Bayesian bandwidth estimation for a semi-functional partial linear regression model with unknown error density

Han Lin Shang*

Abstract

In the context of semi-functional partial linear regression model, we study the problem of error density estimation. The unknown error density is approximated by a mixture of Gaussian densities with means being the individual residuals, and variance a constant parameter. This mixture error density has a form of a kernel density estimator of residuals, where the regression function, consisting of parametric and nonparametric components, is estimated by the ordinary least squares and functional Nadaraya-Watson estimators. The estimation accuracy of the ordinary least squares and functional Nadaraya-Watson estimators jointly depends on the same bandwidth parameter. A Bayesian procedure is proposed to simultaneously estimate the bandwidths in the kernel-form error density and in the regression function. Under the kernel-form error density, we derive a kernel likelihood and posterior for the bandwidth parameters. For estimating the regression function and error density, a series of simulation studies show that the Bayesian procedure yields better accuracy than the benchmark functional cross validation. Illustrated by a spectroscopy data set, we found that the Bayesian procedure gives better point forecast accuracy of the regression function than the functional cross validation, and it is capable of producing prediction intervals nonparametrically.

Key Words: functional Nadaraya-Watson estimator, functional regression, Gaussian kernel mixture, error density estimation, Markov chain Monte Carlo

1. Introduction

The standard error distributional assumptions, such as normality, typically have advantage of simplicity in deriving theoretical results and inference in functional regression models. However, it is often the case that such parametric assumptions do not apply to the real data. In this case, we will need to sacrifice the analytical convenience in order to obtain an acceptable fit to the data by using flexible error distributions. This paper proposes a method to address error distribution (or density) estimation in functional regression models, and demonstrates its usage with the semi-functional partial linear regression model.

Since the introductory work by Ansley & Wecker (1983), the partial linear model has been widely studied (see for example, Heckman 1986, Rice 1986, Chen 1988, Robinson 1988, Speckman 1988, Severini & Staniswalis 1994, Härdle et al. 2000) and implemented in many fields of applied statistics, such as to study the impact of weather on electricity demand (Engle et al. 1986), to model household gasoline consumption in the United State (Schmalensee & Stoker 1999), and to model hedonic price function (Anglin & Gencay 1996). The advantage of this regression model is to allow some of explanatory variables to act in a nonparametric manner, while others are governed in a parametric manner. Thus, it reconciles the flexibility of nonparametric models and fast convergence rate of parametric models. However, the explanatory variables in the classical partial linear regression are limited to be real values (see Härdle et al. 2000, for detail).

In recent years, the presence of functional variables is common in many applied fields, ranging from minute-by-minute electricity demand forecasting (Shang 2013*b*) to high-frequency financial stock returns (Gabrys et al. 2010). These practical challenges demand

*University of Southampton, University Road, Highfield, Southampton, SO17 1BJ, United Kingdom. Tel: +44 (0)2380 595 796, Fax: +44 (0)2380 593 858. Email: H.Shang@soton.ac.uk

the innovations of new statistical data analysis techniques for analyzing ever-increasingly high-dimensional data. This leads to the development of functional data analysis from a parametric viewpoint (Ramsay & Silverman 2005) and a nonparametric viewpoint (Ferraty & Vieu 2006). While Horváth & Kokoszka (2012) studied statistical properties of functional time series, Ferraty & Romain (2010) and Ferraty (2011) collected a number of research papers on the latest development of functional data analysis.

To our knowledge, Aneiros-Pérez & Vieu (2006) were among the first to extend the partial linear regression to the functional setting, and demonstrate its superiority over the functional linear regression and functional nonparametric regression. Since then, the semi-functional partial linear regression has been widely studied in theory and practice (see for example, Aneiros-Pérez & Vieu 2008, Dabo-Niang & Guillas 2010, Zhou & Chen 2012, Zhang & Wang 2012). This work aims to contribute to the semi-functional partial linear regression model by incorporating error density into bandwidth estimation.

The estimation of error density is important to understand the residual behaviour and to assess the adequacy of error distribution assumption (see for example, Akritas & Van Keilegom 2001, Cheng & Sun 2008). The estimation of error density is useful to test the symmetry of the residual distribution (see for example, Ahmad & Li 1997, Dette et al. 2002, Neumeyer & Dette 2007). The estimation of error density is vital to statistical inference, prediction and model validation (see for example, Efromovich 2005, Muhsal & Neumeyer 2010). The estimation of error density is also crucial for the estimation of response variable (see for example, Escanciano & Jacho-Chávez 2012). Despite its importance, error density estimation in the semi-functional partial linear regression model remains largely unexplored.

This motivates the investigation of a kernel-form error density for estimating unknown error density in a semi-functional partial linear regression model with function-valued predictor and real-valued response. This kernel-form error density explores data-driven features, such as skewness and multimodality, and it depends on three parameters: (1) the type of semi-metric used in the regression function to measure distances among curves, such as semi-metric based on second-order derivative; (2) residuals fitted through the functional Nadaraya-Watson (NW) estimator of the regression function; and (3) bandwidth of residuals.

In the context of a multivariate nonparametric regression model, this kernel-form error density estimator has been investigated by Cheng (2002, 2004), who proved weak, strong and uniform consistency. Samb (2011) established the optimal convergence rate of this estimator. Recently, Shang (2013*a*) extended this kernel-form error density estimator to a nonparametric functional regression model. Building upon the work of Shang (2013*a*), we also consider this kernel-form error density estimator, in the context of the semi-functional partial linear regression model.

The Bayesian procedure equipped with Markov chain Monte Carlo technique allows us to simultaneously estimate the bandwidth parameters. Conditional on the bandwidth parameter in the regression function, we sample the bandwidth parameter in the kernel-form error density that maximizes the posterior (also known as cost or utility function in econometrics). Based on the sampled bandwidth parameter in the kernel-form error density, we then sample the bandwidth parameter in the regression function that maximizes the posterior.

The rest of the paper is organized as follows. In Section 2, we introduce the semi-functional partial linear model. The proposed Bayesian procedure is described in Section 3. Illustrated by a series of simulation studies in Section 4, we evaluate and compare the finite-sample estimation accuracy of the regression function and error density between the Bayesian procedure and the functional cross validation (CV). Illustrated by a well-studied

spectroscopy data set, the point forecast accuracy of the regression function is compared between the Bayesian procedure and the functional CV in Section 5. Section 6 concludes the paper, along with some thoughts on how the methods developed here might be further extended.

