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 in Kernel Density Estimation

Tingting Cheng, Jiti Gao and Xibin Zhang

May 2014

Working Paper 14/14

Semiparametric Localized Bandwidth Selection in Kernel Density Estimation

TINGTING CHENG, JITI GAO, and XIBIN ZHANG

Department of Econometrics and Business Statistics

Monash University, Australia

May 29, 2014

Authors' footnote: Tingting Cheng is a PhD student, Jiti Gao is Professor, and Xibin Zhang is an Associate Professor. Address: 900 Dandenong Road, Caulfield East, Victoria 3145, Australia. Emails: tingting.cheng@monash.edu; jiti.gao@monash.edu; and xibin.zhang@monash.edu.

Abstract

Since conventional cross–validation bandwidth selection methods do not work for the case where the data considered are serially dependent, 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 serially dependent. Nonetheless, a difficult issue is how we can consistently estimate a localized bandwidth. In this paper, we propose a semiparametric estimation method and establish an asymptotic theory for the proposed estimator. A by–product of this bandwidth estimate is a new sampling–based likelihood approach to hyperparameter estimation. Monte Carlo simulation studies show that the proposed hyperparameter estimation method works very well, and that the proposed bandwidth estimator outperforms its competitors. Applications of the new bandwidth estimator to the kernel density estimation of Eurodollar deposit rate, as well as the S&P 500 daily return under conditional heteroscedasticity, demonstrate the effectiveness and competitiveness of the proposed semiparametric localized bandwidth.

KEYWORDS: hyperparameter estimation; likelihood score; localized bandwidth.

JEL classification: C13, C14, C21.

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 Bayesian approach to bandwidth selection.

There is a large body of literature on bandwidth selection for kernel density estimation. Jones, Marron and Sheather (1996), Sheather (2004) and Heidenreich, Schindler and Sperlich (2013) presented surveys on bandwidth selection methods, including the rule–of–thumb, least squares cross–validation, biased cross–validation, plug–in method and a smoothed bootstrapping method. There are 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 small bandwidths be assigned to the observations in the high–density region and large bandwidths be assigned to those in the low–density region. The localized kernel density estimator attempts to resolve this problem by allowing different levels of smoothness 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 posterior of the bandwidth parameter. The resulting bandwidth estimate is a function of the density point x, 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 localized bandwidth and derived a closed–form Bayes estimate of the bandwidth, where the prior of bandwidth belongs to a conjugate family of the likelihood.[1]

[1] In these papers, as well as in our approach to bandwidth selection, bandwidth is treated as a parameter, whose posterior density is employed to derive a posterior estimate of bandwidth. Therefore, it should be stressed here that this approach is quite 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).

However, to the best of our knowledge, there is no asymptotic theory for the above–derived Bayes bandwidth estimation method in the current literature. This paper aims to fill this gap and present an asymptotic theory for the Bayes bandwidth estimation method. We propose a semiparametric localized bandwidth estimation method and then establish an asymptotically normal distribution with $\sqrt{n}$–rate of convergence. In addition, the semiparametric localized bandwidth estimation method has better out–of–sample performance than a global bandwidth when estimating the density of daily Eurodollar deposit rate, as well as 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 a vector of parameters, known as hyperparameters. In this paper, we assume 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 in the literature that expressing honest prior information can be difficult. From a Bayesian's 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 exist 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 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. We develop 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 subjectively chosen hyperparameter values.

The main contributions of this paper are:

(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 very well in Monte Carlo simulation and 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 likelihood approach to hyperparameter estimation;

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

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 Section 5 and Section 6, two empirical examples are presented to illustrate the application of the proposed semiparametric localized bandwidth selection method. Section 7 concludes the paper. The proofs of the main theorems are given in a supplemental document.

2. SEMIPARAMETRIC LOCALIZED BANDWIDTH

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

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

where $K_h(\cdot) = K(\cdot/h)/h$, and $K(\cdot)$ is the kernel function. 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 density 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

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

which is denoted as $q(x)/p(x)$ with $q(x) = \int hf(x|h)\pi(h)dh$ and $p(x) = \int f(x|h)\pi(h)dh$. As $\pi(h)$ is usually unknown in practice, we consider the situation where $\pi(h)$ belongs to a parametric

family indexed by a vector of parameters, that is,

$$\pi(h) \in \left\{ \pi(h|\theta) : \int \pi(h|\theta)\, dh = 1, \theta \in \Theta \right\},$$

where θ is 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

$$q(x|\theta) = \int h f(x|h)\pi(h|\theta)\, dh, \quad \text{and} \quad p(x|\theta) = \int f(x|h)\pi(h|\theta)\, dh.$$

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 just below Theorem 4, we also discuss the case where $h = a_n \cdot b$ and then the corresponding estimation of b, where $a_n \to 0$ and b is being treated as an unknown parameter.

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 $\{x_1, x_2, \cdots, x_n\}$, 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 each given θ, we estimate $h_0(x|\theta)$ by

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

which we call the empirical Bayes estimator of $h_0(x|\theta)$ and is denoted as $q_n(x|\theta)/p_n(x|\theta)$ with

$$q_n(x|\theta) = \int h\widehat{f}(x|h)\pi(h|\theta)\, dh, \quad \text{and} \quad p_n(x|\theta) = \int \widehat{f}(x|h)\pi(h|\theta)\, dh.$$

We will investigate the asymptotic properties of $h_n(x|\theta)$ in the next section. We now consider a situation where the prior of h, $\pi(h|\theta)$, can be estimated by $\pi(h|\widehat{\theta})$, in which $\widehat{\theta}$ is the maximum likelihood estimate (MLE) of θ and is derived through the following simulation scheme.

In order to 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 an initial value of θ. This random 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 the 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 $\left\{x_j^* : k = 1,2,\cdots,n^*\right\}$, is a random sample. As to the choice of n^*, our asymptotic results show that $n/n^* = o_P(1)$. This means that n^* should be large enough compared with n. In this paper, we choose $n^*/n = 5$ given the computational intensity. The detailed sampling procedure is described as follows.

Step 1: Generate a random sample $\{x_i : i = 1,2,\cdots,n\}$ from $f(x)$ and denote $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)'$.

Step 2: For a given initial value of θ denoted as θ_0, we choose an arbitrary initial value of x_0^*. The choice of θ_0 will be discussed in the next section.

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, and τ is a tuning parameter. τ is tuning constant 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

$$\min\left(\frac{p_n\left(x_j^*|\theta_0\right)}{p_n\left(x_{j-1}^*|\theta_0\right)}, 1\right), \tag{5}$$

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 generated sample is denoted as $\left\{x_j^* : j = 1,2,\cdots,n^*\right\}$.

The MLE of θ is obtained as

$$\widehat{\theta} = \arg\max_{\theta\in\Theta} \prod_{i=1}^{n^*} \int \widehat{f}\left(x_i^*|h\right)\pi(h|\theta)\,dh. \tag{6}$$

We then define a semiparametric localized bandwidth (SLB) estimator of $h_0(x|\theta)$ by

$$\widehat{h}_n(x|\widehat{\theta}) = \frac{\int h\widehat{f}(x|h)\pi(h|\widehat{\theta})dh}{\int \widehat{f}(x|h)\pi(h|\widehat{\theta})dh}, \tag{7}$$

which is denoted as $\widehat{q}_n(x|\widehat{\theta})/\widehat{p}_n(x|\widehat{\theta})$ with

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

In the next section, we establish an asymptotic theory for $\widehat{h}_n(x|\widehat{\theta})$.

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

Assumption 2. Kernel function $K(\cdot)$ is continuous and bounded.

Assumption 3. Let $\pi(h;\theta_1)$ satisfy

$$|\pi(h|\theta_1) - \pi(h|\theta_0)| \leq L(h|\theta_0)||\theta_1 - \theta_0||, \tag{8}$$

where $L(h|\theta_0)$ is a positive function such that $\int h^i f(x|h)L(h|\theta_0)dh < \infty$, for $i = 0$ and 1, and $\theta_1 \in \theta(\delta) = \{\theta_1 : ||\theta_1 - \theta_0|| \leq \delta\}$, for some $\delta > 0$. In addition, $\pi(h|\theta)$ satisfies that

$$\int \left|hp(x) - q(x)\right|\pi(h|\theta)dh < \infty.$$

We introduce the following notation:

$$
\begin{aligned}
f_u(x) &= \frac{1}{u}EK\left(\frac{X_i - x}{u}\right) = \frac{1}{u}\int K\Big(\frac{y-x}{u}\Big) f(y)dy, \\
f_{uv}(x) &= \frac{1}{uv}E\left[K\left(\frac{X_i - x}{u}\right)K\left(\frac{X_i - x}{v}\right)\right] = \frac{1}{uv}\int K\Big(\frac{y-x}{u}\Big)K\Big(\frac{y-x}{v}\Big) f(y)dy, \\
R_{uv}(x) &= f_{uv}(x) - f_u(x)f_v(x), \\
g_{uv,s}(x) &= \frac{1}{uv}\int K\Big(\frac{y-x}{u}\Big)K\Big(\frac{z-x}{v}\Big) f_s(y,z)dydz, \\
G_{uv,s}(x) &= g_{uv,s}(x) - f_u(x)f_v(x), \\
m_{uv,s}(x_a,x_b) &= \frac{1}{uv}\int K\Big(\frac{y-x_a}{u}\Big)K\Big(\frac{z-x_b}{v}\Big) f_s(y,z)dydz, \\
S_{uv,s}(x_a,x_b) &= m_{uv,s}(x_a,x_b) - f_u(x_a)f_v(x_b),
\end{aligned}
$$

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

Let $p(x)p(x) = Q(x)$ and $Q_{ij}^{-1} = \text{diag}\big(Q^{-1}(x_i), Q^{-1}(x_j)\big)$. We now present the asymptotic properties of empirical Bayes 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.

