Zhang, Xibin; King, Maxwell L.; Shang, Han Lin

Article

Bayesian bandwidth selection for a nonparametric regression model with mixed types of regressors

Econometrics

Provided in Cooperation with:
MDPI – Multidisciplinary Digital Publishing Institute, Basel

Suggested Citation: Zhang, Xibin; King, Maxwell L.; Shang, Han Lin (2016) : Bayesian bandwidth selection for a nonparametric regression model with mixed types of regressors, Econometrics, ISSN 2225-1146, MDPI, Basel, Vol. 4, Iss. 2, pp. 1-27, https://doi.org/10.3390/econometrics4020024

This Version is available at:
https://hdl.handle.net/10419/171875

econometrics

MDPI

Article

Bayesian Bandwidth Selection for a Nonparametric Regression Model with Mixed Types of Regressors

Xibin Zhang [1,*], Maxwell L. King [1] and Han Lin Shang [2]

[1] Department of Econometrics and Business Statistics, Monash Business School, Monash University, 900 Dandenong Road, Caulfield East, VIC 3145, Australia; max.king@monash.edu

[2] Research School of Finance, Actuarial Studies and Statistics, Australian National University, Canberra, ACT 2601, Australia; hanlin.shang@anu.edu.au

* Correspondence: xibin.zhang@monash.edu; Tel.: +61-3-9903-2130; Fax: +61-3-9903-2007

Academic Editor: Isabel Casas

Received: 8 December 2015; Accepted: 6 April 2016; Published: 22 April 2016

Abstract: This paper develops a sampling algorithm for bandwidth estimation in a nonparametric regression model with continuous and discrete regressors under an unknown error density. The error density is approximated by the kernel density estimator of the unobserved errors, while the regression function is estimated using the Nadaraya-Watson estimator admitting continuous and discrete regressors. We derive an approximate likelihood and posterior for bandwidth parameters, followed by a sampling algorithm. Simulation results show that the proposed approach typically leads to better accuracy of the resulting estimates than cross-validation, particularly for smaller sample sizes. This bandwidth estimation approach is applied to nonparametric regression model of the Australian All Ordinaries returns and the kernel density estimation of gross domestic product (GDP) growth rates among the organisation for economic co-operation and development (OECD) and non-OECD countries.

Keywords: cross-validation; Nadaraya-Watson estimator; posterior predictive density; random-walk Metropolis; unknown error density; value-at-risk

JEL: C11, C14, C35

1. Introduction

Nonparametric regression is an important tool for exploring the unknown relationship between a response variable and a set of explanatory variables also known as regressors. A simple and commonly used estimator of the regression function is the Nadaraya-Watson (NW) estimator proposed by [1,2]. In many empirical applications of nonparametric regression models, regressors are often of mixed types such as continuous and categorical. In such a situation, [3] proposed estimating the regression function by the NW-type estimator with different types of regressors being assigned different kernel functions. Since their seminal work, there have been many theoretical and methodological investigations on nonparametric regression with mixed types of regressors (see for example, [4–11]). It has been generally accepted that the performance of the NW estimator is mainly determined by its bandwidths. In the current literature, the cross-validation (CV) technique is often used for choosing bandwidths. Following the recent work by [12], this paper aims to investigate a Bayesian sampling approach to bandwidth estimation for the NW estimator in a nonparametric regression model, where regressors can be continuous and discrete variables.

The popularity of CV is accredited to its simplicity and reasonably good performance. However, this bandwidth selection method has some limitations. First, the CV technique tends to choose a too small bandwidth. As discussed by [13], there are necessary and sufficient conditions to ensure the

optimality of CV for bandwidth selection in nonparametric regression models. It may not be possible to examine whether these conditions hold in practice. Second, even though the CV method does not require an assumption about the error density, it provides no direct solutions to error density estimation. However, error density estimation is important for assessing the goodness of fit of a specified error distribution [14–16]; for testing symmetry, skewness and kurtosis of the residual distribution [17–19]; for statistical inference, prediction and model validation [20,21]; and for estimating the density of the response variable [22,23]. Therefore, being able to estimate the error density is as important as being able to estimate the regression function.

We present a Bayesian sampling approach to bandwidth estimation for the NW estimator involving mixed types of regressors, where the errors are independent identically distributed (iid) and follow a kernel-form error density studied previously by [24] for GARCH models, and [12] for nonparametric regression models with only continuous regressors. We develop a sampling algorithm, which is an extension to the algorithm proposed by [12] in the sense that the regressors are of mixed types. It leads to a posterior estimate of the response density, where the response variable is modeled as an unknown function of continuous and discrete explanatory variables. The importance of such models can be explained through finance and economic data examples.[1]

Suppose we are interested in the distribution of daily returns of the All Ordinaries (Aord) index of the Australian stock market. Many analysts believe that since the beginning of the global financial crisis, the Australian stock market generally follows the overnight performance of the US and several European markets. As the US stock market is thought to have a leading effect on other markets worldwide, we may choose the daily return of the S&P 500 index as a regressor, and a binary variable indicating the sign of the daily return of FTSE index as another regressor. With our proposed sampling algorithm, we are able to not only estimate bandwidths for the NW estimator and the kernel-form error density, but also derive a posterior estimate of the response density. When CV is used to choose bandwidths for the NW estimator, one might have to apply likelihood cross-validation to the residuals so as to derive a kernel density estimator of residuals. This would result in a two-stage procedure for choosing bandwidths in the regression and error-density estimators, and we show that it is empirically inferior to our sampling procedure through a series of simulation studies.

We conduct Monte Carlo simulation studies to compare the in-sample and out-of-sample performances between Bayesian sampling and CV in choosing bandwidths for the regression estimator and error density estimator. Our Bayesian sampling approach to bandwidth estimation leads to more accurate estimators than CV in most situations and in particular for smaller sample sizes, while the latter performs as well as the former in only a few occasions.

Our sampling algorithm is empirically validated through an application to bandwidth estimation in the nonparametric regression of the Aord daily return on the overnight S&P 500 return and a binary variable showing the sign of overnight FTSE return. An important and very useful output from this sampling algorithm is the one-day-ahead posterior predictive density of the Aord daily return, which we use to calculate value-at-risk (VaR). Given the close relationship between conditional mean regression and conditional density estimation, we also modify the sampling algorithm for the purpose of choosing bandwidths in kernel condition density estimation of GDP growth rate of a country given its OECD status and the year value of growth-rate observations.

The rest of the paper is organized as follows. Section 2 presents a brief description of the NW estimator when the regressors include continuous and discrete variables. In Section 3, we derive the likelihood and posterior for bandwidth parameters. A sampling algorithm is also presented. Section 4 presents Monte Carlo simulation studies that examine the performance of the proposed sampling method for bandwidth estimation. In Section 5, we use the sampling method to estimate bandwidths

[1] A former version of this paper is [25], based on which [26] studied bandwidth selection for a nonparametric functional regression model with mixed types of regressors, where the distance metric is not well defined in a function space.

for a nonparametric regression model of stock returns. We modify the proposed sampling algorithm to estimate bandwidths in kernel conditional density estimation of a country's GDP growth rate in Section 6. Section 7 concludes the paper.

2. Nadaraya-Watson Estimator with Mixed Types of Regressors

We present a brief description of the NW estimator of the unknown regression function that contains continuous and discrete explanatory variables. More details can be found in [6]. We consider the nonparametric regression model given by

$$y_i = m(\boldsymbol{x}_i) + \varepsilon_i, \qquad i = 1, 2, \cdots, n, \tag{1}$$

where y_i is an observation of a scalar response, $\boldsymbol{x}_i = \left(\boldsymbol{x}_i^{(c)}, \boldsymbol{x}_i^{(d)}\right)$ with $\boldsymbol{x}_i^{(c)}$ being a vector of p continuous variables and $\boldsymbol{x}_i^{(d)}$ a vector of q discrete variables, and $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ are iid errors with an unknown probability density function denoted as $f(\varepsilon)$. The discrete variables can be either ordered or unordered, which in turn affects the choice of kernel functions.

The flexibility of model (1) stems from the fact that the unknown regression function $m(\cdot)$ does not need to have a specific functional form. With some smoothness properties, $m(\cdot)$ is estimated by the NW estimator of the form given by

$$\widehat{m}\left(\boldsymbol{x}; \boldsymbol{h}, \boldsymbol{\lambda}\right) = \frac{\sum_{i=1}^{n} \mathcal{K}_{\boldsymbol{h},\boldsymbol{\lambda}}\left(\boldsymbol{x}, \boldsymbol{x}_i\right) y_i}{\sum_{i=1}^{n} \mathcal{K}_{\boldsymbol{h},\boldsymbol{\lambda}}\left(\boldsymbol{x}, \boldsymbol{x}_i\right)}, \tag{2}$$

where $\mathcal{K}_{\boldsymbol{h},\boldsymbol{\lambda}}\left(\boldsymbol{x}, \boldsymbol{x}_i\right) = K_{\boldsymbol{h}}^{(c)}\left(\boldsymbol{x}^{(c)} - \boldsymbol{x}_i^{(c)}\right) \times K_{\boldsymbol{\lambda}}^{(d)}\left(\boldsymbol{x}^{(d)}, \boldsymbol{x}_i^{(d)}\right)$ is a generalized product kernel that admits continuous and discrete regressors. There is a wide range of kernel functions for continuous type of variables (see for example, [6]), and kernel functions for discrete type of variables (see for example, [27]). Here, the kernel function for continuous regressors is a product of p identical Gaussian kernel functions expressed as

$$K_{\boldsymbol{h}}^{(c)}\left(\boldsymbol{x}^{(c)} - \boldsymbol{x}_i^{(c)}\right) = \prod_{j=1}^{p} \frac{1}{h_j} \phi\left(\frac{x_j^{(c)} - x_{i,j}^{(c)}}{h_j}\right),$$

where $x_j^{(c)}$ and $x_{i,j}^{(c)}$ are respectively, the jth elements of $\boldsymbol{x}^{(c)}$ and $\boldsymbol{x}_i^{(c)}$, $\boldsymbol{h} = (h_1, h_2, \cdots, h_p)'$ is a vector of bandwidths associated with the p continuous regressors, and $\phi(\cdot)$ is the standard Gaussian density being used as the kernel function for a continuous variable throughout this paper.

The kernel function for discrete regressors is a product of q identical discrete kernel functions expressed as

$$K_{\boldsymbol{\lambda}}^{(d)}\left(\boldsymbol{x}^{(d)}, \boldsymbol{x}_i^{(d)}\right) = \prod_{j=1}^{q} K_{\lambda_j}^{(d)}\left(x_j^{(d)}, x_{ij}^{(d)}\right).$$

If the jth element of $\boldsymbol{x}^{(d)}$ is nominal, the kernel function is Aitchison and Aitken's kernel [28] given by

$$K_{\lambda_j}^{(d)}\left(x_j^{(d)}, x_{i,j}^{(d)}\right) = \begin{cases} 1 - \lambda_j & \text{if } x_j^{(d)} = x_{i,j}^{(d)} \\ \lambda_j / (\mathrm{c} - 1) & \text{otherwise} \end{cases}, \tag{3}$$

where $x_j^{(d)}$ and $x_{i,j}^{(d)}$ are respectively, the jth elements of $\boldsymbol{x}^{(d)}$ and $\boldsymbol{x}_i^{(d)}$, and $\lambda_j \in (0, (\mathrm{c} - 1)/\mathrm{c})$ is the bandwidth, for $j = 1, 2, \cdots, q$, and c denotes the number of discrete outcomes. Note that this kernel function can be used for either unordered categorical variables or unequal-interval ordered variables.

If the jth element of $\boldsymbol{x}^{(d)}$ is ordinal, its kernel is Li and Racine's kernel [6] kernel expressed as

$$K_{\lambda_j}^{(d)}\left(x_j^{(d)}, x_{i,j}^{(d)}\right) = \begin{cases} 1 & \text{if } x_j^{(d)} = x_{i,j}^{(d)} \\ \lambda_j^{\left|x_j^{(d)} - x_{i,j}^{(d)}\right|} & \text{otherwise} \end{cases}, \tag{4}$$

for $j = 1, 2, \cdots, q$, and $i = 1, 2, \cdots, n$, where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \cdots, \lambda_q)'$ is a vector of bandwidths assigned to the corresponding q discrete regressors. These bandwidths are restricted to be within $(0, 1)$.

For the case where discrete explanatory variables are included in a nonparametric regression model, there exists a conventional frequency estimator of the regression function in the literature. This frequency approach involves splitting data into cells based on different values of the discrete variables and using the data in each cell to derive an estimator of the regression function. Li and Racine [6] found that the NW estimator given by (2) is strongly supported against the frequency estimator for both theoretical and practical reasons.

It is noteworthy that if $\lambda_j = 0$ for all $j = 1, 2, \cdots, q$, then $K_{\lambda_j}^{(d)}\left(x_j^{(d)}, x_{i,j}^{(d)}\right)$ becomes an indicator function taking values 1 and 0, corresponding to the conventional frequency estimator. When $\lambda_j = 1$, this indicates that $x_j^{(d)}$ is "smoothed out" and becomes an irrelevant variable [5]. Similarly, if h_j is very large in the kernel function given by (2), it indicates that $x_j^{(c)}$ is smoothed out and has no explanatory effect on the response variable. As a by-product, the ability of distinguishing irrelevant variables from relevant variables makes the resulting NW estimator very attractive, in comparison to the conventional frequency estimator.

