{f}(x|h)\pi(h|\theta)dh$. We will investigate the asymptotic properties of $h_n(x) = h_n(x|\theta)$ for each given θ in the next section.

Since θ involved is a vector of hyperparameters, to the best of our knowledge, its estimation theory remains difficult and unavailable from the literature. This is the reason why specific values are often chosen for θ when they are needed in empirical examples. This paper imposes a kind of general condition in Assumption 4(i) below on the existence of a consistent estimator, $\widehat{\theta}$, of θ_0, the true value of θ. Meanwhile, some discussion about how to construct such consistent estimator is given in Appendix A after the proposed estimation method for the hyperparameters is used in both the simulations and the empirical applications in Sections 4–6, respectively.

We consider the case where the prior of h, $\pi(h|\theta)$, can be estimated by $\pi(h|\widehat{\theta})$. We then define the semiparametric localized bandwidth (SLB) estimator of $h_0(x|\theta)$ as

$$\widehat{h}_n(x|\widehat{\theta}) = \frac{\int h\widehat{f}(x|h)\pi(h|\widehat{\theta})dh}{\int \widehat{f}(x|h)\pi(h|\widehat{\theta})dh} \equiv \frac{\widehat{q}_n(x|\widehat{\theta})}{\widehat{p}_n(x|\widehat{\theta})}, \tag{5}$$

which $\widehat{q}_n(x|\widehat{\theta}) = \int h\widehat{f}(x|h)\pi(h|\widehat{\theta})dh$ and $\widehat{p}_n(x|\widehat{\theta}) = \int \widehat{f}(x|h)\pi(h|\widehat{\theta})dh$.

The resulting kernel density estimate is

$$\widehat{f}_n^*(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\widehat{h}_n(x|\widehat{\theta})} K\left(\frac{X_i - x}{\widehat{h}_n(x|\widehat{\theta})}\right).$$

In the next section, we establish asymptotic properties for $\widehat{h}_n(x|\widehat{\theta})$ and then the corresponding kernel density estimate $\widehat{f}_n^*(x)$.

3 Asymptotic theory

The following conditions are required to establish some asymptotic results, although some of them might not be the weakest possible.

Assumption 1. (i) $\{X_i\}$ is a strictly stationary and α–mixing process with mixing coefficient $\alpha(\cdot)$ satisfying $\sum_{j=1}^{\infty} \alpha(j)^{1-2/\delta} < \infty$; (ii) $\mathbb{E}(|X_i|^{\delta}) < \infty$ for some constant $\delta > 2$.

Assumption 2. $K(\cdot)$ is continuous, bounded and symmetric probability kernel function with $\int u^2 K(u)\,du < \infty$.

Assumption 3. Let $\pi(h|\theta)$ satisfy

$$|\pi(h|\theta) - \pi(h|\theta_0)| \leq L(h|\theta_0)||\theta - \theta_0||, \tag{6}$$

where $L(h|\theta_0)$ is some positive function such that $\int h^i f(x|h) L(h|\theta_0)\,dh < \infty$, for $i = 0, 1$, and $\theta \in \theta(\epsilon) = \{\theta : ||\theta - \theta_0|| \leq \epsilon\}$ for some $\epsilon > 0$. In addition, $\pi(h|\theta)$ satisfies

$$\int h^{1-\delta} \left| hp(x) - q(x) \right| \Lambda(x,h)\pi(h|\theta)\,dh < \infty \text{ and } \int \left| hp(x) - q(x) \right| \pi(h|\theta)\,dh < \infty$$

for the same $\delta > 2$ as in Assumption 1(i), in which $\Lambda(x,h) = \int K^{\delta}(v) f(x + vh)\,dv$.

Assumption 4. (i) There is some consistent estimator $\widehat{\theta}$ of θ_0 such that $\sqrt{n}\left(\widehat{\theta} - \theta_0\right) = o_P(1)$.

(ii) $f(x)$ is twice differentiable and the second–order derivative, $f^{(2)}(x)$, is continuous.

(iii) There is some nonnegative function $L(\cdot)$ such that $|K(y) - K(x)| \leq L(x)\ |y - x|$ for $y \in \mathscr{S}(x) = \{y : |y - x| \leq \epsilon\}$ for some $\epsilon > 0$, in which $L(\cdot)$ satisfies $\int |v|^j L(v)\,dv < \infty$ for $j = 0, 1, 2$.

Except Assumption 4(i), the rest of Assumptions 1–4 are standard and verifiable conditions. Assumption 4(i) imposes a kind of high–level condition on the existence of a super–consistent estimator of θ_0. Since there may be some different ways of constructing a consistent estimate for θ_0, instead of providing one set of low–level conditions, we propose using the general high–level condition to ensure the super–consistency. The basic idea is to find a suitable pseudo sample with n^* being the sample size such that $\frac{n}{n^*} \to 0$ and $\sqrt{n^*}\left(\widehat{\theta} - \theta_0\right) = O_P(1)$ under the probability distribution on the pseudo sample space to imply $\sqrt{n}\left(\widehat{\theta} - \theta_0\right) = \frac{\sqrt{n}}{\sqrt{n^*}} \cdot \sqrt{n^*}\left(\widehat{\theta} - \theta_0\right) = o_P(1)$. Appendix A below provides some detailed discussion about how such super–consistency may be achieved when a likelihood function is constructed based on a pseudo sampling method.

Such an estimation method will be used in both the simulation studies and the empirical applications of this paper.

We introduce the following notation: $f_u(x) = \frac{1}{u}\int K\left(\frac{y-x}{u}\right) f(y)dy$,

$$
\begin{aligned}
f_{uv}(x) &= \frac{1}{uv}\int K\left(\frac{y-x}{u}\right) K\left(\frac{y-x}{v}\right) f(y)dy, \\
R_{uv}(x) &= f_{uv}(x) - f_u(x) f_v(x), \\
g_{uv,s}(x) &= \frac{1}{uv}\int K\left(\frac{y-x}{u}\right) K\left(\frac{z-x}{v}\right) f_s(y,z)dydz, \\
G_{uv,s}(x) &= g_{uv,s}(x) - f_u(x) f_v(x), \\
m_{uv,s}(x_a,x_b) &= \frac{1}{uv}\int K\left(\frac{y-x_a}{u}\right) K\left(\frac{z-x_b}{v}\right) f_s(y,z)dydz, \\
S_{uv,s}(x_a,x_b) &= m_{uv,s}(x_a,x_b) - f_u(x_a) f_v(x_b), \\
\int y^2 K(y)dy &= \mu_2(K) > 0, \quad R(K) = \int K^2(y)dy,
\end{aligned}
$$

where $s = |i-j|$ and $f_{|i-j|}(y,z)$ denotes the joint density of (X_i, X_j).

We now present the asymptotic properties of sample–based bandwidth estimator $h_n(x|\theta)$ and SLB estimator $\widehat{h}_n(x|\widehat{\theta})$. To simplify the notations in the presentation of all the theoretical results, we use $p(x)$, $q(x)$, $p_n(x)$, $q_n(x)$, $\widehat{p}_n(x)$, $\widehat{q}_n(x)$, $h_0(x)$, $h_n(x)$ and $\widehat{h}_n(x)$ to denote $p(x|\theta)$, $q(x|\theta)$, $p_n(x|\theta)$, $q_n(x|\theta)$, $\widehat{p}_n(x|\widehat{\theta})$, $\widehat{q}_n(x|\widehat{\theta})$, $h_0(x|\theta)$, $h_n(x|\theta)$ and $\widehat{h}_n(x|\widehat{\theta})$, respectively. Let $Q(x) = p^2(x)$.

We then establish some new theorems; their proofs are given in Appendix B below.

Theorem 1. *Under Assumptions 1–3, we have as* $n \to \infty$

$$\sqrt{n}(h_n(x) - h_0(x)) \to_D \mathcal{N}(0, \Sigma_0(x)), \tag{7}$$

where $\Sigma_0(x) = Q^{-2}(x)\Sigma_L(x)$ *and* $\Sigma_L(x) = \gamma(0) + 2\sum_{j=1}^{\infty}\gamma(j)$ *with*

$$
\begin{aligned}
\gamma(j) &= \iint \left[up(x) - q(x)\right]\left[vp(x) - q(x)\right] G_{uv,j}(x)\pi(u)\pi(v)dudv, \\
\gamma(0) &= \iint [up(x) - q(x)][vp(x) - q(x)] R_{uv}(x)\pi(u)\pi(v)dudv.
\end{aligned}
$$

Remark 1. This theorem shows that $h_n(x)$ is asymptotically normal and is a consistent estimator of $h_0(x)$ with root–n rate of convergence. This property holds for the localized bandwidth at each point x.

Theorem 2. *Let $N \geq 2$ be an integer and let $x_1, x_2, \cdots, x_N$ be fixed points. Under Assumptions 1–3, we have as $n \to \infty$*

$$\left[\sqrt{n}\left(h_n(x_1) - h_0(x_1)\right), \cdots, \sqrt{n}\left(h_n(x_N) - h_0(x_N)\right)\right] \to_D \mathcal{N}(0, \Sigma_N), \tag{8}$$

where $\Sigma_{N,aa} = \Sigma_0(x_a)$, $\Sigma_{N,ab} = Q^{-1}(x_a)\Sigma_v(x_a, x_b)Q^{-1}(x_b)$, $\Sigma_v(x_a, x_b) = \gamma_{ab}(0) + 2\sum_{s=1}^{\infty}\gamma_{ab}(s)$, $\gamma_{ab}(s) = \mathbb{E}\left[V_i(x_a)V_j(x_b)\right] = \iint \left[up(x_a) - q(x_a)\right]\left[vp(x_b) - q(x_b)\right] S_{uv,s}(x_a, x_b)\pi(u)\pi(v)dudv$ *and* $\gamma_{ab}(0) = \mathbb{E}\left[V_i(x_a)V_i(x_b)\right] = \iint \left[up(x_a) - q(x_a)\right]\left[vp(x_b) - q(x_b)\right] S_{uv,0}(x_a, x_b)\pi(u)\pi(v)dudv$.

Remark 2. This theorem shows that when we consider localized bandwidth around more than one point, the corresponding localized bandwidths are asymptotically jointly normal.

Theorem 3. *Let Assumptions 1–4(i) hold. Then, as $n \to \infty$,*

$$\sqrt{n}(\widehat{h}_n(x) - h_0(x)) \to_D \mathcal{N}(0, \Sigma_0(x)), \tag{9}$$

where $\Sigma_0(x)$ is defined in Theorem 1.

Remark 3. This theorem shows that the SLB estimator $\widehat{h}_n(x)$ is asymptotically normal and is a consistent estimator of $h_0(x)$ with a root–n rate of convergence.

Theorem 4. *Let $N \geq 2$ be an integer and let $x_1, x_2, \cdots, x_N$ be fixed points. Let Assumptions 1–4(i) hold. Then as $n \to \infty$,*

$$\left[\sqrt{n}(\widehat{h}_n(x_1) - h_0(x_1)), \cdots, \sqrt{n}(\widehat{h}_n(x_N) - h_0(x_N))\right] \to_D \mathcal{N}(0, \Sigma_N), \tag{10}$$

where $\Sigma_N(x)$ is defined in Theorem 2.

Remark 4. (i) This theorem shows that when we consider the SLB estimators at different density points, the resulting bandwidths are asymptotically joint normal.

(ii) The discussion in Theorems 1–4 for the estimators of $h_0(x)$ can be extended to the case where $h = a_n \cdot b$, in which $a_n \to 0$ as $n \to \infty$, and b is being treated as an unknown parameter. In this case, we have

$$h_0(x) = \mathbb{E}_3[h|x] = a_n \cdot \mathbb{E}_3[b|x] \equiv a_n \cdot b_0(x), \tag{11}$$

where $h_0(x)$ is as defined in (2), and $\mathbb{E}_3[h|x]$ denotes the conditional expectation of h given $X = x$. In this case, we have $b_n(x) = a_n^{-1}h_n(x)$ and $\widehat{b}_n(x) = a_n^{-1} \cdot \widehat{h}_n(x)$. As a consequence, the rate of convergence of $b_n(x) - b_0(x)$ and $\widehat{b}_n(x) - b_0(x)$ becomes $n^{-\frac{1}{2}}a_n^{-1}$. In the univariate case where $a_n = n^{-\frac{1}{5}}$, the rate of convergence of $b_n(x)$ to $b_0(x)$ reduces to $n^{-\frac{3}{10}}$.

Meanwhile, the rate of $n^{-5/10}$ of $\widehat{h}_n(x) - h_0(x)$ is faster than that of $n^{-3/10}$ of $\widehat{h}_{\text{cv}} - h_0$ obtained by Härdle et al. (1988) for the case where the bandwidth $h_0 \sim n^{-1/5}$ was treated as an unknown fixed value and selected by the conventional cross–validation method.

Remark 5. For a multivariate kernel density setting, we can still use the proposed semiparametric localized bandwidth selection method to choose an optimal bandwidth matrix H following de Lima and Atuncar (2011) and obtain similar asymptotic results to those in the univariate case, although the technical details are complicated.

Recall $\mu_2(K) = \int v^2 K(v)dv$ and $R(K) = \int K^2(v)dv$. We then have the following asymptotic normality.

Theorem 5. *Let Assumption 1–4 hold. If, in addition,* $h_0(x) = a_n\ b_0(x) \to 0$ *and* $nh_0(x) \to \infty$ *as* $n \to \infty$*, then for each given* x*, we have as* $n \to \infty$

$$\sqrt{nh_0(x)}\left(\widehat{f}_n^*(x) - f(x) - \frac{h_0^2(x)}{2}\ \mu_2(K)\ f^{(2)}(x)\right) \to_D \mathcal{N}(0, R(K)f(x)).$$

Theorem 5 shows that the conventional normality is still achievable for $\widehat{f}_n^*(x)$. In Sections 4–6 below, we evaluate the applicability and the finite–sample performance of $\widehat{f}_n^*(x)$.

4 Monte Carlo simulation results

We present several examples to examine the finite–sample performance of the semiparametric localized bandwidth estimator. Assume that the prior of h^2 is the density of an inverse Gamma distribution denoted as $h^2 \sim IG(\alpha, \beta)$ with its probability density function (pdf) expressed

$$\pi\left(h^2\right) = \frac{\beta^\alpha}{\Gamma(\alpha)}\left(\frac{1}{h^2}\right)^{\alpha+1} \exp\left\{-\frac{\beta}{h^2}\right\}, \tag{12}$$

where $\theta = (\alpha, \beta)'$ is a vector of two hyperparameters. Therefore, the prior of h is

$$\pi\left(h|\theta\right) = \frac{2\beta^\alpha}{\Gamma(\alpha)}\left(\frac{1}{h}\right)^{2\alpha+1} \exp\left\{-\frac{\beta}{h^2}\right\}. \tag{13}$$

In both the simulation studies and empirical applications of this paper, we propose to estimate the true value of θ in the same way as proposed in Appendix A below. Previously in the literature, the true value of θ has often been chosen arbitrarily in both simulation studies and empirical applications.

4.1 Computational aspects

In most situations, a closed form expression of $h_0(x|\theta)$, $h_n(x|\theta)$ and $\widehat{h}_n(x|\widehat{\theta})$ is not available. We introduce the following Lemmas to approximate $h_0(x|\theta)$, $h_n(x|\theta)$ and $\widehat{h}_n(x|\widehat{\theta})$. Their proofs are given in the appendix.

Let

$$
\begin{aligned}
p_m(x) &= \frac{1}{m}\sum_{i=1}^{m} f(x+u_i v_i), & q_m(x) &= \frac{1}{m}\sum_{i=1}^{m} f(x+u_i v_i)v_i, \\
p_{nm}(x) &= \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{1}{v_j}K\left(\frac{X_i-x}{v_j}\right), & q_{nm}(x) &= \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}K\left(\frac{X_i-x}{v_j}\right),
\end{aligned} \tag{14}
$$

where u_i and v_i are drawn from respectively, $K(u)$ and $\pi(v|\theta)$, and $K(u)$ is the Gaussian kernel.

Lemma 1. As $m\to\infty$, $p_m(x)-p(x)=o_P(1)$ and $q_m(x)-q(x)=o_P(1)$.

Lemma 2. As $m\to\infty$, $p_{nm}(x)-p_n(x)=o_P(1)$ and $q_{nm}(x)-q_n(x)=o_P(1)$.