2. Model and estimator

There are a growing literature on the development of parametric functional regression models, such as functional linear regression model (Ramsay & Silverman 2005) and functional quadratic regression model (Yao & Müller 2010). As well, there are an increasing amount of literature on the development of functional estimators in nonparametric functional regression models, such as functional NW estimator (Ferraty & Vieu 2006), functional local linear estimator (Berlinet et al. 2011), functional k-nearest neighbour estimator (Burba et al. 2009), and distance-based local linear estimator (Boj et al. 2010). In this paper, we focus on the semi-functional partial linear regression model because it combines the advantages of both parametric and nonparametric models.

2.1 Semi-functional partial linear regression

Let $\{(y_i, X_{i1}, \cdots, X_{ip}, T_i)\}_{i=1}^n$ be $n(p+2)$-variate random vectors identically distributed as $(y, X_1, \cdots, X_p, T)$, where y represents a real-valued response variable, which can be predicted by a set of real-valued explanatory variable $(X_1, \ldots, X_p)$ and a function-valued explanatory variable T in a semi-functional partial linear regression model given by:

$$y_i = \boldsymbol{X}_i\boldsymbol{\beta} + m(T_i) + \varepsilon_i, \qquad i = 1, 2, \ldots, n, \tag{1}$$

where $\boldsymbol{X}_i$ represents the ith observation of the real-valued explanatory variables, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is a vector of unknown regression coefficients in the parametric component, m is an unknown smooth real function, and $(\varepsilon_1, \ldots, \varepsilon_n)$ are identically distributed random error satisfying

$$\mathrm{E}\left(\varepsilon_i | \boldsymbol{X}_i, T_i\right) = 0. \tag{2}$$

Following Aneiros-Pérez & Vieu (2008), it is natural to estimate the vector $\boldsymbol{\beta}$ and the function m by the ordinary least squares and functional NW estimator, given by

$$\widehat{\boldsymbol{\beta}}_{h_n} = \left(\tilde{\boldsymbol{X}}_{h_n}^\top \tilde{\boldsymbol{X}}_{h_n}\right)^{-1} \tilde{\boldsymbol{X}}_{h_n}^\top \tilde{\boldsymbol{y}}_{h_n}, \tag{3}$$

$$\widehat{m}_{h_n}(t) = \sum_{i=1}^{n} w_{h_n}(t, T_i)\left(y_i - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}}_{h_n}\right) \tag{4}$$

where $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^\top$ and $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$, $\tilde{\boldsymbol{X}}_{h_n} = (\boldsymbol{I} - \boldsymbol{W}_{h_n})\boldsymbol{X}$ and $\tilde{\boldsymbol{y}}_{h_n} = (\boldsymbol{I} - \boldsymbol{W}_{h_n})\boldsymbol{y}$, $\boldsymbol{W}_{h_n} = (w_{h_n}(T_i, T_j))_{i,j}$ is a weight matrix with $w_{h_n}(t, T_i)$ being the NW-type weights,

$$w_{h_n}(t, T_i) = \frac{K(d(t, T_i)/h_n)}{\sum_{j=1}^n K(d(t, T_j)/h_n)}, \tag{5}$$

where K is a kernel function, such as the Gaussian kernel function considered in this paper; h_n is a positive real-valued estimated bandwidth parameter, controlling the trade-off between squared bias and variance in the mean squared error; and $d(\cdot, \cdot)$ is a semi-metric used to quantify differences among curves. Since our simulated and real data in Sections 4

and 5 are quite smooth, we follow the early work of Goutis (1998), Ferraty & Vieu (2002, 2009) and chose a semi-metric based on second derivative, which is given by

$$d_q^{\text{deriv}}(T_i, T) = \sqrt{\int_t \left[T_i^{(q)}(t) - T^{(q)}(t)\right]^2 dt}, \qquad q = 2. \tag{6}$$

For a non-smooth functional data set, a semi-metric based on functional principal component analysis is advocated (see Ferraty & Vieu 2006, Chapters 3 and 13 for detail on the choice of semi-metric from the practical and theoretical aspects, respectively).

The regression coefficients $\widehat{\boldsymbol{\beta}}_{h_n}$ in (3) can be seen as the ordinary least squares estimator obtained by regressing the partial residual vector $\tilde{\boldsymbol{y}}_h$ on the partial residual matrix $\tilde{\boldsymbol{X}}_h$ (Aneiros-Pérez & Vieu 2008). Alternatively, $\widehat{m}_{h_n}(t)$ in (4) can be seen as the functional NW estimator with the partial residual vector as the response variable. Therefore, the estimation accuracy of $\widehat{\boldsymbol{\beta}}_{h_n}$ and $\widehat{m}_{h_n}(t)$ depends crucially on the optimal choice of h.

In the statistical literature, the bandwidth parameter is commonly determined by functional CV in many nonparametric functional regression models (see, for example, Burba et al. 2009, Ferraty & Vieu 2009, Barrientos-Marin et al. 2010, Ferraty et al. 2010). In the semi-functional partial linear regression model, Aneiros-Pérez & Vieu (2006, 2008) considered functional CV to select optimal bandwidth. However, the functional CV fails to incorporate error density of a regression model into bandwidth estimation. Due to the importance of error density, this motivates us to propose a Bayesian procedure to estimate bandwidth parameters in regression function and error density, simultaneously.

3. Bayesian bandwidth estimation

3.1 Estimation of error density

We assume that the unknown error density $f(\varepsilon)$ is approximated by a location-mixture Gaussian density, given by

$$f(\varepsilon; b) = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{b}\phi\left(\frac{\varepsilon - \varepsilon_j}{b}\right), \tag{7}$$

where $\phi(\cdot)$ is the probability density function of the standard Gaussian distribution, and Gaussian densities have means at ε_j, for $j = 1, 2, \ldots, n$ and a common standard deviation b. Note that our proposed kernel-form error density has only one parameter to be estimated from data, as opposite to many parameters in the location-scale mixture (see McLachlan & Peel 2000, for detail on finite mixture models). Although $\boldsymbol{\beta}$ and m are unknown, they can be estimated by the ordinary least squares and functional NW estimators, respectively. Based on the residuals, the density of ε_j is approximated by the estimated error density $\widehat{f}(\varepsilon; b_n)$, expressed as

$$f(\varepsilon; b) \approx \widehat{f}(\varepsilon; b_n) = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{b_n}\phi\left(\frac{\varepsilon - \widehat{\varepsilon}_j}{b_n}\right), \tag{8}$$

where b_n represents the residual bandwidth in a finite sample, and $b_n \to 0$ as $n \to \infty$. As pointed out by Samb (2011), the kernel estimator $\widehat{f}(\varepsilon; b_n)$ is a feasible estimator that does not depend on any unknown quantity, in contrast to (7).