We then establish some new theorems; their proofs are given in a supplemental document.

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

$$\sqrt{n}(h_n(x) - h_0(x)) \to_D N(0, \Sigma_0(x)), \tag{9}$$

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

$$\gamma(j) = \iint \big[up(x) - q(x)\big]\big[vp(x) - q(x)\big] G_{uv,j}(x)\pi(u)\pi(v)dudv,$$

and

$$\gamma(0) = \iint [up(x) - q(x)][vp(x) - q(x)] R_{uv}(x)\pi(u)\pi(v)dudv.$$

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 around each point x.

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

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

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

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

Theorem 3. Let Assumptions 1–3 hold. Then as $n \to \infty$,

$$\sqrt{n}(\widehat{h}_n(x) - h_0(x)) \to_D N(0, \Sigma_0(x)), \tag{11}$$

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 random chosen points. Under Assumptions 1–3, as $n \to \infty$, we have

$$\left[\sqrt{n}(\widehat{h}_n(x_1) - h_0(x_1)), \cdots, \sqrt{n}(\widehat{h}_n(x_N) - h_0(x_N))\right] \to_D N(0, \Sigma_N), \tag{12}$$

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 1–3 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

$$b_0(x) = \int b\pi(b|x)db = \frac{\int bf(x|b)\pi(b)db}{\int f(x|b)\pi(b)db} = a_n^{-1} \cdot h_0(x), \tag{13}$$

where $h_0(x)$ is as defined in (2). 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 reduces to $n^{-\frac{3}{10}}$.

Remark 5. For multivariate kernel density estimator, 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.

4. MONTE CARLO SIMULATION RESULTS

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

$$\pi\left(h^2\right) = \frac{\beta^\alpha}{\Gamma(\alpha)}\left(\frac{1}{h^2}\right)^{\alpha+1} \exp\left\{-\frac{\beta}{h^2}\right\}, \tag{14}$$

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

$$\pi\left(h|\theta\right) = \frac{2\beta^\alpha}{\Gamma(\alpha)}\left(\frac{1}{h}\right)^{2\alpha+1} \exp\left\{-\frac{\beta}{h^2}\right\}. \tag{15}$$

We provide several Lemmas for computation and summarize simulation results as follows.

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 $\widehat{h}_n(x|\widehat{\theta})$. The proofs are relegated to Appendix.

Let

$$\begin{aligned} p_m(x) &= \frac{1}{m}\sum_{i=1}^{m} f(x+u_i v_i), & q_m(x) &= \frac{1}{m}\sum_{i=1}^{m} f(x+u_i v_i)v_i, \\ p_{nm}(x) &= \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{1}{v_j}K\left(\frac{x_i-x}{v_j}\right), & q_{nm}(x) &= \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} K\left(\frac{x_i-x}{v_j}\right), \end{aligned} \tag{16}$$

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

$$p_n(x|\theta) = \frac{1}{n}\frac{\beta^\alpha \Gamma(\alpha+1/2)}{\sqrt{2\pi}\Gamma(\alpha)} \sum_{i=1}^{n} \left(\frac{(x_i - x)^2}{2} + \beta \right)^{-(\alpha+1/2)}. \tag{17}$$

Thus, for the simulated random sample $\left\{ x_j^* : j = 1, 2, \cdots, n^* \right\}$, we have

$$p_n\left(x_j^*|\theta\right) = \int_0^\infty \widehat{f}\left(x_j^*|h\right)\pi(h|\theta)dh = \frac{1}{n}\frac{\beta^\alpha \Gamma(\alpha+1/2)}{\sqrt{2\pi}\Gamma(\alpha)} \sum_{i=1}^{n} \left(\frac{\left(x_i - x_j^*\right)^2}{2} + \beta \right)^{-(\alpha+1/2)}. \tag{18}$$

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

$$\ell_*\left(x_1^*, \cdots, x_{n^*}^*|\theta\right) = \prod_{j=1}^{n^*} p_n\left(x_j^*|\theta\right) = \prod_{j=1}^{n^*} \left\{ \frac{1}{n}\frac{\beta^\alpha \Gamma(\alpha+1/2)}{\sqrt{2\pi}\Gamma(\alpha)} \sum_{i=1}^{n} \left(\frac{(x_i - x_j^*)^2}{2} + \beta \right)^{-(\alpha+1/2)} \right\}, \tag{19}$$

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 (16) 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 θ_0, which we use to simulate a random sample from $p_n\left(x_j^*|\theta_0\right)$. In Monte Carlo simulation studies, we specify hyperparameter values, which can be used as the initial values of the components of θ, because the purpose of this simulation study is to examine the performance of our proposed likelihood approach to hyperparameter estimation. In empirical studies, θ is unknown and will be estimated. We propose that the initial value of θ be estimated based on the original sample, $\{x_i : i = 1, 2, \cdots, n\}$. The resulting estimate is denoted as $\widehat{\theta}_0$, and thus, we can simulate random samples from $p_n\left(x_j^*|\widehat{\theta}_0\right)$.

In this simulation study, we consider two scenarios, which are classified by the way that the initial value of θ specified. In the first scenario, we choose $\theta_0 = (1, 0.05)$, whose components are the hyperparameter values of the inverse Gamma prior for h^2. In the second scenario, we estimate θ using the original sample of $\boldsymbol{x}$, and use the estimated θ denoted as $\widehat{\theta}_0$, as the initial value of θ. In the first scenario, we calculate $h_0(x|\theta_0)$ given by (2), $h_n(x|\theta_0)$ given by (4) and $\widehat{h}_n(x|\widehat{\theta})$ given by (7). In the second scenario, we calculate $h_0(x|\widehat{\theta}_0)$, $h_n(x|\widehat{\theta}_0)$ and $\widehat{h}_n(x|\widehat{\theta})$.

Scenario 1: The initial value of θ is known.