Based on Lemmas 1 and 2, we approximate $h_0(x|\theta)$ and $h_n(x|\theta)$ by respectively, $q_m(x)/p_m(x)$ and $q_{nm}(x)/p_{nm}(x)$, which involves unknown hyperparameters, α and β. Therefore, we need to estimate these two hyperparameters, so as to estimate $h_0(x|\theta)$ and $h_n(x|\theta)$.

As the prior of h given by (13) belongs to a conjugate family of $\widehat{f}(x|h)$, we can work out the denominator of (4), and it turns out to be

$$
p_n(x|\theta) = \frac{1}{n}\frac{\beta^{\alpha}\Gamma(\alpha+1/2)}{\sqrt{2\pi}\Gamma(\alpha)}\sum_{i=1}^{n}\left(\frac{(X_i-x)^2}{2}+\beta\right)^{-(\alpha+1/2)}. \tag{15}
$$

Thus, for the simulated random sample $\left\{X_j^* : j=1,2,\cdots,n^*\right\}$ generated by the sampling method proposed in Appendix A, we have

$$
p_n\left(X_j^*|\theta\right) = \int_0^{\infty}\widehat{f}\left(X_j^*\middle|h\right)\pi(h|\theta)dh = \frac{1}{n}\frac{\beta^{\alpha}\Gamma(\alpha+1/2)}{\sqrt{2\pi}\Gamma(\alpha)}\sum_{i=1}^{n}\left(\frac{\left(X_i-X_j^*\right)^2}{2}+\beta\right)^{-(\alpha+1/2)}. \tag{16}
$$

Therefore, the likelihood function of X_j^*, for $j=1,2,\cdots,n^*$, given α and β, is

$$
\ell_*\left(X_1^*,X_2^*,\cdots,X_{n^*}^*|\theta\right) = \prod_{j=1}^{n^*}p_n\left(X_j^*\middle|\theta\right) = \prod_{j=1}^{n^*}\left\{\frac{1}{n}\frac{\beta^{\alpha}\Gamma(\alpha+1/2)}{\sqrt{2\pi}\Gamma(\alpha)}\sum_{i=1}^{n}\left(\frac{(X_i-X_j^*)^2}{2}+\beta\right)^{-(\alpha+1/2)}\right\}, \tag{17}
$$

which is maximized to derive the MLE of θ denoted as $\widehat{\theta}$. Therefore, $h_n(x|\theta)$ is estimated by

$\widehat{h}_n(x|\widehat{\theta}) = q_{nm}(x)/p_{nm}(x)$ defined in (14) with v_j drawn from $\pi(v|\widehat{\theta})$.

4.2 Simulation results

A key issue to the estimation of θ is how to choose its initial value, which will be needed before we simulate a random sample from $p_n\left(X_j^*|\theta_0\right)$.

While we may not be able to construct $\widehat{\theta}$ purely based on $\{X_i : i = 1,2,\cdots,n\}$, we should still be able to find an initial value for θ by using the original sample $\{X_i : i = 1,2,\cdots,n\}$. The density of X_i is approximately $p_n(X_i|\theta)$, for $i = 1,2,\cdots,n$. The likelihood of X_i, for $i = 1,2,\cdots,n$, given θ is

$$\ell(X_1,X_2,\cdots,X_n|\theta) = \prod_{i=1}^{n}\left\{\frac{1}{n}\frac{\beta^\alpha\Gamma(\alpha+1/2)}{\sqrt{2\pi}\Gamma(\alpha)}\sum_{j=1;j\neq i}^{n}\left(\frac{\left(X_j-X_i\right)^2}{2}+\beta\right)^{-(\alpha+1/2)}\right\}. \tag{18}$$

which we maximize with respect to θ to obtain an initial estimate of θ denoted as θ_0.

As proposed in Appendix A below, we simulate a random sample from $p_n(x|\theta_0)$, and then use this sample to derive the MLE of θ. The details of this simulation–based likelihood approach to hyperparameter estimation are given as follows.

Step 1: Obtain an initial value for θ by maximizing (18) using the original sample $\{X_i : i = 1,2,\cdots,n\}$.

Step 2: Generate a random sample, $\left\{X_j^* : j = 1,2,\cdots,n^*\right\}$, from $p_n\left(X^*|\theta_0\right)$ given by (16) through the random–walk Metropolis algorithm.

Step 3: Estimate θ based on the generated sample by maximizing the likelihood function given by (17); and denote the resulting estimate as $\widehat{\theta}$.

Step 4: Derive the SLB estimate of $h_0(x|\theta_0)$ as $\widehat{h}_n(x|\widehat{\theta})$.

We aim to compare the performance of $h_0(x|\theta_0)$, $h_n(x|\theta_0)$ and $\widehat{h}_n(x|\widehat{\theta})$ through simulated random samples, where sample sizes are respectively, 250, 750 and 1500, and the number of replications for each sample size is 500. To generate such random samples, we consider the following four data generating processes: normal distribution $N(0,1)$, the mixture of two normal distributions $0.4\mathcal{N}(-0.6,0.16)+0.6\mathcal{N}(0.4,0.49)$, the Weibull distribution with shape parameter 1.5 and scale parameter 1 denoted as $\mathcal{W}(1.5,1)$, and a stationary AR(1) process $x_t = 0.2x_{t-1}+u_t$ where $u_t \sim \mathcal{N}(0,1)$.

For each sample size considered, we calculated $h_0(x|\theta_0)$, $h_n(x|\theta_0)$ and $\widehat{h}_n(x|\widehat{\theta})$ using each of the 500 random samples generated from each distribution, where x takes values on 100 equally spaced grid points on a finite interval. Such intervals are chosen to be $(-3,3)$ for $\mathcal{N}(0,1)$, $(-3,4)$ for the mixture density, $(0,4.1)$ for $\mathcal{W}(1.5,1)$ and $(-4,4)$ for the AR(1) process $x_t = 0.2x_{t-1} + u_t$. We then took the average of each bandwidth curve over 500 replications. We plotted these three averaged bandwidth curves in Figure 1.

Parts (1)–(3) of Figure 1 present the bandwidth curves averaged over 500 samples, which were generated from $\mathcal{N}(0,1)$. These three averaged bandwidth curves clearly differ from each other when the size of the original sample is 250, but are almost the same when the sample size is 1500. Each averaged bandwidth curve indicates that a global bandwidth is inappropriate. However, these bandwidth curves may indicate that bandwidth may be approximately treated as a constant within a certain interval, which may choose for example, $(-2,2)$. As the underlying true density is the standard Gaussian, this interval is approximately the 95% high density region. Nonetheless, localized bandwidths should be used in both tails of the underlying density.

Parts (4)–(6) of Figure 1 present the bandwidth curves averaged over 500 samples that were generated from the mixture density of two Gaussians. These three averaged bandwidth curves are almost the same regardless size of the original sample. However, the shape of each bandwidth curve indicates that localized bandwidth should be used in the kernel estimator of the underlying mixture density. When x is within a certain interval, which for example, is from -1.5 to 2, a high density region, bandwidth can be approximately treated as a constant.

Parts (7)–(9) of Figure 1 present the bandwidth curves averaged over 500 samples, which were generated from the Weibull distribution. Regardless the size of original samples, $\widehat{h}_n(x|\widehat{\theta})$ is almost the same as the other two bandwidth curves when x is less than 3, but is clearly different from the other two when x is greater than 3. The shape of each bandwidth curve indicates that localized bandwidth should be used when x is greater than a threshold value. The graphs show that this threshold can be chosen as 2.5 for sample size of 250, 3 for sample size of 750, and 3.5 for sample size of 1500. Bandwidth can be treated as a constant when x is less the threshold value, which classifies a high density region.

Parts (10)–(12) of Figure 1 present the bandwidth curves averaged over 500 samples, which were generated from the AR(1) process. The shape of each bandwidth curve indicates that localized bandwidth should be used in the kernel estimator of the underlying density. When x is within a certain interval, which for example, is from -2 to 2, a high density region, bandwidth can be approximately treated as a constant.

To sum up, the estimated bandwidth curve indicates that localized bandwidth should be used for kernel density estimation, although bandwidth can be approximately treated as a constant in the high density region. Moreover, we found that our proposed semiparametric localized bandwidth curve, $\widehat{h}_n(x|\widehat{\theta})$, clearly differs from $h_n(x|\widehat{\theta}_0)$ when random samples are generated respectively, from the Gaussian and Weibull distributions.

We now calculate the bias and standard deviation (std) of each bandwidth curve over 500 replications. Let $h_{n,r}(x_i|\theta_0)$ and $\widehat{h}_{n,r}(x_i|\widehat{\theta})$ denote $h_n(x_i|\theta_0)$ and $\widehat{h}_n(x_i|\widehat{\theta})$ calculated at x_i using the rth sample, for $r = 1,2,\cdots,500$, and $i = 1,2,\cdots,100$. For each sample size, we calculate the bias and standard deviation measures as follows:

$$\begin{aligned}
\text{bias}_1 &= \frac{1}{100}\frac{1}{500}\sum_{i=1}^{100}\sum_{r=1}^{500}\left(h_{n,r}(x_i|\theta_0) - h_0(x_i|\theta_0)\right),\\
\text{bias}_2 &= \frac{1}{100}\frac{1}{500}\sum_{i=1}^{100}\sum_{r=1}^{500}\left(\widehat{h}_{n,r}(x_i|\widehat{\theta}) - h_0(x_i|\theta_0)\right),\\
\text{std}_1 &= \sqrt{\frac{1}{100}\frac{1}{500}\sum_{i=1}^{100}\sum_{r=1}^{500}\left(h_{n,r}(x_i|\theta_0) - \overline{h}_n(x_i|\theta_0)\right)^2},\\
\text{std}_2 &= \sqrt{\frac{1}{100}\frac{1}{500}\sum_{i=1}^{100}\sum_{r=1}^{500}\left(\widehat{h}_{n,r}(x_i|\widehat{\theta}) - \overline{\widehat{h}}_n(x_i|\widehat{\theta})\right)^2},
\end{aligned}$$

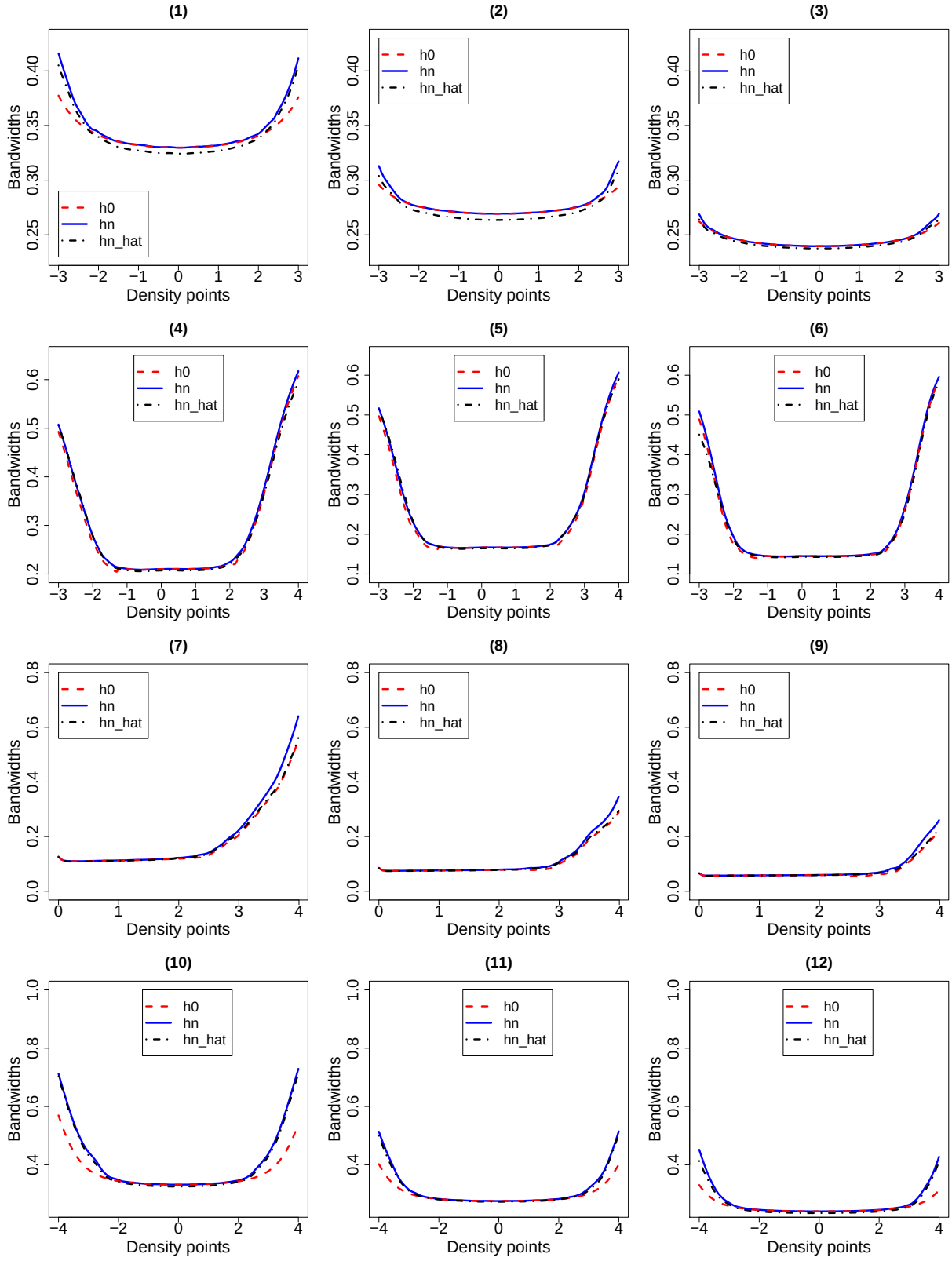
where $\overline{h}_n(x_i|\theta_0) = 1/500\sum_{r=1}^{500} h_{n,r}(x_i|\theta_0)$ and $\overline{\widehat{h}}_n(x_i|\widehat{\theta}) = 1/500\sum_{r=1}^{500}\widehat{h}_{n,r}(x_i|\widehat{\theta})$. The MSEs of $h_n(x_i|\theta_0)$ and $\widehat{h}_n(x_i|\widehat{\theta})$ are calculated as

$$\text{MSE}_j = \text{bias}_j^2 + \text{std}_j^2,$$

for $j = 1$ and 2. As shown by (14), the computation of $h_n(x|\theta_0)$ and $\widehat{h}_n(x|\widehat{\theta})$ involves simulating m random numbers from each of $K(u)$ and $\pi(v|\theta_0)$, respectively. We choose $m = 5000$ in this study. Table 1 presents the results of bias and standard deviation values obtained based on random samples generated from $\mathcal{N}(0,1)$, $0.4\mathcal{N}(-0.6,0.16) + 0.6\mathcal{N}(0.4,0.49)$, $\mathcal{W}(1.5,1)$ and the AR(1) process $x_t = 0.2x_{t-1} + u_t$, respectively.

We found the following evidence. First, when random samples are generated from the Gaussian density, $\widehat{h}_n(x|\widehat{\theta})$ has larger bias and variation than $h_n(x|\theta_0)$ for each sample size considered. Consequently, under the MSE measure, $\widehat{h}_n(x|\widehat{\theta})$ performs poorer than $h_n(x|\theta_0)$. However, when samples are generated from the mixture density of two Gaussians and the AR(1) process, $\widehat{h}_n(x|\widehat{\theta})$ has smaller bias and variation than $h_n(x|\theta_0)$, and thus, $\widehat{h}_n(x|\widehat{\theta})$ outperforms $h_n(x|\theta_0)$ under the

Figure 1: *Averaged* $h_0(x|\theta_0)$, $h_n(x|\theta_0)$ *and* $\widehat{h}_n(x|\widehat{\theta})$ *with random samples generated from* $\mathcal{N}(0,1)$ *(first row),* $0.4\mathcal{N}(-0.6,0.16)+0.6\mathcal{N}(0.4,0.49)$ *(second row),* $\mathcal{W}(1.5,1)$ *(third row) and AR(1) process (fourth row), where* θ_0 *is estimated using the original sample of* $\boldsymbol{x}$*. The three columns correspond to sample sizes of 250, 750 and 1500, respectively.*



MSE measure. When samples are generated from the Weibull distribution, $\widehat{h}_n(x|\widehat{\theta})$ has smaller bias and variation than $h_n(x|\theta_0)$, except for the sample size of 750 in the latter measure. Under this distribution, the MSE measure shows that $\widehat{h}_n(x|\widehat{\theta})$ performs slightly better than $h_n(x|\theta_0)$ for large samples, while the latter performs slightly better than the former for small samples.