Jaki & West (2008, p.989) and Jaki & West (2011) also proposed to estimate the error density by a kernel density estimator given by (8), and estimate parameters by maximizing the so-called kernel likelihood. The kernel likelihood of $\boldsymbol{y}$ is defined as the product of the

density given by (8), based on the independent and identically distributed (iid) assumption of the error term. In the context of bandwidth estimation, it is however impossible to estimate b by maximizing the kernel likelihood without a slight modification, because it contains un-desirable term $\phi(0)/b_n$. The kernel likelihood would approach infinity as b_n approaches to zero. To address this issue, a leave-one-out version of the kernel likelihood is used and expressed by

$$\widehat{f}(\widehat{\varepsilon}_{i,h_n}; b_n) = \frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{1}{b_n} \phi \left(\frac{\widehat{\varepsilon}_{i,h_n} - \widehat{\varepsilon}_{j,h_n}}{b_n} \right), \tag{9}$$

where $\widehat{\varepsilon}_{i,h_n} = y_i - \widehat{m}_{h_n}(t) - \boldsymbol{X}_i \widehat{\boldsymbol{\beta}}_{h_n}$ is the ith residual for $i = 1, 2, \ldots, n$, and h_n represents the bandwidth estimate in the ordinary least squares and functional NW estimators. Given h_n and b_n, the kernel likelihood of $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^\top$ can be approximated by

$$\widehat{L}(\boldsymbol{y}|h_n, b_n) = \prod_{i=1}^{n} \left[\frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{1}{b_n} \phi \left(\frac{\widehat{\varepsilon}_{i,h_n} - \widehat{\varepsilon}_{j,h_n}}{b_n} \right) \right]. \tag{10}$$

3.2 Prior density

We now discuss the issue of prior density for the bandwidths. Let $\pi(h^2)$ and $\pi(b^2)$ be the prior of squared bandwidths h and b. Since h^2 and b^2 can be considered as a variance parameter in the Gaussian distribution, we consider the conjugate prior of h^2 and b^2. Let the prior densities of h^2 and b^2 be inverse Gamma densities, denoted as IG(α_h, β_h) and IG(α_b, β_b), respectively. The prior densities can be expressed as

$$\pi(h^2) = \frac{(\beta_h)^{\alpha_h}}{\Gamma(\alpha_h)} (\frac{1}{h^2})^{\alpha_h + 1} \exp(-\frac{\beta_h}{h^2}), \qquad \pi(b^2) = \frac{(\beta_b)^{\alpha_b}}{\Gamma(\alpha_b)} (\frac{1}{b^2})^{\alpha_b + 1} \exp(-\frac{\beta_b}{b^2}),$$

where $\alpha_h = \alpha_b = 1$ and $\beta_h = \beta_b = 0.05$ are hyperparameters. Notice that IG$(1, 0.05)$ has previously been used as a prior density in Geweke (2010).

3.3 Posterior density and Bayesian sampler

According to Bayes theorem, the posterior of h_n^2 and b_n^2 is approximated by (up to a normalizing constant)

$$\pi(h_n^2, b_n^2|\boldsymbol{y}) \propto \widehat{L}(\boldsymbol{y}|h_n^2, b_n^2) \pi(h^2) \pi(b^2), \tag{11}$$

where $\widehat{L}(\boldsymbol{y}|h_n^2, b_n^2)$ is the approximate kernel likelihood function with squared bandwidths. The parameters are sampled from its posterior density and estimated by the means of Markov chain Monte Carlo (MCMC). In essence, the MCMC method sets up a Markov chain so that its stationary distribution is the same as the posterior density. When the Markov chain converges, the ergodic averages of the simulated realizations are treated as the estimated parameter values. For a detailed exposition on MCMC method, consult Geweke (1999), Gilks et al. (1996) and Robert & Casella (2010).

From (11), we use a generic algorithm known as the adaptive random-walk Metropolis algorithm of Garthwaite et al. (2010) to sample (h_n^2, b_n^2) jointly. The sampling algorithm is described below.

Step 0 Specify a Gaussian proposal density, with an arbitrary starting point $b_{n,(0)}^2$ and $h_{n,(0)}^2$. The starting points can be drawn from a uniform distribution $U(0, 1)$.

Step 1 At the kth iteration, the current state $b^2_{n,(k)}$ is updated as $b^2_{n,(k)} = b^2_{n,(k-1)} + \tau_{(k-1)}\varepsilon$, where $\varepsilon \sim N(0,1)$, and $\tau_{(k-1)}$ is an adaptive tuning parameter with an arbitrary initial value $\tau_{(0)}$.

Step 2 The updated $b^2_{n,(k)}$ is accepted with probability $\min\left\{\frac{\pi(b^2_{n,(k)},h^2_{n,(k-1)}|\boldsymbol{y})}{\pi(b^2_{n,(k-1)},h^2_{n,(k-1)}|\boldsymbol{y})}, 1\right\}$, where π represents the posterior density.

Step 3 By using the stochastic search algorithm of Robbins & Monro (1951), the tuning parameter is

$$\tau_{(k)} = \begin{cases} \tau_{(k-1)} + c(1-p)/k & \text{if } b^2_{n,(k)} \text{ is accepted;} \\ \tau_{(k-1)} - cp/k & \text{if } b^2_{n,(k)} \text{ is rejected,} \end{cases}$$

where $c = \frac{\tau_{(k-1)}}{p(1-p)}$ is a varying constant, and $p = 0.44$ is the optimal acceptance probability for drawing one parameter (Roberts & Rosenthal 2009).

Step 4 Repeat Steps 1-3 for $h^2_{n,(k)}$, conditional on $b^2_{n,(k)}$ and $\boldsymbol{y}$.