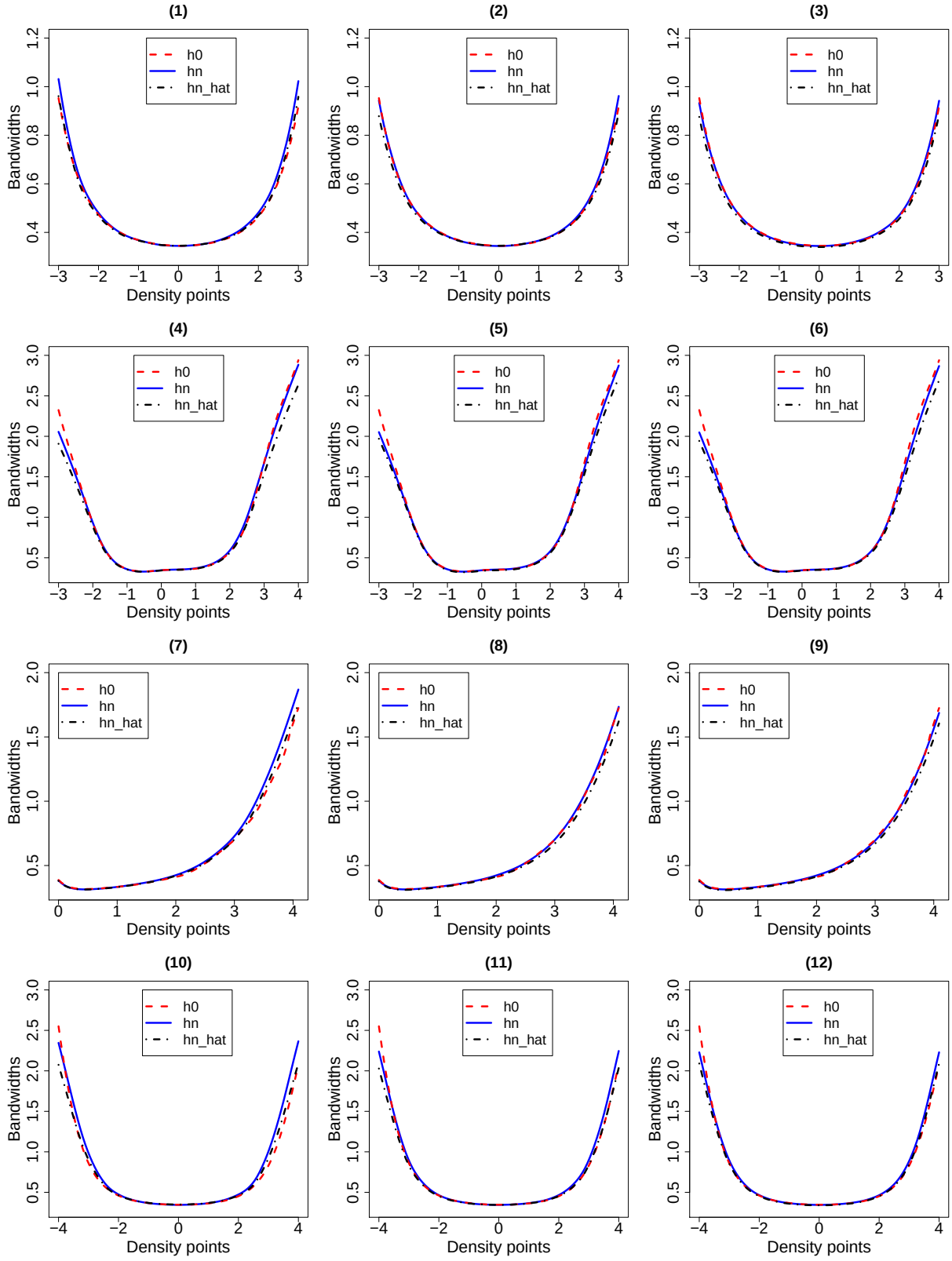
We aim to compare the performance of $h_0(x|\theta_0)$, $h_n(x|\theta_0)$ and $\widehat{h}_n(x|\widehat{\theta})$ through simulated random samples, where sample sizes are respectively, 250, 750 and 1500, and the number of replications for each sample size is 500. To generate such random samples, we consider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

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 $N(0,1)$, $(-3,4)$ for the mixture density, $(0,4.1)$ for $W(1.5,1)$ 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. For random samples generated from $N(0,1)$, such averaged bandwidth curves are plotted in Figure 1 (1)–(3); for random samples generated from the mixture density of two Gaussians, the averaged bandwidth curves are plotted in Figure 1 (4)–(6); for random samples generated from the weibull distribution, the averaged bandwidth curves are plotted in Figure 1 (7)–(9); for random samples generated from the specified AR(1) process, the averaged bandwidth curves are plotted in Figure 1 (10)–(12).

All twelve graphs presented in Figure 1 show that localized bandwidth should be used for kernel density estimation, Nonetheless, when random samples are generated from the mixture density of two Gaussians, bandwidth can be approximately treated as a constant when x is inside $(-1,1)$, which is a high density region of the mixture density. When random samples are generated from the AR(1) process, bandwidth can be approximately treated as a constant when x is inside $(-2,2)$, which is a high density region of the AR(1) process. We also found that our proposed semiparametric localized bandwidth curve, $\widehat{h}_n(x|\widehat{\theta})$, differs from $h_n(x|\theta_0)$ and $h_0(x|\theta_0)$ when x is located outside the high density region of the underlying density function to be estimated.

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

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* $N(0,1)$ *(first row),* $0.4N(-0.6,0.16)+0.6N(0.4,0.49)$ *(second row),* $W(1.5,1)$ *(third row) and AR(1) process (fourth row), where* θ_0 *is known. The three columns correspond to sample sizes of 250, 750 and 1500, respectively.*



bias and standard deviation measures as follows:

$$\text{bias}_1 = \frac{1}{100}\frac{1}{500}\sum_{i=1}^{100}\sum_{r=1}^{500}\left(h_{n,r}\left(x_i|\theta_0\right) - h_0\left(x_i|\theta_0\right)\right),$$

$$\text{bias}_2 = \frac{1}{100}\frac{1}{500}\sum_{i=1}^{100}\sum_{r=1}^{500}\left(\widehat{h}_{n,r}\left(x_i|\widehat{\theta}\right) - h_0\left(x_i|\theta_0\right)\right),$$

$$\text{std}_1 = \sqrt{\frac{1}{100}\frac{1}{500}\sum_{i=1}^{100}\sum_{r=1}^{500}\left(h_{n,r}(x_i|\theta_0) - \overline{h}_n(x_i|\theta_0)\right)^2},$$

$$\text{std}_2 = \sqrt{\frac{1}{100}\frac{1}{500}\sum_{i=1}^{100}\sum_{r=1}^{500}\left(\widehat{h}_{n,r}(x_i|\widehat{\theta}) - \overline{\widehat{h}}_n(x_i|\widehat{\theta})\right)^2},$$

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 mean squared errors (MSE) is calculated as

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

for $j = 1$ and 2, which correspond to $h_n(x|\theta_0)$ and $\widehat{h}_n(x|\widehat{\theta})$, respectively.

As shown by (16), 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 $N(0,1)$, $0.4N(-0.6,0.16) + 0.6N(0.4,0.49)$, $W(1.5,1)$ and the AR(1) process $x_t = 0.2x_{t-1} + u_t$, respectively.

We found the following evidence. First, when we generate random samples from all the four data generating processes, $\widehat{h}_n(x|\widehat{\theta})$ has larger bias and variation than $h_n(x|\theta_0)$ for sample sizes of 750 and 1500. The MSE measure shows that $h_n(x|\theta_0)$ performs slightly better than $\widehat{h}_n(x|\widehat{\theta})$ under each sample size.

Second, as the sample size increases, the bias measures, as well as the standard deviation measures and MSE 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 under each of the two scenarios.*

	Sample	**Scenario 1**				**Scenario 2**			
	size	Gaussian	Mixture	Weibull	AR(1)	Gaussian	Mixture	Weibull	AR(1)
		$h_n(x\|\theta_0) - h_0(x\|\theta_0)$				$h_n(x\|\widehat{\theta}_0) - h_0(x\|\widehat{\theta}_0)$			
	250	0.0149	−0.0210	0.0204	0.0613	0.0055	0.0261	0.0574	0.0364
bias_1	750	0.0037	−0.0136	0.0035	0.0183	0.0020	0.0188	0.0281	0.0162
	1500	0.0014	−0.0098	−0.0021	0.0137	0.0007	0.0129	0.0184	0.0141
	250	0.0962	0.0996	0.1273	0.2229	0.0929	0.0984	0.1370	0.2556
std_1	750	0.0533	0.0841	0.0771	0.1638	0.0536	0.0743	0.0698	0.1527
	1500	0.0361	0.0498	0.0555	0.1275	0.0370	0.0567	0.0388	0.1236
	250	0.0095	0.0104	0.0166	0.0534	0.0087	0.0104	0.0221	0.0666
MSE_1	750	0.0029	0.0073	0.0060	0.0272	0.0029	0.0059	0.0057	0.0236
	1500	0.0013	0.0026	0.0031	0.0164	0.0014	0.0034	0.0018	0.0155
		$\widehat{h}_n(x\|\widehat{\theta}) - h_0(x\|\theta_0)$				$\widehat{h}_n(x\|\widehat{\theta}) - h_0(x\|\widehat{\theta}_0)$			
	250	0.0081	−0.0493	0.0101	−0.0546	0.0079	0.0224	0.0465	0.0288
bias_2	750	−0.0070	−0.0256	−0.0097	−0.0435	−0.0033	0.0174	0.0230	0.0121
	1500	−0.0054	−0.0168	−0.0071	−0.0315	−0.0015	0.0082	0.0142	0.0064
	250	0.1273	0.1165	0.1548	0.2664	0.0999	0.0867	0.1360	0.2319
std_2	750	0.0706	0.0949	0.0887	0.2125	0.0585	0.0505	0.0732	0.1376
	1500	0.0495	0.0570	0.0657	0.1806	0.0440	0.0338	0.0338	0.1114
	250	0.0163	0.0160	0.0241	0.0740	0.0100	0.0080	0.0207	0.0546
MSE_2	750	0.0050	0.0097	0.0080	0.0470	0.0034	0.0029	0.0059	0.0191
	1500	0.0025	0.0035	0.0044	0.0336	0.0019	0.0012	0.0013	0.0125

We are also interested in the performance of our proposed likelihood approach to hyperparameter estimation, where the size of the simulated sample $\left\{x_j^* : j = 1,2,\cdots,n^*\right\}$ is chosen, such that $n^*/n = 5$. Therefore, the sample sizes are $n^* = 1250, 3750$ and 7500 that correspond to sizes of original samples, $n = 250, 750$ and 1500, respectively. For each sample size, we calculated the following bias and standard deviation measures for the vector of hyperparameters $\theta = (\alpha, \beta)'$,

whose true value is $\theta_0 = (1, 0.05)'$. These measures are

$$\text{bias}_\theta = \frac{1}{500}\sum_{r=1}^{500}\left(\widehat{\theta}_r - \theta_0\right),$$

$$\text{std}_\theta = \sqrt{\frac{1}{500}\sum_{r=1}^{500}\left(\widehat{\theta}_r - \overline{\widehat{\theta}}\right)^2},$$

where $\widehat{\theta}_r$ is the estimate of θ derived at the rth replication, for $r = 1, 2, \cdots, 500$, and $\overline{\widehat{\theta}} = \frac{1}{/500}\sum_{r=1}^{500}\widehat{\theta}_r$. Moreover, we calculated the MSE as $\text{MSE}_\theta = \text{bias}_\theta^2 + \text{std}_\theta^2$.