Second, as the sample size increases, all three measures of $h_n(x|\theta_0)$ and $\widehat{h}_n(x|\widehat{\theta})$ decreases, respectively.

Table 1: *Bias, standard deviation and MSE of SLB estimates with random samples generated from four different data generating processes.*

	Sample	SLB			
	size	Gaussian	Mixture	Weibull	AR(1)
$h_n(x\|\theta_0) - h_0(x\|\theta_0)$					
	250	0.0055	0.0261	0.0574	0.0364
bias_1	750	0.0020	0.0188	0.0281	0.0162
	1500	0.0007	0.0129	0.0184	0.0141
	250	0.0929	0.0984	0.1370	0.2556
std_1	750	0.0536	0.0743	0.0698	0.1527
	1500	0.0370	0.0567	0.0388	0.1236
	250	0.0087	0.0104	0.0221	0.0666
MSE_1	750	0.0029	0.0059	0.0057	0.0236
	1500	0.0014	0.0034	0.0018	0.0155
$\widehat{h}_n(x\|\widehat{\theta}) - h_0(x\|\theta_0)$					
	250	0.0079	0.0224	0.0465	0.0288
bias_2	750	−0.0033	0.0174	0.0230	0.0121
	1500	−0.0015	0.0082	0.0142	0.0064
	250	0.0999	0.0867	0.1360	0.2319
std_2	750	0.0585	0.0505	0.0732	0.1376
	1500	0.0440	0.0338	0.0338	0.1114
	250	0.0100	0.0080	0.0207	0.0546
MSE_2	750	0.0034	0.0029	0.0059	0.0191
	1500	0.0019	0.0012	0.0013	0.0125

We also examine the performance of our proposed likelihood approach to hyperparameter estimation by the resulting kernel density estimates. The estimation of the density $f(\cdot)$ is carried out at m equally spaced points $x_1, x_2, \cdots, x_m$ with the corresponding estimates given by

Table 2: *ASE of kernel density estimates using different localized and global bandwidth* ($\times 10^{-3}$).

			SLB		CV	Bayes
DGP	n	ASE($h_0(x\|\theta_0)$)	ASE($h_n(x\|\theta_0)$)	ASE($\widehat{h}_n(x\|\widehat{\theta})$)	ASE(h_{cv})	ASE(h_{bayes})
Normal	250	0.0903	0.0933	0.0896	0.5660	0.1086
	750	0.0453	0.0457	0.0430	0.2436	0.0602
	1500	0.0268	0.0269	0.0264	0.1344	0.0389
Mixture	250	0.1403	0.1527	0.1413	0.7750	0.2175
	750	0.0601	0.0707	0.0649	0.3296	0.1073
	1500	0.0450	0.0537	0.0441	0.2124	0.0801
Weibull	250	0.5709	0.6418	0.6011	2.2265	1.0902
	750	0.3605	0.3783	0.3628	1.2450	0.6654
	1500	0.2681	0.2747	0.2675	0.7973	0.5039
AR(1)	250	0.0699	0.0804	0.0751	0.4307	0.0864
	750	0.0427	0.0441	0.0430	0.1980	0.0557
	1500	0.0199	0.0205	0.0189	0.1090	0.0292

$\widehat{f}_n^*(x_1), \cdots, \widehat{f}_n^*(x_m)$. We define the goodness of fit criterion by the average squared error (ASE)

$$ASE = \frac{1}{m}\sum_{i=1}^{m}\left(\overline{\widehat{f}}_n^*(x_i) - f(x_i)\right)^2, \tag{19}$$

where $\overline{\widehat{f}}_n^*(x_i) = 1/500\sum_{r=1}^{500}\widehat{f}_n^*(x_{ri})$ and $\widehat{f}_n^*(x_{ri})$ is the density estimate at point x_i in the r–th replication. We use ASE($h_0(x|\theta_0)$), ASE($h_n(x|\theta_0)$), ASE($\widehat{h}_n(x|\widehat{\theta})$), ASE($h_{\text{cv}}$) and ASE($h_{\text{bayes}}$) to denote the ASE of the kernel density estimate associated with each of the bandwidths $h_0(x|\theta_0)$, $h_n(x|\theta_0)$, $\widehat{h}_n(x|\widehat{\theta})$, h_{cv} and h_{bayes}, respectively.

The results are summarized in Table 2. From Table 2, we found the following evidence. (1) Under each DGP and for each bandwidth choice, with the increase of sample size, ASE of each of the kernel density estimates decreases. (2) For all the cases above, the kernel density estimates associated with the localized bandwidth have much smaller ASE than those based on the global bandwidth chosen by either CV or Bayesian method. (3) For all the cases above, ASE of each of the kernel density estimates associated with $\widehat{h}_n(x|\widehat{\theta})$ is smaller than that based on $\widehat{h}_n(x|\theta_0)$, which indicates that our estimated hyperparameters can result in better kernel density estimates.

5 Estimating and forecasting the density of Eurodollar deposit rate

We estimate localized bandwidth for the kernel density estimator of Eurodollar deposit rate using our proposed SLB method. The performance of the estimated bandwidth is measured by the resulting kernel density estimator. For comparison purposes, we also examine the performance of a global bandwidth selected or estimated by cross–validation (CV) and a Bayesian sampling approach of Zhang, King, and Hyndman (2006). The performance of these methods are also examined via the forecasting performance of the resulting kernel density estimates.

The data consists of daily Eurodollar deposit rates in London with deposit maturities of 1, 3 and 6 months, respectively. The data are collected from the website of Federal reserve bank of St. Louis in the U.S. The sample period is from the 4th January 1971 to the 3rd January 1995. Time series plots of these three Eurodollar deposit rates are presented in Figure 2.

5.1 Full sample density estimation

The kernel density estimator is given by (3), where we use localized bandwidth in the sense that the bandwidth depends on the density point. We assume an inverse Gamma prior of the bandwidth parameter given by (12), in which the hyperparameter vector, $\theta = (\alpha, \beta)'$, is estimated through our proposed likelihood–based approach.

For comparison purpose, a global bandwidth is also used in the kernel density estimator given by (3). Such a bandwidth is obtained through CV and Bayesian sampling, respectively. The resulting three kernel density estimates as well as the histogram of daily Eurodollar deposit rate with each maturity are plotted in Figure 3.

The estimated density with localized bandwidth clearly differs from the estimated density with a global bandwidth except for their right–tail areas, where both types of bandwidth lead to almost the same density estimate. Moreover, we find that the estimated density with our proposed semiparametric localized bandwidth is more close to the histogram than that with global bandwidths selected or estimated through either Bayesian sampling or cross–validation method. This finding indicates that our proposed semiparametric localized bandwidth may have better performance than the use of global bandwidth.

As in practice, the true density is unknown, we employ the scoring rule introduced by Amisano and Giacomini (2007) to examine the out–of–sample performance of an estimated density function. Using this scoring rule, we are able to decide the best performer among a group

Figure 2: *Eurodollar daily deposit rates with maturities of 1, 3 and 6 months: (1) 1–month maturity; (2) 3–month maturity; and (3) 6–month maturity.*

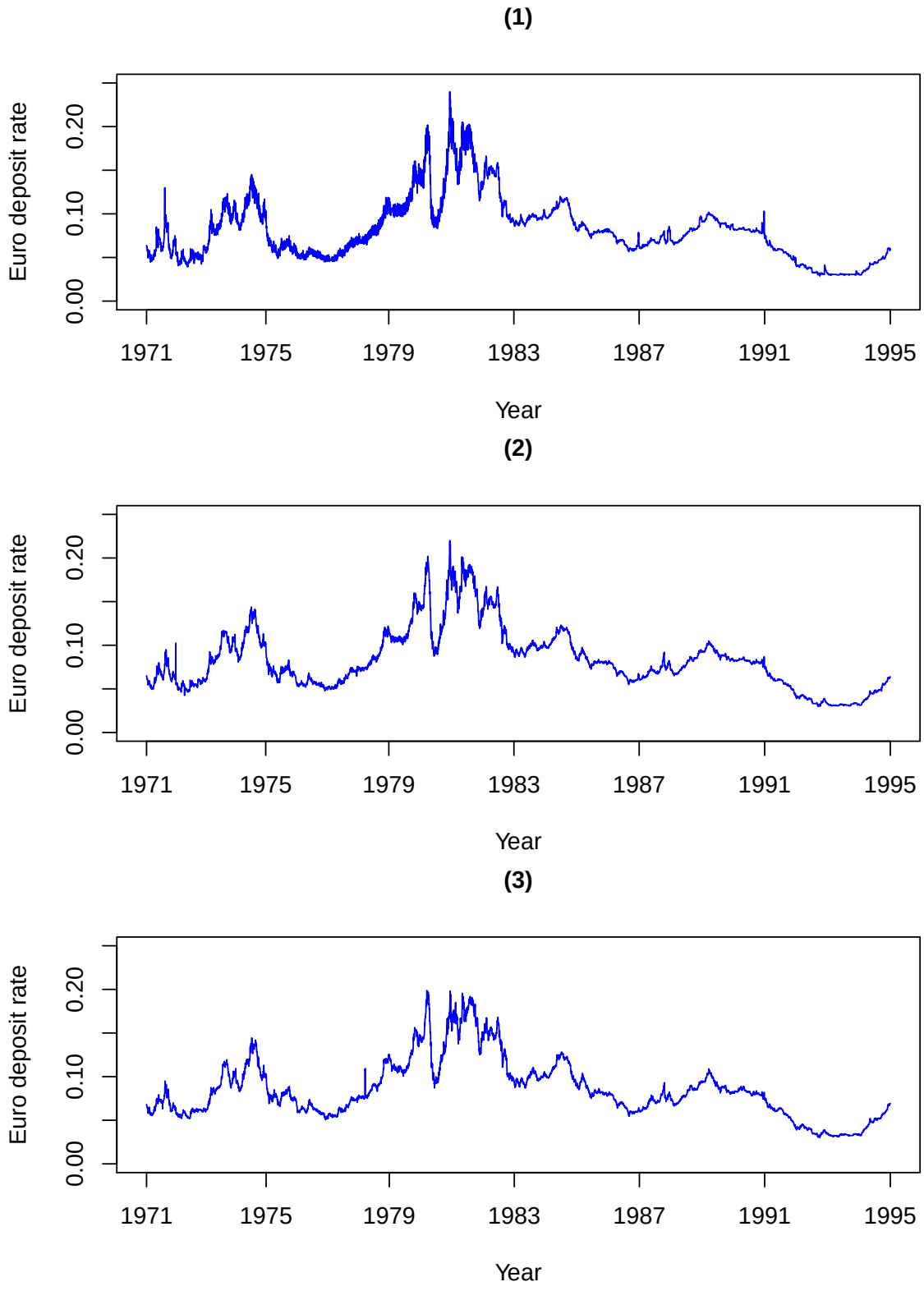
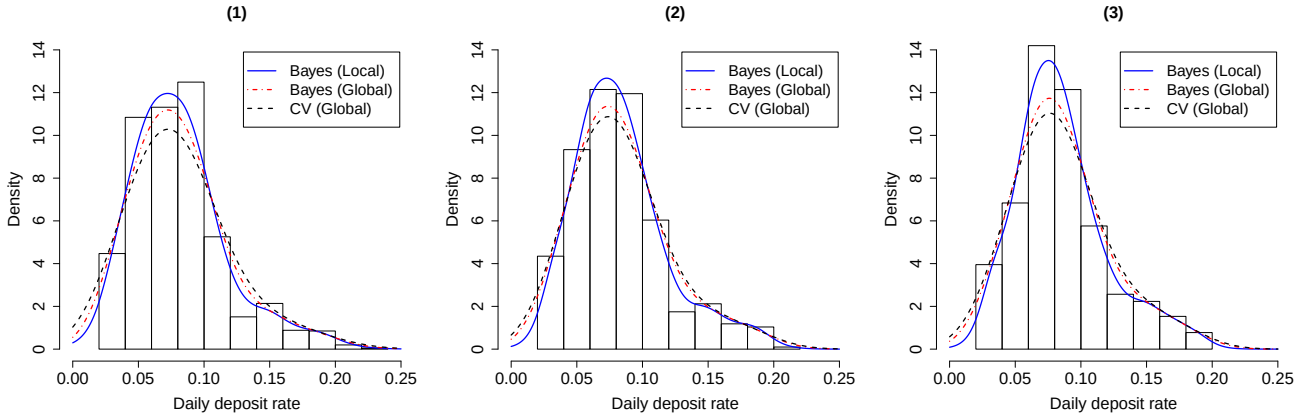


Figure 3: *Density estimates of Eurodollar daily deposit rates with maturities of 1, 3 and 6 months: (1) 1–month maturity; (2) 3–month maturity; and (3) 6–month maturity.*



of competing density estimates.

5.2 Forecasting results

We conducted a rolling–sample procedure to evaluate the performance of each density estimate resulted from each of the three bandwidth estimation methods. For each maturity, let T denote the number of all observations, and let y_t denote the observed deposit rate at day t, for $t = 1, 2, \cdots, T$. The first sample contains the first n observations, $y_1, y_2, \cdots, y_n$, and is used to derive the bandwidth through the above–mentioned three methods. The resulting three density estimates are used to forecast the density of y_{n+1}. The second sample contains $y_2, y_3, \cdots, y_{n+1}$, which are obtained by rolling the first sample forward for one step. Using this sample, we repeat what was done based on the previous sample and forecast the density of y_{n+2}. This rolling procedure continues until the density of y_T is forecasted.

At the rth iteration, for $r = 1, 2, \cdots, T-n$, we denote the SLB estimator as $h_{\text{SLB}}(y)$, the global bandwidth chosen through CV as h_{CV}, and the global bandwidth estimated through Bayesian sampling of Zhang, King, and Hyndman (2006) as h_{Bayes}. The corresponding density estimates are $\widehat{f}\left(y_{n+r}|h_{\text{SLB}}(y_{n+r})\right)$, $\widetilde{f}\left(y_{n+r}|h_{\text{CV}}\right)$ and $\widetilde{f}\left(y_{n+r}|h_{\text{Bayes}}\right)$.

We calculated the average logarithmic scores over the out–of–sample period:

$$\begin{aligned}
&\frac{1}{T-n}\sum_{r=1}^{T-n}\log\widehat{f}\left(y_{n+r}|h_{\mathrm{SLB}}(y_{n+r})\right)=\frac{1}{T-n}\sum_{r=1}^{T-n}\log\left[\frac{1}{n}\sum_{i=r}^{n+r-1}\frac{1}{h_{\mathrm{SLB}}(y_{n+r})}\phi\left(\frac{y_{n+r}-y_i}{h_{\mathrm{SLB}}(y_{n+r})}\right)\right],\\
&\frac{1}{T-n}\sum_{r=1}^{T-n}\log\widetilde{f}\left(y_{n+r}|h_{\mathrm{Bayes}}\right)=\frac{1}{T-n}\sum_{r=1}^{T-n}\log\left[\frac{1}{n}\sum_{i=r}^{n+r-1}\frac{1}{h_{\mathrm{Bayes}}}\phi\left(\frac{y_{n+r}-y_i}{h_{\mathrm{Bayes}}}\right)\right],\\
&\frac{1}{T-n}\sum_{r=1}^{T-n}\log\widetilde{f}\left(y_{n+r}|h_{\mathrm{CV}}\right)=\frac{1}{T-n}\sum_{r=1}^{T-n}\log\left[\frac{1}{n}\sum_{i=r}^{n+r-1}\frac{1}{h_{\mathrm{CV}}}\phi\left(\frac{y_{n+r}-y_i}{h_{\mathrm{CV}}}\right)\right],
\end{aligned}\tag{20}$$

where $\phi(\cdot)$ is the density function of the standard Gaussian distribution. In terms of the average logarithmic score, the larger it is, the better the corresponding density performs. Thus, we first rank the density forecasts by comparing their average scores and further select the forecast yielding the highest score.

Table 3 presents a summary of the total number of observations, size of the rolling sample, and number of rolling samples for each maturity.

Table 3: *A summary of rolling sample facts.*

Maturity	T	Size of rolling sample (n)	Number of rolling samples
1 month	6128	4059	2069
3 months	6129	4060	2069
6 months	6135	4066	2069

Table 4 presents the average logarithmic scores derived through the three density estimates. At the maturities of 1 month and 3 months, the use of localized bandwidth leads to a slightly better performance of the density estimator than the use of global bandwidth estimated through Bayesian sampling. The former performs clearly better than the latter at the maturity of 6 months. Moreover, at each maturity, the use of localized bandwidth clearly outperforms the use of a global bandwidth selected through CV. We can draw a conclusion that our proposed SLB estimator performs better than its competitors, which are Bayesian sampling and CV for estimating a global bandwidth.