The performance of the NW estimator is mainly determined by its bandwidths, and in the current literature, bandwidths are often chosen through CV with the CV function defined as

$$\mathrm{CV}(\boldsymbol{h}, \boldsymbol{\lambda}) = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \widehat{m}_{(-i)}\left(\boldsymbol{x}_i; \boldsymbol{h}, \boldsymbol{\lambda}\right)\right]^2 w(\boldsymbol{x}_i),$$

where $\widehat{m}_{(-i)}(\boldsymbol{x}_i; \boldsymbol{h}, \boldsymbol{\lambda})$ is the leave-one-out NW estimator given by

$$\widehat{m}_{(-i)}(\boldsymbol{x}_i; \boldsymbol{h}, \boldsymbol{\lambda}) = \frac{\sum_{j=1; j\neq i}^{n} \mathcal{K}_{\boldsymbol{h},\boldsymbol{\lambda}}\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) y_j}{\sum_{j=1; j\neq i}^{n} \mathcal{K}_{\boldsymbol{h},\boldsymbol{\lambda}}\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)},$$

and $w(\cdot)$ is a weight function taking values in $[0, 1]$. The purpose of the weight function in $\mathrm{CV}(\boldsymbol{h}, \boldsymbol{\lambda})$ is to avoid difficulties caused either by division by zero or by the slow convergence rate when $\boldsymbol{x}_i$ is near the boundary of the support of $\boldsymbol{x}$ (see [13]). We follow [10] and choose

$$w(\boldsymbol{x}_i) = \prod_{j=1}^{p+q} \boldsymbol{I}\left(\left|x_{i,j} - \overline{x}_j\right| \leq 1.5 s_j\right), \tag{5}$$

where $\boldsymbol{I}(\cdot)$ is an indicator function, and $\overline{x}_j$ and s_j are the sample mean and standard deviation of $\{x_{i,j} : 1 \leq i \leq n\}$.

In some empirical studies, CV tends to choose too small a bandwidth. Li and Zhou [13] observed that there are some conditions to ensure the optimality of the CV function. However, it is impossible to check whether all these conditions hold due to the unknown regression function. This problem has motivated us to investigate an alternative approach to bandwidth estimation, namely a Bayesian sampling approach.

3. Bayesian Estimation of Bandwidths

We consider the nonparametric regression model given by (1) and assume that the iid errors follow an unknown distribution with its density approximated by

$$\widetilde{f}(\varepsilon; b) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{b}\phi\left(\frac{\varepsilon-\varepsilon_i}{b}\right), \tag{6}$$

where $\phi(\cdot)$ is the standard Gaussian probability density function. Being previously introduced by [24] in GARCH models, this density function of errors is a mixture of n identical standard Gaussian densities with different means located at individual errors and a common standard deviation b. Moreover, from the view of kernel smoothing, this error density can be regarded as a kernel density based on independent errors, where b plays the role of bandwidth or smoothing parameter. Note that the error density given by (6) is different from the one proposed by [29], who used residuals as proxies of errors and employed the kernel density estimator of residuals to approximate the true error density.

In order to conduct Bayesian sampling for the purpose of estimating bandwidths, we treat the bandwidths in the NW estimator of the regression function and the kernel-form error density as parameters. Even though chosen bandwidths depend on sample size as revealed by existing asymptotic results, such a treatment will not cause problems for a fixed-size sample (see also [30–32]).

Let $(y_i, \boldsymbol{x}_i)$, for $i = 1, 2, \cdots, n$, denote the observations of $(y, \boldsymbol{x})$. Under the error density given by (6), we have

$$y_i \sim \widetilde{f}\left(\{y_i - m(\boldsymbol{x}_i)\}; b\right),$$

for $i = 1, 2, \cdots, n$. As $m(\cdot)$ is unknown, we replace it with its leave-one-out NW estimator. Thus, the density of y_i is approximated by

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

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

3.1. An Approximate Likelihood

The vector of all bandwidths denoted as $(\boldsymbol{h}', \boldsymbol{\lambda}', b)'$ are treated as parameters, given which we wish to derive an approximate likelihood of $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)'$. For this purpose, we cannot use the approximate density of y_i given by (7) directly because it contains $\phi(0)/b$ for $j = i$. When b is a parameter and is estimated based on a fixed-size sample, numerical optimization of the likelihood built up through (7) may end up with an arbitrarily small bandwidth due to the existence of $\phi(0)/b$. Therefore, any pair of observations that leads to a zero argument of $\phi(\cdot)$ should be excluded from the summation given by (7). Following the suggestion of [12], we exclude the jth term when $\left\{y_j - \widehat{m}_{(-j)}(\boldsymbol{x}_j; \boldsymbol{h}, \boldsymbol{\lambda})\right\} = \left\{y_i - \widehat{m}_{(-i)}(\boldsymbol{x}_i; \boldsymbol{h}, \boldsymbol{\lambda})\right\}$, from the summation given in (7). Let

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

and n_i is the number of terms excluded from the sum given in (7). Although the response variable is continuous, by chance, the set J_i may contain more than one element. Therefore, the density of y_i is approximated by

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

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

Let $\boldsymbol{h}^2$ denote a row vector whose elements are the squared elements of $\boldsymbol{h}$. Thus, given $(\boldsymbol{h}^2, \boldsymbol{\lambda}', b^2)'$, the likelihood of $\boldsymbol{y}$ is approximated by

$$\ell\left(\boldsymbol{y}|\boldsymbol{h}^2, \boldsymbol{\lambda}, b^2\right) \approx \prod_{i=1}^{n} \left\{ \frac{1}{n-n_i} \sum_{j \in J_i} \frac{1}{b} \phi \left(\frac{\left\{y_i - \widehat{m}_{(-i)}(\boldsymbol{x}_i; \boldsymbol{h}, \boldsymbol{\lambda})\right\} - \left\{y_j - \widehat{m}_{(-j)}(\boldsymbol{x}_j; \boldsymbol{h}, \boldsymbol{\lambda})\right\}}{b} \right) \right\}. \quad (8)$$

Note that if $n_i = 1$, this likelihood function reduces to the leave-one-out likelihood.

3.2. Priors

We follow [12] choices of priors of bandwidths for continuous regressors and the kernel-form error density. Let $\pi(h_j^2)$ denote the prior of h_j^2, for $j = 1, 2, \cdots, p$, and $\pi(b^2)$ the prior of b^2. The priors of h_j^2 and b^2 can be chosen from a variety of densities with a positive domain, such as log normal and inverse Gamma densities. As the Gaussian kernel is used for each variate in this situation, each bandwidth is also considered as the standard deviation of the corresponding Gaussian density. Therefore, the prior of each squared bandwidth can be chosen to be an inverse Gamma density, which is a frequently used proper prior for the variance parameter. Thus, the prior of h_j^2 is

$$\pi\left(h_j^2\right) = \frac{(\beta_h)^{\alpha_h}}{\Gamma(\alpha_h)} \left(\frac{1}{h_j^2}\right)^{\alpha_h+1} \exp\left\{-\frac{\beta_h}{h_j^2}\right\}, \quad \text{for } j = 1, 2, \cdots, p,$$

and the prior of b^2 is

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

where the hyperparameters are chosen as $\alpha_h = \alpha_b = 1$ and $\beta_h = \beta_b = 0.05$ (see for example, [33] and [34] pp. 38–39).

Let $\pi(\lambda_j)$ denote the prior of λ_j, the bandwidth assigned to the jth discrete regressor, for $j = 1, 2, \cdots, q$. The prior of λ_j is assumed to be a uniform density defined on (z_a, z_b). For nominal regressors, $z_a = 0$ and $z_b = (c-1)/c$; and for ordered categorical regressors, $z_a = 0$ and $z_b = 1$. See for example, [8], for discussion of restrictions on these smoothing parameters.

The joint prior of $\left(\boldsymbol{h}^2, \boldsymbol{\lambda}', b^2\right)'$ is the product of all the marginal priors and is denoted as $\pi\left(\boldsymbol{h}^2, \boldsymbol{\lambda}, b^2\right)$.

3.3. An Approximate Posterior

An approximate posterior of $\left(\boldsymbol{h}^2, \boldsymbol{\lambda}', b^2\right)'$ is obtained as the product of the approximate likelihood given by (8) and the joint prior, and is expressed as (up to a normalizing constant)

$$\pi\left(\boldsymbol{h}^2, \boldsymbol{\lambda}, b^2|\boldsymbol{y}\right) \propto \ell\left(\boldsymbol{y}|\boldsymbol{h}^2, \boldsymbol{\lambda}, b^2\right) \times \pi\left(\boldsymbol{h}^2, \boldsymbol{\lambda}, b^2\right). \quad (9)$$

The random-walk Metropolis algorithm can be used to carry out the simulation, where the acceptance rate of random-walk Metropolis algorithm is targeted at 0.234 for multivariate draws and 0.44 for univariate draws [35,36]. In order to achieve similar levels of acceptance rates, we use the adaptive random-walk Metropolis algorithm proposed by [36]. This algorithm is capable of selecting appropriate scales, and achieves the targeted acceptance rates without manual adjustment. The sampling procedure is described as follows.

Step 1: Specify a Gaussian proposal distribution, and start the sampling iteration process by choosing an arbitrary value of $(\boldsymbol{h}^2, \boldsymbol{\lambda}', b^2)'$ and denoting it as $(\boldsymbol{h}_{(0)}^2, \boldsymbol{\lambda}_{(0)}', b_{(0)}^2)'$. For example, the elements of $\boldsymbol{h}_{(0)}^2$ and $b_{(0)}^2$ can be any values on $(0, 1)$ and the elements of $\boldsymbol{\lambda}_{(0)}$ can be any values on $(0, (c-1)/c)$ for nominal regressors and $(0, 1)$ for categorical regressors.

Step 2: At the kth iteration, the current state $\boldsymbol{h}^2_{(k)}$ is updated as $\boldsymbol{h}^2_{(k)} = \boldsymbol{h}^2_{(k-1)} + \gamma_{(k-1)}\boldsymbol{u}/||\boldsymbol{u}||$, where $\boldsymbol{u}$ is drawn from the proposal density which is the p dimensional standard Gaussian density, and $\gamma_{(k-1)}$ is an adaptive tuning parameter with an arbitrary initial value $\gamma_{(0)}$. The updated $\boldsymbol{h}^2_{(k)}$ is accepted with a probability given by

$$\min\left\{\frac{\pi\left(\boldsymbol{h}^2_{(k)}, \boldsymbol{\lambda}_{(k-1)}, b^2_{(k-1)}|\boldsymbol{y}\right)}{\pi\left(\boldsymbol{h}^2_{(k-1)}, \boldsymbol{\lambda}_{(k-1)}, b^2_{(k-1)}|\boldsymbol{y}\right)}, 1\right\}.$$

Step 3: The tuning parameter for the next iteration is set to

$$\gamma_{(k)} = \begin{cases} \gamma_{(k-1)} + c\left(1-\xi\right)/k & \text{if } \boldsymbol{h}^2_{(k)} \text{ is accepted} \\ \gamma_{(k-1)} - c\xi/k & \text{if } \boldsymbol{h}^2_{(k)} \text{ is rejected} \end{cases},$$

where $c = \gamma_{(k-1)}/\left(\xi - \xi^2\right)$ is a constant, and ξ is the optimal target acceptance probability, which is 0.234 for multivariate updating and 0.44 for univariate updating (see for example, [35,36]).

Step 4: Update $\boldsymbol{\lambda}_{(k-1)}$ and $b^2_{(k-1)}$ in the same way as described by Steps 2 and 3.

Step 5: Repeat Steps 2–4, discard the burn-in period of iterations, and the draws after the burn-in period are recorded and denoted as $\left\{\left(\boldsymbol{h}_{(k)}, \boldsymbol{\lambda}'_{(k)}, b_{(k)}\right)' : k = 1, 2, \cdots, M\right\}$.

Upon completing the above iterations, we use the ergodic mean (or posterior mean) of each simulated chain as an estimate of each bandwidth. The mixing performance of each simulated chain is monitored by the simulation inefficiency factor (SIF). As each simulated chain is a Markov chain, its SIF can be interpreted as the number of draws required so as to derive independent draws from the simulated chain (see for example, [33,37–39]).

In the following analyses, the burn-in period is taken as 1000 iterations and the number of recorded iterations after the burn-in period is 10,000. The number of batches is 200, and there are 50 draws within each batch.

4. Monte Carlo Simulation Study

A Monte Carlo simulation study was conducted to investigate the properties of the proposed Bayesian sampling approach to bandwidth estimation in comparison to the CV method for bandwidth selection in the NW regression estimator. The conventional frequency estimator, which is equivalent to setting the bandwidths for all discrete and continuous regressors to zero in the NW estimator [6] (Chapter 3) was also included in the comparison.

A range of different simulation experiments were conducted using seven different data generating processes (DGPs), five of which were discussed in [4]. To assess the performance of each approach, at each iteration we generated $2n$ observations denoted as $\left\{\left(y_i, x_i^{(d)}, x_i^{(c)}\right) : 1 \leq i \leq 2n\right\}$, where y_i is calculated via the DGP, $x_i^{(d)}$ is the vector of discrete regressors and $x_i^{(c)}$ is the vector of continuous regressors. The first n observations were used for estimation and in-sample evaluation, and the last n observations for out-of-sample evaluation (see also [10]). We used the average squared error (ASE) as an evaluation measure for both in-sample and out-of-sample evaluation:

$$\mathrm{ASE}_{\mathrm{in}} = \frac{1}{n}\sum_{i=1}^{n}\left[m\left(\boldsymbol{x}_i^{(d)}, x_i^{(c)}\right) - \widehat{m}\left(\boldsymbol{x}_i^{(d)}, x_i^{(c)}; \widehat{\boldsymbol{\theta}}\right)\right]^2 w\left(\boldsymbol{x}_i\right),$$

$$\mathrm{ASE}_{\mathrm{out}} = \frac{1}{n}\sum_{i=1}^{n}\left[m\left(\boldsymbol{x}_{n+i}^{(d)}, x_{n+i}^{(c)}\right) - \widehat{m}\left(\boldsymbol{x}_{n+i}^{(d)}, x_{n+i}^{(c)}; \widehat{\boldsymbol{\theta}}\right)\right]^2 w\left(\boldsymbol{x}_i\right),$$

where $\boldsymbol{\theta} = (\boldsymbol{h}, \boldsymbol{\lambda}, b)$ under the Bayesian method, $\boldsymbol{\theta} = (\boldsymbol{h}, \boldsymbol{\lambda})$ under CV, and $\boldsymbol{\theta} = \boldsymbol{h}$ for the frequency estimator; $\widehat{m}\left(\boldsymbol{x}_i^{(d)}, x_i^{(c)}; \boldsymbol{\theta}\right)$ is the NW estimator of m based on the sample of the first n observations, and $w(\boldsymbol{x}_i)$ is the weight function given by (5) for the in-sample and out-of-sample evaluation. The purpose of the weight function is to trim those extreme observations that lie outside most of the data points in the continuous regressors [4]. We calculated the mean, median, standard deviation (SD) and interquartile range (IQR) of each ASE averaged over the 1000 Monte Carlo replications.