Table 2: *Bias, standard deviation and MSE of hyperparameter estimates with random samples generated from the four different data generating processes, where $\widehat{\theta}$ of $\widehat{h}_n(x|\widehat{\theta})$ is derived via an initial value of $\theta_0 = (1, 0.05)'$.*

	Sample	Gaussian		Mixture		Weibull		AR(1)	
	size	α	β	α	β	α	β	α	β
bias_θ	1250	0.0985	0.0187	0.0700	0.0114	0.0322	0.0059	0.1119	0.0215
	3750	0.0512	0.0101	0.0062	0.0016	0.0199	0.0029	0.0549	0.0104
	7500	0.0229	0.0048	−0.0035	0.0008	−0.0027	0.0000	0.0111	0.0036
std_θ	1250	0.1868	0.0510	0.1835	0.0300	0.1610	0.0201	0.2639	0.0486
	3750	0.1094	0.0240	0.1190	0.0159	0.0963	0.0120	0.1438	0.0261
	7500	0.0962	0.0184	0.0762	0.0112	0.0741	0.0088	0.1221	0.0196
MSE_θ	1250	0.0446	0.0030	0.0386	0.0010	0.0270	0.0004	0.0822	0.0028
	3750	0.0146	0.0007	0.0142	0.0003	0.0097	0.0002	0.0237	0.0008
	7500	0.0098	0.0004	0.0058	0.0001	0.0055	0.0001	0.0150	0.0004

Table 2 shows that our proposed likelihood approach to hyperparameter estimation works very well, and that bias, standard deviation and MSE of the estimated hyperparameters decrease as the sample size increases.

Scenario 2: The initial value of θ is estimated using the original sample.

We now consider to specify the initial value of θ by estimating it through 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

$$\ell(x_1, x_2, \cdots, x_n|\theta) = \prod_{i=1}^{n} \left\{ \frac{1}{n} \frac{\beta^\alpha \Gamma(\alpha + 1/2)}{\sqrt{2\pi}\Gamma(\alpha)} \sum_{j=1; j\neq i}^{n} \left(\frac{(x_j - x_i)^2}{2} + \beta \right)^{-(\alpha+1/2)} \right\}. \tag{20}$$

which we maximize with respect to θ to obtain an initial estimate of θ denoted as $\widehat{\theta}_0$. As this sample contains observations of x rather than h, this initial estimate is likely to be inaccurate. Therefore, we simulate a random sample from $p_n(x|\widehat{\theta}_0)$, and then use this sample to derive the MLE of θ. The details of this simulation–based likelihood approach to hyperparameter estimation are as follows.

Step 1: Estimate θ_0 using the original sample by maximizing (20) and denote the resulting estimate as $\widehat{\theta}_0$.

Step 2: Generate a random sample, $\left\{x_j^* : j = 1, 2, \cdots, n^*\right\}$, from $p_n\left(x^*|\widehat{\theta}_0\right)$ given by (18) through the random–walk Metropolis algorithm.

Step 3: Estimate θ based on the generated sample by maximizing the likelihood function given by (19); and denote the resulting estimate as $\widehat{\theta}$.

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

For each sample size and under each distribution, we used the same 500 random samples as those generated in **Scenario 1**. Using each sample, we estimated θ through the above likelihood approach and calculated the bandwidth curve $\widehat{h}_n(x|\widehat{\theta})$ with x taking values on 100 equally spaced grid points on a finite interval. Moreover, we used this sample to calculate $h_0(x|\widehat{\theta}_0)$ and $h_n(x|\widehat{\theta}_0)$ on the same set of grid points. After averaging these three bandwidth curves over 500 Monte Carlo replications, we plotted these three averaged bandwidth curves in Figure 2.

Figure 2 (1)–(3) present the bandwidth curves averaged over 500 samples, which were generated from $N(0, 1)$. 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.

Figure 2 (4)–(6) 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.

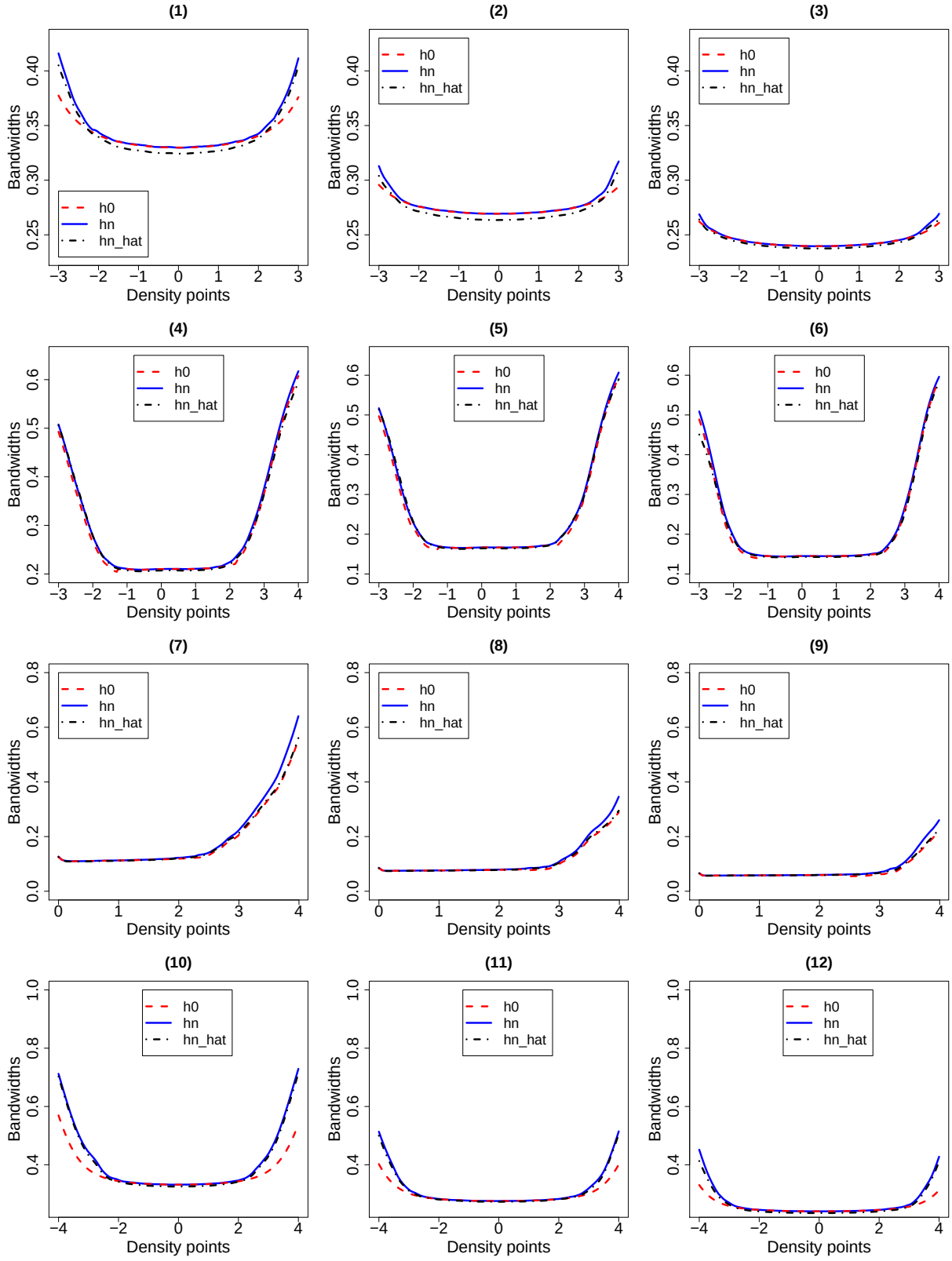
Figure 2 (7)–(9) present the bandwidth curves averaged over 500 samples, which were generated from the Weibull distribution. Regardless the size of original samples, $\widehat{h}_n(x|\widehat{\theta})$ is almost the same as the other two bandwidth curves when x is less than 3, but is clearly different 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.