Figure 4: *Time series plot of S&P 500 daily returns.*

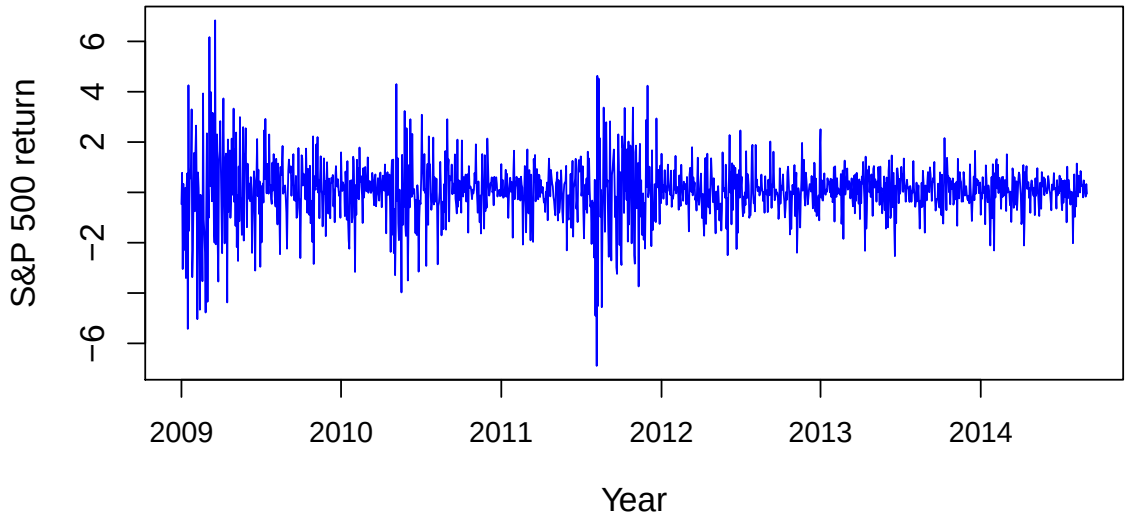


Table 4: *Out–of–sample average logarithmic scores of density estimates with bandwidths estimation through three different methods.*

	Average logarithmic score		
Maturity	Localized bandwidths	Global bandwidth (Bayesian)	Global bandwidth (CV)
1 month	2.4579	2.4561	2.3469
3 months	2.4981	2.4906	2.4471
6 months	2.6484	2.5253	2.4958

6 Estimating and forecasting the density of S&P 500 daily returns

It is important for being able to estimate the density of an asset return in finance. In this section, we estimate the density of the continuously compounded return of the S&P 500 daily index. We downloaded the S&P 500 daily closing prices, p_t, during the period from the 2nd January 2009 to the 2nd September 2014 from http://finance.yahoo.com. The date t return is calculated as $y_t = \log(p_t/p_{t-1})$, and there are $T = 1425$ observations of the return. The time series plot of the return series is presented in Figure 4.

6.1 Density estimation of S&P 500 daily return under conditional heteroscedasticity

Let $\boldsymbol{y} = (y_1, y_2, \cdots, y_T)'$ be a vector of T observations of S&P 500 daily returns. We consider a semiparametric GARCH (1,1) model given by

$$\begin{aligned} y_t &= \sigma_t \varepsilon_t, \\ \sigma_t^2 &= b_0 + b_1 y_{t-1}^2 + b_2 \sigma_{t-1}^2, \end{aligned} \tag{21}$$

where ε_t, for $t = 1, 2, \cdots, T$, are independent and follow an unknown distribution with its density denoted as $f(\varepsilon)$. Zhang and King (2013) proposed approximating $f(\varepsilon)$ by a Gaussian kernel density given by

$$\widehat{f}_\varepsilon(\varepsilon_t) = \frac{1}{T}\sum_{i=1}^{T} \frac{1}{h}\phi\Big(\frac{\varepsilon_t - \varepsilon_i}{h}\Big). \tag{22}$$

This density has the form of kernel density estimator of errors. They suggested using a global bandwidth, as well as a localized version with $h(1 + h_\varepsilon|\varepsilon_i|)$ being assigned to ε_i as its bandwidth, for $i = 1, 2, \cdots, T$.

In this section, we use our SLB approach to bandwidth estimation, which is now based on residuals rather than a sample of observations in Section 5. We also assume that the prior of h^2 is the inverse Gamma density with its hyperparameter vector, $\theta = (\alpha, \beta)'$, being estimated through our likelihood–based approach described in Appendix A. The estimation procedure is described as follows.

Step 1: Estimate the GARCH model given by (21) using the quasi maximum likelihood method under the normality assumption of ε_t; and calculate residuals.

Step 2: Obtain an initial estimate of θ, denoted as $\theta_0 = (\alpha_0, \beta_0)'$, by maximizing the marginal likelihood given by (15) with x and X_i being replaced by ε and ε_i.

Step 3: Simulate a random sample, denoted as $\{\varepsilon_i : i = 1, 2, \cdots, 5T\}$, from the marginal likelihood given by (15) with α and β being their initial estimates. Obtain the MLE of θ, denoted as $\widehat{\theta}$, by maximizing the likelihood function given by (17) with X_i^* being replaced by ε_i^*.

Step 4: Derive the estimate of localized bandwidth, denoted as $h_n(\varepsilon_t|\widehat{\theta})$, according to (5) with x

being replaced by ε. Thus, the Gaussian kernel error density is approximated as

$$\widehat{f}_{\varepsilon}(\varepsilon_t) = \frac{1}{T}\sum_{i=1}^{T}\frac{1}{h_n(\varepsilon_t|\widehat{\theta})}\phi\left(\frac{\varepsilon_t-\varepsilon_i}{h_n(\varepsilon_t|\widehat{\theta})}\right). \tag{23}$$

Step 5: With the derived estimate of bandwidth, we express the density of y_t as

$$\widehat{f}_Y(y_t|\lambda) = \frac{1}{T\sigma_t}\sum_{i=1}^{T}\frac{1}{h_n\left(y_t/\sigma_t|\widehat{\theta}\right)}\phi\left(\frac{y_t/\sigma_t-y_i/\sigma_i}{h_n\left(y_t/\sigma_t|\widehat{\theta}\right)}\right), \tag{24}$$

for $t = 1,2,\cdots,T$, where $\lambda = (b_1,b_2)'$.[‡] Therefore, the likelihood of $\boldsymbol{y} = (y_1,y_2,\cdots,y_T)'$ given λ is approximately $\ell\left(\boldsymbol{y}|\lambda\right) = \prod_{t=1}^{T}\widehat{f}_Y(y_t|\lambda)$.

Step 6: Derive a semiparametric estimate of $\lambda = (b_1,b_2)'$ by maximizing $\ell\left(\boldsymbol{y}|\lambda\right)$.

After completing these steps, we derived the SLB estimate and a semiparametric estimate of λ, which are denoted as $\widehat{h}_n(y_t/\widehat{\sigma}_t|\widehat{\theta})$ and $\widehat{\lambda} = \left(\widehat{b}_1,\widehat{b}_2\right)'$.

It is of great interest to estimate the one–day–ahead density of the S&P 500 daily return, y_{T+1}. Under the Gaussian kernel GARCH model, the density of y_{T+1} is estimated as

$$\widehat{f}_Y(y_{T+1}|\widehat{\lambda}) = \frac{1}{T\widehat{\sigma}_{T+1}}\sum_{t=1}^{T}\frac{1}{\widehat{h}_n\left(y_{T+1}/\widehat{\sigma}_{T+1}\right)}\phi\left(\frac{y_{T+1}/\widehat{\sigma}_{T+1}-y_t/\widehat{\sigma}_t}{\widehat{h}_n\left(y_{T+1}/\widehat{\sigma}_{T+1}|\widehat{\theta}\right)}\right), \tag{25}$$

which was at 1000 grid points, where $\widehat{\sigma}_t^2 = \widehat{b}_0 + \widehat{b}_1 y_{t-1}^2 + \widehat{b}_2\widehat{\sigma}_{t-1}^2$, for $t = 1,2,\cdots,T+1$.

For comparison purpose, we considered a global bandwidth for the kernel density estimator of y_{T+1}, where the bandwidth is estimated through Bayesian sampling of Zhang and King (2013) and CV, respectively. In the context of Bayesian sampling, the global bandwidth is treated as a parameter. Therefore, the vector of parameters is $\lambda_{\text{Bayes}} = (h,\sigma_0^2,b_1,b_2)'$. The sampling algorithm of Zhang and King (2013) was carried out to derive the estimate of λ_{Bayes} denoted as $\widetilde{\lambda}_{\text{Bayes}} = \left(\widetilde{h},\widetilde{\sigma}_0^2,\widetilde{b}_1,\widetilde{b}_2\right)'$. We then calculated the conditional variance $\widetilde{\sigma}_t^2$ as

$$\widetilde{\sigma}_t^2 = \widetilde{b}_0 + \widetilde{b}_1 y_{t-1}^2 + \widetilde{b}_2\widetilde{\sigma}_{t-1}^2,$$

for $t = 1,2,\cdots,T+1$, where $\widetilde{b}_0 = 1-\widetilde{b}_1-\widetilde{b}_2$. The density of y_{T+1} is then calculated according to (25) at 1000 grid points, where h and σ_t are replaced by respectively, $\widetilde{h}$ and $\widetilde{\sigma}_t$.

[‡]When $\{y_t : t = 1,2,\cdots,n\}$ is pre–standardized, Zhang and King (2013) suggested choosing the value of b_0 as $(1-b_1-b_2)$ due to identification reasons.

We also used the likelihood cross–validation method to choose a global bandwidth for the Gaussian kernel density estimator of y_{T+1}. First, we estimated b_0, b_1 and b_2 of (21) through the quasi maximum likelihood method under the normality assumption of ε_t, and the resulting estimates are denoted as $\widetilde{b}_0^{(\mathrm{cv})}$, $\widetilde{b}_1^{(\mathrm{cv})}$ and $\widetilde{b}_2^{(\mathrm{cv})}$. We then computed the residuals as $\widetilde{\varepsilon}_t = y_t/\widetilde{\sigma}_{t,\mathrm{cv}}$, where

$$\widetilde{\sigma}^2_{t,\mathrm{cv}} = \widetilde{b}_0^{(\mathrm{cv})} + \widetilde{b}_1^{(\mathrm{cv})} y^2_{t-1} + \widetilde{b}_2^{(\mathrm{cv})} \widetilde{\sigma}^2_{t-1,\mathrm{cv}}, \tag{26}$$

for $t = 1, 2, \cdots, T$. Second, we chose a global bandwidth denoted as $\widetilde{h}_{cv}$, for $\{\widetilde{\varepsilon}_t : t = 1, 2, \cdots, T\}$ using the likelihood CV.

Third, with the selected bandwidth for the Gaussian kernel density estimator, we derived the likelihood function, which is constructed through (22) and expressed as

$$\widetilde{f}_Y(y_t|\lambda) = \frac{1}{T\sigma_t} \sum_{i=1}^{T} \frac{1}{\widetilde{h}_{\mathrm{cv}}} \phi\left(\frac{y_t/\sigma_t - y_i/\sigma_i}{\widetilde{h}_{\mathrm{cv}}}\right), \tag{27}$$

It was maximized with respect to b_1 and b_2, where $b_0 = 1 - b_1 - b_2$. Thus, a semiparametric estimate of $(b_1, b_2)'$ was derived and denoted as $\left(\widetilde{b}_1^{(\mathrm{cv})}, \widetilde{b}_2^{(\mathrm{cv})}\right)'$. This estimation procedure so far is similar to the semiparametric estimation of ARCH models proposed by Engle and González-Rivera (1991).

As we are interested in not only the bandwidth of the Gaussian kernel error density, but also the parameter estimates, we used the derived parameter estimates to calculate residuals again. Then, the likelihood CV method is applied to the updated residuals to derive a bandwidth, which is also denoted as $\widetilde{h}_{\mathrm{cv}}$. With the updated bandwidth being substituted into the likelihood function given by (27), we maximized the likelihood function and obtained an updated estimate of $(b_1, b_2)'$, denoted as $\left(\widetilde{b}_1^{(\mathrm{cv})}, \widetilde{b}_2^{(\mathrm{cv})}\right)$, which is a semiparametric estimate of the parameter vector.

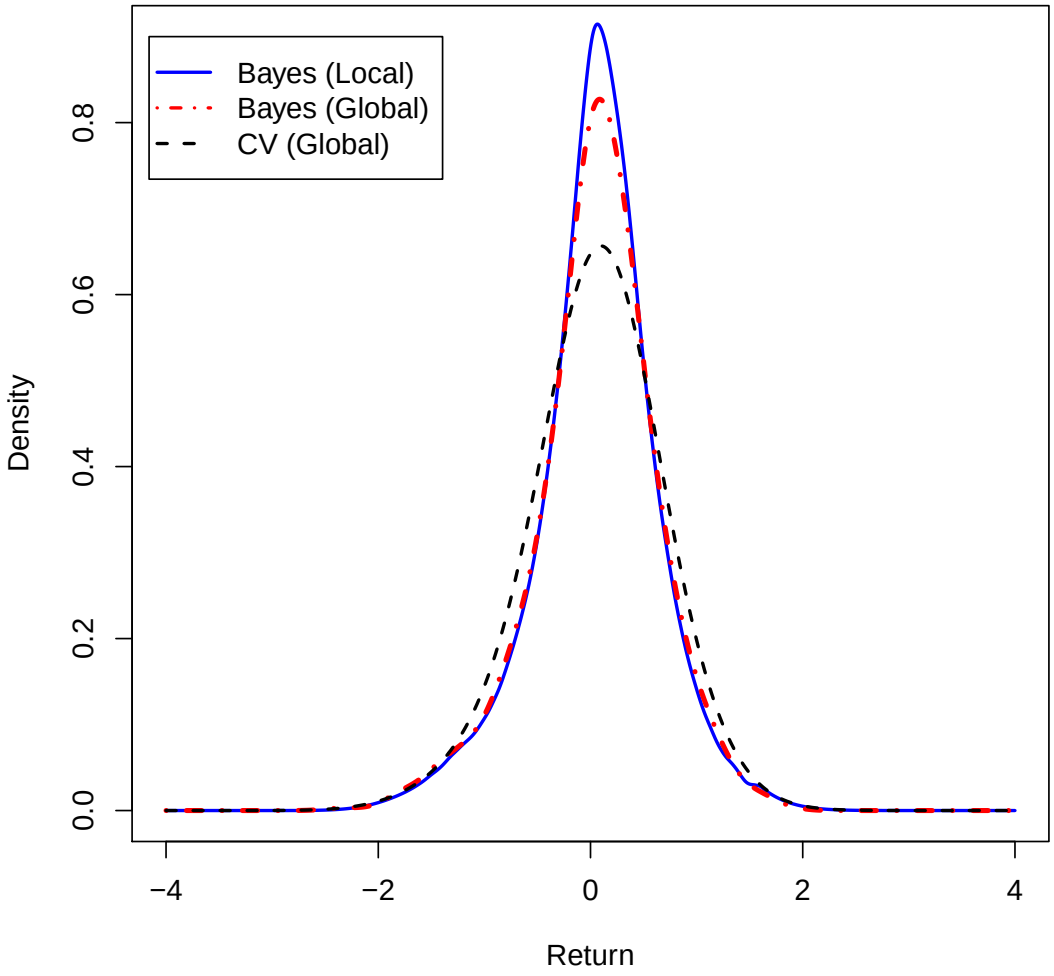
After completing these steps, we calculated the kernel density estimate of y_{T+1},

$$\widetilde{f}_Y(y_{T+1}|\lambda) = \frac{1}{T\widetilde{\sigma}_{T+1,\mathrm{cv}}} \sum_{t=1}^{T} \frac{1}{\widetilde{h}_{\mathrm{cv}}} \phi\left(\frac{y_{T+1}/\widetilde{\sigma}_{T+1,\mathrm{cv}} - y_t/\widetilde{\sigma}_{t,\mathrm{cv}}}{\widetilde{h}_{\mathrm{cv}}}\right), \tag{28}$$

at 1000 grid points, where $\widetilde{\sigma}_{T+1,\mathrm{cv}}$ is calculated through (26).

