ISSN 1440-771X

MONASH University

Australia

Department of Econometrics and Business Statistics

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

A sampling algorithm for bandwidth estimation in a nonparametric regression model with a flexible error density

Xibin Zhang, Maxwell L. King and Han Lin Shang

September 2013

Working Paper 20/13 (Revised 10/11)

A sampling algorithm for bandwidth estimation in a nonparametric regression model with a flexible error density

Xibin Zhang[1], Maxwell L. King

Department of Econometrics and Business Statistics, Monash University, Australia

Han Lin Shang

ESRC Centre for Population Change, University of Southampton, United Kingdom

March 5, 2013

Abstract: We propose to approximate the unknown error density of a nonparametric regression model by a mixture of Gaussian densities with means being the individual error realizations and variance a constant parameter. This mixture density has the form of a kernel density estimator of error realizations. We derive an approximate likelihood and posterior for bandwidth parameters in the kernel–form error density and the Nadaraya–Watson regression estimator and develop a sampling algorithm. A simulation study shows that when the true error density is non–Gaussian, the kernel–form error density is often favored against its parametric counterparts including the correct error density assumption. Our approach is demonstrated through a nonparametric regression model of the Australian All Ordinaries daily return on the overnight FTSE and S&P 500 returns. Using the estimated bandwidths, we derive the one–day–ahead density forecast of the All Ordinaries return, and a distribution–free value–at–risk is obtained. The proposed algorithm is also applied to a nonparametric regression model involved in state–price density estimation based on S&P 500 options data.

Key words: Bayes factors, kernel–form error density, Metropolis–Hastings algorithm, state–price density, value–at–risk.

[1] Corresponding author. Address: 900 Dandenong Road, Caulfield East, Victoria 3145, Australia. Telephone: +61-3-99032130. Fax: +61-3-99032007. Email: xibin.zhang@monash.edu.

1 Introduction

A simple and commonly used estimator of the regression function in a nonparametric regression model is the Nadaraya–Watson (NW) estimator, whose performance is mainly determined by the choice of bandwidths. A large literature exists on bandwidth selection for the NW estimator, and the most popular approaches are the rule–of–thumb, cross–validation (CV), plug–in and bootstrapping methods (see for example, Härdle, 1990; Herrmann, Engel, Wand, and Gasser, 1995; Hall, Lahiri, and Polzehl, 1995). Even though the NW estimator does not require an assumption on the analytical form of the error density, it is often of great interest to investigate the distribution of the response around the estimated mean. Such a distribution is characterized by the error density, estimation of which is a fundamental issue in statistical inference for any regression model. This issue was extensively discussed by Efromovich (2005), who developed a nonparametric approach to error–density estimation in a nonparametric regression model using residuals as proxies of errors.

A simple approach to error density estimation is the kernel density estimator of residuals, whose performance is mainly determined by the choice of bandwidth. This density estimator depends on residuals fitted through the NW estimator of the regression function. Moreover, the resulting density estimator of residuals provides no information for the purpose of choosing bandwidths in the NW regression estimator, although bandwidth selection in this situation depends on the error distribution (see for example, Zhang, Brooks, and King, 2009). Therefore, there is a lack of a data–driven procedure for choosing bandwidths for the two estimators simultaneously. This motivates the study reported in this paper.

Our investigation of error density estimation is also motivated by its practical applications. In financial econometrics, an important use of the estimated error density in modeling an asset return is to estimate the value–at–risk (VaR) for holding the asset. In such a model, any mis-specification of the error density may produce an inaccurate estimate of the VaR and make the asset holder unable to control risk. Therefore, being able to estimate the error density can be just as important as being able to estimate the mean of the regression model.

Let $\boldsymbol{y}$ denote the response and $\boldsymbol{x} = (x_1, x_2, \dots, x_d)'$ a set of explanatory variables or regressors. Given observations $(y_i, \boldsymbol{x}_i)$, for $i = 1, 2, \dots, n$, a nonparametric regression model is expressed as

$$y_i = m(\boldsymbol{x}_i) + \varepsilon_i, \tag{1}$$

where ε_i, for $i = 1, 2, \dots, n$, are assumed to be independent and identically distributed (iid) with an unknown density denoted as $f(\varepsilon)$. Let the NW estimator of the regression function be denoted as $\widehat{m}(\boldsymbol{x}; \boldsymbol{h})$ with $\boldsymbol{h}$ a vector of bandwidths. In this paper, we assume that the unknown $f(\varepsilon)$ is approximated by a kernel–form density given by

$$f(\varepsilon; b) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{b} \phi\left(\frac{\varepsilon - \varepsilon_i}{b}\right), \tag{2}$$

where $\phi(\cdot)$ is the probability density function of the standard Gaussian distribution.

The density function given by (2) is a mixture of n Gaussian densities, and the component densities have a common standard deviation b and means ε_i, for $i = 1, 2, \dots, n$. From the viewpoint of kernel smoothing, this error density is of the form of a kernel density estimator of the errors (rather than residuals) with $\phi(\cdot)$ the kernel function and b the bandwidth. Consequently, it is reasonable to expect that $f(\varepsilon; b)$ can approximate $f(\varepsilon)$ well when $f(\varepsilon)$ is unknown. We call (2) the kernel–form error density, and b is referred to as the bandwidth.

We aim to develop a sampling algorithm, through which the bandwidths, $\boldsymbol{h}$ and b, can be simultaneously estimated. We treat bandwidths as parameters and conduct our investigation in a parametric way although the underlying model is nonparametric. Our main contribution is to construct an approximate likelihood and therefore, the posterior of bandwidth parameters for the nonparametric regression model with its unknown error density approximated by the kernel–form error density given by (2).

When the iid errors follow a Gaussian distribution, Zhang et al. (2009) derived an approximate posterior of $\boldsymbol{h}$ for given $\boldsymbol{y} = (y_1, y_2, \dots, y_n)'$, where the likelihood of $\boldsymbol{y}$ for given $\boldsymbol{h}$ is the product of the Gaussian densities of y_i with its mean approximated by the leave–one–out NW estimator denoted as $\widehat{m}_i(\boldsymbol{x}_i; \boldsymbol{h})$, for $i = 1, 2, \dots, n$. The error density can be assumed to be of other parametric forms such as a mixture of Gaussian densities. However, any parametric assumption of the error density is likely to be wrong, and subsequent inference might be misleading. The contribution of this paper is not only a relaxation of the Gaussian error assumption of Zhang et al. (2009), but also a novel sampling algorithm under a flexible error density in regression models.

There is a growing literature on the estimation of the error density in a nonparametric regression model. Efromovich (2005) presented the so–called Efromovich–Pinsker estimator of the error density and showed that this estimator is asymptotically as accurate as an oracle that knows the underlying errors. Cheng (2004) showed that the kernel density estimator of residuals is uniformly, weakly and strongly consistent. When the regression function is estimated by the NW estimator and the error density is estimated by the kernel estimator of residuals, Samb (2011) proved the asymptotic normality of the bandwidths in both estimators and derived the optimal convergence rates of the two types of bandwidths. Linton and Xiao (2007) proposed a kernel estimator based on residuals obtained through local polynomial fitting of the unknown regression function. They showed that their estimator is adaptive and concluded that adaptive estimation is possible in local polynomial fitting, which includes the NW estimator as a special case. In a class of nonlinear regression models, Yuan and de Gooijer (2007) constructed an approximate likelihood through the kernel density estimator of pre-fitted residuals with its bandwidth pre-chosen by the rule–of–thumb. They proved that under some regularity conditions, the resulting maximum likelihood estimates of parameters are consistent, asymptotically normal and efficient. Jaki and West (2008) proposed using the kernel density estimator of the pre-fitted residuals to construct an approximate likelihood, which they called the kernel likelihood.

In all these investigations, residuals were commonly used as proxies of errors, and the bandwidth for the kernel density estimator of residuals was pre–chosen. To our knowledge, there is no method that can simultaneously estimate the bandwidths for the NW estimator of the regression function and the kernel–form error density.

Our proposed kernel–form error density is robust in terms of different specifications of the error density in a nonparametric regression model. In order to understand the relative gains and losses that result from this robust assumption against other parametric assumptions, we conduct simulation studies by simulating samples through a nonlinear regression function, where the error densities are respectively, the Gaussian and several mixture densities of two Gaussians. We find that the proposed sampling approach to bandwidth estimation outperforms its traditional counterparts, the rule–of–thumb, plug–in and bootstrapping methods. Moreover, within the sampling framework, when the true error density is non–Gaussian, the kernel–form error density is often favored against its parametric competitors, including the correct error–density assumption.

We apply the proposed sampling algorithm to the estimation of the bandwidths for the nonparametric regression of the Australian All Ordinaries (Aord) daily return on the overnight FTSE and S&P 500 returns with its error density being the kernel–form. With the estimated bandwidths and overnight FTSE and S&P 500 returns, we derive the one–day–ahead density forecast of the Aord return to use it to compute a distribution–free VaR. Moreover, the kernel–form error density is favored with very strong evidence against the Gaussian and a location–scale mixture density of two Gaussians according to Bayes factors. Our second application is the one discussed by Zhang et al. (2009) who estimated the bandwidths for a nonparametric regression model so as to estimate the state–price density (SPD) of the S&P 500 index at the maturity of its call option. In this application, we assume that the unknown error density is approximated by the kernel–form density and find that this robust error density is favored with very strong evidence against the Gaussian error density.

The rest of this paper is organized as follows. In Section 2, we derive an approximate posterior of bandwidth parameters in the NW estimator and kernel-form error density. Section 3 presents simulations to evaluate the performance of Bayesian estimation of bandwidths under the Gaussian and several mixture densities of two Gaussians. In Section 4, we present an empirical investigation of the nonparametric relationship between stock index returns across three stock markets. Our proposed method is also validated through a nonparametric regression model involved in the SPD estimation. Section 6 concludes the paper.

2 Bayesian estimation of bandwidths

The bandwidths in the NW estimator of the regression function and the kernel-form error density estimator play an important role in controlling the smoothness of the regression function and the error density estimator. In this paper, we treat these bandwidths as parameters. This is not new in the context of kernel density estimation based on direct observations (see for example, Brewer, 2000; Gangopadhyay and Cheung, 2002; de Lima and Atuncar, 2011). In nonparametric and semiparametric regression models, bandwidths are also treated as parameters (see Härdle, Hall, and Ichimura, 1993; Rothe, 2009, among others). Given observations $(y_i, \boldsymbol{x}_i)$, for $i = 1, 2, \ldots, n$, we aim to construct an approximate likelihood, as well as an approximate posterior of the parameters.