Figure 2 (10)–(12) 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 conclude the simulation study in **Scenario 2**, 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 calculated the bias, standard deviation and MSE of $h_n(x|\widehat{\theta}_0)$ and $\widehat{h}_n(x|\widehat{\theta})$ as follows. Let $h_{n,r}(x_i|\widehat{\theta}_0)$ and $\widehat{h}_{n,r}(x_i|\widehat{\theta})$ denote $h_n(x_i|\widehat{\theta}_0)$ 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

Figure 2: *Averaged $h_0(x|\widehat{\theta}_0)$, $h_n(x|\widehat{\theta}_0)$ and $\widehat{h}_n(x|\widehat{\theta})$ with random samples generated from $N(0,1)$ (first row), $0.4N(-0.6,0.16)+0.6N(0.4,0.49)$ (second row), $W(1.5,1)$ (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.*



deviation measures as follows:

$$\text{bias}_1 = \frac{1}{100}\frac{1}{500}\sum_{i=1}^{100}\sum_{r=1}^{500}\left(h_{n,r}(x_i|\widehat{\theta}_0) - h_0(x_i|\widehat{\theta}_0)\right),$$

$$\text{bias}_2 = \frac{1}{100}\frac{1}{500}\sum_{i=1}^{100}\sum_{r=1}^{500}\left(\widehat{h}_{n,r}(x_i|\widehat{\theta}) - h_0(x_i|\widehat{\theta}_0)\right),$$

$$\text{std}_1 = \sqrt{\frac{1}{100}\frac{1}{500}\sum_{i=1}^{100}\sum_{r=1}^{500}\left(h_{n,r}(x_i|\widehat{\theta}_0) - \overline{h}_n(x_i|\widehat{\theta}_0)\right)^2},$$

$$\text{std}_2 = \sqrt{\frac{1}{100}\frac{1}{500}\sum_{i=1}^{100}\sum_{r=1}^{500}\left(\widehat{h}_{n,r}(x_i|\widehat{\theta}) - \overline{\widehat{h}}_n(x_i|\widehat{\theta})\right)^2},$$

where $\overline{h}_n(x_i|\widehat{\theta}_0) = 1/500\sum_{r=1}^{500} h_{n,r}(x_i|\widehat{\theta}_0)$ 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 $h_n(x_i|\widehat{\theta}_0)$ 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. Table 1 presents the bias, standard deviation and MSE of the SLB estimates under **Scenario 2** in the last four columns.

First, when random samples are generated from the Gaussian density, $\widehat{h}_n(x|\widehat{\theta})$ has larger bias and variation than $h_n(x|\widehat{\theta}_0)$ for each sample size considered. Consequently, under the MSE measure, $\widehat{h}_n(x|\widehat{\theta})$ performs poorer than $h_n(x|\widehat{\theta}_0)$. 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 $h_n(x|\widehat{\theta}_0)$, and thus, $\widehat{h}_n(x|\widehat{\theta})$ outperforms $h_n(x|\widehat{\theta}_0)$ under the MSE measure. When samples are generated from the Weibull distribution, $\widehat{h}_n(x|\widehat{\theta})$ has smaller bias and variation than $h_n(x|\widehat{\theta}_0)$, 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 $h_n(x|\widehat{\theta}_0)$ 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 $h_n(x|\widehat{\theta}_0)$ and $\widehat{h}_n(x|\widehat{\theta})$ decreases, respectively.

5. ESTIMATING AND FORECASTING THE DENSITY OF EURODOLLAR DEPOSIT RATE

We aim to 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 per-

formance 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 3.

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 x. We assume an inverse Gamma prior of the bandwidth parameter given by (14), in which the hyperparameter vector, $\theta = (\alpha, \beta)'$, is estimated through our proposed likelihood–based approach. As this approach requires an initial estimate of θ, we employ the sampling procedure described in **Scenario 2** of Section 4.

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 4.

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 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 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 x_t denote the observed deposit rate at day t, for

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

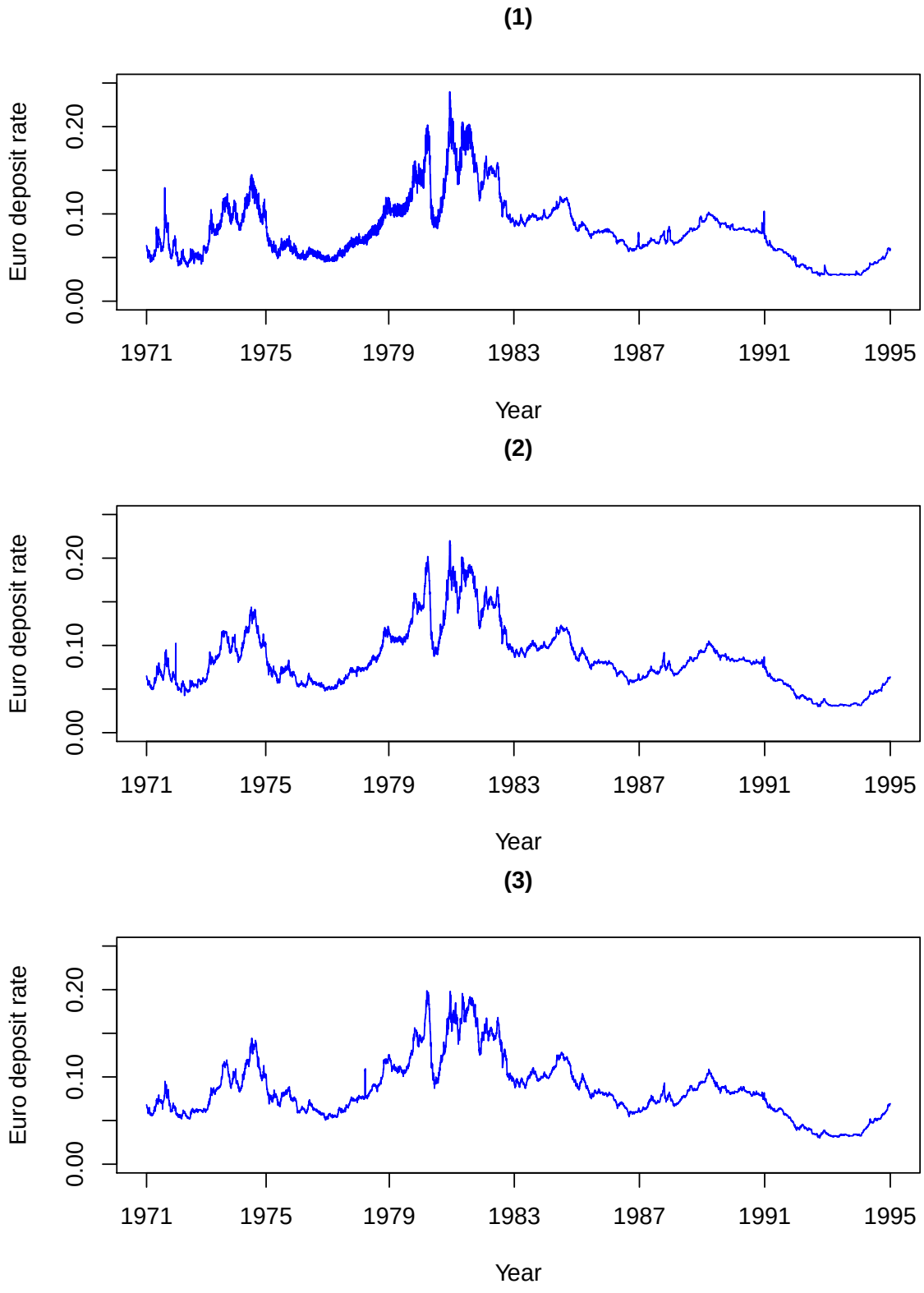
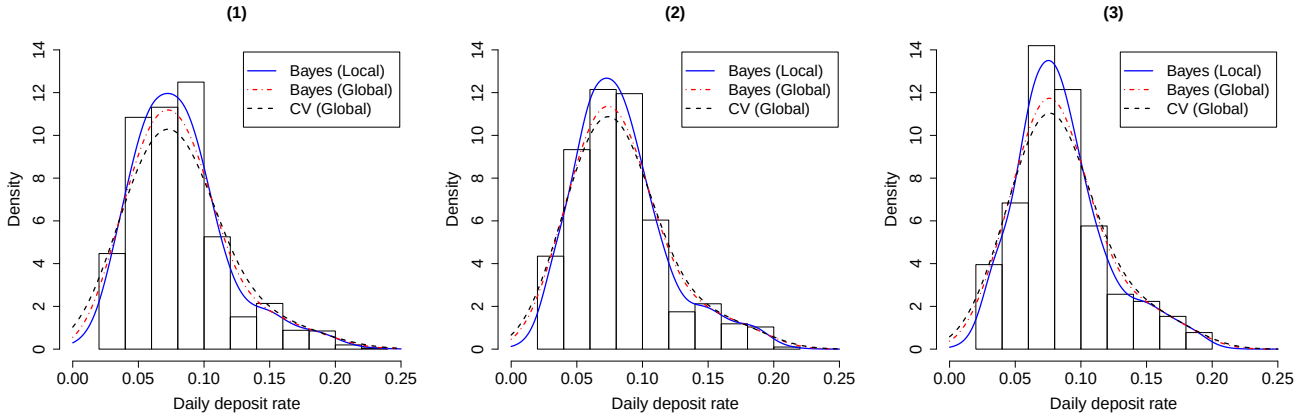