The resulting three kernel estimates of the density of the one–day–ahead S&P 500 daily return are plotted in Figure 5, where the density estimate with our SLB estimate clearly differs from its competitor whose bandwidth was estimated via Bayesian sampling, in their left–tail and peak areas. However, both of them are almost the same in their right–tail areas. Moreover, the density estimate with our SLB method is clearly different from its competitor with bandwidth chosen via

Figure 5: *Density estimates of the one–day–ahead out–of–sample S&P 500 daily return.*



CV in tail and peak areas.

6.2 Forecasting results

We conducted the same rolling–sample procedure as we did in Section 5.2 to evaluate the out–of–sample performance of each density estimate resulted from each of these three bandwidth estimation methods. The number of all observations is $T = 1425$, and the size of a rolling sample is $n = 1005$, where the first rolling sample is from 2nd January 2013 to 2nd September 2014.

At the rth iteration of the rolling sample procedure, we denote the SLB estimate as $h_{\text{SLB}}(y)$, the global bandwidth chosen through CV as h_{CV}, and the global bandwidth estimated through Bayesian sampling of Zhang, King, and Hyndman (2006) as h_{Bayes}. The corresponding density estimates are $\widehat{f}\left(y_{n+r}|h_{\text{SLB}}(y_{n+r})\right)$, $\widetilde{f}\left(y_{n+r}|h_{\text{CV}}\right)$ and $\widetilde{f}\left(y_{n+r}|h_{\text{Bayes}}\right)$.

We calculated the average logarithmic scores over the out–of–sample period:

$$\frac{1}{T-n}\sum_{r=1}^{T-n}\log\widehat{f}\left(y_{n+r}|h_{\mathrm{SLB}}(y_{n+r})\right)=-0.99,$$
$$\frac{1}{T-n}\sum_{r=1}^{T-n}\log\widetilde{f}\left(y_{n+r}|h_{\mathrm{Bayes}}\right)=-1.03,$$
$$\frac{1}{T-n}\sum_{r=1}^{T-n}\log\widetilde{f}\left(y_{n+r}|h_{\mathrm{CV}}\right)=-1.09.$$

By comparing their average scores, we find that the forecast with the use of our proposed SLB estimate leads the highest score. This means that the use of localized bandwidths outperforms the use of a global bandwidth selected through either Bayesian sampling or likelihood CV. Thus, we can conclude that our proposed SLB estimator performs better than its competitors, which are Bayesian sampling and likelihood CV for estimating a global bandwidth.

7 Conclusions

In this paper, we have investigated the asymptotic properties of a semiparametric localized bandwidth (SLB) estimator for kernel density estimation for stationary time series data. We have proved that the SLB estimator is asymptotically normally distributed with root–n rate of convergence. To carry out the computation of the SLB estimator for a given sample of data, we have proposed a sampling–based likelihood approach to hyperparameter estimation. Monte Carlo simulation studies have shown that the proposed hyperparameter estimation approach works very well, and that the proposed SLB estimator outperforms its competitors.

When estimating the density of Eurodollar deposit rate through the kernel method, we have found that our proposed SLB method leads to a better performance of the resulting density estimator than a global bandwidth either estimated through Bayesian sampling of Zhang, King, and Hyndman (2006) or selected through likelihood CV. In the kernel estimator of the density of S&P 500 daily return under conditional heteroscedasticity, our proposed SLB method leads to a clearly better performance than the global bandwidth estimated through Bayesian sampling and likelihood CV. These results show that our proposed bandwidth estimator has better out–of–sample performance than its competitors.

Acknowledgements

The authors acknowledge constructive comments from the seminar participants at Monash University, University of Bergen, University of York in England and University of Science and Technology of China in Hefei, in particular to Anastasios Panagiotelis, Dag Tjøstheim, Yaohua Wu, Wenyang Zhang and Lincheng Zhao. This research was supported under the Australian Research Council's *Discovery Projects* Scheme under Grant Numbers: DP1095838 and DP130104229.

References

Aït-Sahalia, Y., 1996. Testing continuous–time models of the spot interest rate. Review of Financial studies 9 (2), 385–426.

Amemiya, T., 1985. Advanced Econometrics. Harvard University Press, Cambridge.

Amisano, G., Giacomini, R., 2007. Comparing density forecasts via weighted likelihood ratio tests. Journal of Business & Economic Statistics 25 (2), 177–190.

Atchadé, Y. F., 2011. A computational framework for empirical Bayes inference. Statistics and Computing 21 (4), 463–473.

Bithell, J., 1990. An application of density estimation to geographical epidemiology. Statistics in Medicine 9 (6), 691–701.

Brewer, M. J., 2000. A Bayesian model for local smoothing in kernel density estimation. Statistics and Computing 10 (4), 299–309.

Casella, G., 2001. Empirical Bayes Gibbs sampling. Biostatistics 2 (4), 485–500.

Chib, S., Greenberg, E., 1995. Understanding the Metropolis–Hastings algorithm. The American Statistician 49 (4), 327–335.

de Lima, M. S., Atuncar, G. S., 2011. A Bayesian method to estimate the optimal bandwidth for multivariate kernel estimator. Journal of Nonparametric Statistics 23 (1), 137–148.

Elgammal, A., Duraiswami, R., Harwood, D., Davis, L. S., 2002. Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. Proceedings of the IEEE 90 (7), 1151–1163.

Engle, R. F., González-Rivera, G., 1991. Semiparametric ARCH models. Journal of Business and Economic Statistics 9 (4), 345–359.

Fan, J., Yao, Q., 2003. Nonlinear Time Series: Non– and Parametric Methods. Springer, New York.

Gangopadhyay, A., Cheung, K., 2002. Bayesian approach to the choice of smoothing parameter in kernel density estimation. Journal of Nonparametric Statistics 14 (6), 655–664.

Gao, J., 2007. Nonlinear Time Series: Semi– and Non–Parametric Methods. Chapman & Hall/CRC, London.

Garthwaite, P. H., Fan, Y., Sisson, S. A., 2010. Adaptive optimal scaling of Metropolis-Hastings algorithms using the Robbins-Monro process. Working paper, University of New South Wales.
URL http://arxiv.org/pdf/1006.3690v1.pdf

Ghosh, J. K., Ramamoorthi, R. V., 2003. Bayesian Nonparametrics. Springer, New York.

Hamilton, J., 1994. Time Series Analysis. Princeton University Press, New Jersey.

Härdle, W., Hall, P., Marron, J. S., 1988. How far are automatically chosen regression smoothing parameters from their optimum? Journal of the American Statistical Association 83 (401), 86–95.

Heidenreich, N.-B., Schindler, A., Sperlich, S., 2013. Bandwidth selection for kernel density estimation: a review of fully automatic selectors. AStA Advances in Statistical Analysis 97 (4), 403–433.

Hjort, N. L., Holmes, C., Müller, P., Walker, S. G., 2010. Bayesian Nonparametrics. Cambridge University Press, Cambridge.

Jones, M. C., Marron, J. S., Sheather, S. J., 1996. A brief survey of bandwidth selection for density estimation. Journal of the American Statistical Association 91 (433), 401–407.

Kulasekera, K. B., Padgett, W. J., 2006. Bayes bandwidth selection in kernel density estimation with censored data. Nonparametric Statistics 18 (2), 129–143.

Lo, A. Y., 1984. On a class of Bayesian nonparametric estimates: I. Density estimates. Annals of Statistics 12 (1), 351–357.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., Teller, E., 1953. Equation of state calculations by fast computing machines. Journal of Chemical Physics 21 (6), 1087–1092.

Sain, S., Scott, D., 1996. On locally adaptive density estimation. Journal of the American Statistical Association 91 (436), 1525–1534.

Seaman, D. E., Powell, R. A., 1996. An evaluation of the accuracy of kernel density estimators for home range analysis. Ecology 77 (7), 2075–2085.

Sheather, S. J., 2004. Density estimation. Statistical Science 19 (4), 588–597.

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

Zhang, X., King, M. L., 2013. Gaussian kernel GARCH models. Working paper, Monash University.
URL http://www.buseco.monash.edu.au/ebs/pubs/wpapers/2013/19-13.php

Zhang, X., King, M. L., Hyndman, R. J., 2006. A Bayesian approach to bandwidth selection for multivariate kernel density estimation. Computational Statistics and Data Analysis 50 (11), 3009–3031.

Appendix A

This appendix provides an example and some discussion about how the existence of a super–consistent estimator for a vector of hyperparameters imposed in Assumption 4(i) may be achievable. While the hyperparameter estimation method proposed in Casella (2001) is not applicable, the idea of using the maximum likelihood estimation (MLE) method is being extended to deal with our case.

Let us have a look at whether we may use $\{X_i : i = 1,2,\cdots,n\}$ to construct $\widehat{\theta}$. Observe that

$$\begin{aligned} p(x|\theta) &= \int f(x|h)\pi(h|\theta)dh = \iint f(y)\frac{1}{h}K\left(\frac{y-x}{h}\right)\pi(h|\theta)dydh \\ &= \iint f(x+sh)K(s)\pi(h|\theta)dhds = \iint \left(f(x)+shf^{(1)}(x)+\frac{s^2h^2\,f^{(2)}(\xi)}{2}\right)K(s)\pi(h|\theta)dhds \\ &= f(x)+\int s^2K(s)ds\cdot\int\frac{f^{(2)}(\xi)h^2}{2}\pi(h|\theta)dh, \end{aligned} \tag{29}$$

where $f^{(1)}(\cdot)$ and $f^{(2)}(\cdot)$ denote the first–order and second–order derivatives of $f(x)$, respectively, and ξ lies between x and $x+h$.

It can be seen from (29) that θ is not directly involved in $f(x)$. In other words, the information available from $\{X_i : i = 1,2,\cdots,n\}$ may not be sufficient to construct a consistent estimate for θ. Therefore, in order to consistently estimate θ, we propose to generate a random sample, $\left\{X_j^* : j = 1,2,\cdots,n^*\right\}$, from $p_n(x|\theta_0)$ with θ_0 being estimated through the original sample. As to the choice of n^*, our derivation given in equation (32) below suggests that $n/n^* = o(1)$. This means that n^* should be large enough compared with n. In the finite sample implementation, we thus choose $n^*/n = 5$ given the computational intensity. This simulated sample is then used to construct the likelihood function and derive the MLE of θ.

To simulate a random sample from $p_n(x|\theta_0)$, we use the random–walk Metropolis algorithm (see Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller, 1953; Chib and Greenberg, 1995, among others). To check the mixing performance of a simulated chain, we calculate the simulation inefficiency factor (SIF), which can be approximately interpreted as the number of draws required in order to obtain independent draws. The resulting SIF value is around 5 for each simulated chain, and this indicates a very good mixing performance. For each simulated chain, we retain one draw for every 20 draws. Therefore, the resulting simulated sample, denoted as $\left\{X_j^* : j = 1,2,\cdots,n^*\right\}$, is a random sample. The detailed sampling procedure about how to generate a sequence of stationary Markov chains is described as follows.

Step 1: Generate a random sample $\{X_i : i = 1,2,\cdots,n\}$ from $f(x)$.

Step 2: For a given initial value of θ denoted as θ_0, we choose an arbitrary initial value of X_0^*.

Step 3: At the jth iteration, the current state X_j^* is updated as $X_j^* = X_{j-1}^* + \tau u$, where u is drawn from a proposal density which is the standard Gaussian density in this paper. τ is a tuning constant which is chosen such that the acceptance rate is targeted at 44% (see for example, Garthwaite, Fan, and Sisson, 2010). The updated X_j^* is accepted with a probability given by

$$\min\left(\frac{p_n\left(X_j^*|\theta_0\right)}{p_n\left(X_{j-1}^*|\theta_0\right)},1\right), \tag{30}$$

where $p_n\left(X_j^*|\theta_0\right) = \int_0^\infty \widehat{f}\left(X_j^*|h\right)\pi(h|\theta_0)dh$.

Step 4: Repeat Step 3 and discard the burn–in period of iterations, after which we retain one draw for every 20 draws. The resulting sample, still denoted as $\left\{X_j^* : j = 1,2,\cdots,n^*\right\}$, can be considered as a sequence of stationary Markov chains.

The MLE of θ is obtained as

$$\widehat{\theta} = \arg\max_{\theta\in\Theta}\prod_{i=1}^{n^*}\int \widehat{f}\left(X_i^*|h\right)\pi(h|\theta)dh. \tag{31}$$

Under some regularity conditions (similarly to those of Theorems 4.1.1–4.1.3 of Amemiya (1985); Section 14.4 of Hamilton (1994)) on $p_n(x|\theta)$, one may show that $\sqrt{n^*}\left(\widehat{\theta}-\theta_0\right) = O_P(1)$ as $n^* \to \infty$. Therefore, Assumption 4(i) follows from

$$\sqrt{n}\left(\widehat{\theta}-\theta_0\right) = \frac{\sqrt{n}}{\sqrt{n^*}}\cdot\sqrt{n^*}\left(\widehat{\theta}-\theta_0\right) = o_P(1) \tag{32}$$

when n^* is chosen such that $\frac{n}{n^*} = o(1)$. Note that the convergence in probability is under the probability distribution of $\{X_t^*\}$ given the original probability distribution of $\{X_i\}$. This shows that Assumption 4(i) may be verifiable.

Appendix B

We now provide the proofs of Lemmas 1 and 2 and Theorems 1–5. Throughout this appendix, we use $p(x)$, $q(x)$, $p_n(x)$, $q_n(x)$, $\widehat{p}_n(x)$, $\widehat{q}_n(x)$, $h_0(x)$, $h_n(x)$ and $\widehat{h}_n(x)$ to denote $p(x|\theta)$, $q(x|\theta)$, $p_n(x|\theta)$, $q_n(x|\theta)$, $\widehat{p}_n(x|\widehat{\theta})$, $\widehat{q}_n(x|\widehat{\theta})$, $h_0(x|\theta)$, $h_n(x|\theta)$ and $\widehat{h}_n(x|\widehat{\theta})$, respectively.

Proof of Lemma 1.

We have

$$
\begin{aligned}
p(x) &= \int f(x|h)\pi(h|\theta)dh = \iint f(y)K_h(y-x)\pi(h|\theta)dhdy \\
&= \iint f(x+uh)K(u)\pi(h|\theta)dhdu = \iint f(x+uv)K(u)\pi(v|\theta)dvdu = \mathbb{E}\left[f(x+uv)\right], \\
q(x) &= \int hf(x|h)\pi(h|\theta)dh = \iint hf(x+uh)K(u)\pi(h|\theta)dhdu \\
&= \iint vf(x+uv)K(u)\pi(v|\theta)dvdu = \mathbb{E}\left[f(x+uv)v\right].
\end{aligned}
$$

As $\{u_i, v_i; i = 1,2,\cdots,m\}$ are independent and identically distributed (iid), $f(x+u_iv_i)$ is also iid. Therefore, by the law of large numbers, as $m \to \infty$, $\frac{1}{m}\sum_{i=1}^{m} f(x+u_iv_i) - \mathbb{E}\left[f(x+uv)\right] = o_P(1)$. Hence $p_m(x) - p(x) = o_P(1)$. Similarly, $f(x+u_iv_i)v_i$ is also iid. Therefore, by the law of large numbers, as $m \to \infty$, $\frac{1}{m}\sum_{i=1}^{m} f(x+u_iv_i)v_i - \mathbb{E}\left[f(x+uv)v\right] = o_P(1)$. Hence $q_m(x) - q(x) = o_P(1)$.

Proof of Lemma 2.

We have

$$
\begin{aligned}
p_n(x) &= \int \widehat{f}(x|h)\pi(h|\theta)dh = \int \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)\pi(h|\theta)dh \\
&= \frac{1}{n}\sum_{i=1}^{n}\int \frac{1}{v}K\left(\frac{X_i - x}{v}\right)\pi(v|\theta)dv = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_2\left[\frac{1}{v}K\left(\frac{X_i - x}{v}\right)\right] \\
q_n(x) &= \int h\widehat{f}(x|h)\pi(h|\theta)dh = \int h\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)\pi(h|\theta)dh \\
&= \frac{1}{n}\sum_{i=1}^{n}\int K\left(\frac{X_i - x}{v}\right)\pi(v|\theta)dv = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_2\left[K\left(\frac{X_i - x}{v}\right)\right],
\end{aligned}
$$

where $\mathbb{E}_2(\cdot)$ denotes the conditional expectation of h given X_1.