2.1 Parametric assumptions about the error density

Before introducing a flexible error density into the nonparametric regression model given by (1), we briefly describe how an approximate likelihood function could be constructed under parametric assumptions of the error density. Zhang et al. (2009) considered the nonparametric regression model given by (1), where ε_i, for $i = 1, 2, \ldots, n$, are iid and follow $N(0, \sigma^2)$ with σ^2 an unknown parameter. The model implies that

$$\frac{y_i - m(\boldsymbol{x}_i)}{\sigma} \sim N(0,1).$$

As the analytical form of $m(\boldsymbol{x}_i)$ is unknown, it is estimated by the leave–one–out NW estimator,

$$\widehat{m}_{(-i)}(\boldsymbol{x}_i; \boldsymbol{h}) = \frac{(n-1)^{-1} \sum_{j=1; j\neq i}^{n} K_{\boldsymbol{h}}(\boldsymbol{x}_i - \boldsymbol{x}_j) y_j}{(n-1)^{-1} \sum_{j=1; j\neq i}^{n} K_{\boldsymbol{h}}(\boldsymbol{x}_i - \boldsymbol{x}_j)}, \tag{3}$$

where $K_{\boldsymbol{h}}(\boldsymbol{z}) = (h_1 h_2 \cdots h_d)^{-1} K(\boldsymbol{z}./\boldsymbol{h})$ with $K(\cdot)$ being a kernel function and "./" division by elements. Let $\boldsymbol{h}^2 = \left(h_1^2, h_2^2, \ldots, h_d^2\right)'$. Treating σ^2 and the elements of $\boldsymbol{h}^2$ as parameters, one can derive an approximate likelihood of $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$ as

$$L_G\left(\boldsymbol{y} \middle| \boldsymbol{h}^2, \sigma^2\right) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left[y_i - \widehat{m}_i(\boldsymbol{x}_i, \boldsymbol{h})\right]^2\right). \tag{4}$$

Zhang et al. (2009) derived the posterior of $\boldsymbol{h}^2$ and σ^2, which is proportional to the product of (4) and pre-chosen priors of $\boldsymbol{h}^2$ and σ^2. A posterior simulation algorithm was also presented for estimating $\boldsymbol{h}^2$ and σ^2.

One may also assume a location–scale mixture of two Gaussian densities for the iid errors in (1) to reflect a wider range of error distributions. Such a mixture density is given by

$$\phi_M(\varepsilon) = \frac{w}{\sigma_1}\phi\left(\frac{\varepsilon-\mu_1}{\sigma_1}\right) + \frac{(1-w)}{\sigma_2}\phi\left(\frac{\varepsilon-\mu_2}{\sigma_2}\right), \tag{5}$$

where $w \in (0,1)$ is a weight parameter, μ_1 and μ_2 are location parameters, and σ_1 and σ_2 are scale parameters. As this mixture density is assumed to be the error density, a restriction of the mean $w\mu_1 + (1-w)\mu_2 = 0$ leads to $\mu_2 = -w\mu_1/(1-w)$. Thus, there are four parameters to be estimated, namely $(w, \sigma_1, \sigma_2, \mu_1)'$.

The location–scale mixture density given by (5) can be simplified to either a scale mixture of two Gaussians by restricting $\mu_1 = 0$, or a location mixture of two Gaussians by restricting $\sigma_1 = \sigma_2$. Consequently, the number of parameters under each simplified mixture is less than that under the location–scale mixture.

A limitation of these parametric assumptions about the error density is that any wrong assumption may lead to a poor choice of bandwidths, leading to a more inaccurate estimate of the regression function. In what follows, we will investigate a very flexible specification of the error density, namely the kernel–form error density.

2.2 A kernel–form error density

A standard distributional assumption of regression errors has the benefit of simplicity in obtaining theoretical results, but may suffer from the problem of being a poor fit to the data. Consequently, one may wish to sacrifice some analytical convenience so as to improve the fit to the data through a more flexible distribution. There have been some advances along these lines. For example, in a smoothly mixing regression model, Geweke and Keane (2007) used a finite Gaussian mixture to derive the likelihood. In a linear regression model, Leslie, Kohn, and Nott (2007) proposed to model the error distribution through a Dirichlet process mixture prior. Griffin, Quintana, and Steel (2011) provided a survey of some of the recent work in this area.

In the nonparametric regression model given by (1), we assume that the iid errors follow an unknown distribution with its density approximated by the kernel–form density given by (2). In the current literature, when the error density of a regression model is non–Gaussian, a location–scale mixture density of several Gaussian components is often used, where the component Gaussian densities have different means and variances. However, the use of such a mixture density is at the cost of increasing the number of parameters. In contrast, the kernel–form error density has only one bandwidth parameter to be estimated, and this is one of its advantages against the location–scale mixture density. Moreover, due to its form of the kernel density estimator of errors, this kernel–form error density will be very close to the true error density when the sample size is large. Zhang and King (2011) demonstrated the validity of this kernel–form density as a density of iid errors in a family of univariate GARCH models. Their derived approximate likelihood is well defined, and subsequent posterior simulation is meaningful.

We investigate the construction of an approximate likelihood and posterior for (1) with its unknown error density approximated by the kernel–form error density given by (2). If $m(\boldsymbol{x})$ is known, this kernel–form density is a well–defined density function of the iid errors. Therefore, we can derive the density of the response variable as

$$y_i \sim f\left(\{y_i - m(\boldsymbol{x}_i)\}; b\right) = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{b}\,\phi\left(\frac{\{y_i - m(\boldsymbol{x}_i)\} - \{y_j - m(\boldsymbol{x}_j)\}}{b}\right), \tag{6}$$

for $i = 1, 2, \ldots, n$.

The regression function in (1) is typically unknown, but can be estimated by the NW estimator for the purpose of constructing the likelihood. The realized errors or residuals are used as proxies of errors. We propose to plug–in the leave–one–out NW estimator of $m(\boldsymbol{x})$ given by (3) into (6). Therefore, the density of y_i is approximated by $\widetilde{f}\left(\{y_i - \widehat{m}_{(-i)}(\boldsymbol{x}_i; \boldsymbol{h})\}; b\right)$, which is expressed as

$$\widetilde{f}\left(\{y_i - \widehat{m}_{(-i)}(\boldsymbol{x}_i;\boldsymbol{h})\};b\right) = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{b}\,\phi\left(\frac{\{y_i - \widehat{m}_{(-i)}(\boldsymbol{x}_i;\boldsymbol{h})\} - \{y_j - \widehat{m}_{(-j)}(\boldsymbol{x}_j;\boldsymbol{h})\}}{b}\right). \qquad (7)$$

The use of $\widehat{m}_{(-i)}(\boldsymbol{x}_i;\boldsymbol{h})$ as an approximation to the mean of y_i was also proposed by Zhang et al. (2009) in the nonparametric regression model with Gaussian errors. They constructed the likelihood through the Gaussian density of y_i with its mean approximated by $\widehat{m}_{(-i)}(\boldsymbol{x}_i;\boldsymbol{h})$.

2.2.1 An approximate likelihood

The likelihood of $\boldsymbol{y}$ for given $\boldsymbol{h}$ and b, is approximated by the product of the density of y_i given by (7), for $i = 1,2,\ldots,n$. However, it is impossible to estimate b by maximizing the resulting approximate likelihood, because it contains at least one unwanted term $\phi(0)/b$. The resulting approximate likelihood would approach infinity as b tends to zero. Exclusion of the ith term only from the summation of (7) is not enough, because if $y_j - \widehat{m}_{(-j)}(\boldsymbol{x}_j;\boldsymbol{h}) = y_i - \widehat{m}_{(-i)}(\boldsymbol{x}_i;\boldsymbol{h})$, for $j \neq i$, the jth term in the summation becomes $\phi(0)/b$. Nonetheless, a remedy to this problem is to exclude the jth term that makes $\{y_j - \widehat{m}_{(-j)}(\boldsymbol{x}_j;\boldsymbol{h})\} = \{y_i - \widehat{m}_{(-i)}(\boldsymbol{x}_i;\boldsymbol{h})\}$, from the summation given by (7). Let

$$J_i = \left\{j : y_j - \widehat{m}_{(-j)}(\boldsymbol{x}_j;\boldsymbol{h}) \neq y_i - \widehat{m}_{(-i)}(\boldsymbol{x}_i;\boldsymbol{h}),\ \text{for } j = 1,2,\ldots,n\right\},$$

for $i = 1,2,\ldots,n$, and let n_i denote the number of terms excluded from the summation in (7). The density of y_i is therefore, approximated as

$$\widetilde{f}\left(\{y_i - \widehat{m}_{(-i)}(\boldsymbol{x}_i;\boldsymbol{h})\};b\right) = \frac{1}{n-n_i}\sum_{j\in J_i}\frac{1}{b}\,\phi\left(\frac{\{y_i - \widehat{m}_{(-i)}(\boldsymbol{x}_i;\boldsymbol{h})\} - \{y_j - \widehat{m}_{(-j)}(\boldsymbol{x}_j;\boldsymbol{h})\}}{b}\right), \qquad (8)$$

where the subscript of $\widehat{m}$ means that this is the leave–one–out estimate. Let $\widehat{\varepsilon}_i$ denote $y_i - \widehat{m}_{(-i)}(\boldsymbol{x}_i;\boldsymbol{h})$, for $i = 1,2,\ldots,n$.