Step 5 Repeat Steps 1-4 for $M+N$ times, discard $(h^2_{n,(0)}, b^2_{n,(0)}), (h^2_{n,(1)}, b^2_{n,(1)}), \ldots, (h^2_{n,(M)}, b^2_{n,(M)})$ for burn-in in order to let the effects of the transients wear off, estimate $\widehat{h}^2_n = \frac{\sum_{k=M+1}^{M+N} h^2_{n,(k)}}{N}$ and $\widehat{b}^2_n = \frac{\sum_{k=M+1}^{M+N} b^2_{n,(k)}}{N}$. The burn-in period is taken to be $M = 1,000$ iterations, and the number of iterations after burn-in period is $N = 10,000$ iterations. The analytical form of the kernel-form error density can be derived based on $\widehat{h}^2_n$ and $\widehat{b}^2_n$. It is noteworthy that a similar error density result can be obtained by taking the average of the kernel-form error densities computed at all iterations, but at the cost of much slower computational speed.

The mixing performance of the sample paths can be measured by total standard error (SE) and batch-mean SE, from which we can also calculate simulation inefficiency factor (see also Kim et al. 1998, Meyer & Yu 2000, Tse et al. 2004, Zhang et al. 2009, Shang 2013*a*). The simulation inefficiency factor can be interpreted as the number of draws needed to have iid observations.

3.4 Localized estimation of error density

In the kernel density estimation, it has been observed that the leave-one-out estimator, such as (9), can be heavily affected by extreme observations in the data sample (see for example, Bowman 1984, Zhang & King 2011). When the error density has sufficient long tails or heavy skewness, the leave-one-out kernel density estimation with its bandwidth estimated under the Kullback-Leibler criterion, is likely to overestimate the tail density. This may cause by the use of a global bandwidth. To rectify this problem, localized bandwidth method proposed by Zhang & King (2011) is implemented.

The idea of localized bandwidths is to assign small bandwidths to the observations in the high density region, while large bandwidths to the observations in the low density region. The localized error density estimator proposed by Zhang & King (2011) is given by

$$\widehat{f}(\widehat{\varepsilon}_{i,h_n}; b_n, \tau_\varepsilon) = \frac{1}{n-1}\sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{1}{b_n(1+\tau_\varepsilon|\widehat{\varepsilon}_{j,h_n}|)}\phi\left[\frac{\widehat{\varepsilon}_{i,h_n} - \widehat{\varepsilon}_{j,h_n}}{b_n\,(1+\tau_\varepsilon|\widehat{\varepsilon}_{j,h_n}|)}\right], \qquad (12)$$

where $b_n(1+\tau_\varepsilon|\widehat{\varepsilon}_{j,h_n}|)$ is the bandwidth assigned to $\widehat{\varepsilon}_j$, for $j = 1, 2, \ldots, n$. Together, τ_ε and the magnitude of the residuals control the amount of adjustment to the localized bandwidths. The vector of parameters is now $(h_n, b_n, \tau_\varepsilon)$.

4. Simulation study

The main goal of this section is to illustrate the proposed methodology through simulated data. One way to do that consists in comparing the true regression function with the estimated regression function, and comparing the true error density with the estimated error density. To measure the estimation accuracy between $\gamma = m(T) + \mathbf{X}\boldsymbol{\beta}$ and $\widehat{\gamma} = \widehat{m}(T) + \mathbf{X}\widehat{\boldsymbol{\beta}}$, we first calculate the mean absolute error and mean squared error, given by

$$\text{MAE} = \text{E}(|\gamma - \widehat{\gamma}|), \qquad \text{MSE} = \text{E}[(\gamma - \widehat{\gamma})^2].$$

Averaged across 100 replications, the averaged mean absolute error (AMAE) and the averaged mean squared error (AMSE) are used to assess the estimation accuracy of regression function. They are defined as

$$\text{AMAE} = \frac{1}{B}\sum_{b=1}^{B}\text{MAE}_b, \qquad \text{AMSE} = \frac{1}{B}\sum_{b=1}^{B}\text{MSE}_b.$$

where $B = 100$ represents the number of replications.

To measure the discrepancy between $f(\varepsilon)$ and $\widehat{f}(\varepsilon)$, we first calculate the mean integrated absolute error (MIAE) and mean integrated squared error (MISE), defined by

$$\text{MIAE}[\widehat{f}(\varepsilon)] = \int_a^b |f(\varepsilon) - \widehat{f}(\varepsilon)|d\varepsilon, \qquad \text{MISE}[\widehat{f}(\varepsilon)] = \int_a^b [f(\varepsilon) - \widehat{f}(\varepsilon)]^2 d\varepsilon,$$

for $\varepsilon \in [a, b]$. For each replication, the MIAE and MISE can be approximated at 1,001 grid points bounded between an interval, such as $[-5, 5]$. These can be expressed as

$$\text{MIAE}[\widehat{f}(\varepsilon)] \approx \frac{1}{100}\sum_{i=1}^{1001}\left|f\Big[-5 + \frac{(i-1)}{100}\Big] - \widehat{f}\Big[-5 + \frac{(i-1)}{100}\Big]\right|, \tag{13}$$

$$\text{MISE}[\widehat{f}(\varepsilon)] \approx \frac{1}{100}\sum_{i=1}^{1001}\left\{f\Big[-5 + \frac{(i-1)}{100}\Big] - \widehat{f}\Big[-5 + \frac{(i-1)}{100}\Big]\right\}^2. \tag{14}$$

Averaged across 100 replications, the approximated mean integrated absolute error (AMIAE) and the approximated mean integrated squared error (AMISE) are used to assess the estimation accuracy of error density. They are defined as

$$\text{AMIAE} = \frac{1}{B}\sum_{b=1}^{B}\text{MIAE}_b, \qquad \text{AMISE} = \frac{1}{B}\sum_{b=1}^{B}\text{MISE}_b.$$

Building the simulated samples We briefly describe the construction of the simulated data. First of all, one builds simulated discretized curves

$$T_i(t_j) = a_i\cos(2t_j) + b_i\sin(4t_j) + c_i(t_j^2 - \pi t_j + \frac{2}{9}\pi^2), \qquad i = 1, 2\ldots, n, \tag{15}$$

where $0 \leq t_1 \leq t_2 \cdots \leq t_{100} \leq \pi$ are equispaced points, a_i, b_i, c_i are independently drawn from a uniform distribution on $[0, 1]$, and n represents the sample size. The functional form of (15) is taken from Ferraty et al. (2010).