As $\{v_j; j = 1,2,\cdots,m\}$ is identically independent distributed (iid), $\frac{1}{v_j}K\left(\frac{X_i - x}{v_j}\right)$ is also iid. Therefore, by the law of large numbers, as $m \to \infty$, $\frac{1}{m}\sum_{j=1}^{m}\frac{1}{v_j}K\left(\frac{X_i - x}{v_j}\right) - \mathbb{E}_2\left[\frac{1}{v}K\left(\frac{X_i - x}{v}\right)\right] = o_P(1)$.

Consequently, $\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{1}{v_j}K\left(\frac{X_i - x}{v_j}\right) - \mathbb{E}\left[\frac{1}{v}K\left(\frac{X_1 - x}{v}\right)\right] = o_P(1)$. Hence, $p_{nm}(x) - p_n(x) = o_P(1)$. Similarly, $K\left(\frac{X_i - x}{v_j}\right)$ is also iid. Therefore, by the law of large numbers, as $m \to \infty$, $\frac{1}{m}\sum_{j=1}^{m} K\left(\frac{X_i - x}{v_j}\right) - \mathbb{E}\left[K\left(\frac{X_1 - x}{v}\right)\right] = o_P(1)$. Consequently, $\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} K\left(\frac{X_i - x}{v_j}\right) - \mathbb{E}\left[K\left(\frac{X_1 - x}{v}\right)\right] = o_P(1)$. Therefore, $q_{nm}(x) - q_n(x) = o_P(1)$.

Proof of Theorem 1.

According to (2) and (4), we have

$$\begin{aligned}
h_n(x) - h_0(x) &= \frac{q_n(x)}{p_n(x)} - \frac{q(x)}{p(x)} = \frac{1}{p_n(x)p(x)}\left[q_n(x)p(x) - q(x)p_n(x)\right] \\
&= \frac{1}{p_n(x)p(x)}\left[q_n(x)p(x) - q(x)p(x) + q(x)p(x) - q(x)p_n(x)\right] \\
&= \frac{1}{p_n(x)p(x)}\left[p(x)\left(q_n(x) - q(x)\right) - q(x)\left(p_n(x) - p(x)\right)\right] \\
&= \frac{1}{p_n(x)p(x)} L_n(x),
\end{aligned}$$

where $L_n(x) = p(x)\left(q_n(x) - q(x)\right) - q(x)\left(p_n(x) - p(x)\right)$. Note that $\mathbb{E}_1(\widehat{f}(x|h)) = f(x|h)$ and that

$$\begin{aligned}
\widehat{f}(x|h) - f(x|h) &= \widehat{f}(x|h) - \mathbb{E}_1[\widehat{f}(x|h)] = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) - \frac{1}{nh}\sum_{i=1}^{n}\mathbb{E}_1\left[K\left(\frac{X_i - x}{h}\right)\right] \\
&= \frac{1}{nh}\sum_{i=1}^{n}\left(K\left(\frac{X_i - x}{h}\right) - \mathbb{E}_1\left[K\left(\frac{X_i - x}{h}\right)\right]\right) = \frac{1}{n}\sum_{i=1}^{n} U_i(x;h),
\end{aligned}$$

where $U_i(x;h) = \frac{1}{h}\left[K\left(\frac{X_i - x}{h}\right) - \mathbb{E}_1 K\left(\frac{X_i - x}{h}\right)\right]$. Therefore, we have

$$\begin{aligned}
L_n(x) &= p(x)\left(q_n(x) - q(x)\right) - q(x)\left(p_n(x) - p(x)\right) \\
&= \int hp(x)\pi(h)\left[\widehat{f}(x|h) - f(x|h)\right]dh - \int q(x)\pi(h)\left[\widehat{f}(x|h) - f(x|h)\right]dh \\
&= \int \left[hp(x) - q(x)\right]\pi(h)\left[\widehat{f}(x|h) - f(x|h)\right]dh = \int \left[hp(x) - q(x)\right]\pi(h)\left[\widehat{f}(x|h) - \mathbb{E}(\widehat{f}(x|h))\right]dh \\
&= \int \left[hp(x) - q(x)\right]\pi(h)\left[\frac{1}{n}\sum_{i=1}^{n} U_i(x;h)\right]dh = \frac{1}{n}\sum_{i=1}^{n}\int \left[hp(x) - q(x)\right]U_i(x;h)\pi(h)dh \\
&= \frac{1}{n}\sum_{i=1}^{n} V_i(x),
\end{aligned}$$

where $V_i(x) = \int \left[hp(x) - q(x)\right]U_i(x;h)\pi(h)dh$.

It is easy to check that

$$\mathbb{E}_1[V_i(x)] = \mathbb{E}_1\left[\int \left[hp(x) - q(x)\right]U_i(x;h)\pi(h)dh\right] = \int \left[hp(x) - q(x)\right]\mathbb{E}_1\left[U_i(x;h)\right]\pi(h)dh = 0.$$

Note that

$$\begin{aligned}
V_i^2(x) &= \left\{\int \left[up(x) - q(x)\right]U_i(x;u)\pi(u)du\right\}\left\{\int \left[vp(x) - q(x)\right]U_i(x;v)\pi(v)dv\right\} \\
&= \iint \left[up(x) - q(x)\right]\left[vp(x) - q(x)\right]U_i(x;u)U_i(x;v)\pi(u)\pi(v)dudv.
\end{aligned}$$

Therefore, the variance of $V_i(x)$ is given by

$$\mathrm{Var}\left[V_i(x)\right] = \mathbb{E}\left[V_i^2(x)\right] = \iint \left[up(x) - q(x)\right]\left[vp(x) - q(x)\right]\mathbb{E}_1\left[U_i(x;u)U_i(x;v)\right]\pi(u)\pi(v)\,du\,dv,$$

where

$$\begin{aligned}
&\mathbb{E}_1\left[U_i(x;u)U_i(x;v)\right] = \frac{1}{uv}\mathbb{E}_1\left\{\left[K\left(\frac{X_i - x}{u}\right) - \mathbb{E}_1 K\left(\frac{X_i - x}{u}\right)\right]\left[K\left(\frac{X_i - x}{v}\right) - \mathbb{E}_1 K\left(\frac{X_i - x}{v}\right)\right]\right\} \\
&= \frac{1}{uv}\left(\mathbb{E}_1\left[K\left(\frac{X_i - x}{u}\right)K\left(\frac{X_i - x}{v}\right)\right] - \mathbb{E}_1\left[K\left(\frac{X_i - x}{u}\right)\right]\mathbb{E}_1\left[K\left(\frac{X_i - x}{v}\right)\right]\right) \\
&= \frac{1}{uv}\left[uvf_{uv}(x) - uvf_u(x)f_v(x)\right] = f_{uv}(x) - f_u(x)f_v(x) = R_{uv}(x).
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{E}\left[V_i^2(x)\right] &= \iint \left[up(x) - q(x)\right]\left[vp(x) - q(x)\right]\mathbb{E}_1\left[U_i(x;u)U_i(x;v)\right]\pi(u)\pi(v)\,du\,dv \\
&= \iint [up(x) - q(x)][vp(x) - q(x)]R_{uv}(x)\pi(u)\pi(v)\,du\,dv.
\end{aligned}$$

We denote $\mathrm{Var}\left[V_i(x)\right]$ as $\gamma(0) = \iint [up(x) - q(x)][vp(x) - q(x)]R_{uv}(x)\pi(u)\pi(v)\,du\,dv$.

The covariance of $V_i(x)$ and $V_j(x)$ is given by

$$\begin{aligned}
&\mathrm{Cov}\left[V_i(x), V_j(x)\right] = \mathbb{E}\left[V_i(x)V_j(x)\right] \\
&= \mathbb{E}\left\{\int \left[up(x) - q(x)\right]U_i(x;u)\pi(u)\,du\right\}\left\{\int \left[vp(x) - q(x)\right]U_j(x;v)\pi(v)\,dv\right\} \\
&= \mathbb{E}\left\{\iint \left[up(x) - q(x)\right]\left[vp(x) - q(x)\right]U_i(x;u)U_j(x;v)\pi(u)\pi(v)\,du\,dv\right\} \\
&= \iint \left[up(x) - q(x)\right]\left[vp(x) - q(x)\right]\left\{\mathbb{E}_1\left[U_i(x;u)U_j(x;v)\right]\right\}\pi(u)\pi(v)\,du\,dv,
\end{aligned}$$

where

$$\begin{aligned}
&\mathbb{E}_1\left[U_i(x;u)U_j(x;v)\right] = \frac{1}{uv}\mathbb{E}_1\left\{\left[K\left(\frac{X_i - x}{u}\right) - \mathbb{E}_1 K\left(\frac{X_i - x}{u}\right)\right]\left[K\left(\frac{X_j - x}{v}\right) - \mathbb{E}_1 K\left(\frac{X_j - x}{v}\right)\right]\right\} \\
&= \frac{1}{uv}\left(\mathbb{E}_1\left[K\left(\frac{X_i - x}{u}\right)K\left(\frac{X_j - x}{v}\right)\right] - \mathbb{E}_1\left[K\left(\frac{X_i - x}{u}\right)\right]\mathbb{E}_1\left[K\left(\frac{X_j - x}{v}\right)\right]\right).
\end{aligned}$$

By Assumption 1, $\{X_i\}$ is strictly stationary, we have $\mathbb{E}_1 K\left(\frac{X_j - x}{v}\right) = \mathbb{E}_1 K\left(\frac{X_i - x}{v}\right) = vf_v(x)$. Therefore $\mathbb{E}_1\left[K\left(\frac{X_i - x}{u}\right)\right]\mathbb{E}_1\left[K\left(\frac{X_j - x}{v}\right)\right] = uvf_u(x)f_v(x)$.

$$\mathbb{E}_1\left[K\left(\frac{X_i - x}{u}\right)K\left(\frac{X_j - x}{v}\right)\right] = \int K\left(\frac{y - x}{u}\right)K\left(\frac{z - x}{v}\right)f_s(y,z)\,dy\,dz = uvg_{uv,s}(x),$$

where $s = |i-j|$ and $f_{|i-j|}(y,z)$ denotes the joint density of (X_i, X_j).

$$\mathbb{E}_1\left[U_i(x;u)U_j(x;v)\right] = \frac{1}{uv}\left(\mathbb{E}_1\left[K\left(\frac{X_i-x}{u}\right)K\left(\frac{X_j-x}{v}\right)\right] - \mathbb{E}_1\left[K\left(\frac{X_i-x}{u}\right)\right]\mathbb{E}_1\left[K\left(\frac{X_j-x}{v}\right)\right]\right)$$
$$= \frac{1}{uv}\left[uvg_{uv,s}(x) - uvf_u(x)f_v(x)\right] = g_{uv,s}(x) - f_u(x)f_v(x) = G_{uv,s}(x).$$

Therefore,

$$\mathrm{Cov}\left[V_i(x), V_j(x)\right] = \iint \left[up(x)-q(x)\right]\left[vp(x)-q(x)\right]\left\{\mathbb{E}_1\left[U_i(x;u)U_j(x;v)\right]\right\}\pi(u)\pi(v)dudv$$
$$= \iint \left[up(x)-q(x)\right]\left[vp(x)-q(x)\right]G_{uv,s}(x)\pi(u)\pi(v)dudv.$$

Denote $\mathrm{Cov}\left[V_i(x), V_{i+j}(x)\right]$ as $\gamma(j) = \iint \left[up(x)-q(x)\right]\left[vp(x)-q(x)\right]G_{uv,j}(x)\pi(u)\pi(v)dudv$.

Before we establish the central limit theorem, we need to verify the following condition: $\mathbb{E}|V_i(x)|^\delta < \infty$ for some constant $\delta > 2$ that satisfies $\sum_{j\geq 1}\alpha(j)^{1-2/\delta} < \infty$.

As $V_i(x)$ is a measurable function of X_i, the process $V_i(x)$ possesses the mixing property of $\{X_i\}$. This indicates that $\{V_i\}$ is a sequence with α–mixing coefficient satisfying $\sum_{j\geq 1}\alpha(j)^{1-2/\delta} < \infty$. Denote $g_1(h) = \left[hp(x)-q(x)\right]\pi(h)$ and $g_2(h) = U_i(x;h)$. Under Hölder's inequality that for $\delta > 1$, $1/\delta + 1/q = 1$, we have

$$|V_i(x)|^\delta = \left|\int \left[hp(x)-q(x)\right]U_i(x;h)\pi(h)dh\right|^\delta = \left|\int g_1(h)g_2(h)dh\right|^\delta = \left|\int g_1^{1-\frac{1}{\delta}}(h)\cdot g_1^{\frac{1}{\delta}}(h)g_2(h)dh\right|^\delta$$
$$\leq \left(\int g_1(h)dh\right)^{\frac{\delta}{q}}\int g_1(h)|g_2(h)|^\delta dh = \left(\int g_1(h)dh\right)^{\delta-1}\int g_1(h)|g_2(h)|^\delta dh.$$

By Assumption 3, we have for some $0 < C_\delta < \infty$

$$\mathbb{E}\left[|V_i(x)|^\delta\right] = \left(\int g_1(h)dh\right)^{\delta-1}\int g_1(h)\cdot\mathbb{E}_1\left[|g_2(h)|^\delta\right]dh$$
$$\leq C_\delta\left(\int g_1(h)dh\right)^{\delta-1}\cdot\int g_1(h)h^{1-\delta}\left(\int K^\delta(v)f(x+vh)dv\right)dh < \infty.$$

Lemma A.1 of Gao (2007) implies that

$$|\gamma(j)| = |\mathrm{Cov}(V_t(x), V_s(x))| \leq 10\alpha(j)^{1-2/\delta}\left\{\mathbb{E}|V_t(x)|^\delta\right\}^{\frac{1}{\delta}}\left\{\mathbb{E}|V_s(x)|^\delta\right\}^{\frac{1}{\delta}},$$

where $|t-s| = j$. By the condition $\sum_{j\geq 1}\alpha(j)^{1-2/\delta} < \infty$, we have

$$\sum_{j=1}^{\infty}|\gamma(j)| = 2\sum_{1\leq t<s\leq n}|\mathrm{Cov}(V_t(x), V_s(x))| \leq \sum_{j=1}^{\infty}10\alpha(j)^{1-2/\delta}\left\{\mathbb{E}|V_t(x)|^\delta\right\}^{\frac{1}{\delta}}\left\{\mathbb{E}|V_s(x)|^\delta\right\}^{\frac{1}{\delta}} < \infty.$$

Assumptions 1–3 ensure that condition (i) is satisfied, so by Theorem 2.20 of Fan and Yao (2003), as

$n \to \infty$, we have

$$\begin{aligned}
&\mathrm{Var}\left[\sqrt{n}L_n(x)\right] = \mathrm{Var}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_i(x)\right] = \frac{1}{n}\mathrm{Var}\left[\sum_{i=1}^{n} V_i(x)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}\left[V_i(x)\right] + \frac{2}{n}\sum_{1\le i<j\le n}\mathrm{Cov}\left[V_i(x), V_j(x)\right] \\
&= \gamma(0) + 2\sum_{l=1}^{n-1}\left(1-\frac{l}{n}\right)\gamma(l) \to \gamma(0) + 2\sum_{j=1}^{\infty}\gamma(j) = \Sigma_L(x).
\end{aligned}$$

Theorem 2.21 of Fan and Yao (2003) implies that

$$\sqrt{n}L_n(x) \to_D \mathcal{N}(0, \Sigma_L(x)), \tag{33}$$

where $\Sigma_L(x) = \gamma(0) + 2\sum_{j=1}^{\infty}\gamma(j)$. We have

$$h_n(x) - h_0(x) = \frac{1}{p_n(x)p(x)}L_n(x).$$

By Assumption 1 and Proposition 2.8 of Fan and Yao (2003), as $n \to \infty$, we have

$$\begin{aligned}
&p_n(x) = \int \widehat{f}(x|h)\pi(h)dh = \int \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)\pi(h)dh \\
&= \frac{1}{n}\sum_{i=1}^{n}\int \frac{1}{h}K\left(\frac{X_i - x}{h}\right)\pi(h)dh \to_P \int \frac{1}{h}\mathbb{E}_1\left[K\left(\frac{X_1 - x}{h}\right)\right]\pi(h)dh \\
&= \iint \frac{1}{h}K\left(\frac{y-x}{h}\right)\pi(h)f(y)dhdy = p(x).
\end{aligned}$$

Denote $p^2(x) = Q(x)$. As $n \to \infty$, we have

$$p_n(x)p(x) \to_P Q(x), \qquad \text{and} \qquad \frac{1}{p_n(x)p(x)} \to_P Q^{-1}(x).$$

Therefore, under Assumptions 1–3, as $n \to \infty$,

$$\sqrt{n}(h_n(x) - h_0(x)) \to_D \mathcal{N}(0, \Sigma_0(x)), \tag{34}$$

where $\Sigma_0(x) = Q^{-2}(x)\Sigma_L(x)$. Therefore, we have proved Theorem 1.