Given $\boldsymbol{h}^2$ and b^2, the likelihood of $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$ is approximated by

$$L_K\left(\boldsymbol{y}|\boldsymbol{h}^2, b^2\right) = \prod_{i=1}^{n}\left\{\frac{1}{n-n_i}\sum_{j\in J_i}\frac{1}{b}\,\phi\left(\frac{\{y_i - \widehat{m}_{(-i)}(\boldsymbol{x}_i;\boldsymbol{h})\} - \{y_j - \widehat{m}_{(-j)}(\boldsymbol{x}_j;\boldsymbol{h})\}}{b}\right)\right\}. \qquad (9)$$

2.2.2 Priors

We now discuss the issue of prior choices for the bandwidths. Let $\pi\left(h_k^2\right)$ denote the prior of h_k^2, for $k=0,1,\ldots,d$. As b^2 and h_k^2, for $k=1,2,\ldots,d$, which are respectively, the squared bandwidths for the kernel–form error density and the NW estimator, play the same role as a variance parameter, we assume that the priors of b^2 and h_k^2 are the inverse Gamma density denoted as $\mathrm{IG}(\alpha_b,\beta_b)$ and $\mathrm{IG}(\alpha_h,\beta_h)$, respectively. Therefore, the priors of b^2 and h_k^2 are expressed explicitly as

$$\pi\left(b^2\right)=\frac{(\beta_b)^{\alpha_b}}{\Gamma(\alpha_b)}\left(\frac{1}{b^2}\right)^{\alpha_b+1}\exp\left\{-\frac{\beta_b}{b^2}\right\}, \tag{10}$$

$$\pi\left(h_k^2\right)=\frac{(\beta_h)^{\alpha_h}}{\Gamma(\alpha_h)}\left(\frac{1}{h_k^2}\right)^{\alpha_h+1}\exp\left\{-\frac{\beta_h}{h_k^2}\right\},\qquad \text{for } k=1,2,\ldots,d, \tag{11}$$

where α_b, β_b, α_h and β_h are hyperparameters.

2.2.3 An approximate posterior

The joint posterior of $\boldsymbol{h}^2$ and b^2 is approximately expressed as (up to a normalizing constant)

$$\pi\left(\boldsymbol{h}^2,b^2|\boldsymbol{y}\right)\propto L_K\left(\boldsymbol{y}|\boldsymbol{h}^2,b^2\right)\pi(b^2)\prod_{k=1}^{d}\pi\left(h_k^2\right). \tag{12}$$

Note that unlike the joint approximate posterior derived by Zhang et al. (2009), the posterior given by (12) does not suggest a closed form for either a marginal posterior or a conditional posterior. Therefore, we use an adaptive version of the random–walk Metropolis algorithm to sample the elements of $\boldsymbol{h}^2$ and b^2 from (12) (see Garthwaite, Fan, and Sisson, 2011).

At the jth iteration, the current value of $\boldsymbol{h}^2$ denoted as $\boldsymbol{h}^2_{(j)}$ is updated by $\boldsymbol{h}^2_{(j+1)}=\boldsymbol{h}^2_{(j)}+\tau_j\boldsymbol{u}/||\boldsymbol{u}||$, where $\boldsymbol{u}$ is simulated from a proposal density that is the multivariate standard Gaussian, and τ_j is an adaptive tuning coefficient computed according to Garthwaite et al. (2011). $\boldsymbol{h}^2_{(j+1)}$ is accepted or rejected according to the Metropolis–Hastings rule. At this iteration, the current value of b^2 is also updated through the random–walk Metropolis algorithm, where the proposal density is the univariate standard Gaussian. The optimal target value of the acceptance probability is 0.234 for multivariate updating and 0.44 for univariate updating (see for example, Roberts and Rosenthal, 2009; Garthwaite et al., 2011).

Upon completing the sampling procedure, we use the ergodic average of the sampled chain of $(b, \boldsymbol{h}')'$ as the estimate of $(b, \boldsymbol{h}')'$. Therefore, the analytical form of the kernel–form error density can be derived based on the estimated b and $\boldsymbol{h}$.

3 Monte Carlo simulation

The purpose of this simulation study is three–fold. First, with one simulated sample, we illustrate the use and effectiveness of our Bayesian sampling algorithm for estimating the bandwidths in the NW regression estimator and the kernel–form error density. As the true error density is unknown in practice, the proposed method is expected to be more flexible than its parametric counterparts.

Second, we generate 1,000 samples from the same nonparametric regression model with its error densities assumed to be respectively, Gaussian, a scale mixture of two Gaussians, a location mixture of two Gaussians, and a location–scale mixture of two Gaussians. We examine the performance of the proposed Bayesian sampling in estimating the bandwidths of the NW estimator, in comparison with the performance of some existing bandwidth selection methods, such as the rule–of–thumb (ROT) discussed in Scott (1992), CV and bootstrapping. Moreover, we examine the performance of the bandwidth in the kernel–form error density estimated through Bayesian sampling, in comparison to the performance of the bandwidth selected through CV based on residuals.

Finally, we compare the results derived through Bayesian sampling under the assumptions of kernel–form error density, the Gaussian and a mixture of two Gaussians, where Bayes factors are used for comparison purposes. We briefly describe Bayes factors below.

3.1 Bayes factors

In Bayesian inference, model selection is often conducted through the Bayes factor of the model of interest against a competing model. The Bayes factor reflects a summary of evidence provided by the data supporting the model as opposed to its competing model. The Bayes factor is defined as the ratio of the marginal likelihoods derived under the model of interest and its competing model, respectively. The marginal likelihood is the expectation of the likelihood with respect to the prior of parameters. It is seldom calculated as the integral of the product of the likelihood and prior of parameters, but instead, is often computed numerically (Gelfand and Dey, 1994; Newton and Raftery, 1994; Chib, 1995; Kass and Raftery, 1995; Geweke, 1999, among others). In this paper, we employed the method proposed by Chib (1995) to compute the marginal likelihood.

Let θ denote the parameter vector and $\boldsymbol{y}$ the data. Chib (1995) showed that the marginal likelihood under a model $\mathscr{A}$ is expressed as

$$P_{\mathscr{A}}(\boldsymbol{y}) = \frac{\ell_{\mathscr{A}}(\boldsymbol{y}|\theta)\pi_{\mathscr{A}}(\theta)}{\pi_{\mathscr{A}}(\theta|\boldsymbol{y})}, \tag{13}$$

where $\ell_{\mathscr{A}}(\boldsymbol{y}|\theta)$, $\pi_{\mathscr{A}}(\theta)$ and $\pi_{\mathscr{A}}(\theta|\boldsymbol{y})$ denote respectively, the likelihood, prior and posterior under model $\mathscr{A}$. $P_{\mathscr{A}}(\boldsymbol{y})$ is usually computed at the posterior estimate of θ. The numerator has a closed form and can be computed analytically. The denominator is the posterior of θ, which is often replaced by its kernel density estimator based on the simulated chain of θ through a posterior sampler. The Bayes factor of model $\mathscr{A}$ against model $\mathscr{B}$ is defined as

$$\text{BF} = \frac{P_{\mathscr{A}}(\boldsymbol{y})}{P_{\mathscr{B}}(\boldsymbol{y})},$$

which is used to make a decision on whether $\mathscr{A}$ is favored against $\mathscr{B}$, according to the Jeffreys (1961) scales modified by Kass and Raftery (1995). A Bayes factor value between 1 and 3 indicates that the evidence supporting $\mathscr{A}$ against $\mathscr{B}$ is not worth more than a bare mention. When the Bayes factor is between 3 and 20, $\mathscr{A}$ is favored against $\mathscr{B}$ with positive evidence; when the Bayes factor is between 20 and 150, $\mathscr{A}$ is favored against $\mathscr{B}$ with strong evidence; and when the Bayes factor is above 150, $\mathscr{A}$ is favored against $\mathscr{B}$ with very strong evidence.

3.2 Performance of the proposed bandwidth estimation methods

Consider the relationship between $\boldsymbol{y}$ and $\boldsymbol{x} = (x_1, x_2, x_3)'$ given by

$$y_i = \sin(2\pi x_{1,i}) + 4(1 - x_{2,i})(1 + x_{2,i}) + \frac{2x_{3,i}}{1 + 0.8x_{3,i}^2} + \varepsilon_i, \tag{14}$$

for $i = 1, 2, \ldots, n$. A sample of 1,000 observations was generated by drawing $x_{1,i}$, $x_{2,i}$ and $x_{3,i}$ independently from the uniform density on $(0,1)$ and ε_i from the mixture of two Gaussian densities defined as $0.7N(1, 0.7^2) + 0.3N(-7/3, 1.5^2)$, and calculating y_i according to (14).

The relationship between y_i and $(x_{1,i}, x_{2,i}, x_{3,i})'$ was modeled by the nonparametric regression model given as

$$y_i = m(x_{1,i}, x_{2,i}, x_{3,i}) + \varepsilon_i, \tag{15}$$

where $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are assumed to be iid.

Assuming that the error density of (15) is unknown and is approximated by the kernel–form density (2), we applied our sampling algorithm to (15) using the generated sample. The hyperparameters were chosen as $\alpha_h = \alpha_b = 1$ and $\beta_h = \beta_b = 0.05$. These values are often used as hyperparameter values of an inverse Gamma density when it is chosen as the prior of a variance parameter (see for example, Geweke, 2009). The adaptive random–walk Metropolis algorithm was employed, where the proposal density is the multivariate Gaussian with an identity variance–covariance matrix for updating $\boldsymbol{h}^2$, but each vector simulated from the proposal density was scaled by its length. The proposal density for updating b^2 is the standard Gaussian. The burn–in period contains the first 1,000 draws, and the following 10,000 draws were recorded. The acceptance rate was controlled to be around 0.234 for multivariate draws and 0.44 for univariate draws through the adaptive random–walk Metropolis algorithm of Garthwaite et al. (2011). The posterior means of the bandwidths are presented in Table 1.

The mixing performance of this posterior sampler is examined by the simulation inefficiency factor (SIF), which can be loosely interpreted as the number of draws needed so as to obtain independent draws from the simulated Markov chain. For example, a SIF value of 20 indicates that approximately, we would need to keep one draw for every 20 draws so as to derive independent draws (see for example, Roberts, 1996; Kim, Shephard, and Chib, 1998; Tse, Zhang, and Yu, 2004; Nott and Kohn, 2005).