In order to test whether variations in in-sample ASE between the different estimation methods are significantly different, we conducted the Kruskal-Wallis test [40]. When the Kruskal-Wallis test rejects the null hypothesis of no difference among the methods considered, we implement a posthoc multiple comparison, which is Dunn's test [41] with Bonferroni correction, to examine which method differs significantly from others. Differences in out-of-sample ASEs were assessed using the model confidence set (MCS) procedure of [42].

4.1. Accuracy of Regression Estimator

4.1.1. Experiment 1: Binary and Continuous Regressors

This experiment involves three DGPs. The first DGP is given by

$$y_i = \sum_{j=1}^{4} x_{i,j}^{(d)} + \sum_{j=1}^{4} \sum_{\substack{k=1 \\ k \neq j}}^{4} \frac{1}{2} x_{i,j}^{(d)} x_{i,k}^{(d)} + \sum_{j=1}^{4} x_{i,j}^{(d)} m_1\left(x_i^{(c)}\right) + m_2\left(x_i^{(c)}\right) + u_i,$$

for $i = 1, 2, \cdots, n$, where $x_{i,j}^{(d)}$ takes values of 0 and 1 with an equal probability of 0.5, $j = 1, 2, \cdots, 4$, $x_i^{(c)}$ is drawn from the uniform distribution on $[0, 2]$, $m_1\left(x_i^{(c)}\right) = \sin\left(x_i^{(c)}\pi\right)$; $m_2\left(x_i^{(c)}\right) = x_i^{(c)} - 0.5\left(x_i^{(c)}\right)^2 + 0.3\left(x_i^{(c)}\right)^3$, and u_i is drawn from $N(0, 1)$. Sample sizes of $n = 100$, 200 and 500 were used, and the results are presented in Table 1.

Table 1. Mean, median and variation measures of the average squared error (ASE) values derived through 1000 samples simulated according to the first data generating processes (DGP). The bolded numbers represent the minimum summary statistics of ASEs.

n	Method	In-Sample ASE				Out-of-Sample ASE			
		Mean	**Median**	**SD**	**IQR**	**Mean**	**Median**	**SD**	**IQR**
100	Bayesian	**0.5562**	0.5390	**0.1301**	0.1520	**1.0740**	**0.9692**	**0.4451**	**0.4827**
	CV	0.5877	**0.5192**	0.3017	**0.1320**	1.0990	0.9927	0.4525	0.5230
	Frequency	0.6098	0.5707	0.1931	0.1590	1.3680	1.2840	0.4858	0.5990
200	Bayesian	**0.3571**	**0.3512**	0.0571	0.0750	**0.5233**	**0.4998**	0.1567	**0.1683**
	CV	0.3577	0.3520	**0.0567**	**0.0727**	0.5355	0.5070	0.1617	0.1740
	Frequency	0.4115	0.3952	0.0865	0.0877	0.6132	0.5927	**0.1533**	0.1947
500	Bayesian	**0.1994**	**0.1978**	0.0255	**0.0332**	**0.2158**	**0.2120**	0.0415	**0.0489**
	CV	0.1997	0.1989	**0.0253**	0.0333	0.2175	0.2146	**0.0411**	0.0506
	Frequency	0.2298	0.2169	0.0562	0.0436	0.2332	0.2300	0.0420	0.0529

The NW estimator with bandwidths estimated through Bayesian sampling slightly outperforms the same estimator with bandwidths selected through CV, and these methods outperform the conventional frequency estimator in terms of all four summary statistics. The differences decline as n increases. The Kruskal-Wallis test has a p value of zero for each value of n implying that there are significant differences in performance between at least one pair of methods. According to Dunn's test with Bonferroni correction, the Bayesian method differs significantly from the frequency method for all sample sizes, but differs significantly from the CV method for $n = 100$ only. For $n = 200$ and 500,

differences between the Bayesian and CV methods are insignificant. Based on the out-of-sample ASE values, the MCS procedure determines that the Bayesian method performs best for all values of n.

The second and third DGPs are given by

$$y_i = x_{i,1}^{(d)} + x_{i,2}^{(d)} + x_{i,1}^{(c)} + x_{i,2}^{(c)} + u_i,$$

and

$$y_i = x_{i,1}^{(d)} + x_{i,2}^{(d)} + x_{i,1}^{(d)} x_{i,2}^{(d)} + x_{i,1}^{(c)} + x_{i,2}^{(c)} + x_{i,1}^{(c)} x_{i,2}^{(c)} + u_i,$$

for $i = 1, 2, \cdots, n$, where $x_{i,1}^{(d)}$ and $x_{i,2}^{(d)}$ take values from $\{0, 1\}$ with equal probabilities of 0.5, $x_{i,1}^{(c)}$ and $x_{i,2}^{(c)}$ are drawn from $N(0, 1)$, and u_i is drawn from $N(0, 1)$. The third DGP differs from the second by the inclusion of two interaction terms. Again, sample sizes of $n = 100$, 250 and 500 were used. The results are presented in Table 2.

Table 2. Mean, median and variation measures of the ASE values derived through 1000 samples simulated according to the second and third DGPs. The bolded numbers represent the minimum summary statistics of ASEs.

n	Method	In-Sample ASE				Out-of-Sample ASE			
		Mean	**Median**	**SD**	**IQR**	**Mean**	**Median**	**SD**	**IQR**
Second DGP									
100	Bayesian	**0.3082**	**0.2947**	**0.0884**	**0.1142**	**0.3296**	**0.3111**	**0.1220**	**0.1455**
	CV	0.3188	0.3019	0.1018	0.1297	0.3380	0.3147	0.1413	0.1623
	Frequency	0.3192	0.3004	0.1022	0.1275	0.3846	0.3653	0.1361	0.1644
200	Bayesian	**0.2098**	**0.2037**	**0.0492**	**0.0668**	0.1990	0.1932	**0.0606**	**0.0791**
	CV	0.2114	0.2050	0.0520	0.0682	**0.1987**	**0.1903**	0.0608	0.0835
	Frequency	0.2116	0.2050	0.0515	0.0677	0.2228	0.2135	0.0646	0.0807
500	Bayesian	**0.1277**	0.1258	**0.0236**	**0.0305**	0.1100	**0.1033**	0.0642	**0.0331**
	CV	0.1282	0.1258	0.0243	0.0315	**0.1068**	**0.1033**	**0.0269**	**0.0331**
	Frequency	0.1283	**0.1257**	0.0247	0.0318	0.1165	0.1126	0.0278	0.0344
Third DGP									
100	Bayesian	**0.3751**	**0.3643**	**0.0920**	**0.1172**	**0.4715**	**0.4438**	0.1866	0.2401
	CV	0.6132	0.5809	0.2045	0.2438	0.4931	0.4619	0.1925	**0.2161**
	Frequency	0.6136	0.5795	0.2021	0.2453	0.4917	0.4600	**0.1797**	0.2310
200	Bayesian	**0.2707**	**0.2647**	**0.0557**	**0.0696**	**0.2829**	**0.2691**	0.0870	0.1139
	CV	0.4218	0.4085	0.1069	0.1329	0.2953	0.2836	0.0974	0.1171
	Frequency	0.4226	0.4086	0.1091	0.1325	0.2917	0.2785	**0.0847**	**0.1056**
500	Bayesian	**0.1727**	**0.1706**	**0.0289**	**0.0367**	**0.1533**	**0.1490**	0.0357	0.0481
	CV	0.2619	0.2567	0.0495	0.0647	0.1550	0.1526	**0.0348**	**0.0480**
	Frequency	0.2619	0.2567	0.0495	0.0644	0.1557	0.1522	0.0355	0.0490

To estimate the regression functions for the two DGPs, we employed the NW estimator with bandwidths estimated through Bayesian sampling and CV, as well as the conventional frequency estimator. The summary statistics for the in-sample and out-of-sample ASE values for both DGPs are tabulated in Table 2.

For the second DGP, Bayesian sampling leads to a slightly better NW estimator than CV in most cases based on the summary statistics of ASE values, but these differences are not significant because the Kruskal-Wallis test cannot reject the null hypothesis of no differences between the three methods for all three sample sizes. Using the out-of-sample ASE values, the MCS procedure finds in favour of the Bayesian approach for $n = 100$ and the CV method for $n = 200$ and 500.

For the third DGP, the Kruskal-Wallis test clearly rejects the null hypothesis of no difference between the three methods for all values of n. Using Dunn's test with Bonferroni correction, we found that the Bayesian method differs significantly from the CV and frequency methods respectively. However, the CV method does not differ significantly from the frequency method. The MCS procedure finds that under the out-of-sample ASE, the Bayesian method performs better than the CV method, and both methods outperform the frequency approach.

4.1.2. Experiment 2: Ordered and Unordered Categorical Regressors

The fourth and fifth DGPs were used to examine the difference between kernel functions for ordered categorical variables given by (4) and those for unordered categorical variables given by (3). The fourth DGP includes ordered categorical and continuous variables and is expressed by

$$y_i = x_{i,1}^{(c)} + x_{i,2}^{(c)} + x_{i,1}^{(d)} + x_{i,2}^{(d)} + u_i,$$

where $x_{i,j}^{(c)}$ is drawn from $N(0,1)$, for $j = 1$ and 2, $x_{i,j}^{(d)}$ takes values from $\{0, 1, \cdots, 5\}$ with an equal probability of $1/6$, for $j = 1$ and 2, and the error term u_i is drawn from $N(0,1)$, for $i = 1, 2, \cdots, n$.

Let $\boldsymbol{x}^{(c)}$ denote the vector of two continuous regressors and $\boldsymbol{x}^{(d)}$ the vector of two discrete regressors. The relationship between the response and regressors is modeled by

$$y_i = m\left(x_i^{(c)}, \boldsymbol{x}_i^{(d)}\right) + u_i, \tag{10}$$

where u_i, for $i = 1, 2, \cdots, n$, are assumed to be independent. Bandwidths in the NW regression estimator are estimated through the proposed Bayesian method and the CV approach each applied in two different ways; first using the kernel for ordered categorical variables given by (4) and then using the kernel for unordered categorical variables given by (3). Together with the conventional frequency estimator, this means that we are now evaluating five different estimators. We expect that the use of the ordered kernel should dominate the use of the unordered kernel because the observations of the discrete regressors have a natural order. Sample sizes of $n = 100$, 200, 500 and 1000 were used and the results are presented in Table 3.

The Kruskal-Wallis test has a p value of zero for all values of n thus rejecting the null hypothesis of no difference in performance among the five estimators. According to the in-sample and out-of-sample ASE measures, the Bayesian method performs slightly better than the corresponding CV method, and the use of the kernel function for the ordered variables outperforms the use of the kernel function for the unordered variables. The Bayesian method using the ordered kernel almost always has the lowest values of all four summary statistics for both in-sample and out-of-sample ASEs. Applying Dunn's test with Bonferroni correction to the in-sample ASEs, we found that the Bayesian method with ordered kernel is similar to the Bayesian method with unordered kernel and CV with ordered kernel, but differs from the CV method with unordered kernel and frequency method for all sample sizes. The Bayesian method with unordered kernel differs from the CV method with ordered and unordered kernels and frequency method when $n = 100$. As n increases, the Bayesian method differs only from the CV with unordered kernel and frequency method. There are significant differences between the CV method with ordered and unordered kernels and frequency method. The MCS procedure finds that the ordered kernel should be used for ordered categorical variables, and the Bayesian method performs better than the CV method.