Proof of Theorem 2.

Without loss of generality, it suffices to prove the theorem for $N = 2$.

$$\sqrt{n}(h_n(x_1) - h_0(x_1)) = \frac{1}{p_n(x_1)p(x_1)}\sqrt{n}L_n(x_1) = \frac{1}{p_n(x_1)p(x_1)}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_i(x_1),$$
$$\sqrt{n}(h_n(x_2) - h_0(x_2)) = \frac{1}{p_n(x_2)p(x_2)}\sqrt{n}L_n(x_2) = \frac{1}{p_n(x_2)p(x_2)}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_i(x_2).$$

Note that

$$\begin{aligned}&\mathrm{Cov}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_i(x_1), \frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_i(x_2)\right) = \mathbb{E}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_i(x_1)\frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_i(x_2)\right)\\ &= \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n} V_i(x_1)\sum_{i=1}^{n} V_i(x_2)\right) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left(V_i(x_1)V_j(x_2)\right),\end{aligned}$$

where

$$\begin{aligned}&\mathbb{E}\left[V_i(x_1)V_j(x_2)\right] = \mathbb{E}\left\{\int\left[up(x_1) - q(x_1)\right]U_i(x_1;u)\pi(u)du\right\}\left\{\int\left[vp(x_2) - q(x_2)\right]U_j(x_2;v)\pi(v)dv\right\}\\ &= \mathbb{E}\left\{\iint\left[up(x_1) - q(x_1)\right]\left[vp(x_2) - q(x_2)\right]U_i(x_1;u)U_j(x_2;v)\pi(u)\pi(v)dudv\right\}\\ &= \iint\left[up(x_1) - q(x_1)\right]\left[vp(x_2) - q(x_2)\right]\left\{\mathbb{E}_1\left[U_i(x_1;u)U_j(x_2;v)\right]\right\}\pi(u)\pi(v)dudv.\\ &\mathbb{E}_1\left[U_i(x_1;u)U_j(x_2;v)\right] = \frac{1}{uv}\mathbb{E}_1\left\{\left[K\left(\frac{X_i - x_1}{u}\right) - \mathbb{E}_1 K\left(\frac{X_i - x_1}{u}\right)\right]\left[K\left(\frac{X_j - x_2}{v}\right) - \mathbb{E}_1 K\left(\frac{X_j - x_2}{v}\right)\right]\right\}\\ &= \frac{1}{uv}\left(\mathbb{E}_1\left[K\left(\frac{X_i - x_1}{u}\right)K\left(\frac{X_j - x_2}{v}\right)\right] - \mathbb{E}_1\left[K\left(\frac{X_i - x_1}{u}\right)\right]\mathbb{E}_1\left[K\left(\frac{X_j - x_2}{v}\right)\right]\right).\end{aligned}$$

As $\mathbb{E}_1\left[K\left(\frac{X_i - x_1}{u}\right)\right]\mathbb{E}_1\left[K\left(\frac{X_j - x_2}{v}\right)\right] = uvf_u(x_1)f_v(x_2)$, we have

$$\mathbb{E}_1\left[K\left(\frac{X_i - x_1}{u}\right)K\left(\frac{X_j - x_2}{v}\right)\right] = \int K\left(\frac{y - x_1}{u}\right)K\left(\frac{z - x_2}{v}\right)f_s(y,z)dydz = uvm_{uv,s}(x_1,x_2),$$

where $s = |i - j|$ and $f_{|i-j|}(y,z)$ denotes the joint density of (X_i, X_j).

$$\begin{aligned}&\mathbb{E}_1\left[U_i(x_1;u)U_j(x_2;v)\right] = \frac{1}{uv}\left(\mathbb{E}_1\left[K\left(\frac{X_i - x_1}{u}\right)K\left(\frac{X_j - x_2}{v}\right)\right] - \mathbb{E}_1\left[K\left(\frac{X_i - x_1}{u}\right)\right]\mathbb{E}_1\left[K\left(\frac{X_j - x_2}{v}\right)\right]\right)\\ &= \frac{1}{uv}\left[uvm_{uv,s}(x_1,x_2) - uvf_u(x_1)f_v(x_2)\right] = m_{uv,s}(x_1,x_2) - f_u(x_1)f_v(x_2) = S_{uv,s}(x_1,x_2).\end{aligned}$$

Consequently, we have

$$\mathbb{E}\left[V_i(x_1)V_j(x_2)\right] = \iint\left[up(x_1) - q(x_1)\right]\left[vp(x_2) - q(x_2)\right]S_{uv,s}(x_1,x_2)\pi(u)\pi(v)dudv.$$

Let $\gamma_2(0) = \mathbb{E}[V_i(x_1)V_i(x_2)] = \iint \left[up(x_1) - q(x_1)\right]\left[vp(x_2) - q(x_2)\right] S_{uv,0}(x_1, x_2)\pi(u)\pi(v)dudv$ and $\gamma_2(s) = \mathbb{E}\left[V_i(x_1)V_j(x_2)\right] = \iint \left[up(x_1) - q(x_1)\right]\left[vp(x_2) - q(x_2)\right] S_{uv,s}(x_1, x_2)\pi(u)\pi(v)dudv$. Thus, as $n \to \infty$, we have

$$\begin{aligned}
&\mathrm{Cov}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_i(x_1), \frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_i(x_2)\right) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left(V_i(x_1)V_j(x_2)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\iint \left[up(x_1) - q(x_1)\right]\left[vp(x_2) - q(x_2)\right] S_{uv,s}(x_1, x_2)\pi(u)\pi(v)dudv \\
&= \gamma_2(0) + \frac{2}{n}\sum_{1\le i<j\le n}\gamma_2(|i-j|) = \gamma_2(0) + 2\sum_{s=1}^{n-1}\gamma_2(s)\left(1 - \frac{s}{n}\right) \\
&\to \gamma_2(0) + 2\sum_{s=1}^{\infty}\gamma_2(s) = \Sigma_v(x_1, x_2).
\end{aligned}$$

Define $S = C_1\frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_i(x_1) + C_2\frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_i(x_2)$, where C_1 and C_2 are constants. We have

$$S = C_1\frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_i(x_1) + C_2\frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_i(x_2) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{C_1V_i(x_1) + C_2V_i(x_2)\} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i.$$

where $Y_i = C_1V_i(x_1) + C_2V_i(x_2)$. It is easy to show that $\mathbb{E}(Y_i) = 0$.

$$\begin{aligned}
&\mathrm{Var}(Y_i) = \mathbb{E}(Y_i^2) = \mathbb{E}\left[C_1^2V_i^2(x_1) + C_2^2V_i^2(x_2) + 2C_1C_2V_i(x_1)V_i(x_2)\right] \\
&= C_1^2\mathbb{E}\left[V_i^2(x_1)\right] + C_2^2\mathbb{E}\left[V_i^2(x_2)\right] + 2C_1C_2\mathbb{E}[V_i(x_1)V_i(x_2)] \\
&= a_1 + a_2 + a_3 = \gamma_y(0),
\end{aligned}$$

where

$$\begin{aligned}
&a_1 = C_1^2\mathbb{E}\left[V_i^2(x_1)\right] = C_1^2\iint [up(x_1) - q(x_1)][vp(x_1) - q(x_1)]R_{uv}(x_1)\pi(u)\pi(v)dudv, \\
&a_2 = C_2^2\mathbb{E}\left[V_i^2(x_2)\right] = C_2^2\iint [up(x_2) - q(x_2)][vp(x_2) - q(x_2)]R_{uv}(x_2)\pi(u)\pi(v)dudv, \\
&a_3 = 2C_1C_2\mathbb{E}[V_i(x_1)V_i(x_2)] = 2C_1C_2\iint \left[up(x_1) - q(x_1)\right]\left[vp(x_2) - q(x_2)\right] S_{uv,0}(x_1, x_2)\pi(u)\pi(v)dudv, \\
&\mathrm{Cov}(Y_i, Y_j) = \mathbb{E}\left\{[C_1V_i(x_1) + C_2V_i(x_2)]\left[C_1V_j(x_1) + C_2V_j(x_2)\right]\right\} \\
&= C_1^2\mathbb{E}\left\{V_i(x_1)V_j(x_1)\right\} + C_1C_2\mathbb{E}\left\{V_i(x_2)V_j(x_1)\right\} + C_1C_2\mathbb{E}\left\{V_i(x_1)V_j(x_2)\right\} + C_2^2\mathbb{E}\left\{V_i(x_2)V_j(x_2)\right\} \\
&= b_1 + 2b_2 + b_3 = \gamma_y(s),
\end{aligned}$$

where $s = |i-j|$ and

$$b_1 = C_1^2 \mathbb{E}\{V_i(x_1)V_j(x_1)\} = C_1^2 \iint [up(x_1) - q(x_1)][vp(x_1) - q(x_1)] G_{uv,s}(x_1)\pi(u)\pi(v)dudv,$$
$$b_2 = C_1 C_2 \mathbb{E}\{V_i(x_2)V_j(x_1)\} = C_1 C_2 \iint [up(x_1) - q(x_1)][vp(x_2) - q(x_2)] S_{uv,s}(x_1,x_2)\pi(u)\pi(v)dudv,$$
$$b_3 = C_2^2 \mathbb{E}\{V_i(x_2)V_j(x_2)\} = C_2^2 \iint [up(x_2) - q(x_2)][vp(x_2) - q(x_2)] G_{uv,s}(x_2)\pi(u)\pi(v)dudv.$$

In univariate case, we have verified that $\mathbb{E}|V_i(x)|^\delta < \infty$ and $V_i(x)$ sequence is α–mixing with the mixing coefficient satisfying $\sum_{j\geq 1}\alpha(j)^{1-2/\delta} < \infty$ for some constant $\delta > 2$.

$$\mathbb{E}|Y_i|^\delta = \mathbb{E}|C_1 V_i(x_1) + C_2 V_i(x_2)|^\delta \leq |C_1|^\delta \mathbb{E}|V_i(x_1)|^\delta + |C_2|^\delta \mathbb{E}|V_i(x_2)|^\delta < \infty.$$

In addition, Y_i is a measurable function of $V_i(x_1)$ and $V_i(x_2)$, therefore $\{Y_i\}$ is α–mixing with the mixing coefficient satisfying $\sum_{j\geq 1}\alpha(j)^{1-2/\delta} < \infty$ for some constant $\delta > 2$. By Theorem 2.20 of Fan and Yao (2003), we can obtain that $\mathrm{Var}(S) \to \gamma_y(0) + 2\sum_{s=1}^{\infty}\gamma_y(s) = \Sigma_S$.
Therefore, by Theorem 2.21 of Fan and Yao (2003), we obtain that
$S = C_1 \frac{1}{\sqrt{n}}\sum_{i=1}^n V_i(x_1) + C_2 \frac{1}{\sqrt{n}}\sum_{i=1}^n V_i(x_2) \to_D \mathcal{N}(0,\Sigma_S)$. Therefore, as $n \to \infty$, we have

$$\left[\frac{1}{\sqrt{n}}\sum_{i=1}^n V_i(x_1), \frac{1}{\sqrt{n}}\sum_{i=1}^n V_i(x_2)\right] \to_D \mathcal{N}(0, \Sigma(x_1,x_2)), \tag{35}$$

where

$$\Sigma(x_1,x_2) = \begin{pmatrix} \Sigma_L(x_1) & \Sigma_v(x_1,x_2) \\ \Sigma_v(x_1,x_2) & \Sigma_L(x_2) \end{pmatrix}.$$

As $n \to \infty$, we have,

$$\frac{1}{p_n(x_1)p(x_1)} \to_P Q^{-1}(x_1), \quad \text{and} \quad \frac{1}{p_n(x_2)p(x_2)} \to_P Q^{-1}(x_2),$$

Let

$$Q_{n,12}^{-1} = \mathrm{diag}\left(\frac{1}{p_n(x_1)p(x_1)}, \frac{1}{p_n(x_2)p(x_2)}\right), \quad \text{and} \quad Q_{12}^{-1} = \mathrm{diag}\left(Q^{-1}(x_1), Q^{-1}(x_2)\right).$$

Therefore, as $n \to \infty$,

$$\left[\sqrt{n}(h_n(x_1) - h_0(x_1)), \sqrt{n}(h_n(x_2) - h_0(x_2))\right] = Q_{n,12}^{-1}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^n V_i(x_1), \frac{1}{\sqrt{n}}\sum_{i=1}^n V_i(x_2)\right]$$
$$\to_D Q_{12}^{-1}\mathcal{N}(0,\Sigma(x_1,x_2)) = \mathcal{N}(0,\Sigma_{12}), \tag{36}$$

where $\Sigma_{12} = Q_{12}^{-1}\Sigma(x_1, x_2)Q_{12}^{-1}$.

More generally,

$$\left[\sqrt{n}(h_n(x_1) - h_0(x_1)), \cdots, \sqrt{n}(h_n(x_N) - h_0(x_N))\right] \to_D \mathcal{N}(0, \Sigma_N), \tag{37}$$

where $\Sigma_{N,aa} = Q^{-1}(x_a)\Sigma_L(x_a)Q^{-1}(x_a) = \Sigma_0(x_a)$, $\Sigma_{N,ab} = Q^{-1}(x_a)\Sigma_v(x_a, x_b)Q^{-1}(x_b)$, $\Sigma_v(x_a, x_b) = \gamma_{ab}(0) + 2\sum_{s=1}^{\infty}\gamma_{ab}(s)$,
$\gamma_{ab}(s) = \mathbb{E}\left[V_i(x_a)V_j(x_b)\right] = \iint \left[up(x_a) - q(x_a)\right]\left[vp(x_b) - q(x_b)\right] S_{uv,s}(x_a, x_b)\pi(u)\pi(v)dudv$ and $\gamma_{ab}(0) = \mathbb{E}[V_i(x_a)V_i(x_b)] = \iint \left[up(x_a) - q(x_a)\right]\left[vp(x_b) - q(x_b)\right] S_{uv,0}(x_a, x_b)\pi(u)\pi(v)dudv$. Thus, we have proved Theorem 2.

Proof of Theorem 3.

According to (4) and (5), we have

$$\begin{aligned}
\widehat{h}_n(x) - h_n(x) &= \frac{\widehat{q}_n(x)}{\widehat{p}_n(x)} - \frac{q_n(x)}{p_n(x)} = \frac{\widehat{q}_n(x)p_n(x) - \widehat{p}_n(x)q_n(x)}{\widehat{p}_n(x)p_n(x)} \\
&= \frac{\left(\widehat{q}_n(x) - q_n(x)\right)p_n(x) - \left(\widehat{p}_n(x) - p_n(x)\right)q_n(x)}{\widehat{p}_n(x)p_n(x)} \\
&= \frac{\left(\widehat{q}_n(x) - q_n(x)\right)p_n(x) - \left(\widehat{p}_n(x) - p_n(x)\right)q_n(x)}{\left(\widehat{p}_n(x) - p_n(x)\right)p_n(x) + p_n^2(x)}.
\end{aligned}$$

Note that

$$\begin{aligned}
\widehat{q}_n(x) - q_n(x) &= \int h\widehat{f}(x|h)\pi(h;\widehat{\theta})dh - \int h\widehat{f}(x|h)\pi(h;\theta_0)dh \\
&= \int h\widehat{f}(x|h)\left(\pi(h;\widehat{\theta}) - \pi(h;\theta_0)\right)dh.
\end{aligned}$$

Thus, by Assumption 3, we have

$$\begin{aligned}
|\widehat{q}_n(x) - q_n(x)| &\le ||\widehat{\theta} - \theta_0|| \int h\widehat{f}(x|h)L(h;\theta_0)dh, \\
&= ||\widehat{\theta} - \theta_0|| \left[\int h\left(\widehat{f}(x|h) - f(x|h)\right)L(h;\theta_0)dh + \int hf(x|h)L(h;\theta_0)dh\right].
\end{aligned}$$

As X_i is strictly stationary, we have $f(x|h) = \frac{1}{h}\mathbb{E}_1\left[K\left(\frac{X_1 - x}{h}\right)\right] = \frac{1}{nh}\sum_{i=1}^{n}\mathbb{E}_1\left[K\left(\frac{X_i - x}{h}\right)\right]$. Therefore,

$$\begin{aligned}
\widehat{f}(x|h) - f(x|h) &= \frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{X_i - x}{h}\right) - \frac{1}{nh}\sum_{i=1}^{n}\mathbb{E}_1\left[K\left(\frac{X_i - x}{h}\right)\right] \\
&= \frac{1}{nh}\sum_{i=1}^{n}\left(K\left(\frac{X_i - x}{h}\right) - \mathbb{E}_1\left[K\left(\frac{X_i - x}{h}\right)\right]\right) = \frac{1}{n}\sum_{i=1}^{n}U_i(x;h),
\end{aligned}$$

where $U_i(x;h) = \frac{1}{h}\left(K\left(\frac{X_i-x}{h}\right) - \mathbb{E}_1\left[K\left(\frac{X_i-x}{h}\right)\right]\right)$. It is easy to show that $\mathbb{E}_1[U_i(x;h)] = 0$. Therefore, we have

$$\begin{aligned}
&\mathbb{E}\left[\widehat{f}(x|h) - f(x|h)\right]^2 = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} U_i(x;h)\right]^2 \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}U_i(x;h)^2 + \frac{2}{n^2}\sum_{1\le i<j\le n}\mathbb{E}\left[U_i(x;u)U_j(x;v)\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{var}(U_i(x;h)) + \frac{2}{n^2}\sum_{1\le i<j\le n}\mathrm{Cov}\left(U_i(x;u)U_j(x;v)\right) \\
&= \frac{1}{n}R_{uv}(x) + \frac{2}{n}\sum_{s=1}^{n-1}\left(1-\frac{s}{n}\right)\mathrm{Cov}\left(U_i(x;u)U_j(x;v)\right) = \frac{1}{n}R_{uv}(x) + \frac{2}{n}\sum_{s=1}^{n-1}\left(1-\frac{s}{n}\right)\gamma_u(s),
\end{aligned}$$

where $\gamma_u(s) = \mathrm{Cov}\left(U_i(x;u), U_j(x;v)\right)$.