Once the curves are defined, one simulates a functional regression model to compute the responses:

- construct a nonparametric component $m(T_i) = 10 \times (a_i^2 - b_i^2)$, which performs the mapping from function-valued space to real-valued space.

- construct a parametric component, where X_{i1} and X_{i2} are independently drawn from a uniform distribution $U(0,1)$, for $i = 1, 2, \ldots, n$. The true values of regression coefficients are set to be $\boldsymbol{\beta} = (-1, 2)$. The regression function is constructed as

$$\gamma_i = m(T_i) + \boldsymbol{X}_i\boldsymbol{\beta}, \qquad \text{for} \quad i = 1, 2, \ldots, n. \tag{16}$$

- generate $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ from a student-t distribution with df $= 5$, as an illustration. For further comparison, we have also simulated error densities from 15 different normal mixtures (see Marron & Wand 1992, Table 1 for detail).

- compute the corresponding response:

$$y_i = \gamma_i + \varepsilon_i = m(T_i) + \boldsymbol{X}_i\boldsymbol{\beta} + \varepsilon_i, \qquad \text{for} \quad i = 1, 2, \ldots, n.$$

Estimating the regression function For a fixed bandwidth h_n and regression coefficient $\boldsymbol{\beta}$, one compute the discrepancy between γ_i and $\widehat{\gamma}_i$, for $i = 1, 2, \ldots, n$. To do that, one uses the following Monte-Carlo scheme:

1) Build 100 replications $\{(y_i^s, \boldsymbol{X}_i^s, T_i^s)_{i=1,\ldots,n}\}_{s=1,\ldots,100}$;

2) Compute 100 estimates $[(\gamma_i - \widehat{\gamma}_i^s)_{i=1,\ldots,n}]_{s=1,\ldots,100}$, where $\widehat{\gamma}_i^s$ represents the regression function estimated by the functional NW estimator for the nonparametric component and the ordinary least squares estimator for the parametric component;

3) For each replication, we calculate the MAE and MSE between the true regression function γ and the estimated regression function $\widehat{\gamma}$. Averaged over 100 replications, we use the AMAE and AMSE to assess the estimation accuracy of the regression function.

Table 1 presents the AMAE and AMSE for the parametric part, nonparametric part and regression function in the semi-functional partial linear regression, where the bandwidth was estimated by the functional CV and Bayesian methods. When the error density is simulated from t_5, the Bayesian methods perform better than the functional CV across inference sample sizes.

Estimating the error density With a set of residuals and a fixed residual bandwidth, one can apply a univariate kernel density estimator and compute the discrepancy between $f(\varepsilon)$ and $\widehat{f}(\varepsilon)$. The Monte-Carlo scheme can be described as follows:

1) Compute 100 replications of residuals $[(y_i^s - \gamma_i^s)_{i=1,\ldots,n}]_{s=1,\ldots,100}$;

2) The Bayesian approach jointly estimates the bandwidths in the regression function and error density, but we also consider a two-stage CV method for the purpose of comparison. Based on 100 replications of residuals and the corresponding residual bandwidths estimated by likelihood CV (Bowman 1984), the two-stage CV method applies a univariate kernel density to estimate error density. The asymptotic optimality of the bandwidth selected by the likelihood CV has been studied by Hall (1987).

3) For $s = 1, \ldots, 100$, compute the MIAE in (13) and MISE in (14) between the true error density $f(\varepsilon)$ and the estimated error density $\widehat{f}^s(\varepsilon)$.

4) Averaged over 100 replications, we calculate the AMIAE and AMISE to assess the estimation accuracy of the error density.

Table 1: Estimation accuracy of the parametric component, nonparametric component and regression function in the semi-functional partial linear model, between the functional CV and Bayesian methods when the error density is t_5. The number in parenthesis represents the sample standard deviation of the mean absolute errors and mean squared errors. The bolded text represents the minimal AMAE and minimal AMSE.

Method	n	Parametric part		Nonparametric part		Regression function	
		$\lvert\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\rvert$	$(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})^2$	$\lvert m-\widehat{m}\rvert$	$(m-\widehat{m})^2$	$\lvert\gamma-\widehat{\gamma}\rvert$	$(\gamma-\widehat{\gamma})^2$
CV	50	1.3976	2.9927	1.6428	4.4863	1.3898	3.2232
		(0.7018)	(2.9283)	(0.4839)	(2.4719)	(0.1577)	(0.6816)
	150	0.8203	1.0558	1.0540	1.8659	0.9066	1.4555
		(0.4495)	(0.9858)	(0.2575)	(0.8124)	(0.0851)	(0.2479)
	250	0.5770	0.5159	0.8495	1.2669	0.7583	1.0605
		(0.3098)	(0.5350)	(0.1769)	(0.4607)	(0.0650)	(0.1523)
Bayesian *global*	50	0.9448	1.2690	1.4132	3.4261	1.3122	2.8930
		(0.4768)	(1.2514)	(0.2578)	(1.1962)	(0.1608)	(0.6675)
	150	0.3824	0.2329	0.8776	1.4000	0.8460	1.3159
		(0.2005)	(0.2107)	(0.0973)	(0.2565)	(0.0746)	(0.2195)
	250	0.2726	0.1136	0.7373	1.0013	0.7125	0.9491
		(0.1425)	(0.1173)	(0.0657)	(0.1639)	(0.0485)	(0.1186)
Bayesian *local*	50	0.9457	1.2733	1.4100	3.4248	**1.3001**	**2.8685**
		(0.4698)	(1.2325)	(0.2526)	(1.1723)	(0.1493)	(0.6299)
	150	0.3822	0.2308	0.8746	1.3908	**0.8441**	**1.3089**
		(0.1988)	(0.2013)	(0.0978)	(0.2523)	(0.0735)	(0.2085)
	250	0.2738	0.1140	0.7367	0.9999	**0.7117**	**0.9477**
		(0.1422)	(0.1165)	(0.0651)	(0.1635)	(0.0476)	(0.1176)

Table 2 presents AMISE and AMIAE for the kernel-form error density with bandwidth estimated by the likelihood CV and Bayesian methods. When the error density is simulated from t_5, the Bayesian methods perform uniformly better than the likelihood CV across inference sample sizes.