Table 3. Mean, median and variation measures of the ASE values derived through 1000 samples simulated according to the fourth DGP, where the bolded numbers represent the minimum summary statistics of ASEs. Note that "order" refers to the kernel function for ordered categorical variables given by (4), and "unorder" refers to the kernel function for unordered categorical variables given by (3).

n	Method	In-Sample ASE				Out-of-Sample ASE			
		Mean	**Median**	**SD**	**IQR**	**Mean**	**Median**	**SD**	**IQR**
100	Bayesian (order)	**0.5035**	0.4925	**0.1044**	**0.1306**	**0.6236**	**0.5903**	**0.2255**	0.2993
	Bayesian (unorder)	0.5127	0.4995	0.1096	0.1371	0.6382	0.6003	0.2588	**0.2948**
	CV (order)	0.5535	**0.4828**	0.1055	0.1348	0.6553	0.6163	0.2469	0.3138
	CV (unorder)	0.6333	0.6232	0.1097	0.1411	0.7958	0.7660	0.2902	0.3525
	Frequency	0.8950	0.8516	0.2504	0.3071	1.9830	1.9150	0.6116	0.7890
200	Bayesian (order)	**0.3758**	**0.3689**	0.0615	0.0793	**0.4085**	**0.3978**	**0.1128**	**0.1525**
	Bayesian (unorder)	0.3845	0.3794	0.0619	0.0790	0.4124	0.3998	0.1134	0.1554
	CV (order)	0.3812	0.3755	**0.0611**	**0.0778**	0.4252	0.4098	0.1205	0.1641
	CV (unorder)	0.4968	0.4918	0.0647	0.0844	0.5456	0.5351	0.1497	0.1928
	Frequency	0.6438	0.6261	0.1348	0.1747	1.0660	1.0430	0.2816	0.3550
500	Bayesian (order)	**0.2610**	**0.2580**	**0.0315**	**0.0420**	**0.2466**	**0.2434**	**0.0526**	**0.0726**
	Bayesian (unorder)	0.2646	0.2606	0.0381	0.0441	0.2489	0.2441	0.0600	0.0734
	CV (order)	0.2633	0.2595	0.0322	0.0438	0.2562	0.2529	0.0563	0.0770
	CV (unorder)	0.3458	0.3425	0.0343	0.0455	0.3325	0.3280	0.0663	0.0885
	Frequency	0.4204	0.4146	0.0610	0.0746	0.4895	0.4883	0.0907	0.1220
1000	Bayesian (order)	**0.1948**	**0.1947**	**0.0185**	**0.0249**	**0.1714**	**0.1715**	0.0287	**0.0367**
	Bayesian (unorder)	0.1965	0.1958	0.0230	0.0253	0.1718	0.1717	0.0286	0.0380
	CV (order)	0.1968	0.1955	0.0298	0.0251	0.1731	0.1728	**0.0282**	0.0370
	CV (unorder)	0.2554	0.2557	0.0210	0.0285	0.2300	0.2289	0.0367	0.0466
	Frequency	0.3040	0.3026	0.0353	0.0486	0.2992	0.2997	0.0450	0.0573

The fifth DGP includes ordered categorical and continuous regressors and is given by

$$y_i = 1 + \sqrt{x_i^{(d)}} + x_i^{(c)} + u_i, \text{ for } i = 1, 2, \cdots, n,$$

where $x_i^{(d)}$ is drawn from $\{0, 1, \cdots, 4\}$ with an equal probability of $1/5$, $x_i^{(c)}$ is drawn from $N(0,1)$, and the error term u_i is drawn from $N(0,1)$. As the response variable is affected by the discrete regressor through its square root, the actual distance between categories 0 and 1 is 1, between categories 1 and 2 is $\sqrt{2} - 1 \approx 0.41$, between categories 2 and 3 is $\sqrt{3} - \sqrt{2} \approx 0.32$, and between categories 3 and 4 is $\sqrt{4} - \sqrt{3} \approx 0.27$ (see also [4]).

The purpose of this simulation is to investigate the five estimators in a situation where the distance between any pair of successive observations is not a fixed constant. Sample sizes of $n = 50$, 100 and 200 were used, and the results are presented in Table 4.

Again the Kruskal-Wallis test has a p value of zero for all values of n thus rejecting the null hypothesis of no difference in performance among the five estimators. This time the Bayesian method using the unordered kernel almost always has the lowest values of all four summary statistics for both in-sample and out-of-sample ASEs. Differences diminish as n increases. According to Dunn's test with Bonferroni criterion, the Bayesian method with ordered and unordered kernels behave similarly, while the CV method with unordered kernel performs similarly to the frequency method. There are significant differences among the Bayesian method with unordered kernel, CV with ordered and unordered kernels, and frequency methods. Based on the in-sample ASEs, the ordered Bayesian method is also better than both CV methods and the frequency approach for all values of n. Not surprisingly, the MCS procedure finds in favour of the Bayesian method using the unordered kernel for all n values.

What is surprising is that differences between the CV method with unordered kernel and the frequency approach are not significant. In order to examine the role of the choice of kernel in the relative performance of the various estimators, we repeated the last simulation using Wang and Van Ryzin's kernel function [43], where bandwidths were estimated through Bayesian, CV and frequency approaches. Table 5 presents a summary of descriptive statistics of the in-sample and out-of-sample ASE values. Under the criterion of in-sample ASE, the frequency approach is comparable to the Bayesian method, but both perform poorer than CV. However, under the criterion of out-of-sample ASE, the Bayesian approach is comparable to CV, but both perform better than the frequency approach. The Kruskal-Wallis test reveals no significant differences between the three methods for all values of n.

Table 4. Mean, median and variation measures of the ASE values derived through 1000 samples simulated according to the fifth DGP. The bolded numbers represent the minimum summary statistics of ASEs.

n	Method	In-Sample ASE				Out-of-Sample ASE			
		Mean	**Median**	**SD**	**IQR**	**Mean**	**Median**	**SD**	**IQR**
50	Bayesian (order)	0.2134	0.1979	0.0919	0.1255	0.2410	0.2110	0.1343	0.1570
	Bayesian (unorder)	**0.2090**	**0.1939**	**0.0884**	0.1203	**0.2291**	**0.2006**	**0.1266**	**0.1494**
	CV (order)	0.2311	0.2135	0.0994	**0.1170**	0.2670	0.2260	0.1674	0.1780
	CV (unorder)	0.2785	0.2628	0.1116	0.1416	0.3023	0.2540	0.1996	0.2267
	Frequency	0.2797	0.2638	0.1130	0.1423	0.3886	0.3525	0.1982	0.2371
100	Bayesian (order)	0.1349	0.1282	0.0530	0.0723	0.1375	0.1228	0.0702	0.0801
	Bayesian (unorder)	**0.1340**	**0.1272**	**0.0521**	0.0700	**0.1342**	**0.1216**	**0.0679**	**0.0776**
	CV (order)	0.1605	0.1523	0.0530	**0.0684**	0.1480	0.1300	0.0795	0.0915
	CV (unorder)	0.1755	0.1676	0.0628	0.0793	0.1610	0.1412	0.0898	0.1041
	Frequency	0.1758	0.1677	0.0628	0.0794	0.2002	0.1852	0.0890	0.1069
200	Bayesian (order)	0.0842	**0.0808**	0.0283	0.0401	0.0790	0.0746	0.0332	0.0404
	Bayesian (unorder)	**0.0841**	0.0810	**0.0282**	0.0393	**0.0782**	**0.0737**	**0.0328**	**0.0398**
	CV (order)	0.1178	0.1151	0.0298	**0.0383**	0.0846	0.0786	0.0382	0.0434
	CV (unorder)	0.1075	0.1044	0.0348	0.0450	0.0884	0.0815	0.0390	0.0478
	Frequency	0.1077	0.1048	0.0349	0.0450	0.1155	0.1033	0.0697	0.0538

Table 5. Mean, median and variation measures of the ASE values derived through 1000 samples simulated according to the fifth DGP, where Wang and Van Ryzin's kernel function [43] is used. The bolded numbers represent the minimum summary statistics of ASEs.

n	Method	In-Sample ASE				Out-of-Sample ASE			
		Mean	**Median**	**SD**	**IQR**	**Mean**	**Median**	**SD**	**IQR**
50	Bayesian	0.2315	0.2202	**0.0865**	**0.1153**	**0.2601**	0.2296	**0.1417**	**0.1660**
	CV	**0.2158**	**0.1958**	0.0970	0.1172	0.2625	0.2245	0.1614	0.1715
	Frequency	0.2302	0.2132	0.0982	0.1174	0.3992	0.3519	0.2197	0.2497
100	Bayesian	0.1639	0.1554	**0.0516**	**0.0682**	0.1492	0.1322	**0.0781**	**0.0920**
	CV	**0.1439**	**0.1369**	0.0553	0.0687	**0.1465**	**0.1309**	0.0783	0.0924
	Frequency	0.1610	0.1531	0.0539	0.0701	0.2176	0.1954	0.1118	0.1215
200	Bayesian	0.1191	0.1162	**0.0288**	**0.0383**	**0.0830**	**0.0772**	**0.0354**	**0.0429**
	CV	**0.0983**	**0.0948**	0.0309	0.0385	0.0843	0.0778	0.0387	0.0437
	Frequency	0.1181	0.1151	0.0304	0.0384	0.1208	0.1093	0.0564	0.0568

4.1.3. Experiment 3: Time Series Regression with Continuous and Binary Regressors

The sixth DGP includes a time series response variable explained by its lagged variable, a continuous time series regressor x_t and a binary regressor I_t, which is generated based on another time series. This model is expressed as

$$\begin{aligned} y_t &= 0.25y_{t-1} + 0.6x_t + r(I_t - 0.5) + \varepsilon_t, \\ \varepsilon_t &= \sigma_{y,t}e_t, \\ \sigma^2_{y,t} &= 0.02 + 0.05\varepsilon^2_{t-1} + 0.9\sigma^2_{y,t-1}, \end{aligned} \tag{11}$$

for $t = 1, 2, \cdots, n$, where r is a random number that follows the uniform distribution on $(0, 1)$, and $e_1, e_2, \cdots, e_n$ are iid $N(0, 0.49)$. The continuous regressor x_t is generated as

$$\begin{aligned} x_t &= \sigma_{x,t}u_t, \\ \sigma^2_{x,t} &= 0.02 + 0.05x^2_{t-1} + 0.9\sigma^2_{x,t-1}, \end{aligned}$$

where $u_1, u_2, \cdots, u_n$ are iid $N(0, 1)$. The binary regressor I_t is an indicator of z_t and equals one with a positive z_t and zero otherwise, where z_t is generated as

$$\begin{aligned} z_t &= \sigma_{z,t}v_t, \\ \sigma^2_{z,t} &= 0.02 + 0.05z^2_{t-1} + 0.9\sigma^2_{z,t-1}, \end{aligned}$$

where $v_1, v_2, \cdots, v_n$ are iid $N(0, 1)$. The variance-covariance matrix of $(e_t, u_t, v_t)'$, for $t = 1, 2, \cdots, n$, is

$$\begin{pmatrix} 0.49 & 0.25 & 0.25 \\ 0.25 & 1 & 0.25 \\ 0.25 & 0.25 & 1 \end{pmatrix}.$$

Sample sizes of $n = 100$, 200 and 500 were used in this experiment and the results are presented in Table 6.

Table 6. Mean, median and variation measures of the ASE values derived through 1,000 samples simulated according to the sixth DGP. The bolded numbers represent the minimum summary statistics of ASEs.

n	Method	In-Sample ASE				Out-of-Sample ASE			
		Mean	**Median**	**SD**	**IQR**	**Mean**	**Median**	**SD**	**IQR**
100	Bayesian	**0.3825**	**0.3464**	**0.1677**	**0.1693**	**0.3558**	**0.3233**	**0.1652**	**0.1656**
	CV	0.4107	0.3567	0.2078	0.1959	0.4449	0.3733	0.2752	0.2405
	Frequency	0.4123	0.3573	0.2075	0.1952	0.4372	0.3657	0.2622	0.2264
200	Bayesian	**0.2930**	**0.2773**	**0.0854**	**0.0954**	**0.2769**	**0.2568**	**0.0925**	**0.0950**
	CV	0.3044	0.2790	0.1085	0.1064	0.3249	0.2936	0.1491	0.1396
	Frequency	0.3032	0.2786	0.1043	0.1045	0.3239	0.2939	0.1388	0.1400
500	Bayesian	**0.2367**	0.2309	**0.0417**	**0.0464**	**0.2289**	**0.2216**	**0.0424**	**0.0484**
	CV	0.2397	**0.2298**	0.0476	0.0516	0.2528	0.2390	0.0624	0.0658
	Frequency	0.2397	**0.2298**	0.0480	0.0520	0.2554	0.2433	0.0629	0.0684

The Kruskal-Wallis test rejects the null hypothesis of no difference between the ASEs of the three methods for $n = 100$ but not for $n = 200$ and 500. The Bayesian method always has the smallest mean ASE and almost always has the smallest median across both in-sample and out-of-sample ASEs. According to Dunn's test with Bonferroni correction, the Bayesian method significantly differs from the frequency method, while the CV method performs similarly to the Bayesian and frequency methods for $n = 100$. For this sample size, the in-sample and out-of-sample mean ASEs obtained through

CV are respectively, 8.8% and 23.2% larger than those for the Bayesian method. Based on the MCS procedure applied to the out-of-sample ASEs, the Bayesian method is significantly better than its two counterparts.

4.1.4. Experiment 4: The Presence of Irrelevant Regressors

In a nonparametric regression model, the NW estimator has the ability to smooth out irrelevant regressors through a data-driven bandwidth selection method such as CV by choosing very large bandwidth parameters for those regressors. This has important implications for variable selection (see [5]). This experiment compares the Bayesian and CV methods in the presence of two irrelevant regressors.

The seventh DGP generates variables as follows. Binary variables $z_{1,i}$ and $z_{2,i}$ were generated so that $\Pr\{z_{1,i} = 1\} = 0.69$ and $\Pr\{z_{2,i} = 1\} = 0.73$ and continuous variables $x_{1,i}$ and $x_{2,i}$ were generated as independent standard normal values for $i = 1, 2, \ldots, n$. They were generated so there is no correlation between any pair of regressors. The response variable was generated via

$$y_i = z_{1,i} + x_{1,i} + \epsilon_i,$$

where $\epsilon_i \sim N(0, 1)$, for $i = 1, 2, \ldots, n$. In order to examine whether irrelevant regressors can be smoothed out by assigning them large bandwidths, we estimate the following model:

$$y_i = z_{1,i} + z_{2,i} + x_{1,i} + x_{2,i} + \epsilon_i,$$

for $i = 1, 2, \ldots, n$. This model means that z_2 and x_2 are irrelevant regressors. Samples sizes of $n = 100$ and 250 were used. Summary statistics of the estimated bandwidths for 1000 iterations are presented in Table 7 with the summary statistics for the ASEs of the two methods given in Table 8.