The standard deviation of the posterior mean is approximated by the batch–mean standard deviation. It becomes smaller and smaller when the number of simulation iterations

Table 1: *Parameter estimates and their statistics for Bayesian bandwidth estimation under the kernel–form, Gaussian and location–scale mixture error densities for a simulated sample with sample size* $n = 1,000$. *LML refers to log marginal likelihood.*

Error density	Parameter	Estimate	95% Bayesian credible interval	Standard deviation	Batch–mean standard dev	SIF
Kernel–form	b	0.2514	(0.1613, 0.3520)	0.1582	0.0061	14.9
	h_1	0.2132	(0.1790, 0.2645)	0.0539	0.0026	24.0
	h_2	0.1984	(0.1711, 0.2296)	0.0411	0.0015	13.2
	h_3	0.4982	(0.4252, 0.5749)	0.1072	0.0033	9.6
	LML	-1832.49				
Gaussian	σ	1.9217	(1.8405, 2.0037)	0.0221	0.0002	1.2
	h_1	0.1994	(0.1472, 0.2806)	0.0266	0.0008	9.1
	h_2	0.2052	(0.1618, 0.2685)	0.0217	0.0007	10.1
	h_3	0.4409	(0.3027, 0.6285)	0.0767	0.0032	17.1
	LML	-2079.48				
Location–scale mixture	w	0.6804	(0.6381, 0.7191)	0.0226	0.0008	13.9
	h_1	0.1068	(0.0874, 0.1288)	0.0282	0.0009	10.5
	h_2	0.1003	(0.0823, 0.1202)	0.0247	0.0008	9.7
	h_3	0.2489	(0.2058, 0.3017)	0.0642	0.0021	10.3
	σ_1	0.7313	(0.6764, 0.7851)	0.0290	0.0009	8.6
	σ_2	1.6889	(1.4677, 1.9279)	0.1173	0.0038	10.4
	μ_1	1.0562	(0.9873, 1.1206)	0.0356	0.0011	10.4
	LML	-1848.89				

increases, if the sampler achieves a reasonable mixing performance. The SIF and batch–mean standard deviation were used to monitor the mixing performance. Table 1 presents the values of these two indicators, which show that the sampler has mixed very well.

For comparison purposes, we also report the estimates of bandwidths and parameter(s) derived through Zhang et al.'s (2009) sampling algorithm with the error density assumed to be respectively, the Gaussian and a location–scale mixture of two Gaussians given by (5). Under both assumptions of the error density, the priors of bandwidths are the same as those under the kernel–form error density. Under Gaussian errors, the prior of σ^2 is the inverse Gamma density with hyperparameters $\alpha_\sigma = 1$ and $\beta_\sigma = 0.05$. The same prior was chosen for σ_1^2 and σ_2^2 under the assumption of location–scale mixture density, where in addition, the prior of w is the uniform density on $(0,1)$, and the prior of μ_1 is $N(0,9)$. The results are also given in Table 1.

We calculated the marginal likelihood at the estimates of parameters in each situation and found the following evidence. First, the marginal likelihood obtained under the assumption of kernel–form error density is larger than that obtained under either the Gaussian or the location–scale mixture density. The Bayes factors of the kernel–form density are respectively, exp(246.99) against the Gaussian and exp(16.4) against the location–scale mixture. Therefore, the kernel–form error density is favored against its parametric counterparts with very strong evidence. Second, we obtained an estimate of the bandwidth for the kernel–form error density estimator, whose graph is plotted in Figure 1, along with the graphs of the true error density, the Gaussian and location–scale mixture with plugged–in parameter estimates. Among these three estimated error density functions, the one derived under the kernel–form error density is the closest to the true density.

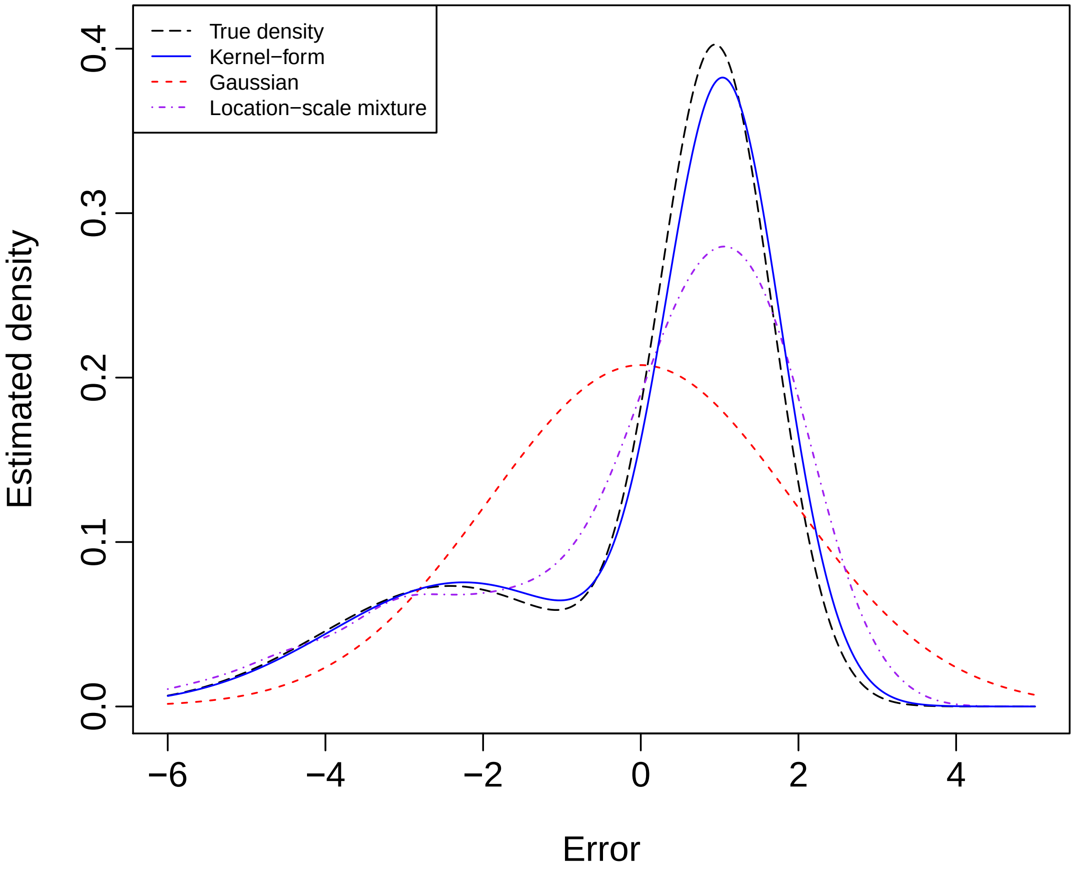

Figure 1: *Graphs of the true and estimated densities of regression errors.*

For each of the three error–density assumptions in the nonparametric regression model given by (15), we computed the marginal likelihood given by (13) and the average squared errors (ASE) defined as

$$\text{ASE} = \frac{1}{n}\sum_{i=1}^{n}\left[\widehat{m}(\boldsymbol{x}_i, \boldsymbol{h}_n) - m(\boldsymbol{x}_i)\right]^2.$$

3.3 Accuracy of the estimated bandwidths

We examine the accuracy of the estimated bandwidths for the NW estimator and the kernel–form error density through the proposed Bayesian sampling algorithm. Therefore, we generated 1,000 samples through the model given by (14), where the error density was assumed to be respectively, Gaussian, a scale mixture of two Gaussians, a location mixture of two Gaussians and a location–scale mixture of two Gaussians. As both $m(\boldsymbol{x})$ and the error density are known, we are able to examine the performance of the proposed Bayesian sampling procedure for estimating bandwidths.

3.3.1 Accuracy of the estimated bandwidths for the NW estimator

The accuracy of the estimated/selected bandwidths for the NW estimator is measured by the ASE of the resulting NW estimator of the regression function. In kernel density estimation of directly observed data, the rule–of–thumb (ROT) is often used for bandwidth selection (see for example, Silverman, 1986; Scott, 1992; Bowman and Azzalini, 1997). Härdle and Müller (2000) indicated that methods for bandwidth selection in nonparametric regression are the same as those for kernel density estimation. Therefore, we considered the ROT as a bandwidth selection method for comparison purposes.

The CV for bandwidth selection has been extensively discussed (Wahba and Wold, 1975; Härdle and Marron, 1985; Härdle and Müller, 2000, among others). In addition, the bootstrapping approach to bandwidth selection in nonparametric regression was presented by Hall et al. (1995), where two pilot bandwidths have to be specified before bootstrapping begins. The purpose of the first bandwidth is to generate a bootstrapping sample, while the second aims to obtain an initial estimate of the regression function. In our simulation study, the two pilot bandwidths were chosen using the ROT and CV.

We generated 1,000 samples through the model given by (14) under each of the four error densities. The error densities we used are given in Table 2. For each sample, we estimated bandwidths through the proposed Bayesian sampling with the error density assumed to be respectively, the correct one from which we simulated random errors, and the kernel–

form density. We also chose bandwidths for the NW estimator through the ROT, CV and bootstrapping. The ASE of the NW estimator with its bandwidth vector derived through each of the aforementioned methods was calculated.[2]

Table 2: *Choices of the error density.*

Gaussian density	$N(0,0.9^2)$
Scale mixture of two Gaussians	$0.7N(0,0.7^2)+0.3N(0,1.5^2)$
Location mixture of two Gaussians	$0.7N(-1,1)+0.3N(7/3,1)$
Location-scale mixture of two Gaussians	$0.7N(1,0.7^2)+0.3N(-7/3,1.5^2)$

We calculated the mean and standard deviation (sd) of the 1,000 ASE values obtained under each error density and through each bandwidth estimation/selection method. These results are presented in Table 3. No matter which error density was used to generate samples, the mean ASE derived through Bayesian sampling with any error–density assumption is clearly smaller than that derived through either the ROT, CV or bootstrapping. The standard deviations of the corresponding ASE values are comparable among different methods, except that the bootstrapping results vary for different error densities.

Table 3: *Mean and standard deviation (sd) of 1,000 ASE derived through each bandwidth estimation method based on 1,000 generated samples with errors simulated from four densities.*

Source of simulated errors		ROT	CV	Bootstrap	Bayesian		
					Correct	Kernel–form	Inefficiency (%)
$N(0,0.9^2)$	Mean	0.0721	0.0701	0.0900	0.0582	0.0585	100.52
	sd	0.0089	0.0107	0.0161	0.0090	0.0091	101.11
$0.7N(0,0.7^2)+0.3N(0,1.5^2)$	Mean	0.0794	0.0754	0.1136	0.0653	0.0656	100.46
	sd	0.0103	0.0120	0.0244	0.0104	0.0105	100.96
$0.7N(-1,1)+0.3N(7/3,1)$	Mean	0.1534	0.1452	0.4223	0.1159	0.1176	101.47
	sd	0.0237	0.0247	0.1364	0.0225	0.0228	101.33
$0.7N(1,0.7^2)+0.3N\left(-7/3,1.5^2\right)$	Mean	0.1744	0.1647	0.4307	0.1152	0.1174	101.91
	sd	0.0238	0.0249	0.1501	0.0222	0.0227	102.25

In terms of Bayesian sampling for bandwidth estimation, the kernel–form error density leads to a slightly inaccurate NW estimator in comparison to the correct error density, which is unknown in practice. Regardless of the type of error density, Bayesian sampling with the kernel–form error density performs slightly worse than its Bayesian counterpart with the

[2] As with Bayesian sampling, the bandwidth for the kernel–form error density, together with the parameter(s) in each parametric assumption of the error density, were also estimated, but they were not used for the purpose of calculating the ASE because only the mean function was used.

correct error–density assumption. Under each of the four error densities, the inefficiency factor of the kernel–form density in comparison to the correct one, which we define as the ratio of the mean ASE of the former over the mean ASE of the latter, is reported in the last column of Table 3. In terms of the mean of ASE, the kernel–form error–density assumption is slightly inefficient than the correct assumption of the error density by a value that is between 0.46% and 1.91%. Moreover, in terms of the standard deviation of ASE, the former assumption is inefficient than the latter by a value that is between 0.96% and 2.25%.