Table 2: Estimation accuracy of the error density in the semi-functional partial linear model, between the likelihood CV and Bayesian methods when the error density is t_5. The number in parenthesis represents the sample standard deviation of the integrated absolute and squared errors. The bolded text represents the minimal MIAE and minimal MISE.

n	Likelihood CV		Bayesian *global*		Bayesian *local*	
	MIAE	MISE	MIAE	MISE	MIAE	MISE
50	0.8002	0.1160	0.6551	0.0785	**0.5098**	**0.0479**
	(0.0915)	(0.0277)	(0.1181)	(0.0279)	(0.1098)	(0.0215)
150	0.6633	0.0787	0.5148	0.0507	**0.4082**	**0.0300**
	(0.0988)	(0.0236)	(0.0838)	(0.0153)	(0.0816)	(0.0129)
250	0.5866	0.0636	0.4786	0.0450	**0.3904**	**0.0280**
	(0.1058)	(0.0221)	(0.0828)	(0.0147)	(0.0786)	(0.0117)

5. Application to food quality control

Let us consider a food quality control application, previously studied by Ferraty & Vieu (2006) and Aneiros-Pérez & Vieu (2006). The data set was obtained from `http://lib.stat.cmu.edu/datasets/tecator`. Each food sample contains finely chopped pure meat with different percentages of the fat, protein and moisture contents. For each unit i (among 215 pieces of finely chopped meat), we observe one spectrometric curve, denoted by T_i, which corresponds to the absorbance measured at a grid of 100 wavelengths (i.e., $T_i = (T_i(t_1), \ldots, T_i(t_{100}))$). For each i, we also observe its fat, protein and moisture contents $\boldsymbol{X} \in R^3$, obtained by analytical chemical processing. Each content will be in turn treated as the response variable, hence the data set contains the pairs $(y_i, X_{i1}, X_{i2}, T_i)_{i=1,\ldots,215}$. For example, given a new spectrometric curve T_{new} and new measurements of protein and moisture contents $X_{\text{new}1}$ and $X_{\text{new}2}$, we aim to predict the corresponding fat content y_{new}. Similarly, we also examine the point forecast accuracy of the protein and moisture contents.

In order to assess the out-of-sample point forecast accuracy of the nonparametric functional estimator, we split the original samples into two subsamples (see also Ferraty & Vieu 2006, p.105). The first one is called learning sample, which contains the first 160 units $\{y_i, X_{i1}, X_{i2}, T_i\}_{i=1,\ldots,160}$. The second one is called testing sample, which contains the last 55 units $\{y_j, X_{j1}, X_{j2}, T_j\}_{j=161,\ldots,215}$. The learning sample allows us to build the regression model with the optimal smoothing parameter. The testing sample allows us to evaluate the prediction accuracy.

For comparison, we consider a nonparametric functional regression model studied in Shang (2013*a*) and the semi-functional partial linear regression model. Within each regression model, we consider the functional CV and Bayesian methods for estimating bandwidths, using modified R functions in the *npfda* package (`http://www.math.univ-toulouse.fr/staph/npfda`). To measure the performance of each bandwidth estimation method, we consider the mean absolute prediction errors (MAPE) and mean squared prediction errors (MSPE). They are expressed as

$$\text{MAPE} = \frac{1}{55}\sum_{j=161}^{215} |y_j - \widehat{y}_j|, \qquad \text{MSPE} = \frac{1}{55}\sum_{j=161}^{215} (y_j - \widehat{y}_j)^2.$$

For three different bandwidth estimation methods, the corresponding values of MAPE and MSPE obtained from two regression models are shown in Table 3. We found that the semi-functional partial linear regression model gives higher prediction accuracy than the nonparametric functional regression model, regardless the bandwidths are selected by functional CV or Bayesian methods. The Bayesian methods perform better than the functional CV in all three response variables based on the MSPE. When the aim is to predict fat content, the functional CV is superior to the Bayesian methods, as measured by the MAPE.

We are also interested in computing the prediction intervals nonparametrically. To this end, we first compute the cumulative density function (cdf) of the error distribution, over a set of grid points within a range, say -8 and 8. Then, we take the inverse of the cdf and find two grid points that are closest to the 2.5% and 97.5% quantiles. The 95% prediction intervals of the holdout samples are obtained by adding the two grid points to the point forecasts. For instance, the point forecasts of the fat content obtained from the Bayesian method with localized bandwidths are shown by black dots in Figure 1, while the 95% prediction intervals are shown by red parentheses.

Table 3: Out-of-sample MAPE and MSPE in the semi-functional partial linear regression model with bandwidth estimated by the functional CV and Bayesian methods. The number in parenthesis represents the sample standard deviation of the squared and absolute prediction errors. The bolded text represents the minimal MAPE and minimal MSPE.

Method	Response variable					
	Fat		Protein		Moisture	
	MAPE	MSPE	MAPE	MSPE	MAPE	MSPE
Nonparametric functional regression model						
CV	1.8466	5.5331	1.0283	2.5417	1.6389	4.3186
	(1.4705)	(7.4324)	(1.2295)	(7.6842)	(1.2895)	(6.3178)
Bayesian	1.8125	5.3097	1.0282	2.5313	1.6010	4.1125
local	(1.4359)	(7.1241)	(1.2253)	(7.6973)	(1.2561)	(5.9090)
Semi-functional partial linear regression model						
CV	**0.9075**	1.7855	0.8359	1.4319	0.8266	1.5440
	(0.9898)	(4.6592)	(0.8642)	(3.2824)	(0.9363)	(3.7173)
Bayesian	0.9863	1.9044	**0.7896**	**1.3837**	0.8417	1.5627
global	(0.9741)	(4.7294)	(0.8800)	(3.2289)	(0.9328)	(3.7750)
Bayesian	0.9687	**1.6031**	0.8022	1.3887	**0.7944**	**1.4952**
local	(0.8228)	(2.9567)	(0.8712)	(3.2027)	(0.9381)	(3.8744)