Figure 4: *Density estimates of daily Eurodollar deposit rates with maturities of 1, 3 and 6 months: (1) 1–month maturity; (2) 3–month maturity; and (3) 6–month maturity.*



$t = 1,2,\cdots,T$. The first sample contains the first n observations, $x_1, x_2, \cdots, x_n$, 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 x_{n+1}. The second sample contains $x_2, x_3, \cdots, x_{n+1}$, 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 x_{n+2}. This rolling procedure continues until the density of x_T is forecasted.

At the rth iteration, for $r = 1,2,\cdots,T-n$, we denote the SLB estimator as $h_{\text{SLB}}(x)$, 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 $\widehat{f}\,(x_{n+r}|h_{\text{SLB}}(x_{n+r}))$, $\widetilde{f}\,(x_{n+r}|h_{\text{CV}})$ and $\widetilde{f}\,(x_{n+r}|h_{\text{Bayes}})$.

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

$$\begin{aligned}
\frac{1}{T-n}\sum_{r=1}^{T-n}\log\widehat{f}\,(x_{n+r}|h_{\text{SLB}}(x_{n+r})) &= \frac{1}{T-n}\sum_{r=1}^{T-n}\log\left[\frac{1}{n}\sum_{i=r}^{n+r-1}\frac{1}{h_{\text{SLB}}(x_{n+r})}\phi\left(\frac{x_{n+r}-x_i}{h_{\text{SLB}}(x_{n+r})}\right)\right],\\
\frac{1}{T-n}\sum_{r=1}^{T-n}\log\widetilde{f}\,(x_{n+r}|h_{\text{Bayes}}) &= \frac{1}{T-n}\sum_{r=1}^{T-n}\log\left[\frac{1}{n}\sum_{i=r}^{n+r-1}\frac{1}{h_{\text{Bayes}}}\phi\left(\frac{x_{n+r}-x_i}{h_{\text{Bayes}}}\right)\right],\\
\frac{1}{T-n}\sum_{r=1}^{T-n}\log\widetilde{f}\,(x_{n+r}|h_{\text{CV}}) &= \frac{1}{T-n}\sum_{r=1}^{T-n}\log\left[\frac{1}{n}\sum_{i=r}^{n+r-1}\frac{1}{h_{\text{CV}}}\phi\left(\frac{x_{n+r}-x_i}{h_{\text{CV}}}\right)\right],
\end{aligned} \tag{21}$$

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.

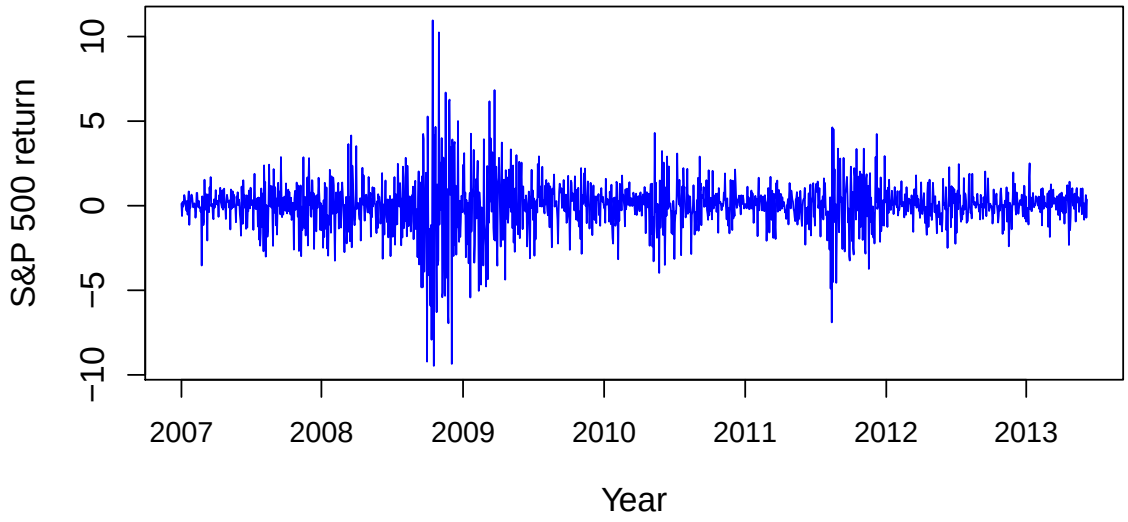
Table 4: *Out–of–sample average logarithmic scores of density estimates with bandwidths estimation through three different methods.*

	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 asset return in finance. In this section, we aim to 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 4th January 2007 to the 30th May 2013 from http://finance.yahoo.com. The date t return is calculated as

Figure 5: *Time series plot of S&P 500 daily returns.*



$y_t = \log(p_t/p_{t-1})$, and there are $T = 1612$ observations of the return. The time series plot of the return series is presented in Figure 5.

6.1 Density estimation of S&P 500 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

$$\begin{aligned} y_t &= \sigma_t \varepsilon_t, \\ \sigma_t^2 &= b_0 + b_1 y_{t-1}^2 + b_2 \sigma_{t-1}^2, \end{aligned} \tag{22}$$

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

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

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 **Scenario 2** of Section 4. The estimation procedure is described as follows.

Step 1: Estimate the GARCH model given by (22) 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 $\widehat{\theta}_0 = (\widehat{\alpha}_0, \widehat{\beta}_0)'$, by maximizing the marginal likelihood given by (17) 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 (17) with α and β being their initial estimates. Obtain the MLE of θ, denoted as $\widehat{\theta}$, by maximizing the likelihood function given by (19) 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 (7) with x being replaced by ε. Thus, the Gaussian kernel error density is approximated as

$$\widehat{f}_\varepsilon(\varepsilon_t) = \frac{1}{T}\sum_{i=1}^{T} \frac{1}{h_n(\varepsilon_t|\widehat{\theta})}\phi\left(\frac{\varepsilon_t - \varepsilon_i}{h_n(\varepsilon_t|\widehat{\theta})}\right). \tag{24}$$

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

$$\widehat{f}_Y(y_t|\lambda) = \frac{1}{T\sigma_t}\sum_{i=1}^{T} \frac{1}{h_n\left(y_t/\sigma_t|\widehat{\theta}\right)}\phi\left(\frac{y_t/\sigma_t - y_i/\sigma_i}{h_n\left(y_t/\sigma_t|\widehat{\theta}\right)}\right), \tag{25}$$

for $t = 1, 2, \cdots, T$, where $\lambda = (b_1, b_2)'$.[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

$$\widehat{f}_Y(y_{T+1}|\widehat{\lambda}) = \frac{1}{T\widehat{\sigma}_{T+1}}\sum_{t=1}^{T} \frac{1}{\widehat{h}_n\left(y_{T+1}/\widehat{\sigma}_{T+1}\right)}\phi\left(\frac{y_{T+1}/\widehat{\sigma}_{T+1} - y_t/\widehat{\sigma}_t}{\widehat{h}_n\left(y_{T+1}/\widehat{\sigma}_{T+1}|\widehat{\theta}\right)}\right), \tag{26}$$

[2]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.

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 (26) at 1000 grid points, where h and σ_t are replaced by respectively, $\widetilde{h}$ and $\widetilde{\sigma}_t$.