3.3.2 Accuracy of the estimated bandwidth for the kernel–form error density

The performance of the estimated bandwidth for the kernel–form error density was examined through the integrated squared errors (ISE) defined as

$$\int_{-\infty}^{\infty} \left[\widehat{f}(\varepsilon;\widehat{b}) - f(\varepsilon)\right]^2 d\varepsilon,$$

where $\widehat{f}(\varepsilon;\widehat{b})$ is the kernel–form error density with bandwidth $\widehat{b}$ estimated through Bayesian sampling. In this simulation study, the ISE was approximated through a large number of grid points on a wide interval that covers the vast majority of the density function.

Our proposed Bayesian sampling method for bandwidth estimation in the kernel–form error density is competing with the likelihood CV for bandwidth selection in the kernel density estimator of the residuals. To compute residuals, CV was used to select bandwidths for the NW estimator because CV is the best performer among the bandwidth selection methods that are alternatives to Bayesian sampling. Thus, this competing approach involves two stages of using the CV method, in which the first stage uses the CV to select bandwidths for the NW estimator, and the second stage uses the likelihood CV to select the bandwidth for the kernel density estimator of residuals. Therefore, we call it the two–stage CV.

Under each of the four error densities, we calculated the approximate ISE of the error density estimator with its bandwidths estimated through Bayesian sampling or the two–stage CV for each simulated sample. Averaging over all 1,000 samples, we derived the mean and standard deviation of ISE for each method under each error density. As shown in Table 4, Bayesian sampling with the errors being assumed to follow either the kernel–form density or the correct density, performs clearly better than the two–stage CV in terms of both mean and standard deviation of ISE.

Table 4: 100× *mean and standard deviation (sd) of 1,000 ISE derived through Bayesian bandwidth estimation with different error–density assumptions, based on 1,000 generated samples with errors simulated from four specified densities.*

Source of simulated errors		Bayesian		Two–stage	Inefficiency (%)	
		Correct	Kernel		Kernel–form	Two–stage
$N(0,0.9^2)$	Mean	0.0430	0.1067	0.1995	248	464
$0.7N(0,0.7^2)+0.3N(0,1.5^2)$	Mean	0.0920	0.1405	0.3821	153	415
	sd	0.0905	0.0959	0.1816	106	201
$0.7N(-1,1)+0.3N(7/3,1)$	Mean	0.0967	0.1695	0.8641	175	894
	sd	0.0660	0.0863	0.3464	131	525
$0.7N(1,0.7^2)+0.3N\left(-7/3,1.5^2\right)$	Mean	0.2237	0.4659	2.8637	208	1280
	sd	0.1313	0.1972	1.2727	150	969

It is not surprising that the kernel–form error density leads to a slightly worse performance than the correct error density in Bayesian sampling. The inefficiency factor of Bayesian sampling with a kernel–form error density against its Bayesian competitor is smaller than that of the two–stage CV method against the same competitor. As the correct error–density assumption is impossible in practice, Bayesian sampling with the kernel–form error density is more appropriate than the two–stage CV for bandwidth estimation/selection for estimating error density.

3.4 Bayesian comparison among error–density assumptions

The benefit of the kernel–form error density assumption is to gain robustness in terms of error–density specifications, because it has the capacity and flexibility to approximate an unknown error density. In the nonparametric regression model given by (1), the kernel–form error density does not outperform its parametric competitors under correct error–density assumptions. However, this kernel–form error density usually outperforms its parametric counterparts when the underlying assumptions of the error density are not met. We conducted a simulation study using the same 1,000 samples generated in Section 3.3 to illustrate this. Under each of the four error densities from which the errors are simulated, we calculate

Bayes factors of one error–density assumption against its competitors for each simulated sample. For all 1,000 simulated samples, the derived Bayes factors are grouped into different categories according to the Jeffreys (1961) scales modified by Kass and Raftery (1995). The relative frequencies of simulated samples falling in these categories are reported, respectively.

3.4.1 Gaussian distribution for simulating errors

The relative frequencies of simulated samples with $N(0,0.9^2)$ errors falling in each category of Bayes factors of the Gaussian error assumption against respectively, the kernel–form error density and the mixture error density, are presented in Table 5.

Table 5: *Relative frequencies of simulated samples falling in different categories of Bayes factors when errors were simulated from Gaussian.*

Category of Bayes factors	Scale mixture of two Gaussians	Kernel–form
$(0,1/150]$	0.0%	1.0%
$(1/150,1/20]$	0.0%	2.1%
$(1/20,1/3]$	0.4%	5.7%
$(1/3,1]$	0.3%	8.7%
$(1,3]$	0.9%	12.1%
$(3,20]$	3.4%	29.6%
$(20,150]$	4.9%	25.1%
$(150,\infty)$	90.1%	15.7%

The kernel–form error density is favoured against the correct Gaussian assumption of the error density with very strong evidence in 1% of simulated samples, with strong evidence in 2.1% of simulated samples and with positive evidence in 5.7% of simulated samples. It means that the kernel–form error density outperforms the correct assumption in 8.8% of simulated samples. On the other hand, the correct error–density assumption is favored against the kernel–form error density with very strong evidence in 15.7% of samples, with strong evidence in 25.1% of samples and with positive evidence in 29.6% of samples. Neither is favored in 20.8% of simulated samples.

In this case, the scale mixture density of two Gaussians is less competitive than the kernel–form error density. It is favored with positive evidence in only 0.4% of samples against the correct error density, while the correct error density is favored against the scale mixture with very strong evidence in 90.1% of samples .

3.4.2 Scale mixture density of two Gaussians for simulating errors

The relative frequencies of simulated samples with a $0.7N(0,0.7^2)+0.3N(0,1.5^2)$ error density falling in each category of Bayes factors of the scale mixture density against respectively, the Gaussian and kernel–form error densities, are reported in Table 6.

Table 6: *Relative frequencies of simulated samples falling in different categories of Bayes factors when errors were simulated from a scale mixture of two Gaussians.*

Category of Bayes factors	Gaussian	Kernel–form
$(0,1/150]$	0.1%	9.2%
$(1/150,1/20]$	0.0%	10.5%
$(1/20,1/3]$	0.1%	16.5%
$(1/3,1]$	0.0%	12.3%
$(1,3]$	0.0%	10.8%
$(3,20]$	0.2%	15.4%
$(20,150]$	1.3%	10.3%
$(150,\infty)$	98.3%	15.0%

The kernel–form error density is favored against the correct assumption of the error density in 36.2% of simulated samples, while the latter is favored against the former in 40.7% of simulated samples. Therefore, the kernel–form error density performs slightly worse than the correct density, which is unknown in practice. The Gaussian error density cannot compete against the correct error density because the latter is favored against the former in 99.8% of simulated samples.

3.4.3 Location mixture density of two Gaussians for simulating errors

The relative frequencies of simulated samples with a $0.7N(-1,1)+0.3N(7/3,1)$ error density falling in each category of Bayes factors of the kernel–form error density against respectively, the Gaussian and location–scale mixture error densities, are reported in Table 7.

The benefit of the kernel–form error density is clearly indicated by its relative performance against its parametric rival, the location–scale mixture of two Gaussians. The kernel–form error density is favored with very strong evidence in 58.8% of simulated samples, strong evidence in 3% of samples and positive evidence in 2.2% of samples. In total, it is favored

Table 7: *Relative frequencies of simulated samples falling in different categories of Bayes factors when errors were simulated from a location mixture of two Gaussians.*

Category of Bayes factors	Gaussian	Location–scale mixture
$(0,1/150]$	0.0%	28.6%
$(1/150,1/20]$	0.0%	2.2%
$(1/20,1/3]$	0.0%	1.9%
$(1/3,1]$	0.0%	2.3%
$(1,3]$	0.0%	0.9%
$(3,20]$	0.0%	2.2%
$(20,150]$	0.0%	3.0%
$(150,\infty)$	100.0%	58.8%

against the location–scale mixture density in 64% of simulated samples. In contrast, the location–scale mixture density is favored against the kernel–form density in 34.7% of samples. The Gaussian error density cannot compete with against kernel–form error density because the latter is favored against the former with very strong evidence in all simulated samples.

3.4.4 Location–scale mixture for simulating errors

The relative frequencies of simulated samples with a $0.7N(-1,0.7^2)+0.3N(7/3,1.5^2)$ error density falling in each category of Bayes factors of the location–scale mixture error density against the Gaussian and kernel–form error densities are respectively, presented in Table 8.

Table 8: *Relative frequencies of simulated samples falling in different categories of Bayes factors when errors were simulated from a location–scale mixture of two Gaussians.*

Category of Bayes factors	Gaussian	Kernel–form
$(0,1/150]$	0.0%	84.5%
$(1/150,1/20]$	0.0%	4.9%
$(1/20,1/3]$	0.0%	2.7%
$(1/3,1]$	0.0%	1.0%
$(1,3]$	0.0%	0.9%
$(3,20]$	0.0%	1.2%
$(20,150]$	0.0%	1.0%
$(150,\infty)$	100.0%	3.8%

The kernel–form error density demonstrated a strong competing capacity against its competitor, the location–scale mixture error density. The former is favored against the latter with very strong evidence in 84.5% of samples, strong evidence in 4.9% of samples and positive evidence in 2.7% of samples. In a total of 92.1% of simulated samples, the kernel–form density is favored against the location–scale mixture. In contrast, the location–scale mixture is only favored in 6% of simulated samples. The Gaussian error–density assumption cannot compete against the location–scale mixture error density because the latter is favored against the former with very strong evidence in all simulated samples.