Table 7. Summary statistics of the estimated bandwidths derived through 1000 repetitions simulated according to the seventh DGP.

n	Method	Bandwidth	Mean	Median	SD	Quantile	
						2.5%	97.5%
100	Bayesian	λ_1	0.1247	0.1095	0.0717	0.0336	0.3117
		λ_2	0.3079	0.3223	0.0768	0.1255	0.4165
		h_1	0.3593	0.3570	0.0938	0.1951	0.5529
		h_2	1.7690	1.5302	1.2477	0.3495	4.2122
	CV	λ_1	0.0612	0.0449	0.0723	0.0000	0.2403
		λ_2	0.3810	0.5000	0.1602	0.0040	0.5000
		h_1	0.3654	0.3796	0.1094	0.1196	0.5472
		h_2	6.53×10^6	1.41×10^6	1.14×10^7	0.3829	3.65×10^7
250	Bayesian	λ_1	0.0676	0.0614	0.0302	0.0265	0.1449
		λ_2	0.3291	0.3466	0.0762	0.1524	0.4300
		h_1	0.3153	0.3181	0.0606	0.1932	0.4419
		h_2	2.2348	2.1492	1.1106	0.5045	4.5167
	CV	λ_1	0.0302	0.0290	0.0235	0.0000	0.0839
		λ_2	0.3966	0.5000	0.1414	0.0573	0.5000
		h_1	0.3209	0.3289	0.0682	0.1614	0.4385
		h_2	4.29×10^6	1.45×10^6	7.36×10^6	0.5236	2.38×10^7

Table 8. Mean, median and variation measures of the ASE values derived through 1000 samples simulated according to the sixth DGP. The bolded numbers represent the minimum summary statistics of ASEs.

n	Method	In-Sample ASE				Out-of-Sample ASE			
		Mean	**Median**	**SD**	**IQR**	**Mean**	**Median**	**SD**	**IQR**
100	Bayesian	**0.1492**	0.1309	**0.0879**	**0.0908**	**0.1291**	0.1068	**0.1017**	**0.0922**
	CV	0.1509	**0.1213**	0.1081	0.1029	0.1304	**0.1017**	0.1057	0.0996
250	Bayesian	0.0694	0.0630	0.0401	**0.0380**	0.0545	0.0481	0.0445	0.0329
	CV	**0.0667**	**0.0590**	**0.0346**	0.0386	**0.0515**	**0.0458**	**0.0275**	**0.0328**

For the proposed Bayesian bandwidth selection method, the average bandwidth chosen for z_2 is larger than that chosen for z_1 by 1.47 times when $n = 100$ (3.87 times when $n = 250$), and the average bandwidth chosen for x_2 is larger than that chosen for x_1 by 3.92 times when $n = 100$ (3.87 times when $n = 250$). This shows that due to their large bandwidths, irrelevant regressors may not explain too much about the response variable in comparison to their corresponding relevant regressors.

When CV is used for bandwidth choice, we find that the average bandwidth chosen for z_2 is larger than the one chosen for z_1 by 5.23 times when $n = 100$ (12.13 times when $n = 250$). Moreover, the average bandwidth chosen for x_2 is larger than that chosen for x_1 by 1.79×10^7 times when $n = 100$ (1.34×10^7 times when $n = 250$). Clearly, CV leads to a larger bandwidth than the Bayesian approach for each irrelevant regressor.

With respect to the ASE measure of accuracy, the proposed Bayesian sampling method leads to slightly better accuracy than CV when $n = 100$ with CV being better when $n = 250$. The Kruskal-Wallis test rejects the null hypothesis at the 5% significance level of no difference between the two methods for both sample sizes. Having implemented Dunn's test with Bonferroni criterion, we found that the Bayesian method differs significantly from the CV method for all sample sizes. The MCS procedure applied to the out-of-sample ASEs finds the Bayesian method is significantly better when $n = 100$ and CV is significantly better when $n = 250$.

In the above seven simulation studies, we only consider certain types of discrete variables such as binary and categorical variables. We have not considered DGPs that allow for other possible types of discrete variables. In terms of other types of discrete variables, [44,45] studied performance of several associate kernels including the binomial kernel for count data in nonparametric regression models.

4.2. Accuracy of the Error Density Estimator

The proposed Bayesian sampling algorithm is based on the assumption of a kernel-form error density given by (6), whose bandwidth is sampled at the same time as when bandwidths of the NW estimator are sampled. Upon completion of the sampling algorithm, we also obtain a kernel density estimator of the error density. However, when CV is used for selecting bandwidths for the NW estimator, one may obtain the kernel density estimator of residuals, but its bandwidth has to be selected based on residuals through an existing bandwidth selector such as likelihood cross-validation. Thus, it requires two stages of using the cross-validation method to select bandwidths for the NW estimator and the kernel density estimator of residuals, and we call it two-stage CV.

The performance of a kernel estimator of the error density denoted as $\widehat{f}(\cdot)$, is examined by its integrated squared errors (ISE). In our Monte Carlo simulation studies, the ISE was numerically approximated at 1001 equally spaced grid points on $[-5, 5]$:

$$\text{ISE} \approx \sum_{i=1}^{1001} \left\{ f\left(-5 + \frac{i-1}{100}\right) - \widehat{f}\left(-5 + \frac{i-1}{100}\right) \right\}^2 \times \frac{10}{1000}.$$

The mean ISE (MISE) was approximated by the mean of ISE values derived from the 1000 Monte Carlo replications for each DGP. The in-sample MISE of the kernel estimator of error density with its bandwidth chosen through Bayesian sampling or the two-stage CV for all seven DGPs are presented in Table 9. For any DGP and any sample size considered, Bayesian sampling obviously outperforms the two-stage CV in estimating the bandwidth for the kernel estimator of the error density, based on the mean of in-sample ISE values. For the third and fifth DGPs, the Bayesian method outperforms the two-stage CV in all four summary statistics.

Table 9. Mean, median and variation measures of the in-sample integrated squared errors (ISE) values derived through 1000 repetitions for all seven DGPs, where the bandwidths of the kernel error density estimator were chosen through Bayesian sampling and two-stage cross-validation (CV). The bolded numbers represent the minimum summary statistics of ISEs.

DGP	*n*	Bayesian				Two-Stage CV			
		Mean	**Median**	**SD**	**IQR**	**Mean**	**Median**	**SD**	**IQR**
1	100	**0.0085**	**0.0055**	**0.0085**	**0.0071**	0.0241	0.0090	0.0421	0.0219
	200	**0.0039**	0.0030	**0.0031**	**0.0031**	0.0047	**0.0028**	0.0057	0.0041
	500	**0.0015**	**0.0012**	**0.0012**	**0.0012**	0.0018	0.0015	**0.0012**	**0.0012**
2	100	**0.0099**	**0.0056**	**0.0295**	**0.0073**	0.0165	0.0067	0.0347	0.0107
	200	**0.0041**	0.0029	**0.0037**	**0.0031**	0.0047	**0.0028**	0.0059	0.0032
	500	**0.0020**	0.0014	**0.0071**	**0.0011**	0.0033	**0.0013**	0.0343	**0.0011**
3	100	**0.0102**	**0.0059**	**0.0161**	**0.0074**	0.0220	0.0086	0.0449	0.0163
	200	**0.0043**	**0.0030**	**0.0058**	**0.0033**	0.0062	0.0036	0.0086	0.0047
	500	**0.0017**	**0.0014**	**0.0014**	**0.0012**	0.0019	**0.0014**	0.0017	0.0014
4	100	**0.0199**	**0.0107**	**0.0305**	**0.0161**	0.0468	0.0181	0.1208	0.0394
	200	**0.0065**	**0.0040**	0.0092	**0.0044**	0.0072	0.0046	**0.0090**	0.0062
	500	**0.0023**	**0.0017**	0.0020	**0.0016**	0.0024	0.0018	**0.0019**	0.0019
	1000	**0.0013**	**0.0011**	**0.0010**	**0.0009**	0.0018	0.0013	0.0056	0.0012
5	50	**0.0101**	**0.0068**	**0.0096**	**0.0080**	0.0156	0.0101	0.0262	0.0111
	100	**0.0056**	**0.0040**	**0.0051**	**0.0041**	0.0082	0.0064	0.0081	0.0059
	200	**0.0030**	**0.0025**	**0.0022**	**0.0021**	0.0048	0.0041	0.0028	0.0032
6	100	**0.0940**	**0.0938**	0.0164	0.0207	0.1138	0.1165	**0.0157**	**0.0173**
	200	**0.0894**	**0.0899**	**0.0111**	0.0147	0.1055	0.1068	0.0107	**0.0129**
	500	**0.0845**	**0.0846**	**0.0068**	**0.0091**	0.0966	0.0972	0.0072	0.0093
7	100	**0.0168**	0.0073	**0.0687**	0.0121	0.0311	**0.0054**	0.2599	**0.0096**
	250	**0.0027**	0.0022	**0.0020**	**0.0024**	0.0031	**0.0020**	0.0041	0.0025

With all the bandwidths chosen based on in-sample observations for each DGP, we calculated the out-of-sample MISE of the kernel estimator of the error density. All the out-of-sample MISE values were tabulated in Table 10. We found that for the second, fifth and sixth DGPs, Bayesian sampling slightly outperforms two-stage CV in terms of all four summary statistics, and for the other three DGPs, results obtained from Bayesian sampling are comparable to those obtained from two-stage CV. Out of all 19 cases of different DGPs and different sample sizes, Bayesian sampling performs better than the two-stage CV for 12 cases, while the latter performs better in 2 cases. Both perform similarly in 5 cases.

Table 10. Mean, median and variation measures of the out-of-sample ISE values derived through 1000 repetitions for all seven DGPs, where the bandwidths of the kernel error density estimator were chosen through Bayesian sampling and two-stage CV. The bolded numbers represent the minimum summary statistics of ISEs.

DGP	n	Bayesian				Two-Stage CV			
		Mean	**Median**	**SD**	**IQR**	**Mean**	**Median**	**SD**	**IQR**
1	100	0.0013	0.0009	**0.0013**	0.0013	**0.0012**	**0.0008**	**0.0013**	**0.0011**
	200	**0.0007**	**0.0005**	**0.0006**	**0.0006**	**0.0007**	**0.0005**	0.0007	0.0006
	500	**0.0004**	**0.0003**	**0.0003**	**0.0004**	0.0006	0.0005	0.0004	0.0005
2	100	**0.0021**	0.0018	**0.0015**	0.0018	0.0023	**0.0017**	0.0021	**0.0017**
	200	**0.0012**	**0.0010**	**0.0009**	**0.0010**	**0.0012**	**0.0010**	**0.0009**	**0.0010**
	500	**0.0006**	**0.0005**	**0.0004**	**0.0005**	**0.0006**	**0.0005**	**0.0004**	**0.0005**
3	100	**0.0017**	**0.0012**	0.0021	**0.0013**	0.0021	0.0018	**0.0015**	0.0016
	200	**0.0008**	**0.0007**	**0.0007**	**0.0007**	0.0013	0.0012	0.0009	0.0010
	500	**0.0004**	**0.0003**	**0.0003**	**0.0003**	0.0008	0.0008	0.0004	0.0005
4	100	**0.0011**	**0.0008**	**0.0009**	**0.0009**	0.0012	0.0009	0.0011	0.0010
	200	**0.0006**	**0.0005**	**0.0004**	**0.0005**	0.0007	0.0006	**0.0004**	**0.0005**
	500	**0.0003**	**0.0003**	**0.0002**	**0.0002**	0.0004	0.0004	**0.0002**	0.0003
	1000	**0.0002**	**0.0002**	**0.0001**	**0.0002**	0.0003	0.0003	0.0002	**0.0002**
5	50	0.0040	0.0030	**0.0032**	0.0030	**0.0038**	**0.0029**	0.0041	**0.0028**
	100	**0.0023**	0.0020	**0.0016**	0.0019	**0.0023**	**0.0019**	0.0018	**0.0018**
	200	**0.0014**	**0.0012**	**0.0009**	0.0010	**0.0014**	**0.0012**	0.0010	**0.0009**
6	100	**0.0914**	**0.0896**	**0.0225**	**0.0298**	0.1085	0.1075	0.0253	0.0315
	200	**0.0892**	**0.0879**	**0.0165**	**0.0215**	0.1031	0.1026	0.0185	0.0253
	500	**0.0886**	**0.0885**	**0.0105**	**0.0137**	0.0991	0.0989	0.0115	0.0155
7	100	**0.0599**	**0.0594**	**0.0160**	**0.0220**	0.0627	0.0624	0.0161	0.0222
	250	**0.0588**	**0.0584**	**0.0108**	0.0148	0.0603	0.0601	**0.0108**	**0.0143**

To summarize, we have found that for all seven DGPs considered, our Bayesian sampling approach outperforms its competitor, the two-stage CV, in estimating bandwidths for the NW estimator of the regression function and kernel estimator of error density.

4.3. Sensitivity of Prior Choices

To examine the sensitivity of prior choices, we change the priors in two ways. First, we keep the same prior densities, namely inverse Gamma, as before but alter the values of hyperparameters. Second, we change the prior densities for squared bandwidth parameters from the inverse Gamma density to log normal density. When we focus on bandwidth parameters, the use of Cauchy prior densities has been considered by [46]. With one sample of size 500 generated through the first DGP, we derived the Markov chain Monte Carlo (MCMC) simulation results using different priors. The results are summarized in Table 11. We use the SIF to monitor the mixing performance of a simulated chain. The last column of Table 11 shows that the mixing performance is not particularly sensitive to different choices of the prior density.

Table 11. Arithmetic mean, 95% Bayesian credible interval, standard deviation, batch-mean standard deviation and simulation inefficiency factor (SIF) of each simulated chain obtained under different choices of priors for the first DGP with a sample size of 500.