We also used the likelihood cross–validation (CV) 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 (22) through the quasi maximum likelihood method under the normality assumption of ε_t, and the resulting estimates are denoted as $\widetilde{b}_0^{(\text{cv})}$, $\widetilde{b}_1^{(\text{cv})}$ and $\widetilde{b}_2^{(\text{cv})}$. We then computed the residuals as $\widetilde{\varepsilon}_t = y_t / \widetilde{\sigma}_{t,\text{cv}}$, where

$$\widetilde{\sigma}_{t,\text{cv}}^2 = \widetilde{b}_0^{(\text{cv})} + \widetilde{b}_1^{(\text{cv})} y_{t-1}^2 + \widetilde{b}_2^{(\text{cv})} \widetilde{\sigma}_{t-1,\text{cv}}^2, \tag{27}$$

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 (23) and expressed as

$$\widetilde{f}_Y(y_t|\lambda) = \frac{1}{T\sigma_t} \sum_{i=1}^{T} \frac{1}{\widetilde{h}_{\text{cv}}} \phi\left(\frac{y_t/\sigma_t - y_i/\sigma_i}{\widetilde{h}_{\text{cv}}}\right), \tag{28}$$

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 $\left(\widetilde{b}_1^{(\text{cv})}, \widetilde{b}_2^{(\text{cv})}\right)'$. 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 $\widetilde{h}_{\text{cv}}$. With the updated bandwidth being substituted into the likelihood

function given by (28), 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},

$$\widetilde{f}_Y(y_{T+1}|\lambda) = \frac{1}{T\widetilde{\sigma}_{T+1,\mathrm{cv}}} \sum_{t=1}^{T} \frac{1}{\widetilde{h}_{\mathrm{cv}}} \phi\left(\frac{y_{T+1}/\widetilde{\sigma}_{T+1,\mathrm{cv}} - y_t/\widetilde{\sigma}_{t,\mathrm{cv}}}{\widetilde{h}_{\mathrm{cv}}}\right), \tag{29}$$

at 1000 grid points, where $\widetilde{\sigma}_{T+1,\mathrm{cv}}$ is calculated through (27).

The resulting three kernel estimates of the density of the one–day–ahead S&P 500 daily return are plotted in Figure 6, 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 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 $T = 1612$, and the size of a rolling sample is $n = 1007$, where the first rolling sample is from 4th January 2007 to 31st December 2010.

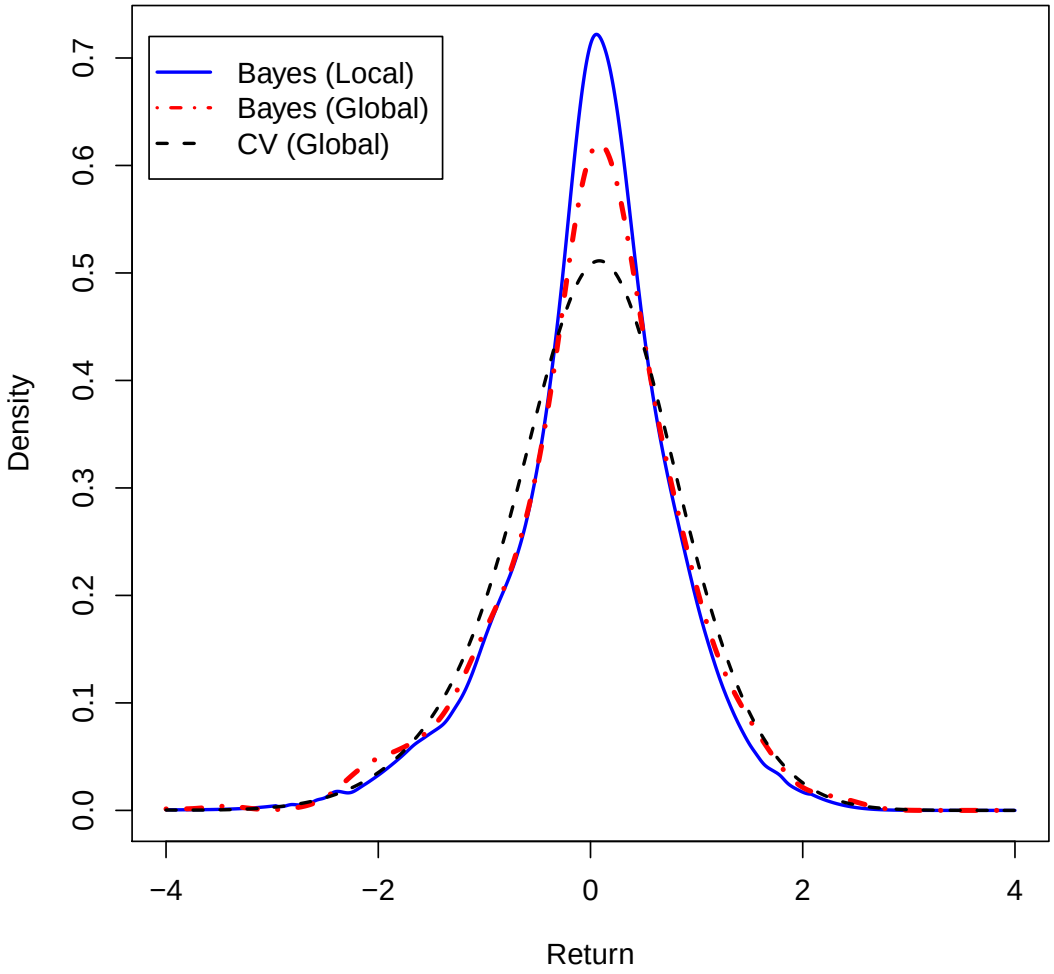
At the rth iteration of the rolling sample procedure, we denote the SLB estimate as $h_{\mathrm{SLB}}(y)$, the global bandwidth chosen through CV as h_{CV}, and the global bandwidth estimated through Bayesian sampling of Zhang, King and Hyndman (2006) as h_{Bayes}. The corresponding density estimates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

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

$$\begin{aligned}
&\frac{1}{T-n}\sum_{r=1}^{T-n} \log \widehat{f}\left(y_{n+r}|h_{\mathrm{SLB}}(y_{n+r})\right) = -1.3347,\\
&\frac{1}{T-n}\sum_{r=1}^{T-n} \log \widetilde{f}\left(y_{n+r}|h_{\mathrm{Bayes}}\right) = -1.4752,\\
&\frac{1}{T-n}\sum_{r=1}^{T-n} \log \widetilde{f}\left(y_{n+r}|h_{\mathrm{CV}}\right) = -1.5415.
\end{aligned}$$

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,

Figure 6: *Density estimates of the one–day–ahead out–of–sample S&P 500 daily return.*



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 and York University in England, in particular to Dr Anastasios Panagiotelis from Monash University and Professor Wenyang Zhang from York University in England. This research was supported under the Australian Research Council's *Discovery Projects* funding scheme (project 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.

Amisano, G. and 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. and Greenberg, E. (1995), 'Understanding the Metropolis–Hastings algorithm', *The American Statistician* **49**(4), 327–335.

de Lima, M. S. and 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. and 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. and González-Rivera, G. (1991), 'Semiparametric ARCH models', *Journal of Business and Economic Statistics* **9**(4), 345–359.

Fan, J. and Yao, Q. (2003), *Nonlinear Time Series: Nonparametric and Parametric Methods*, Springer.

Gangopadhyay, A. and 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: Semiparametric and Nnonparametric Methods*, Chapman & Hall/CRC.

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

Ghosh, J. K. and Ramamoorthi, R. V. (2003), *Bayesian Nonparametrics*, Springer, New York.

Heidenreich, N.-B., Schindler, A. and 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. and Walker, S. G. (2010), *Bayesian Nonparametrics*, Cambridge University Press, Cambridge, U.K.

Jones, M. C., Marron, J. S. and 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. and 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', *The Annals of Statistics* **12**(1), 351–357.

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

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

Seaman, D. E. and 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/CRC.

Zhang, X. and 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. and 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

We now provide the proofs of Lemmas 1 and 2 and Theorems 1–4. 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

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

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

Proof of Lemma 2.

We have

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

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

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

fore, $q_{nm}(x) - q_n(x) = o_P(1)$.

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 $E(\widehat{f}(x|h)) = f(x|h)$ and that

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

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

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

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

It is easy to check that

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

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

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

where

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

Thus,

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

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

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

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

where

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

By Assumption 1, $\{X_t\}$ is α–mixing strictly stationary, we have $EK\left(\frac{X_j - x}{v}\right) = EK\left(\frac{X_i - x}{v}\right) = vf_v(x)$,

therefore $EK\left(\frac{X_i-x}{u}\right)EK\left(\frac{X_j-x}{v}\right)=uvf_u(x)f_v(x)$.

$$EK\left(\frac{X_i-x}{u}\right)K\left(\frac{X_j-x}{v}\right)=\int K\left(\frac{y-x}{u}\right)K\left(\frac{z-x}{v}\right)f_s(y,z)dydz=uvg_{uv,s}(x),$$

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

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

Therefore,

$$\begin{aligned}\mathrm{Cov}\left[V_i(x),V_j(x)\right]&=\iint\left[up(x)-q(x)\right]\left[vp(x)-q(x)\right]\left\{EU_i(x;u)U_j(x;v)\right\}\pi(u)\pi(v)dudv\\&=\iint\left[up(x)-q(x)\right]\left[vp(x)-q(x)\right]G_{uv,s}(x)\pi(u)\pi(v)dudv.\end{aligned}$$

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

Now we need to verify one of the following two conditions to establish the central limit theorem:

(i) $E|V_i(x)|^\delta<\infty$ and $\sum_{j\geq1}\alpha(j)^{1-2/\delta}<\infty$ for some constant $\delta>2$;

(ii) $P(|V_i(x)|<c)=1$ for some constant $c>0$ and $\sum_{j\geq1}\alpha(j)<\infty$.