4 An application to nonparametric regression of stock returns

Many market analysts tend to believe that the Australian stock market usually follows the overnight U.S. stock market since the beginning of the sub–prime mortgage crisis. Therefore, it is of interest to investigate how the Australian stock market is affected by the daily outcome of the U.S. stock market. As the Australian market does not always follow the overnight U.S. market, we might need another variable to explain such discrepancies. That suggests a nonparametric regression model of the All Ordinaries (Aord) daily return on the overnight S&P 500 and FTSE returns, through which we can investigate the empirical relevance of the proposed sampling algorithm for bandwidth estimation. As the opening time for share trading in the Australian stock market is several hours after the closing time of the previous day trading in the UK and USA stock markets, such an investigation can reveal the relationship between the Australian stock market and the other two markets.

4.1 Data

We collected the daily closing indices of Aord, S&P 500 and FTSE during the period from the 3rd January 2007 to the 1st October 2012, excluding non–trading days. The Aord daily index was matched to the overnight FTSE and S&P 500 indices. When one market had a non–trading day, we deleted the trading data (if there were any) in the other two markets on that day. We collected $1,374$ observed vectors of the three indices, from which we computed daily continuously compounded percentage returns. The daily returns of the S&P 500 and FTSE indices on the 1st October 2012 (local time) were not used for estimating bandwidths, but were used for forecasting the Aord return and its density on the next day (local time).

Thus, the sample size is $n = 1,373$.

We used the bivariate nonparametric regression model given by

$$y_i = m(x_{1,i}, x_{2,i}) + \varepsilon_i, \qquad \text{for } i = 1, 2, \ldots, n, \tag{16}$$

where y_i is the Aord daily return, and the two regressors are respectively, the FTSE and S&P 500 returns, and $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are assumed to be iid with their density being assumed be the kernel–form density given by (2). For comparison purposes, we also considered the Gaussian and location–scale mixture density of two Gaussians as the error density, respectively.

4.2 Bandwidth estimates under different error densities

We used the adaptive random–walk Metropolis algorithm to sample bandwidths under the kernel–form error density, as well as bandwidths and other parameter(s) under the Gaussian and location–scale mixture error densities. Under each error density, the ergodic averages of the resulting simulated chains are used as the estimates of the corresponding bandwidths and/or other parameters(s). These estimates, as well as their 95% Bayesian credible intervals and other statistics, are tabulated in Table 9. According to the batch–mean standard deviation and SIF values, all simulated chains have achieved a very good mixing performance.

The empirical finding justifies the validity and usefulness of the kernel–form error density in the nonparametric regression model. The log marginal likelihood values are respectively, -1921.99, -1954.64 and -1925.47 under the kernel–form, Gaussian and location–scale mixture error densities. The Bayes factors of the kernel–form error density are respectively, $\exp(32.65)$ against Gaussian and $\exp(3.48)$ against the location–scale mixture. Therefore, the kernel–form error density is supported against the two parametric competitors with very strong evidence and strong evidence, respectively.

With the available observed FTSE and S&P 500 returns on the 1st October 2012 (local time), we forecasted the Aord return on the 2nd October 2012, which is the day being immediately out of the sample.[3] Under each error density, with the updated bandwidths at each iteration,

[3] Such a forecast could be conducted during the period between the closing time of the U.S. stock market and the opening time of the Australian stock market.

Table 9: *Estimates of bandwidths and parameter(s), their 95% Bayesian credible intervals, and their associated statistics under the kernel–form, Gaussian and location–scale mixture of two Gaussian error densities. $\widehat{y}_{n+1}$ is the one–day–ahead point forecast of the Aord return on the 2nd October 2012.*

Error density	Parameter	Estimate	95% Bayesian credible interval	Standard deviation	Batch–mean standard dev	SIF
Kernel–form	b	0.2072	(0.1104, 0.3306)	0.2718	0.0133	23.8
	h_1	0.4634	(0.4098, 0.5295)	0.1072	0.0061	32.6
	h_2	0.6147	(0.5291, 0.7226)	0.1680	0.0092	30.1
	$\widehat{y}_{n+1}$	0.3455	(0.3353, 0.3540)			
Gaussian	σ	0.9983	(0.9622, 1.0351)	0.0190	0.0002	1.0
	h_1	0.5463	(0.4835, 0.6130)	0.1106	0.0048	18.9
	h_2	0.6884	(0.6091, 0.7742)	0.1419	0.0064	20.2
	$\widehat{y}_{n+1}$	0.3330	(0.3221, 0.3423)			
Location–scale mixture	w	0.7152	(0.5451, 0.8519)	0.0797	0.0045	31.8
	h_1	0.5239	(0.4629, 0.5850)	0.1088	0.0045	17.0
	h_2	0.6363	(0.5627, 0.7103)	0.1283	0.0053	17.3
	σ_1	0.7277	(0.6267, 0.8147)	0.0463	0.0024	26.3
	σ_2	1.4838	(1.2760, 1.7830)	0.1274	0.0067	27.7
	μ_1	0.0443	(-0.0143,0.1142)	0.0320	0.0009	8.1
	$\widehat{y}_{n+1}$	0.3334	(0.3231, 0.3438)			

we calculated the NW estimator and treated it as the one–day–ahead point forecast. Upon completing the sampling procedure, we took the average of the forecasted values made at all iterations. Table 9 presents the averaged forecast and its 95% Bayesian credible interval obtained under each error density. The one–day–ahead point forecasts of the Aord return derived under the kernel–form, Gaussian and a location–scale mixture error densities are respectively, 0.3455%, 0.3330% and 0.3334%, where the forecast under the kernel–form error density is slightly closer to the observed return, which is 0.9842%, than that derived under each competitor.

4.3 One–day–ahead density forecast of the Aord return

Under the kernel–form error density, with the updated bandwidths denoted as $\widetilde{\boldsymbol{h}}$ and $\widetilde{b}$ at each iteration, we calculated the approximate density of y_{n+1} given by

$$\widetilde{f}_Y\left(y_{n+1};\widetilde{\boldsymbol{h}},\widetilde{b}\right)=\frac{1}{n}\sum_{j=1}^{n}\frac{1}{\widetilde{b}}\,\phi\left(\frac{\{y_{n+1}-\widehat{m}(\boldsymbol{x}_{n+1};\widetilde{\boldsymbol{h}})\}-\{y_j-\widehat{m}(\boldsymbol{x}_j;\widetilde{\boldsymbol{h}})\}}{\widetilde{b}}\right), \tag{17}$$

where $\boldsymbol{x}_{n+1}$ is the vector of observed FTSE and S&P 500 returns on the 1st October 2012. Note that the leave–one–out strategy is not required for the purpose of calculating density values. Upon completing all iterations, we took the average of density functions given by (17) forecasted at all iterations. The averaged density forecast of y_{n+1} is presented in Figure 2.

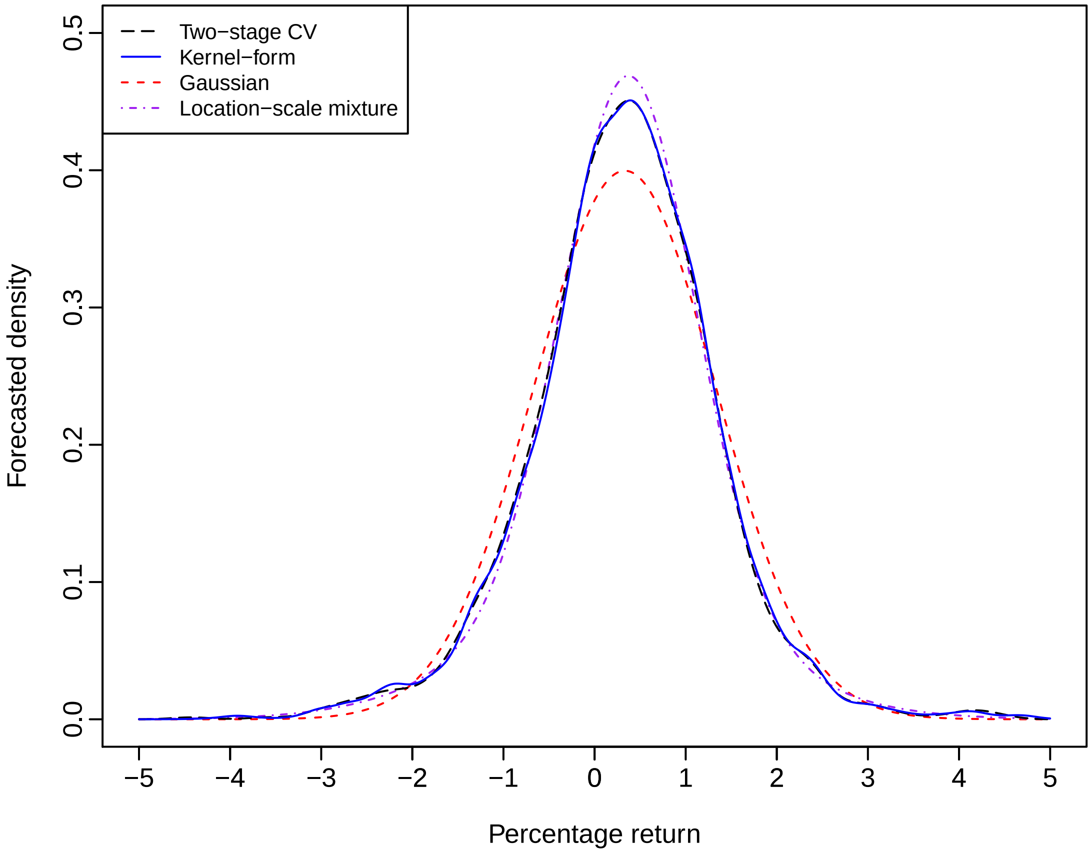

Figure 2: *Estimated densities of the forecasted All Ordinaries return on the 2nd October 2012.*

In a similar way, we calculated the averaged error density functions under the assumptions of Gaussian and location–scale mixture error densities. The graphs of the two forecasted density functions are also presented in Figure 2. In addition, we used the two–stage CV described previously to choose bandwidths for the NW estimator, as well as the bandwidth for the kernel density estimator of residuals, and the chosen bandwidths are $(0.6926, 0.5513)'$ and 0.2070, respectively. The derived error density is also presented in Figure 2.