Prior	Parameter	Mean	95% Bayesian Credible Interval	Standard Deviation	Batch–mean Standard dev.	SIF
IG(1, 0.05)	λ_1	0.0295	(0.0046, 0.0575)	0.0129	0.0007	26.32
	λ_2	0.0205	(0.0007, 0.0428)	0.0117	0.0007	32.49
	λ_3	0.0245	(0.0018, 0.0518)	0.0134	0.0008	33.59
	λ_4	0.0318	(0.0040, 0.0595)	0.0129	0.0008	38.90
	h	0.1349	(0.1135, 0.1576)	0.0120	0.0004	13.18
	b	0.2685	(0.1543, 0.3970)	0.0621	0.0013	4.23
IG(5, 0.25)	λ_1	0.0267	(0.0015, 0.0555)	0.0135	0.0008	36.30
	λ_2	0.0216	(0.0021, 0.0469)	0.0115	0.0007	36.89
	λ_3	0.0244	(0.0020, 0.0522)	0.0132	0.0008	36.01
	λ_4	0.0342	(0.0080, 0.0618)	0.0135	0.0007	27.50
	h	0.1471	(0.1263, 0.1739)	0.0124	0.0004	10.59
	b	0.2572	(0.1742, 0.3546)	0.0481	0.0012	6.52
Cauchy(0, 1)	λ_1	0.0277	(0.0058, 0.0527)	0.0124	0.0008	46.30
	λ_2	0.0210	(0.0016, 0.0457)	0.0118	0.0008	46.55
	λ_3	0.0226	(0.0011, 0.0479)	0.0127	0.0008	38.07
	λ_4	0.0321	(0.0058, 0.0557)	0.0133	0.0008	34.85
	h	0.1338	(0.1093, 0.1608)	0.0130	0.0004	11.20
	b	0.3000	(0.1651, 0.4301)	0.0661	0.0015	5.34
Log N(0,1)	λ_1	0.0268	(0.0006, 0.0528)	0.0123	0.0007	28.91
	λ_2	0.0204	(0.0028, 0.0455)	0.0111	0.0007	38.04
	λ_3	0.0229	(0.0012, 0.0503)	0.0125	0.0005	17.91
	λ_4	0.0332	(0.0065, 0.0614)	0.0134	0.0006	22.93
	h	0.1415	(0.1182, 0.1709)	0.0135	0.0005	12.33
	b	0.3392	(0.2241, 0.4623)	0.0588	0.0013	5.21

5. An Application to Modeling Stock Returns

The purpose of this study is to demonstrate the benefit of the proposed sampling algorithm for bandwidth estimation in comparison with the existing bandwidth selection method. We are interested in modeling the daily return of the All Ordinaries (Aord) index in the Australian stock market, where one explanatory variable is the overnight daily return of the S&P 500 index because from the beginning of 2007 onwards, the US has had a leading effect on other markets worldwide. Such a nonparametric regression model was previously studied by [12] to demonstrate their sampling algorithm for bandwidth estimation.

Although the Australian stock market typically followed the overnight market movement in the US, there are some exceptions where the Australian market moves in the opposite direction. This motivated us to look for another explanatory variable, and one such variable is an indicator of a major stock market in the European zone. The indicator was expected to suggest the market movement in Australia because the US stock market was also supposed to have a leading effect on European stock markets. Therefore, we model the Aord daily return as an unknown function of the overnight S&P 500 return and a binary variable indicating whether the overnight FTSE index went up or down. This nonparametric regression model should better reveal the actual relationship between the Australian stock market and the US market than the model investigated by [12], where only a continuous regressor was considered.

5.1. Data

We downloaded daily closing index values of the Aord, S&P 500 and FTSE between January 3, 2007 and October 1, 2012 from Yahoo Finance. Each value of the Aord index was matched to the

corresponding overnight values of the S&P 500 and FTSE indices. Whenever one market experienced a non-trading day, the trading data collected from all three markets on that day were excluded. The sample contains $n = 1373$ observations of the daily continuously compounded return of each index.

We fitted the nonparametric regression model given by

$$y_i = m(x_{1,i}, x_{2,i}) + \varepsilon_i, \qquad \text{for } i = 1, 2, \cdots, n, \tag{12}$$

to the sample data, where y_i is the Aord daily return, $x_{i,1}$ is the S&P 500 daily return, $x_{i,2}$ is a binary regressor taking the value of one if the FTSE daily return is positive and zero otherwise, and $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ are assumed to be iid and follow a distribution with its density given by (6).

5.2. Results

The proposed Bayesian sampling algorithm was employed to estimate bandwidths in the NW estimator and the kernel–form error density. The first row panel of Table 12 presents the estimates of bandwidths, their 95% Bayesian credible intervals and associated SIF values. These small SIF values indicate that the sampler achieved very good mixing. In our experience, a SIF value of no more than 100 usually indicates reasonable mixing.

Table 12. Estimated bandwidths for the regressors in the Nadaraya-Watson (NW) estimator and the kernel–form error density.

Bandwidth Selector	Parameter	Mean	95% Bayesian Credible Interval	Standard Deviation	Batch-Mean Standard Error	SIF
Bayesian	$h_{\text{S\&P 500}}$	0.6816	(0.5850, 0.7887)	0.0016	0.0533	9.1
	λ_{FTSE}	0.2630	(0.1221, 0.4076)	0.0033	0.0724	20.8
	b	0.4932	(0.4375, 0.5517)	0.0008	0.0297	6.5
Two-stage CV	$h_{\text{S\&P 500}}$	0.3354				
	λ_{FTSE}	0.8249				
	b	0.3181				

The second row panel of Table 12 presents the bandwidths selected through two-stage CV. The bandwidths for the continuous regressor and the kernel-form error density derived through two-stage CV are clearly different from those derived through Bayesian sampling.

With the first pair of out-of-sample observations of S&P 500 and FTSE returns, we can make the one-day-ahead forecast of the Aord return before the Australian stock market opens its trading. We collected observations of the S&P 500 and FTSE returns on October 1, 2012 (local time), and then used the nonparametric regression model given by (12) to make a point forecast of the Aord return on October 2, 2012. Such a point forecast was made at each iteration of the MCMC sampling procedure. Upon finishing the sampling procedure, we derived a posterior point forecast of the Aord return by averaging these forecasts made at all iterations. The point posterior forecast of the Aord return is 0.1984%, and its 95% Bayesian credible interval is $(0.1553\%, 0.2382\%)$. The actual Aord return on October 2, 2012 is 0.9842%. In comparison to the point forecast of the Aord return of 0.0416% obtained by the same nonparametric kernel regression with bandwidth selected by CV, the proposed Bayesian method leads to a closer forecast than CV.

The kernel-form error density given by (6) allows us to forecast the one-day-ahead density of the Aord return. According to (7), the density of y_{n+1} is

$$\widetilde{f}_y\left(y_{n+1}\middle|\boldsymbol{h}^2, \boldsymbol{\lambda}, b^2\right) = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{b}\phi\left(\frac{\{y_{n+1} - \widehat{m}\left(\boldsymbol{x}_{n+1}; \boldsymbol{h}, \boldsymbol{\lambda}\right)\} - \{y_j - \widehat{m}\left(\boldsymbol{x}_j; \boldsymbol{h}, \boldsymbol{\lambda}\right)\}}{b}\right), \tag{13}$$

where $\boldsymbol{x}_{n+1}$ is the vector of first out-of-sample observations of S&P 500 and FTSE returns. The one-day-ahead posterior predictive density is given by

$$f_{\boldsymbol{y}}\left(y_{n+1}|\boldsymbol{y}\right) = \int \widetilde{f}_{\boldsymbol{y}}\left(y_{n+1}|\boldsymbol{h}^2, \boldsymbol{\lambda}, b^2\right) \pi\left(\boldsymbol{h}^2, \boldsymbol{\lambda}, b^2|\boldsymbol{y}\right) d\boldsymbol{h}^2 d\boldsymbol{\lambda} db^2, \tag{14}$$

which we approximate by averaging (13) over the simulated chain of $\left(\boldsymbol{h}^2, \boldsymbol{\lambda}', b^2\right)'$. At each iteration during the sampling procedure, we calculated $\widetilde{f}_{\boldsymbol{y}}\left(y_{n+1}|\boldsymbol{h}^2, \boldsymbol{\lambda}, b^2\right)$ at 25,000 grid points with the simulated values of $\boldsymbol{h}^2$, $\boldsymbol{\lambda}$ and b^2 being plugged-in. Upon finishing the sampling procedure, we calculated the average of these calculated density values at each grid point. The posterior predictive cumulative density function (CDF) of the Aord daily return was obtained similarly. The posterior predictive density of the Aord return and its CDF are plotted in blue solid lines in Figures 1 and 2, respectively.

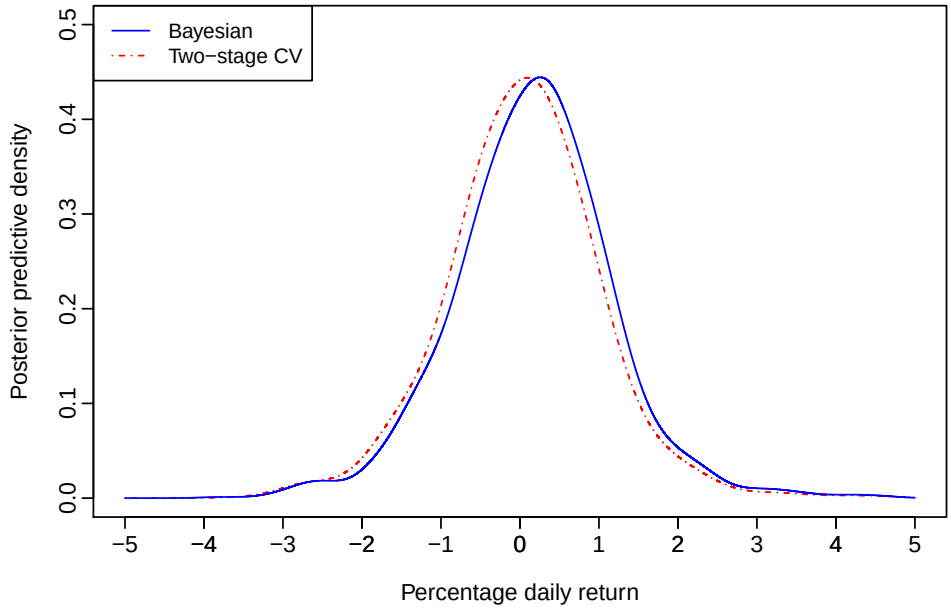

Figure 1. Graphs of posterior predictive density of the forecasted All Ordinaries return on October 2, 2012.

At the 95% and 99% confidence levels, the one-day-ahead VaRs were computed through the posterior predictive CDF. With bandwidths estimated through Bayesian sampling for the nonparametric regression model given by (12), the VaRs for a \$100 investment on the Aord index are respectively, \$1.4861 and \$2.5176 at the 95% and 99% confidence levels.

With bandwidths selected through the two–stage CV for the nonparametric regression model given by (12), the graphs of the one-day-ahead density forecast of the Aord return and its CDF are also plotted in red dot-dash lines in Figures 1 and 2, respectively. The 95% and 99% VaRs for a \$100 investment on the Aord index are respectively, \$1.6048 and \$2.5523, which are larger than the corresponding VaRs derived through Bayesian sampling.

It seems that the two-stage CV leads to an over-estimated VaR in comparison to its Bayesian counterpart. However, the above results were derived based on forecasted densities of one day's Aord return only. To further justify the empirical importance of our method, we checked the relative frequency of exceedance through rolling samples.

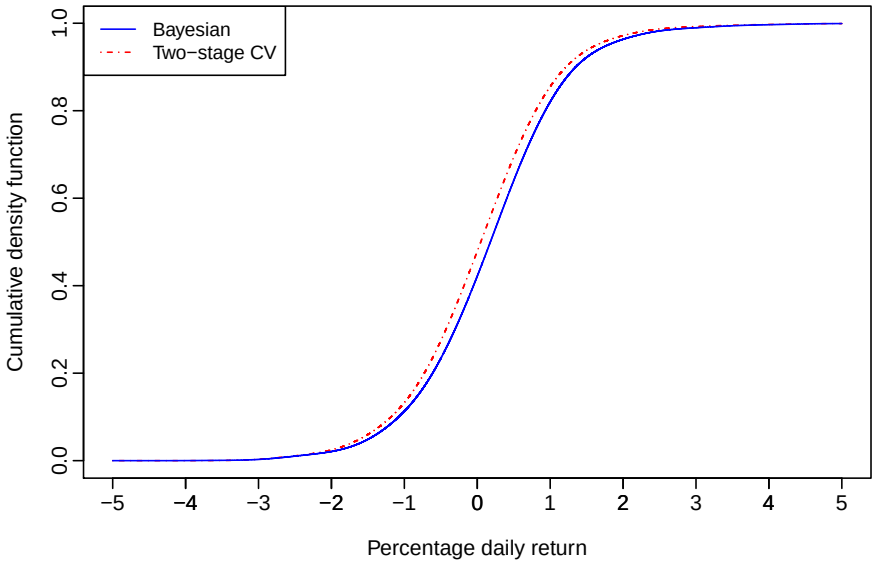

Figure 2. Graphs of the posterior predictive cumulative density of the forecasted All Ordinaries return on October 2, 2012.

5.3. Relative Frequency of Exceedance

The concept of exceedance refers to the phenomenon that the actual daily loss exceeds the estimated daily VaR during the same period of holding the invested asset. The relative frequency of exceedance is a measure of the accuracy of a VaR estimate. Therefore, we use this measure to evaluate the performance of the proposed Bayesian bandwidth estimation method in comparison to the two-stage CV for the nonparametric regression model with both continuous and binary regressors. The performance of the one-day-ahead forecasted VaR was examined by the relative frequency of exceedance derived through rolling samples. Let α denote the confidence level for computing VaRs. If the relative frequency of exceedance is close to $(1-\alpha)$, the underlying method for computing the VaR can be regarded as appropriate. The closer the relative frequency of exceedance is to $(1-\alpha)$, the better the underlying VaR estimation method would be.

In order to calculate the relative frequency of exceedance, the samples have a fixed size of 1000. During the whole sample period, the first sample contains the first 1000 observed vectors of the Aord, S&P 500 and FTSE returns, based on which we estimated bandwidths through Bayesian sampling and computed the VaRs at the 95% and 99% confidence levels. The second sample was derived by rolling the first sample forward one day. Based on the second sample, we did the same as for the previous sample. This procedure continued until the second last observation was included in the sample for estimating bandwidths. There are a total of 373 samples for forecasting the one-day-ahead VaRs.