We have shown that $\mathbb{E}|U_i(x)|^\delta < \infty$. Thus, Lemma A.1 of Gao (2007) implies

$$|\gamma_u(j)| = |\mathrm{Cov}(U_t(x), U_s(x))| \le 10\alpha(j)^{1-2/\delta}\left\{\mathbb{E}|U_t(x)|^\delta\right\}^{\frac{1}{\delta}}\left\{\mathbb{E}|U_s(x)|^\delta\right\}^{\frac{1}{\delta}},$$

where $|t-s| = j$. Because $U_i(x)$ is a measurable function of X_i, the process $U_i(x)$ possesses the mixing property of $\{X_i\}$, which indicates that $\{U_i(x)\}$ sequence with α–mixing coefficient satisfying $\sum_{j\ge1}\alpha(j)^{1-2/\delta} < \infty$. Thus, we have

$$\sum_{j=1}^{\infty}|\gamma_u(j)| = 2\sum_{1\le t<s\le n}|\mathrm{Cov}(U_t(x), U_s(x))| \le \sum_{j=1}^{\infty}10\alpha(j)^{1-2/\delta}\left\{\mathbb{E}|U_t(x)|^\delta\right\}^{\frac{1}{\delta}}\left\{\mathbb{E}|U_s(x)|^\delta\right\}^{\frac{1}{\delta}} < \infty.$$

Thus, $\left|\sum_{s=1}^{n-1}\left(1-\frac{s}{n}\right)\gamma_u(s)\right| \le \sum_{j=1}^{\infty}|\gamma_u(j)| < \infty$.

Therefore, as $n\to\infty$, we have

$$\mathbb{E}\left[\widehat{f}(x|h) - f(x|h)\right]^2 = \frac{1}{n}R_{uv}(x) + \frac{2}{n}\sum_{s=1}^{n-1}\left(1-\frac{s}{n}\right)\gamma_u(s) \to 0.$$

Hence, as $n\to\infty$, we have $\widehat{f}(x|h) - f(x|h) = o_P(1)$. This implies

$$\begin{aligned}
&|\widehat{q}_n(x) - q_n(x)| \le ||\widehat{\theta}-\theta_0||\left[\int h\left|\widehat{f}(x|h) - f(x|h)\right| L(h;\theta_0)dh + \int h f(x|h)L(h;\theta_0)dh\right] \\
&= ||\widehat{\theta}-\theta_0||O_P(1).
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
&|\widehat{p}_n(x) - p_n(x)| \le ||\widehat{\theta}-\theta_0||\int \widehat{f}(x|h)L(h;\theta_0)dh, \\
&\le ||\widehat{\theta}-\theta_0||\left[\int\left|\widehat{f}(x|h) - f(x|h)\right| L(h;\theta_0)dh + \int f(x|h)L(h;\theta_0)dh\right] = ||\widehat{\theta}-\theta_0||O_P(1).
\end{aligned}$$

By Assumption 4(i), we have $||\widehat{\theta}-\theta_0|| = o_P(n^{-1/2})$. Thus, we obtain

$$\widehat{h}_n(x) - h_n(x) = \frac{\left(\widehat{q}_n(x) - q_n(x)\right) p_n(x) - \left(\widehat{p}_n(x) - p_n(x)\right) q_n(x)}{\widehat{p}_n(x)\, p_n(x)} = o_P(n^{-1/2}),$$

which implies as $n \to \infty$, $\sqrt{n}\left(\widehat{h}_n(x) - h_n(x)\right) = o_P(1)$.

Note that

$$\sqrt{n}(\widehat{h}_n(x) - h_0(x)) = \sqrt{n}(\widehat{h}_n(x) - h_n(x) + h_n(x) - h_0(x)) = \sqrt{n}(\widehat{h}_n(x) - h_n(x)) + \sqrt{n}(h_n(x) - h_0(x)).$$

According to Theorem 1, $\sqrt{n}(\widehat{h}_n(x) - h_n(x)) = o_P(1)$ and $\sqrt{n}(h_n(x) - h_0(x)) \to_D \mathcal{N}(0, \Sigma_0(x))$, and thus, we have as $n \to \infty$

$$\sqrt{n}(\widehat{h}_n(x) - h_0(x)) = \sqrt{n}(\widehat{h}_n(x) - h_n(x)) + \sqrt{n}(h_n(x) - h_0(x)) \to_D \mathcal{N}(0, \Sigma_0(x)),$$

which completes the proof of Theorem 3.

Proof of Theorem 4.

The proof of Theorem 4 is similar to that of Theorem 2.

Proof of Theorem 5.

By the definitions of $\widehat{f}_n^*(x)$ and $\widehat{f}_n(x)$, in view of the function $L(\cdot)$ involved in Assumption 4 and the conditions of Theorem 5, we have

$$\mathbb{E}\left[L\left(\frac{X_i - x}{h_0(x)}\right)\right] = h_0(x) f(x) \int L(u)\,du + h_0^2(x) f^{(1)}(x) \int uL(u)\,du + \frac{(1+o(1)) h_0^3(x) f^{(2)}(x)}{2} \cdot \int u^2 L(u)\,du \quad (38)$$

when $h_0(x) = a_n b_0(x) \to 0$ as $n \to \infty$.

We therefore have

$$\begin{aligned}
\left|\widehat{f}_n^*(x) - \widehat{f}_n(x)\right| &= \left|\frac{1}{n}\sum_{i=1}^n \frac{1}{\widehat{h}_n(x)} K\left(\frac{X_i - x}{\widehat{h}_n(x)}\right) - \frac{1}{n}\sum_{i=1}^n \frac{1}{h_0(x)} K\left(\frac{X_i - x}{h_0(x)}\right)\right| \\
&\le \frac{1}{n}\sum_{i=1}^n L\left(\frac{X_i - x}{h_0(x)}\right)\left|\frac{1}{\widehat{h}_n(x)} - \frac{1}{h_0(x)}\right| = \frac{1}{n}\sum_{i=1}^n L\left(\frac{X_i - x}{h_0(x)}\right) \frac{\left|\widehat{h}_n(x) - h_0(x)\right|}{\widehat{h}_n(x) h_0(x)} = O_P(n^{-1/2}).
\end{aligned}$$

It is easy to see that $\widehat{f}_n^*(x) - f(x) = \widehat{f}_n^*(x) - \widehat{f}_n(x) + \widehat{f}_n(x) - f(x)$. Then,

$$\begin{aligned}
\sqrt{nh_0(x)}\left(\widehat{f}_n^*(x) - f(x)\right) &= \sqrt{nh_0(x)}\left(\widehat{f}_n^*(x) - \widehat{f}_n(x)\right) + \sqrt{nh_0(x)}\left(\widehat{f}_n(x) - f(x)\right) \\
&= I_{1n}(x) + I_{2n}(x),
\end{aligned}$$

where $I_{1n}(x) = \sqrt{nh_0(x)}\left(\widehat{f}_n^*(x) - \widehat{f}_n(x)\right)$ and $I_{2n}(x) = \sqrt{nh_0(x)}\left(\widehat{f}_n(x) - f(x)\right)$, in which

$$\begin{aligned} I_{1n}(x) &= \sqrt{nh_0(x)}\left(\widehat{f}_n^*(x) - \widehat{f}_n(x)\right) \leq O_P(n^{-1/2})n^{1/2}\sqrt{h_0(x)} \\ &= O_P(1)\sqrt{h_0(x)} = O_P(1)a_n^{1/2}b_0^{1/2}(x) = o_P(1). \end{aligned}$$

The bias of $\widehat{f}_n(x)$ is given by

$$\begin{aligned} \text{Bias}\{\widehat{f}_n(x)\} &= \mathbb{E}\left\{\frac{1}{nh_0(x)}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h_0(x)}\right)\right\} - f(x) = h_0(x)^{-1}\mathbb{E}\left[K\left(\frac{X_1 - x}{h_0(x)}\right)\right] - f(x) \\ &= h_0(x)^{-1}\int K\left(\frac{y-x}{h_0(x)}\right)f(y)dy - f(x) = h_0(x)^{-1}\int f(x + h_0(x)v)K(v)h_0(x)dv - f(x) \\ &= \int\left\{f(x) + f^{(1)}(x)h_0(x)v + \frac{1}{2}f^{(2)}(x)h_0^2(x)v^2 + O(h_0^3(x))\right\}K(v)dv - f(x) \\ &= \left\{f(x) + 0 + \frac{h_0^2(x)}{2}f^{(2)}(x)\int v^2K(v)dv + O(h_0^3(x))\right\} - f(x) \\ &= \frac{h_0^2(x)}{2}\mu_2(K)f^{(2)}(x) + o(h_0^2(x)). \end{aligned} \tag{39}$$

The variance of $\widehat{f}_n(x)$ is given by

$$\begin{aligned} \text{Var}\{\widehat{f}_n(x)\} &= \text{Var}\left\{\frac{1}{nh_0(x)}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h_0(x)}\right)\right\} = \frac{1}{n^2h_0^2(x)}\sum_{i=1}^{n}\text{Var}\left\{K\left(\frac{X_i - x}{h_0(x)}\right)\right\} \\ &= \frac{1}{nh_0^2(x)}\text{Var}\left\{K\left(\frac{X_1 - x}{h_0(x)}\right)\right\} = \frac{1}{nh_0^2(x)}\left\{\mathbb{E}\left[K^2\left(\frac{X_1 - x}{h_0(x)}\right)\right] - \left[\mathbb{E}K\left(\frac{X_1 - x}{h_0(x)}\right)\right]^2\right\} \\ &= \frac{1}{nh_0^2(x)}\left\{\int K^2\left(\frac{y-x}{h_0(x)}\right)f(y)dy - \left[\int K\left(\frac{y-x}{h_0(x)}\right)f(y)dx_1\right]^2\right\} \\ &= \frac{1}{nh_0^2(x)}\left\{h_0(x)\int f(x + h_0(x)v)K^2(v)dv - \left[h_0(x)\int f(x + h_0(x)v)K(v)dv\right]^2\right\} \\ &= \frac{1}{nh_0^2(x)}\left\{h_0(x)\int [f(x) + f^{(1)}(\xi)h_0(x)v]K^2(v)dv - O(h_0^2(x))\right\} \\ &= \frac{1}{nh_0(x)}\left\{f(x)\int K^2(v)dv + O(h_0(x)\int |v|K^2(v)dv) - O(h_0(x))\right\} \\ &= \frac{1}{nh_0(x)}R(K)f(x) + o\left(\frac{1}{nh_0(x)}\right), \end{aligned} \tag{40}$$

where ξ lies between x and $x + h(x)v$.

Based on (39) and (40), we have that

$$\mathbb{E}\left[\widehat{f}_n(x)\right] - f(x) = ch_0^2(x) + o(h_0^2(x)).$$
$$\mathbb{E}\left[\widehat{f}_n(x) - f(x)\right]^2 = \frac{R(K)f(x)}{nh_0(x)} + \frac{1}{4}\left(\mu_2(K)f^{(2)}(x)\right)^2 h_0^4(x) + O(h_0^4(x) + (nh_0(x))^{-1})$$
$$= \frac{R(K)f(x)}{na_n b_0(x)} + \frac{1}{4}\left(\mu_2(K)f^{(2)}(x)\right)^2 a_n^4 b_0^4(x) + O(h_0^4(x) + (nh_0(x))^{-1}).$$

Then, we can further obtain that

$$\sqrt{nh_0(x)}\left(\widehat{f}_n(x) - f(x) - \frac{h_0^2(x)}{2}\mu_2(K)f^{(2)}(x)\right)$$
$$= \sqrt{nh_0(x)}\left(\widehat{f}_n(x) - \mathbb{E}\left[\widehat{f}_n(x)\right]\right) + \sqrt{nh_0(x)}\left(\mathbb{E}\left[\widehat{f}_n(x)\right] - f(x) - \frac{h_0^2(x)}{2}\mu_2(K)f^{(2)}(x)\right)$$
$$= \sqrt{nh_0(x)}\left(\widehat{f}_n(x) - \mathbb{E}\left[\widehat{f}_n(x)\right]\right) + o_P\left(h_0^2(x)\sqrt{nh_0(x)}\right)$$
$$= \sqrt{nh_0(x)}\left(\widehat{f}_n(x) - \mathbb{E}\left[\widehat{f}_n(x)\right]\right) + o_P(1)$$
$$= \sqrt{nh_0(x)}\frac{1}{nh_0(x)}\sum_{i=1}^{n}\left(K\left(\frac{X_i - x}{h_0(x)}\right) - \mathbb{E}\left[K\left(\frac{X_i - x}{h_0(x)}\right)\right]\right) + o_P(1)$$
$$= \sum_{i=1}^{n}\frac{1}{\sqrt{nh_0(x)}}\left(K\left(\frac{X_i - x}{h_0(x)}\right) - \mathbb{E}\left[K\left(\frac{X_i - x}{h_0(x)}\right)\right]\right) + o_P(1) = \sum_{i=1}^{n} Z_{n,i} + o_P(1),$$

where $Z_{n,i} = \frac{1}{\sqrt{nh_0(x)}}\left(K\left(\frac{X_i - x}{h_0(x)}\right) - \mathbb{E}\left[K\left(\frac{X_i - x}{h_0(x)}\right)\right]\right)$.

From (40), it is easy to show that $\mathrm{Var}\left(\sum_{i=1}^{n} Z_{n,i}\right) = R(K)f(x)$. By Liapunov's CLT, we obtain $\sum_{i=1}^{n} Z_{n,i} \to_D \mathcal{N}(0, R(K)f(x))$. Therefore, we have $\sqrt{nh_0(x)}\left(\widehat{f}_n(x) - f(x) - \frac{h_0^2(x)}{2}\mu_2(K)f^{(2)}(x)\right) \to_D \mathcal{N}(0, R(K)f(x))$, which implies

$$\sqrt{nh_0(x)}\left(\widehat{f}_n^*(x) - f(x) - \frac{h_0^2(x)}{2}\mu_2(K)f^{(2)}(x)\right)$$
$$= \sqrt{nh_0(x)}\left(\widehat{f}_n^*(x) - \widehat{f}_n(x)\right) + \sqrt{nh_0(x)}\left(\widehat{f}_n(x) - f(x) - \frac{h_0^2(x)}{2}\mu_2(K)f^{(2)}(x)\right)$$
$$= \sqrt{nh_0(x)}\left(\widehat{f}_n(x) - f(x) - \frac{h_0^2(x)}{2}\mu_2(K)f^{(2)}(x)\right) + o_P(1).$$
$$\to_D \mathcal{N}(0, R(K)f(x)),$$

We have finally completed the proof of Theorem 5.