The kernel–form error density with its bandwidth estimated through Bayesian sampling is slightly fatter than the density estimator with bandwidths chosen through the two–stage CV in the area with the return value being from -2.5% to -2%. Apart from this area, the two density functions are quite similar to each other. The location–scale mixture error density with its parameters estimated through Bayesian sampling is clearly different from the former two. The Gaussian error density with its variance parameter estimated through Bayesian sampling differs obviously from the former three. It has a lower peak and a slightly thinner left tail than each of the former three density estimators.

According to the four forecasted density functions of the Aord return on the 2nd October 2012, we computed the one–day VaRs for holding a \$100 investment on the Aord index. At the 95% confidence level, the one–day VaRs are respectively, \$1.3450, \$1.3099, \$1.3017 and \$1.3419 through the kernel–form density, Gaussian, location–scale mixture density, and the two–stage method. At the 99% confidence level, the corresponding VaRs are \$2.4513, \$1.9917, \$2.4300 and \$2.4741.

The kernel–form error density leads to the largest VaR regardless whether Bayesian sampling or the two–stage method is used. Because the overnight FTSE and S&P 500 indices dropped by respectively, 0.6475% and 0.4488%, the VaRs derived through the kernel–form error density reflected the high risk in the Australian stock market.

The performance of each error–density assumption in forecasting VaR was examined by the corresponding relative frequency of exceedance derived through rolling samples. The concept of exceedance refers to the fact that the actual daily loss for holding the underlying asset exceeds the estimated VaR. The sample for estimating bandwidths has a fixed size of 1,000 observed vectors of the Aord, FTSE and S&P 500 returns, and the first sample starts from the 5th January 2007 to 17th February 2011. After the one–day–ahead VaR is forecasted, the sample is rolled forward for one day and is used for estimating bandwidths and forecasting VaR. The last sample finishes at the 1st October 2012, and there are a total of 374 samples for calculating the relative frequency of exceedance.

At the 99% confidence level, the relative frequencies of exceedance are respectively, 1.07%, 2.90%, 0.80% and 2.67% under the kernel–form, Gaussian, location–scale mixture error densities, and the kernel density with bandwidths derived trough the two–stage CV. At the 95% confidence level, these relative frequencies are respectively, 5.61%, 7.61%, 5.61% and 7.75%. These calculations suggest that when the confidence level is 99%, the kernel–form error density with bandwidths estimated through Bayesian sampling leads to more accurate VaR estimates than the other three density estimators. When the confidence level is 95%, this error density works well with its bandwidths estimated/chosen through Bayesian sampling or the two–stage CV.

5 An application to SPD estimation

Aït-Sahalia and Lo (1998) showed that in a dynamic equilibrium model, the price of a security is

$$P_t = \exp\left\{r_{t,\lambda}\lambda\right\} E_t^* \left\{Z(S_T)\right\} = \exp\left\{r_{t,\lambda}\lambda\right\} \int_{-\infty}^{\infty} Z(S_T) f_t^*(S_T) dS_T,$$

where $T = t + \lambda$, λ is the length of time to maturity, $r_{t,\lambda}$ is a constant risk–free interest rate between t and T, E_t^* represents the expectation taken conditional on information available at date t, S_T is the price of the security at date T, $Z(S_T)$ is the payoff of the security at the expiry date T, and $f_t^*(S_T)$ is the date–t SPD of S_T for the payoff of the security at date T. When an option is the security of interest, the SPD is the second–order derivative of a call–option pricing formula with respect to strike price calculated at S_T. Aït-Sahalia and Lo (1998) showed that the date–t price of a call option, is a nonlinear function of $(S_t, X_t, \lambda, r_{t,\lambda}, \delta_{t,\lambda})'$, which can be estimated through the nonparametric regression technique, where $\delta_{t,\lambda}$ is the dividend rate at date t.

In order to reduce the number of regressors, Aït-Sahalia and Lo (1998) assumed that the call–option pricing formula is given by the Black–Scholes (BS) formula except that the date–t volatility denoted by σ_t, is estimated by the nonparametric regression of the implied volatility on $\widetilde{z}_t = (F_t, X, \delta)$, where F_t is the futures price of the underlying asset. The kernel estimator of the regression function is

$$\widehat{\sigma}_t(F_t, X, \lambda|\boldsymbol{h}) = \frac{n^{-1}\sum_{j=1}^{n} K_{\boldsymbol{h}}(\widetilde{z}_t - \widetilde{z}_j)\widetilde{\sigma}_j}{n^{-1}\sum_{j=1}^{n} K_{\boldsymbol{h}}(\widetilde{z}_t - \widetilde{z}_j)},$$

where $\widetilde{\sigma}_j$ is the volatility implied by the price of the call option, and $\boldsymbol{h}$ is a vector of bandwidths. According to Aït-Sahalia and Lo (1998) and Huynh, Kervella, and Zheng (2002), the SPD and the risk measures of delta (Δ) and Gamma (Γ) are expressed as

$$f_{BS,t}(S_T) = \frac{1}{S_T\sqrt{2\pi\sigma^2\lambda}}\exp\left\{-\frac{\left[\ln(S_T/S_t)-(r_{t,\lambda}-\delta_{t,\lambda}-\sigma^2/2)\lambda\right]^2}{2\sigma^2\lambda}\right\},$$
$$\Delta_{BS} = \Phi(d_1),$$
$$\Gamma_{BS} = \frac{\phi(d_1)}{S_t\sigma\sqrt{\lambda}}.$$

where $d_1 = \left\{\ln(S_t/X)+(r_{t,\lambda}-\delta_{t,\lambda}+\sigma^2/2)\lambda\right\}/(\sigma\lambda^{1/2})$.

Zhang et al. (2009) assumed that the errors of the nonparametric regression model of $\widetilde{\sigma}_t$ on $\widetilde{z}_t$ are iid and follow the Gaussian distribution with a zero mean and unknown variance. In this paper, we assume that the iid errors follow an unknown distribution with its density given by (2). We fitted the model to the same S&P 500 index options data as those investigated by Aït-Sahalia and Lo (1998) and Zhang et al. (2009). The sample period is from the 4th January to the 31st December 1993, and the sample size is $n = 14,431$. The priors of squared bandwidth parameters are those given by (10) and (11), where hyperparameters are $\alpha_b = \alpha_h = 1$ and $\beta_b = \beta_h = 0.05$. We applied the sampling procedure proposed in Section 2.2 to sample the bandwidths from their posterior. Table 10 presents the estimates of the bandwidths and some associated statistics. The estimated bandwidth vector is clearly different from $(5.6243, 5.4831, 9.7509)'$ derived under the assumption of Gaussian error density, where the priors of squared bandwidths are the same as those under the kernel–form error density, and the prior of the error variance is IG(1,0.05).

Table 10: *Bandwidth estimates obtained through Bayesian sampling for the S&P 500 index options data.*

Parameter	Estimate	95% Bayesian credible interval	Standard deviation	Batch-mean standard error	SIF
b	0.1935	(0.1820, 0.2065)	0.0433	0.0010	5.9
h_1	4.9810	(4.8251, 5.1625)	0.3450	0.0110	10.2
h_2	4.7884	(4.5097, 5.0487)	0.5279	0.0237	20.1
h_3	13.8418	(13.3231, 14.4532)	1.0790	0.0381	12.5

The Bayes factor of the mixture error density against the Gaussian error density is exp(4210.13), which is very strong evidence supporting the former. Using the bandwidth vector derived by Aït-Sahalia and Lo (1998), and the ones estimated through Bayesian sampling under both

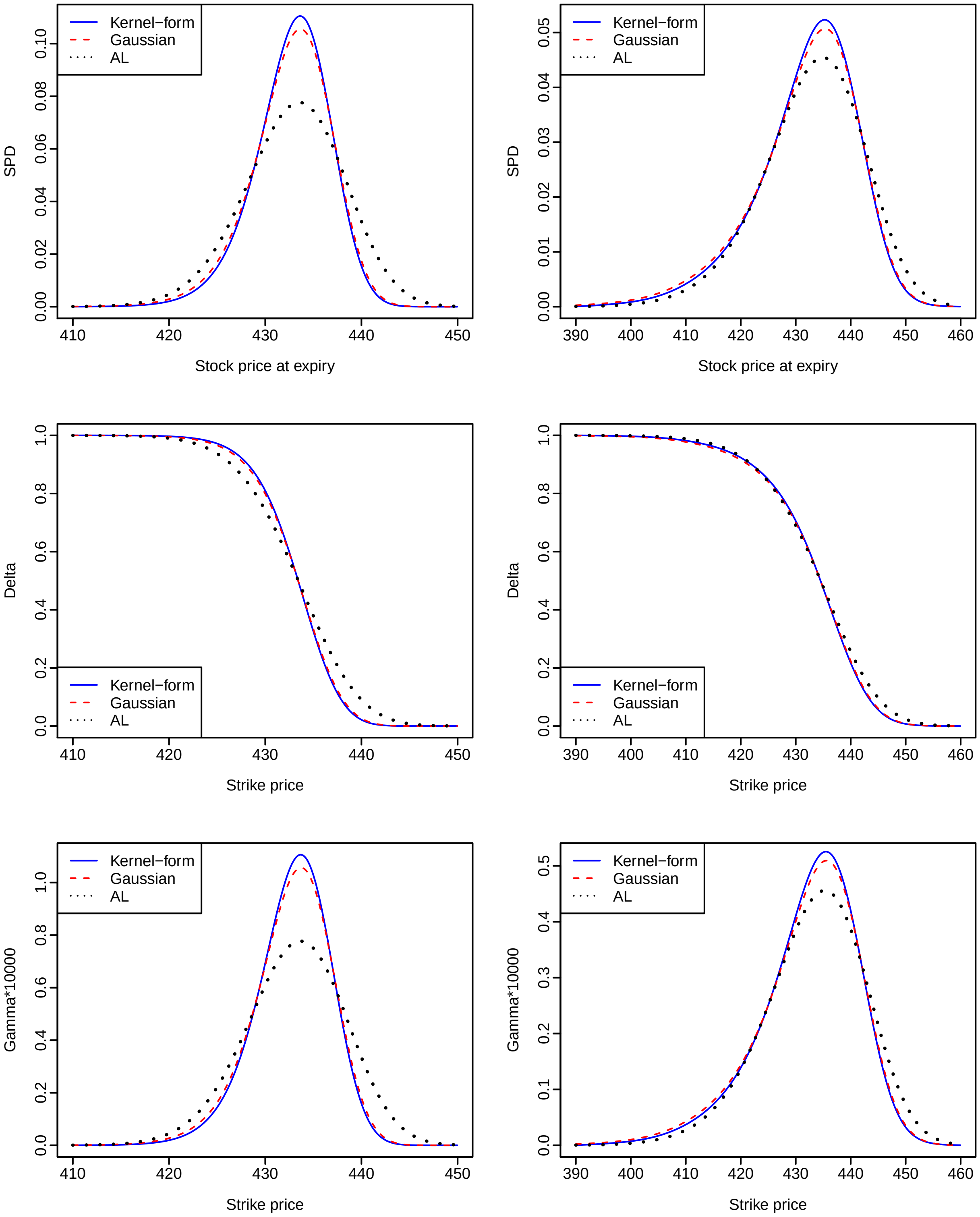

Figure 3: *Graphs of the estimated state–price density, risk measures* Δ *and* Γ *based on S&P 500 index options data. The first column is for the maturity of 2 days, and the second column is for the maturity of 10 days. AL denotes graphs obtained through the bandwidth vector provided by Aït-Sahalia and Lo (1998).*