With the daily VaRs forecasted through rolling samples, we calculated the relative frequency of exceedance at different α values with bandwidths chosen through Bayesian sampling and the two-stage CV. With Bayesian sampling, the resultant relative frequencies are respectively, 0.80% and 4.81% at the 99% and 95% confidence levels. However, with two-stage CV method, the corresponding relative frequencies of exceedance are 1.07% and 6.95%, respectively. It shows that two-stage CV for bandwidth selection leads to under-estimated VaRs in comparison to our proposed sampling method, particularly at the 95% confidence level.

6. Conditional Density Estimation of GDP Growth Rates

Given the close relationship between the conditional mean regression and conditional density estimation, the sampling algorithm proposed in Section 3 can be modified for the purpose of choosing bandwidths in kernel density estimation of continuous and discrete variables. Maasoumi, Racine, and Stengos [7] investigated kernel density estimation of the gross domestic product (GDP) growth rate

among OECD and non-OECD countries from 1965 to 1995, where the OECD status of the country and the year of the observed growth rate are included in the data matrix. They aimed to estimate the trivariate density of growth rates (denoted as $x^{(c)}$), OECD status and year (denoted as $\boldsymbol{x}^{(d)}$), where the last two variables are respectively, binary and ordered categorical. Their primary purpose was to explore the dynamic evolution of OECD and non-OECD countries' distributions of GDP growth rates across different years. Maasoumi *et al.* [7] proposed using the kernel estimator with unordered and ordered discrete kernels assigned to the discrete variables to estimate such trivariate densities, where bandwidths were selected through the likelihood cross-validation (LCV) (see also [7,47]).

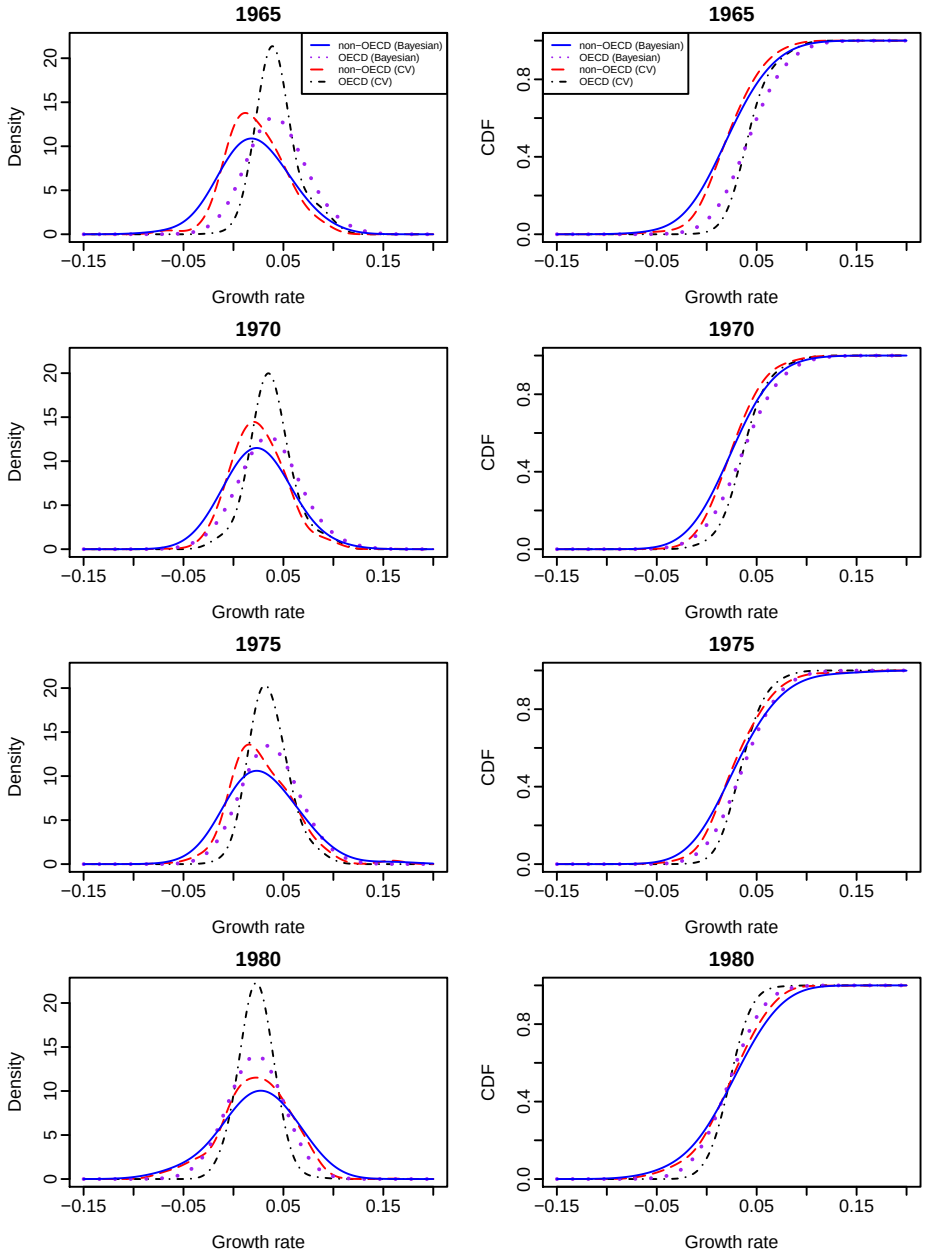

Figure 3. Density and distribution functions of GDP growth rate by year and OECD status.

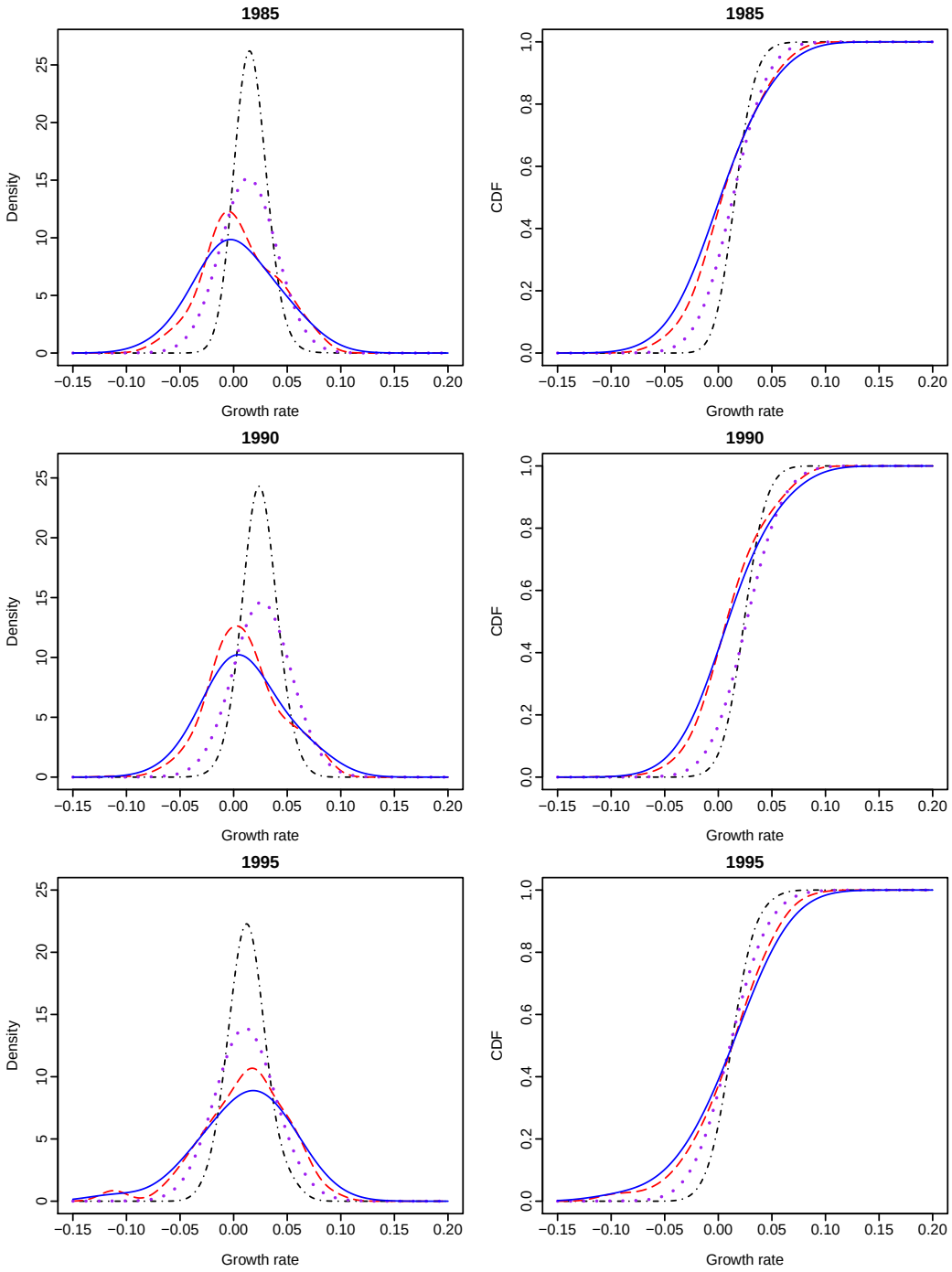

Figure 4. Density and distribution functions of GDP growth rate by year and OECD status (continued).

We are interested in choosing bandwidths for the kernel estimator of $\boldsymbol{x} = \left(x^{(c)}, \boldsymbol{x}^{(d)}\right)'$, which is expressed as

$$\widehat{f}\left(x^{(c)}, \boldsymbol{x}^{(d)}; h, \boldsymbol{\lambda}\right) = \frac{1}{n}\sum_{j=1}^{n} K_h^{(c)}\left(x^{(c)} - x_j^{(c)}\right) \times K_{\boldsymbol{\lambda}}^{(d)}\left(\boldsymbol{x}^{(d)}, \boldsymbol{x}_j^{(d)}\right).$$

where $x_i^{(c)}$ and $\boldsymbol{x}_i^{(d)}$, for $i = 1, 2, \cdots, n$, are observations of $x^{(c)}$ and $\boldsymbol{x}^{(d)}$, respectively. The kernel function for GDP growth rates is the Gaussian kernel, while the OECD status is assigned with a kernel function given by (3), and the kernel for years is given by (4).

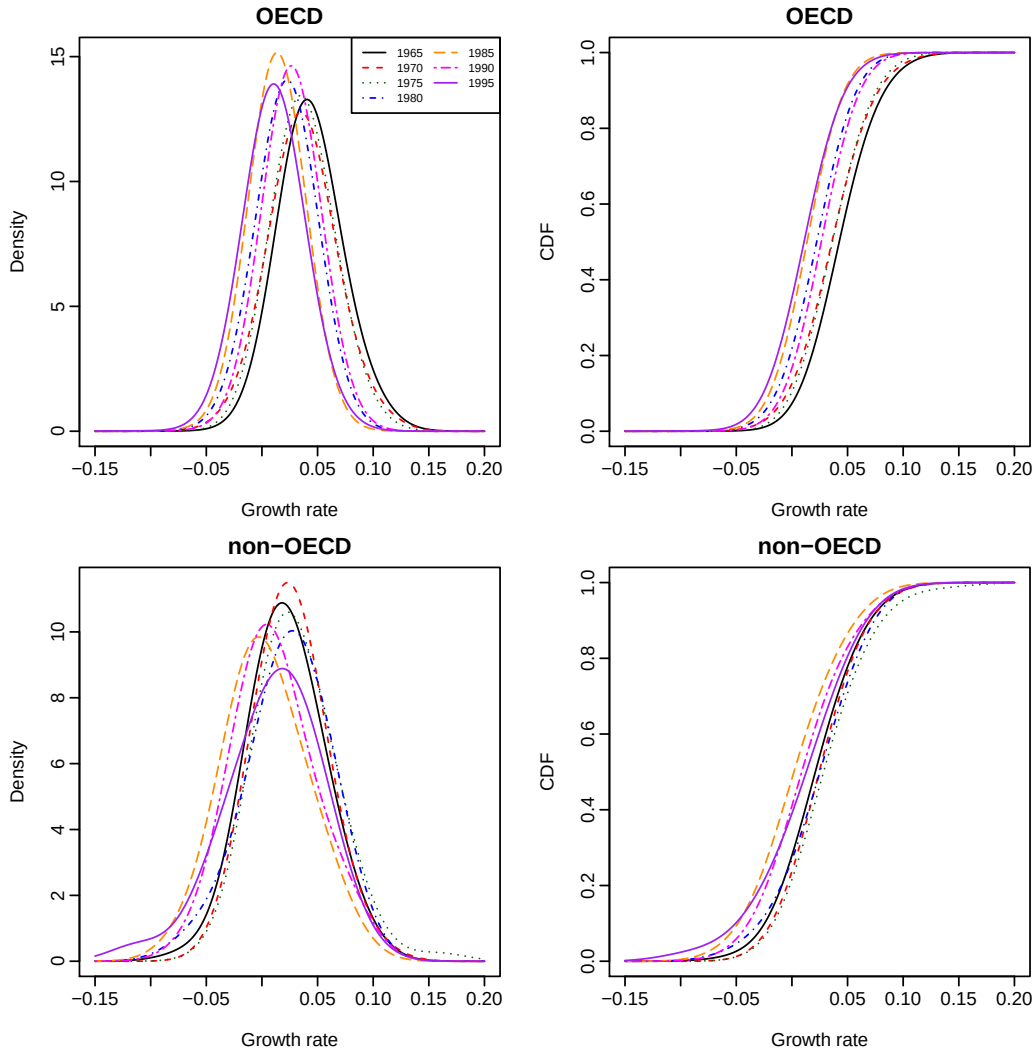

Figure 5. Stacked plots of density and distribution functions of GDP growth rate.