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 $\sum_{j\geq1}\alpha(j)^{1-2/\delta}<\infty$. Denote $g_1(h)=\left[hp(x)-q(x)\right]\pi(h)$ and $g_2(h)=U_i(x;h)$. Under Hödler's inequality that for $\delta>1$, $1/\delta+1/q=1$, we have

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

As $K(\cdot)$ is bounded, we have $|g_2(h)|^\delta=|U_i(x;h)|^\delta=\left|\frac{1}{h}\left[K\left(\frac{X_i-x}{h}\right)-EK\left(\frac{X_i-x}{h}\right)\right]\right|^\delta<\infty$.
According to Assumption 3, we have $\int g_1(h)dh=\int\left(hp(x)-q(x)\right)\pi(h)dh<\infty$, thus,

$$|V_i(x)|^\delta=\left(\int g_1(h)dh\right)^{\delta-1}\int g_1(h)|g_2(h)|^\delta dh<\infty.$$

Therefore, $E|V_i(x)|^\delta < \infty$.

Lemma A.1 of Gao (2007) implies that

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

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

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

Assumption 1–3 ensures 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

$$\sqrt{n}L_n(x) \to_D N(0, \Sigma_L(x)), \tag{30}$$

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

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

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

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

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

$$\sqrt{n}(h_n(x) - h_0(x)) \to_D N(0, \Sigma_0(x)), \tag{31}$$

where $\Sigma_0(x) = Q^{-2}(x)\Sigma_L(x)$. Therefore, we proofed 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

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

where

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

$$\begin{aligned}
&E\left[U_i(x_1;u)U_j(x_2;v)\right] = \frac{1}{uv}E\left\{\left[K\left(\frac{X_i - x_1}{u}\right) - EK\left(\frac{X_i - x_1}{u}\right)\right]\left[K\left(\frac{X_j - x_2}{v}\right) - EK\left(\frac{X_j - x_2}{v}\right)\right]\right\} \\
&= \frac{1}{uv}\left[EK\left(\frac{X_i - x_1}{u}\right)K\left(\frac{X_j - x_2}{v}\right) - EK\left(\frac{X_i - x_1}{u}\right)EK\left(\frac{X_j - x_2}{v}\right)\right].
\end{aligned}$$

As $EK\left(\frac{X_i-x_1}{u}\right)EK\left(\frac{X_j-x_2}{v}\right)=uvf_u(x_1)f_v(x_2)$, we have

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

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

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

Consequently, we have

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

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

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

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 $Y_i = C_1 V_i(x_1) + C_2 V_i(x_2)$. It is easy to show that $E(Y_i) = 0$.

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

where

$$\begin{aligned}
a_1 &= C_1^2 E\left[V_i^2(x_1)\right] = C_1^2 \iint [up(x_1) - q(x_1)][vp(x_1) - q(x_1)] R_{uv}(x_1)\pi(u)\pi(v)dudv, \\
a_2 &= C_2^2 E\left[V_i^2(x_2)\right] = C_2^2 \iint [up(x_2) - q(x_2)][vp(x_2) - q(x_2)] R_{uv}(x_2)\pi(u)\pi(v)dudv, \\
a_3 &= 2C_1 C_2 E\left[V_i(x_1) V_i(x_2)\right] = 2C_1 C_2 \iint \left[up(x_1) - q(x_1)\right]\left[vp(x_2) - q(x_2)\right] S_{uv,0}(x_1, x_2)\pi(u)\pi(v)dudv.
\end{aligned}$$

$$\begin{aligned}
&\mathrm{Cov}(Y_i, Y_j) = E\left\{[C_1 V_i(x_1) + C_2 V_i(x_2)]\left[C_1 V_j(x_1) + C_2 V_j(x_2)\right]\right\} \\
&= C_1^2 E\left\{V_i(x_1) V_j(x_1)\right\} + C_1 C_2 E\left\{V_i(x_2) V_j(x_1)\right\} + C_1 C_2 E\left\{V_i(x_1) V_j(x_2)\right\} + C_2^2 E\left\{V_i(x_2) V_j(x_2)\right\} \\
&= b_1 + 2b_2 + b_3 = \gamma_y(s),
\end{aligned}$$

where $s = |i - j|$ and

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

In univariate case, we have verified that $E|V_i(x)|^\delta < \infty$ 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 $\delta > 2$.

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

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

$S = C_1 \frac{1}{\sqrt{n}} \sum_{i=1}^{n} V_i(x_1) + C_2 \frac{1}{\sqrt{n}} \sum_{i=1}^{n} V_i(x_2) \to_D \mathrm{N}(0, \Sigma_S)$. Therefore, as $n \to \infty$, we have

$$\left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} V_i(x_1), \frac{1}{\sqrt{n}} \sum_{i=1}^{n} V_i(x_2) \right] \to_D \mathrm{N}(0, \Sigma(x_1, x_2)), \tag{32}$$

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 $n \to \infty$,

$$\begin{aligned} &\left[\sqrt{n}(h_n(x_1) - h_0(x_1)), \sqrt{n}(h_n(x_2) - h_0(x_2)) \right] = Q_{n,12}^{-1} \left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} V_i(x_1), \frac{1}{\sqrt{n}} \sum_{i=1}^{n} V_i(x_2) \right] \\ \to_D &Q_{12}^{-1} \mathrm{N}(0, \Sigma(x_1, x_2)) = \mathrm{N}(0, \Sigma_{12}), \end{aligned} \tag{33}$$

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

More generally,

$$\left[\sqrt{n}(h_n(x_1) - h_0(x_1)), \cdots, \sqrt{n}(h_n(x_N) - h_0(x_N)) \right] \to_D N(0, \Sigma_N), \tag{34}$$

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

Proof of Theorem 3.

According to (4) and (7), 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 is strictly stationary, we have $f(x|h) = \frac{1}{h}EK\left(\frac{X-x}{h}\right) = \frac{1}{nh}\sum_{i=1}^n EK\left(\frac{X_i-x}{h}\right)$. Therefore,

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

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

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

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

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

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

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

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

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

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

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

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

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

Similarly, we have

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

Let n^* denote the sample size that used to estimate the hyperparameter θ_0. As $p_n(x_k^*)$ is continuous and twice differentiable, so the maximum likelihood estimate of θ_0, $\widehat{\theta}$ has a $\sqrt{n^*}$ rate of convergence and $||\widehat{\theta} - \theta_0|| = O_P(n^{*-1/2})$. As $n^* \to \infty$, $\widehat{p}_n(x) p_n(x) - p^2(x) = o_P(1)$. Based on this, we could obtain that

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

Thus, 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 $\sqrt{n}(h_n(x) - h_0(x)) \to_D N(0, \Sigma_0(x))$, and thus, we have

$$\sqrt{n}(\widehat{h}_n(x) - h_0(x)) = \sqrt{n}(\widehat{h}_n(x) - h_n(x)) + \sqrt{n}(h_n(x) - h_0(x))$$
$$= \frac{\sqrt{n}}{\sqrt{n^*}} \sqrt{n^*}(\widehat{h}_n(x) - h_n(x)) + \sqrt{n}(h_n(x) - h_0(x))$$
$$= \sqrt{\frac{n}{n^*}} O_P(1) + \sqrt{n}(h_n(x) - h_0(x)) \to_D N(0, \Sigma_0(x)),$$

on the condition that $n/n^* = o(1)$.

Remark. The above results hold true for both scenarios we considered in Section 4. In **Scenario 1**, we set $\theta_0' = c(1, 0.05)$ and then we obtain the maximum likelihood estimate $\widehat{\theta}$ which satisfy that $||\widehat{\theta} - \theta_0|| = O_P(n^{*-1/2})$. In **Scenario 2**, we have the counterpart $||\widehat{\theta}^* - \widehat{\theta}_0|| = O_P(n^{*-1/2})$, which ensures the validity of the above results. Therefore, we proofed Theorem 3.

Proof of Theorem 4.

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