error densities, we plotted in Figure 3, the graphs of the SPD, the risk measures Δ and Γ at maturities of 2 and 10 days, respectively. At the maturity of 2 days, the SPD and Γ produced through the bandwidth vector derived under the mixture error density are respectively, different from those derived under the Gaussian error density. However, as the time to maturity increases to 10 days, both densities lead to similar estimates of the SPD and Γ.

Moreover, the SPD, Δ and Γ derived through Bayesian sampling under each assumption of the error density are clearly different from those derived through the rule–of–thumb reported by Aït-Sahalia and Lo (1998). However, different assumptions of the error density lead to similar estimates of the SPD, risk measures Δ and Γ when maturity is 25 days or more.

6 Conclusion

We have proposed a nonparametric regression model with a flexible kernel–form error densities and presented a sampling algorithm for estimating bandwidths in the Nadaraya–Watson regression estimator and the error density. A series of Monte Carlo simulations reveal that the sampling approach outperforms the traditional bandwidth selection approaches for the Nadaraya–Watson estimator (as measured by ASE). Moreover, within the Bayesian sampling framework, when the true error density is non–Gaussian, the kernel–form error density model seems to fit the data well and is surprisingly competitive with its parametric rival using the true error density. Our Bayesian sampling procedure represents a data–driven solution to the problem of simultaneously estimating bandwidths for the kernel estimators of the regression function and error density.

Applying it to the nonparametric regression of the All Ordinaries daily return on the overnight FTSE and S&P 500 returns, we have obtained the bandwidth estimates for the kernel estimator of the regression under the three error–density assumptions. The assumption of a kernel–form error density is favored with very strong evidence against the assumptions of the Gaussian and a location–scale mixture of two Gaussians. The one–day–ahead density forecast of the All Ordinaries daily return obtained through the kernel–form error density exhibits a more reasonable left–tail behavior than that obtained through the two competitors. Moreover, the kernel–form error density allows for computing a distribution–free value–at–risk, which gains robustness in terms of different specifications of the error density. We also found that the kernel–form error density performs best when the relative frequencies of exceedance in 95% and 99% VaRs are calculated from rolling samples using this model.

The proposed nonparametric model and its sampling algorithm for simultaneously estimating bandwidths has also been validated through the nonparametric regression model involved in the state–price density estimation. The kernel–form error density is favored with very strong evidence against the Gaussian error density. We have also found that the state–price density, risk measures Δ and Γ estimated under this kernel–form error density are different from the corresponding ones estimated using Gaussian error density at short maturities of the underlying asset. The application confirms the usefulness of relaxing the Gaussian assumption of the error density to the kernel–form density in the nonparametric regression model.

Acknowledgements

We extend our sincere thanks to John Geweke and Peter Robinson for their very insightful comments on an earlier draft of this paper. Thanks also go to Victorian Partnership for Advanced Computing (VPAC) for its quality facility. This research was supported under the Australian Research Council's *Discovery Projects* funding scheme (project number DP1095838).

References

Aït-Sahalia, Y., Lo, A. W., 1998. Nonparametric estimation of state–price densities implicit in financial asset prices. The Journal of Finance 53 (2), 499–547.

Bowman, A. W., Azzalini, A., 1997. Applied Smoothing Techniques for Data Analysis. Oxford University Press, London.

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

Cheng, F., 2004. Weak and strong uniform consistency of a kernel error density estimator in nonparametric regression. Journal of Statistical Planning and Inference 119 (1), 95–107.

Chib, S., 1995. Marginal likelihood from the Gibbs output. Journal of the American Statistical Association 90 (432), 1313–1321.

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

Efromovich, S., 2005. Estimation of the density of regression errors. The Annals of Statistics 33 (5), 2194–2227.

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

Garthwaite, P. H., Fan, Y., Sisson, S. A., 2011. Adaptive optimal scaling of Metropolis–Hastings algorithms using the Robbins–Monro process. Working paper, University of New South Wales.
URL http://arxiv.org/abs/1006.3690

Gelfand, A. E., Dey, D. K., 1994. Bayesian model choice: Asymptotics and exact calculations. Journal of the Royal Statistical Society, *Series B* 56 (3), 501–514.

Geweke, J., 2009. Complete and Incomplete Econometric Models. Princeton University Press, New Jersey.

Geweke, J., Keane, M., 2007. Smoothly mixing regressions. Journal of Econometrics 138 (1), 252–290.

Geweke, J. F., 1999. Using simulation methods for Bayesian econometric models: Inference, development, and communication. Econometric Reviews 18 (1), 1–73.

Griffin, J., Quintana, F., Steel, M., 2011. Flexible and nonparametric modelling. In: Geweke, J., Koop, G., van Dijk, H. (Eds.), The Oxford Handbook of Bayesian Econometrics. Oxford University Press, Oxford.

Hall, P., Lahiri, S. N., Polzehl, J., 1995. On bandwidth choice in nonparametric regression with both short-and long-range dependent errors. The Annals of Statistics 23 (6), 1921–1936.

Härdle, W., 1990. Applied Nonparametric Regression. Cambridge University Press, Cambridge.

Härdle, W., Hall, P., Ichimura, H., 1993. Optimal smoothing in single–index models. The Annals of Statistics 21 (1), 157–178.

Härdle, W., Marron, J. S., 1985. Optimal bandwidth selection in nonparametric regression function estimation. The Annals of Statistics 13 (4), 1465–1481.

Härdle, W., Müller, M., 2000. Multivariate and semiparametric kernel regression. In: Schimek, M. G. (Ed.), Smoothing and Regression: Approaches, Computation, and Application. John Wiley & Sons, New York, pp. 357–392.

Herrmann, E., Engel, J., Wand, M. P., Gasser, T., 1995. A bandwidth selector for bivariate kernel regression. Journal of the Royal Statistical Society, *Series B* 57 (1), 171–180.

Huynh, K., Kervella, P., Zheng, J., 2002. Estimating state price densities with nonparametric regression. In: Härdle, W., Kleinow, T., Stahl, T. (Eds.), Applied Quantitative Finance. Springer Verlap, Heidelberg.

Jaki, T., West, R. W., 2008. Maximum kernel likelihood estimation. Journal of Computational and Graphical Statistics 17 (4), 976–993.

Jeffreys, H., 1961. Theory of Probability. Oxford University Press, Oxford, U.K.

Kass, R. E., Raftery, A. E., 1995. Bayes factors. Journal of the American Statistical Association 90 (430), 773–795.

Kim, S., Shephard, N., Chib, S., 1998. Stochastic volatility: Likelihood inference and comparison with ARCH models. Review of Economic Studies 65 (3), 361–393.

Leslie, D. S., Kohn, R., Nott, D. J., 2007. A general approach to heteroscedastic linear regression. Statistics and Computing 17 (2), 131–146.

Linton, O., Xiao, Z., 2007. A nonparametric regression estimator that adapts to error distribution of unknown form. Econometric Theory 23 (3), 371–413.

Newton, M. A., Raftery, A. E., 1994. Approximate Bayesian inference with the weighted likelihood bootstrap. Journal of the Royal Statistical Society, *Series B* 56 (1), 3–48.

Nott, D. J., Kohn, R., 2005. Adaptive sampling for Bayesian variable selection. Biometrika 92 (4), 747–763.

Roberts, G. O., 1996. Markov chain concepts related to sampling algorithms. In: Gilks, W. R., Richardson, S., Spiegelhalter, D. J. (Eds.), Markov Chain Monte Carlo in Practice. Chapman & Hall, London, pp. 45–57.

Roberts, G. O., Rosenthal, J. S., 2009. Examples of adaptive MCMC. Journal of Computational and Graphical Statistics 18 (2), 349–367.

Rothe, C., 2009. Semiparametric estimation of binary response models with endogenous regressors. Journal of Econometrics 153 (1), 51–64.

Samb, R., 2011. Nonparametric estimation of the density of regression errors. Comptes Rendus Mathematique 349 (23–24), 1281–1285.

Scott, D. W., 1992. Multivariate Density Estimation: Theory, Practice, and Visualization. Wiley, New York.

Silverman, B. W., 1986. Density Estimation for Statistics and Data Analysis. Chapman and Hall, New York.

Tse, Y. K., Zhang, X., Yu, J., 2004. Estimation of hyperbolic diffusion using the Markov chain Monte Carlo method. Quantitative Finance 4 (2), 158–169.

Wahba, G., Wold, S., 1975. A completely automatic French curve: Fitting spline functions by cross validation. Communications in Statistics — Theory and Methods 4 (1), 1–17.

Yuan, A., de Gooijer, J. G., 2007. Semiparametric regression with kernel error model. Scandinavian Journal of Statistics 34 (4), 841–869.

Zhang, X., Brooks, R. D., King, M. L., 2009. A Bayesian approach to bandwidth selection for multivariate kernel regression with an application to state–price density estimation. Journal of Econometrics 153 (1), 21–32.

Zhang, X., King, M. L., 2011. Bayesian semiparametric GARCH models. Working paper 24/11, Monash University.
URL http://www.buseco.monash.edu.au/ebs/pubs/wpapers/2011/wp24-11.pdf