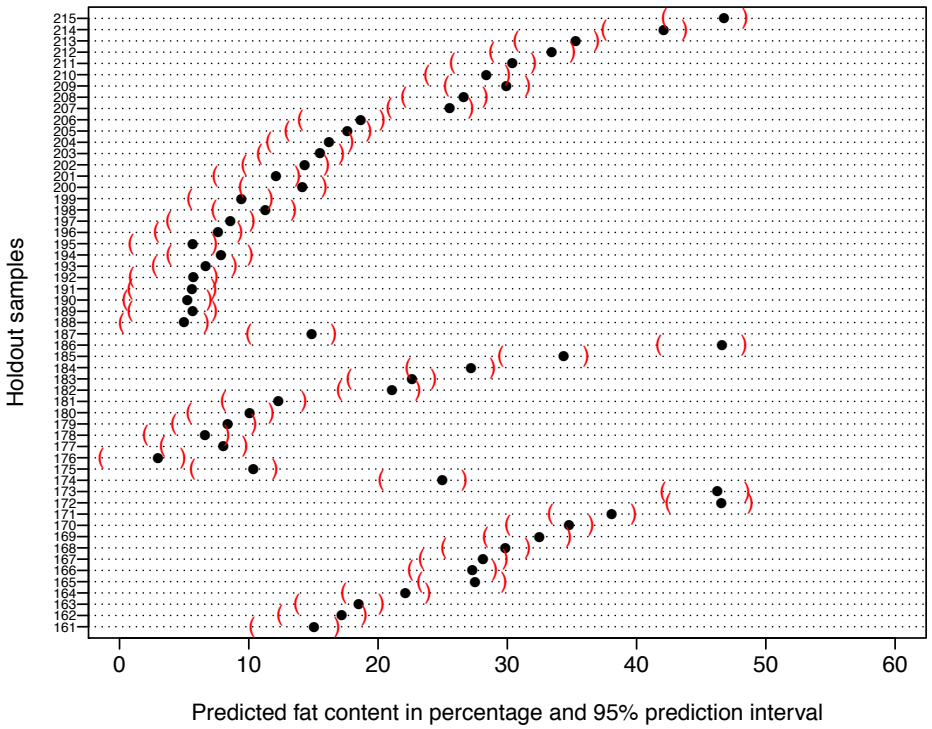

Figure 1: Plot of predicted fat contents in percentage and the 95% prediction intervals. The point forecasts of the fat content obtained from the Bayesian method with localized bandwidths are shown by black dots, while the 95% prediction intervals are shown by red parentheses.

6. Conclusion

This paper addresses the issue of bandwidth estimation in the semi-functional partial linear regression model. The regression function, consisting of a parametric component and a nonparametric component, is estimated by the ordinary least squares and functional NW estimators. However, the estimation accuracy jointly depends on optimal estimation of the same bandwidth parameter h_n. In addition, since we approximate the unknown error density by a kernel-form error density, the estimation accuracy of error density also depends on (at least) another bandwidth parameter b_n. In the context of semi-functional partial linear regression model, we introduce a Bayesian procedure for estimating both bandwidth parameters, simultaneously.

Illustrated by a series of simulation studies, the Bayesian method with localized bandwidths gives the most accurate estimation of regression function, followed closely by the Bayesian approach with a global bandwidth. In contrast, the benchmark functional CV performs the worst for estimating regression function. Based on the residuals, we further compared the estimation accuracy of error density between the Bayesian methods and so-called two-stage CV. The Bayesian method with localized bandwidths generally gives the most accurate estimation of error density, followed closely by the Bayesian method with a global bandwidth. To our surprise, the Bayesian method with a global bandwidth sometimes performs slightly better than the Bayesian method with localized bandwidths for complex error densities (such as, error densities (10)-(15) shown in the supplement material B). In comparison to the Bayesian methods, the two-stage CV gives the worst estimation accuracy across different error densities, except the strongly skewed and outlying error densities.

Illustrated by a well-studied spectroscopy data set, we found that the functional CV performs better than the Bayesian methods in terms of MAPE, when the aim is only to predict fat content. As measured by MSPE, the Bayesian methods are more superior to the functional CV in all three different response variables. In addition, the Bayesian methods are capable of producing nonparametric prediction intervals for quantifying the prediction uncertainty.

References

Ahmad, I. & Li, Q. (1997), 'Testing symmetry of an unknown density function by kernel method', *Journal of Nonparametric Statistics* **7**(3), 279–293.

Akritas, M. G. & Van Keilegom, I. (2001), 'Non-parametric estimation of the residual distribution', *Scandinavian Journal of Statistics* **28**(3), 549–567.

Aneiros-Pérez, G. & Vieu, P. (2006), 'Semi-functional partial linear regression', *Statistics and Probability Letters* **76**(11), 1102–1110.

Aneiros-Pérez, G. & Vieu, P. (2008), 'Nonparametric time series prediction: a semi-functional partial linear modeling', *Journal of Multivariate Analysis* **99**(5), 834–857.

Anglin, P. M. & Gencay, R. (1996), 'Semiparametric estimation of a hedonic price function', *Journal of Applied Econometrics* **11**(6), 633–648.

Ansley, C. F. & Wecker, W. E. (1983), Extensions and examples of the signal extraction approach to regression, *in* A. Zellner, ed., 'Applied time series analysis of economic data', Bureau of the Census, Washington, pp. 181–192.

Barrientos-Marin, J., Ferraty, F. & Vieu, P. (2010), 'Locally modelled regression and functional data', *Journal of Nonparametric Statistics* **22**(5), 617–632.

Berlinet, A., Elamine, A. & Mas, A. (2011), 'Local linear regression for functional data', *Annals of the Institute of Statistical Mathematics* **63**(5), 1047–1075.

Boj, E., Delicado, P. & Fortiana, J. (2010), 'Distance-based local linear regression for functional predictors', *Computational Statistics & Data Analysis* **54**(2), 429–437.

Bowman, A. W. (1984), 'An alternative method of cross-validation for the smoothing of density estimates', *Biometrika* **71**(2), 353–360.

Burba, F., Ferraty, F. & Vieu, P. (2009), 'k-nearest neighbour method in functional nonparametric regression', *Journal of Nonparametric Statistics* **21**(4), 453–469.

Chen, H. (1988), 'Convergence rates for parametric components in a partly linear model', *The Annals of Statistics* **16**(1), 136–146.

Cheng, F. (2002), 'Consistency of error density and distribution function estimators in nonparametric regression', *Statistics & Probability Letters* **59**(3), 257–270.