The likelihood of $\boldsymbol{x} = \left\{ \left(x_i^{(c)}, \boldsymbol{x}_i^{(d)} \right) : i = 1, 2, \cdots, n \right\}$ for given $(h, \boldsymbol{\lambda})$ is approximately

$$\ell\left(\boldsymbol{x} | h, \boldsymbol{\lambda}\right) \approx \prod_{i=1}^{n} \widehat{f}_{(-i)} \left(x_i^{(c)}, \boldsymbol{x}_i^{(d)}; h, \boldsymbol{\lambda} \right),$$

where $\widehat{f}_{(-i)}$ is the leave–one–out version of $\widehat{f}$ (see for example, [48]).

The priors of bandwidth parameters are the same as those discussed in Section 3.2. The posterior of $(h, \boldsymbol{\lambda})$ is proportional to the product of $\ell\left(\boldsymbol{x} | h, \boldsymbol{\lambda}\right)$ and the priors of bandwidths. The adaptive random-walk Metropolis algorithm was used to implement the posterior simulation, where the proposal density is the standard trivariate Gaussian, and the tuning parameter was chosen to make the acceptance rate around 0.234. The posterior mean of each simulated chain is used as an estimate of the corresponding bandwidth.

With the estimated bandwidths, we derived the joint density of GDP growth rate, OECD status and years. The conditional density of GDP growth rates for given values of OECD status and year is the joint density estimator divided by the marginal density estimator of OECD status and year. Note that bandwidths estimated for the joint trivariate kernel density estimator can be used for the kernel conditional density estimator.

Given different values of OECD status and year, we calculated the conditional densities and CDFs of GDP growth rates with their graphs presented in Figures 3 and 4. It shows that Bayesian sampling and LCV lead to clearly different density functions of growth rates. The probability density estimates obtained from LCV have higher peaks than those obtained from the Bayesian method, although they both center at about the same points. The growth-rate distributions of an OECD country and a non-OECD country are very different from 1965 to 1995. Second, the growth-rate density of an OECD country is almost symmetrical and less dispersed than that of a non-OECD country, and this phenomenon becomes obvious over time. The growth-rate density of a non-OECD country is asymmetrical and has larger variation than that of an OECD country. It appears to manifest bimodality and indicate "polarization" within non-OECD countries. The results re-confirm the findings of [7]. Figure 5 presents the stacked plots of OECD and non-OECD density and distribution functions of growth rates for all years, where bandwidths were estimated through Bayesian sampling.

We found the following empirical evidence. First, given the year value at either 1965 or 1970, the conditional growth-rate distribution of an OECD country (purple dotted line on the two top right graphs of Figure 3) stochastically dominates that of a non-OECD country (blue solid line). However, there has been no such stochastic dominance since 1975. Second, given the OECD status of a country, the country's growth-rate distribution in 1965 (black solid line on the top right graph of Figure 5) stochastically dominates its growth-rate distributions in the other years; and its growth-rate distribution in 1990 (pink long- and short-dashed line) stochastically dominates its growth-rate distributions in 1980 (blue dot-dash line), 1985 (brown dashed line) and 1995 (purple solid line), respectively.

7. Conclusions

We have presented a Bayesian sampling approach to the estimation of bandwidths in a nonparametric regression model with continuous and discrete regressors, where the regression function is estimated by the NW estimator and the unknown error density is approximated by a kernel-form error density. Monte Carlo simulation results show that the proposed Bayesian sampling method typically performs better than, or at least on par in only a few occasions with, cross-validation for choosing bandwidths. In particular, the Bayesian method performs better than the CV method when the sample size is small. The advantage of the proposed Bayesian approach over cross-validation is its ability to estimate the error density. As measured by the MISE, the Bayesian method outperforms the two-stage cross-validation method for estimating the bandwidth in the kernel-form error density. Thus, the proposed sampling method is recommended for estimating bandwidths in the regression-function and kernel-form error density estimators.

The proposed Bayesian sampling algorithm is used to estimate bandwidths for the nonparametric regression of All Ordinaries (Aord) daily return on the overnight S&P 500 return and an indicator of the FTSE return. In comparison to cross-validation for bandwidth selection, the proposed sampling method leads to a different one-step-ahead forecasted density of the Aord return. Consequently, the resulting value-at-risk measure, as well as the relative frequency of exceedance, is different from the one derived with bandwidths selected through cross-validation. In this example, Bayesian sampling for bandwidth estimation in the nonparametric regression of mixed regressors leads to better results than cross-validation. In an application that involves of kernel density estimation of a country's GDP growth rate conditional on its OECD status and the year of observations, Bayesian sampling for bandwidth estimation leads to different density estimates from those with bandwidths selected through likelihood cross-validation.

There are several ways, along which this paper can be extended. First, the proposed Bayesian algorithm can be extended to several other models, one of which is the nonparametric regression model with conditional heteroscedastic errors or correlated errors. Second, it is possible to consider using asymmetric kernel functions for continuous variables, such as the beta and gamma kernels

discussed by in [45], as well as other kernels for discrete variables, such as the binomial and Poisson kernels discussed by [44]. We leave these extensions for future research.

Acknowledgments: We extend our sincere thanks to the editor, Isabel Casas, and two reviewers whose comments helped us clearly improve the manuscript. The first author wishes to thank Jeffrey Racine for insightful discussion during an early stage of this paper. This research was supported under Australian Research Council's *Discovery Projects* funding scheme (project numbers DP1095838 and DP130104229).

Author Contributions: All authors contributed equally to the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Nadaraya, E.A. On estimating regression. *Theory Probab. Its Appl.* **1964**, *9*, 141–142.
2. Watson, G. Smooth regression analysis. *Sankhya Indian J. Stat. Ser. A* **1964**, *26*, 359–372.
3. Li, Q.; Racine, J.S. Nonparametric estimation of distributions with categorical and continuous data. *J. Multivar. Anal.* **2003**, *86*, 266–292.
4. Racine, J.; Li, Q. Nonparametric estimation of regression functions with both categorical and continuous data. *J. Econom.* **2004**, *119*, 99–130.
5. Hall, P.; Li, Q.; Racine, J.S. Nonparametric estimation of regression functions in the presence of irrelevant regressors. *Rev. Econ. Stat.* **2007**, *89*, 784–789.
6. Li, Q.; Racine, J.S. *Nonparametric Econometrics: Theory and Practice*; Princeton University Press: Princeton, NJ, USA, 2007.
7. Maasoumi, E.; Racine, J.; Stengos, T. Growth and convergence: A profile of distribution dynamics and mobility. *J. Econom.* **2007**, *136*, 483–508.
8. Hayfield, T.; Racine, J.S. Nonparametric econometrics: The np package. *J. Stat. Softw.* **2008**, *27*, 1–32.
9. Li, C.; Ouyang, D.; Racine, J.S. Nonparametric regression with weakly dependent data: The discrete and continuous regressor case. *J. Nonparametric Stat.* **2009**, *21*, 697–711.
10. Su, L.; Chen, Y.; Ullah, A. Functional coefficient estimation with both categorical and continuous data. In *Advances in Econometrics*; Li, Q., Racine, J., Eds.; Emerald Group Publishing Limited: Bingley, UK, 2009; Volume 25, pp. 131–167.
11. Ma, S.; Racine, J.S. Additive regression splines with irrelevant categorical and continuous regressors. *Stat. Sin.* **2013**, *23*, 515–541.
12. Zhang, X.; King, M.L.; Shang, H.L. A sampling algorithm for bandwidth estimation in a nonparametric regression model with a flexible error density. *Comput. Stat. Data Anal.* **2014**, *78*, 218–234.
13. Li, Q.; Zhou, J. The uniqueness of cross-validation selected smoothing parameters in kernel estimation of nonparametric models. *Econom. Theory* **2005**, *21*, 1017–1025.
14. Loynes, R.M. The empirical distribution function of residuals from generalised regression. *Ann. Stat.* **1980**, *8*, 285–298.
15. Akritas, M.G.; Van Keilegom, I. Non-parameteric estimation of the residual distribution. *Scand. J. Stat.* **2001**, *28*, 549–567.
16. Cheng, F.; Sun, S. A goodness of fit test of the errors in nonlinear autoregressive time series models. *Stat. Probab. Lett.* **2008**, *78*, 50–59.
17. Ahmad, I.A.; Li, Q. Testing symmetry of an unknown density function by kernel method. *J. Nonparametric Stat.* **1997**, *7*, 279–293.
18. Dette, H.; Kusi-Appiah, S.; Neumeyer, N. Testing symmetry in nonparametric regression models. *J. Nonparametric Stat.* **2002**, *14*, 477–494.
19. Neumeyer, N.; Dette, H. Testing for symmetric error distribution in nonparametric regression models. *Stat. Sin.* **2007**, *17*, 775–795.
20. Efromovich, S. Estimation of the density of regression errors. *Ann. Stat.* **2005**, *33*, 2194–2227.
21. Muhsal, B.; Neumeyer, N. A note on residual-based empirical likelihood kernel density estimation. *Electron. J. Stat.* **2010**, *4*, 1386–1401.
22. Samb, R. Nonparametric estimation of the density of regression errors. *Comptes Rendus Math.* **2011**, *349*, 1281–1285.
23. Escanciano, J.C.; Jacho-Chávez, D.T. $\sqrt{n}$ uniformly consistent density estimation in nonparametric regression models. *J. Econom.* **2012**, *167*, 305–316.

24. Zhang, X.; King, M.L. Gaussian kernel GARCH models. Working paper 19/13, Monash University, 2013. Aailable online: https://ideas.repec.org/p/msh/ebswps/2013-19.html (accessed on 1 December 2015).
25. Zhang, X.; King, M.L.; Shang, H.L. Bayesian bandwidth selection for a nonparametric regession model with mixed types of regressors. Working paper 13/13, Monash University, 2013. Available online: https://ideas.repec.org/p/msh/ebswps/2013-13.html (accessed on 1 December 2015).
26. Shang, H.L. Bayesian bandwidth estimation for a functional nonparametric regression model with mixed types of regressors and unknown error density. *J. Nonparametric Stat.* **2014**, *26*, 599–615.
27. Kokonendji, C.C.; Senga Kiessé, T. Discrete associated kernels method and extensions. *Stat. Methodol.* **2011**, *8*, 497–516.
28. Aitchison, J.; Aitken, C.G.G. Multivariate binary discrimination by the kernel method. *Biometrika* **1976**, *63*, 413–420.
29. Yuan, A.; de Gooijer, J.G. Semiparametric regression with kernel error model. *Scand. J. Stat.* **2007**, *34*, 841–869.
30. Kuruwita, C.N.; Kulasekera, K.B.; Padgett, W.J. Density estimation using asymmetric kernels and Bayes bandwidth with censored data. *J. Stat. Plan. Inference* **2010**, *140*, 1765–1774.
31. Zheng, Q.; Kulasekera, K.B.; Gallagher, C. Local adaptive smoothing in kernel regression estimation. *Stat. Probab. Lett.* **2010**, *80*, 540–547.
32. Zougab, N.; Adjabi, S.; Kokonendji, C.C. Binomial kernel and Bayes local bandwidth in discrete function estimation. *J. Nonparametric Stat.* **2012**, *24*, 783–795.
33. Kim, S.; Shephard, N.; Chib, S. Stochastic volatility: Likelihood inference and comparison with ARCH models. *Rev. Econ. Stud.* **1998**, *65*, 361–393.
34. Geweke, J. *Complete and Incomplete Econometric Models*; Princeton University Press: Princeton, NJ, USA, 2009.
35. Roberts, G.O.; Rosenthal, J.S. Examples of adaptive MCMC. *J. Comput. Graph. Stat.* **2009**, *18*, 349–367.
36. Garthwaite, P.H.; Fan, Y.; Sisson, S.A. Adaptive optimal scaling of Metropolis-Hastings algorithms using the Robbins-Monro process. *Commun. Stat. Theory Methods.* **2015**, doi: 10.1080/03610926.2014.936562.
37. Roberts, G.O. Markov chain concepts related to sampling algorithms. In *Markov Chain Monte Carlo in Practice*; Gilks, W.R., Richardson, S., Spiegelhalter, D.J., Eds.; Chapman & Hall: London, UK, 1996; pp. 45–57.
38. Tse, Y.K.; Zhang, X.; Yu, J. Estimation of hyperbolic diffusion using the Markov chain Monte Carlo method. *Quant. Finance* **2004**, *4*, 158–169.
39. Nott, D.J.; Kohn, R. Adaptive sampling for Bayesian variable selection. *Biometrika* **2005**, *92*, 747–763.
40. Kruskal, W.H.; Wallis, W.A. Use of ranks in one-criterion variance analysis. *J. Am. Stat. Assoc.* **1952**, *47*, 583–621.
41. Dunn, O.J. Multiple ccomparison using rank sums. *Technometrics* **1964**, *6*, 241–252.
42. Hansen, P.R.; Lunde, A.; Nason, J.M. The model confidence set. *Econometrica* **2011**, *79*, 453–497.
43. Wang, M.C.; Van Ryzin, J. A class of smooth estimators for discrete distributions. *Biometrika* **1981**, *68*, 301–309.
44. Zougab, N.; Adjabi, S.; Kokonendji, C.C. Bayesian approach in nonparametric count regression with binomial kernel. *Commun. Stat. — Simul. Comput.* **2014**, *43*, 1052–1063.
45. Somé, S.M.; Kokonendji, C.C. Effects of associated kernels in nonparametric multiple regressions. Working paper, University of Franche-Comté, 2015. Available online: arxiv.org/pdf/1502.01488v1.pdf (accessed on 1 December 2015).
46. Zhang, X.; Brooks, R.D.; King, M.L. A Bayesian approach to bandwidth selection for multivariate kernel regression with an application to state-price density estimation. *J. Econom.* **2009**, *153*, 21–32.
47. Hall, P. On Kullback-Leibler loss and density estimation. *Ann. Stat.* **1987**, *15*, 1491–1519.
48. Zhang, X.; King, M.L.; Hyndman, R.J. A Bayesian approach to bandwidth selection for multivariate kernel density estimation. *Comput. Stat. Data Anal.* **2006**, *50*, 3009–3031.