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

Cheng, F. & Sun, S. (2008), 'A goodness-of-fit test of the errors in nonlinear autoregressive time series models', *Statistics and Probability Letters* **78**(1), 50–59.

Dabo-Niang, S. & Guillas, S. (2010), 'Functional semiparametric partially linear model with autoregressive errors', *Journal of Multivariate Analysis* **101**(2), 307–315.

Dette, H., Kusi-Appiah, S. & Neumeyer, N. (2002), 'Testing symmetry in nonparametric regression models', *Nonparametric Statistics* **14**(5), 477–494.

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

Engle, R., Granger, C., Rice, J. & Weiss, A. (1986), 'Semiparametric estimates of the relation between weather and electricity sales', *Journal of the American Statistical Association* **81**(394), 310–320.

Escanciano, J. C. & Jacho-Chávez, D. T. (2012), '$\sqrt{n}$ uniformly consistent density estimation in nonparametric regression models', *Journal of Econometrics* **167**(2), 305–316.

Ferraty, F., ed. (2011), *Recent advances in functional data analysis and related topics*, Springer, Berlin.

Ferraty, F. & Romain, Y., eds (2010), *The Oxford handbook of functional data analysis*, Oxford University Press, Oxford.

Ferraty, F., Van Keilegom, I. & Vieu, P. (2010), 'On the validity of the bootstrap in nonparametric functional regression', *Scandinavian Journal of Statistics* **37**(2), 286–306.

Ferraty, F. & Vieu, P. (2002), 'The functional nonparametric model and application to spectrometric data', *Computational Statistics* **17**(4), 545–564.

Ferraty, F. & Vieu, P. (2006), *Nonparametric functional data analysis: theory and practice*, Springer, New York.

Ferraty, F. & Vieu, P. (2009), 'Additive prediction and boosting for functional data', *Computational Statistics & Data Analysis* **53**(4), 1400–1413.

Gabrys, R., Horváth, L. & Kokoszka, P. (2010), 'Tests for error correlation in the functional linear model', *Journal of the American Statistical Association* **105**(491), 1113–1125.

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

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

Geweke, J. (2010), *Complete and incomplete econometric models*, Princeton University Press, Princeton, USA.

Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. (1996), *Markov chain Monte Carlo in practice*, Chapman & Hall, London.

Goutis, C. (1998), 'Second-derivative functional regression with applications to near infrared spectroscopy', *Journal of the Royal Statistical Society: Series B* **60**(1), 103–114.

Hall, P. (1987), 'On Kullback-Leibler loss and density estimation', *Annals of Statistics* **15**(4), 1491–1519.

Härdle, W., Liang, H. & Gao, J. (2000), *Partially linear models*, Physica-Verlag, New York.

Heckman, N. (1986), 'Spline smoothing in a partly linear model', *Journal of the Royal Statistical Society, Series B* **48**(2), 244–248.

Horváth, L. & Kokoszka, P. (2012), *Inference for Functional Data with Applications*, Springer, New York.

Jaki, T. & West, R. W. (2008), 'Maximum kernel likelihood estimation', *Journal of Computational and Graphical Statistics* **17**(4), 976–993.

Jaki, T. & West, W. (2011), 'Symmetric maximum kernel likelihood estimation', *Journal of Statistical Computation and Simulation* **81**(2), 193–206.

Kim, S., Shephard, N. & Chib, S. (1998), 'Stochastic volatility: likelihood inference and comparison with arch models', *Review of Economic Studies* **65**(3), 361–393.

Marron, J. S. & Wand, M. P. (1992), 'Exact mean integrated squared error', *The Annals of Statistics* **20**(2), 712–736.

McLachlan, G. & Peel, D. (2000), *Finite mixture models*, Wiley, New York.

Meyer, R. & Yu, J. (2000), 'BUGS for a Bayesian analysis of stochastic volatility models', *Econometrics Journal* **3**(2), 198–215.

Muhsal, B. & Neumeyer, N. (2010), 'A note on residual-based empirical likelihood kernel density estimation', *Electronic Journal of Statistics* **4**, 1386–1401.

Neumeyer, N. & Dette, H. (2007), 'Testing for symmetric error distribution in nonparametric regression models', *Statistica Sinica* **17**(2), 775–795.

Ramsay, J. O. & Silverman, B. W. (2005), *Functional data analysis*, 2nd edn, Springer, New York.

Rice, J. (1986), 'Convergence rates for partially splined models', *Statistics and Probability Letters* **4**(4), 203–208.

Robbins, H. & Monro, S. (1951), 'A stochastic approximation method', *The Annals of Mathematical Statistics* **22**(3), 400–407.

Robert, C. P. & Casella, G. (2010), *Introducing Monte Carlo methods with R*, Springer, New York.

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

Robinson, P. (1988), 'Root-N-consistent semiparametric regression', *Econometrica* **56**(4), 931–954.

Samb, R. (2011), 'Nonparametric estimation of the density of regression errors', *C. R. Acad. Sci. Paris, Ser. I* **349**(23-24), 1281–1285.

Schmalensee, R. & Stoker, T. M. (1999), 'Household gasoline demand in the united states', *Econometrica* **67**(3), 645–662.

Severini, T. A. & Staniswalis, J. G. (1994), 'Quasilikelihood estimation in semiparametric models', *Journal of the American Statistical Association* **89**, 501–511.

Shang, H. L. (2013*a*), 'Bayesian bandwidth estimation for a nonparametric functional regression model with unknown error density', *Computational Statistics & Data Analysis* **67**, 185–198.

Shang, H. L. (2013*b*), 'Functional time series approach for forecasting very short-term electricity demand', *Journal of Applied Statistics* **40**(1), 152–168.

Speckman, P. (1988), 'Kernel smoothing in partial linear models', *Journal of the Royal Statistical Society, Series B* **50**(3), 413–436.

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

Yao, F. & Müller, H.-G. (2010), 'Functional quadratic regression', *Biometrika* **97**(1), 49–64.

Zhang, T. & Wang, Q. (2012), 'Semiparametric partially linear regression model for functional data', *Journal of Statistical Planning and Inference* **142**(9), 2518–2529.

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

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

Zhou, J. & Chen, M. (2012), 'Spline estimators for semi-functional linear model', *Statistics and Probability Letters* **82**(3), 505–513.