STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **9**, June 2021, pp 351–367.
Published online in International Academic Press (www.IAPress.org)

Semiparametric Smoothing Spline in Joint Mean and Dispersion Models with Responses from the Biparametric Exponential Family: A Bayesian Perspective

Héctor Zárate*, Edilberto Cepeda

Department of Statistics, Universidad Nacional de Colombia, Bogot, Colombia

Abstract This article extends the fusion among various statistical methods to estimate the mean and variance functions in heteroscedastic semiparametric models when the response variable comes from a two-parameter exponential family distribution. We rely on the natural connection among smoothing methods that use basis functions with penalization, mixed models and a Bayesian Markov Chain sampling simulation methodology. The significance and implications of our strategy lies in its potential to contribute to a simple and unified computational methodology that takes into account the factors that affect the variability in the responses, which in turn is important for an efficient estimation and correct inference of mean parameters without the specification of fully parametric models. An extensive simulation study investigates the performance of the estimates. Finally, an application using the *Light Detection and Ranging technique*, LIDAR, data highlights the merits of our approach.

Keywords Markov chain Monte Carlo, Double generalized linear mixed models, Biparametric exponential family, Spline models.

AMS 2010 subject classifications 60J10, 65C05

DOI: 10.19139/soic-2310-5070-671

1. Introduction

A convenient approach to render a statistical inference in a semiparametric method that combines known parametric structures with flexible nonparametric components is via a mixed model formulation. From this framework, penalized splines can be formulated and fit using standard software such as *PROC MIXED* in *SAS* [1] and *lme*( ) in *R* [2]. Within the mixed model software modeling, the regression function with random effects applies to responses from the exponential family using generalized linear mixed models (*GLMMs*) and provides an automatic smoothing parameter choice by using a restricted maximum likelihood estimation of the variance. The Bayesian equivalent to inference in a semiparametric regression relies on *WinBUGS* codes, which have become standard software for analyses, accompanied by Markov chain Montecarlo *MCMC* methods [3]. Similar in spirit are the Bayesian versions of *TURBO* and *CART* proposed by [4].

In applied modeling, it has been a common practice either to assume a constant variance of the data or to use transformations to reduce the variability. However, frequently, homoscedasticity is violated. A typical example of this issue is given in Figure 1 with *LIDAR* data obtained from the atmospheric monitoring of pollutants, from which an ignorance of the heteroskedasticity led to incorrect inferences and an inefficient estimation, which caused misleading conclusions from hypothesis testing[†]. In addition, an estimation of the variance function is either of

*Correspondence to: Héctor Zárate (Email: hmzarates@unal.edu.co). Department of Statistics, Universidad Nacional de Colombia.
[†]Notably, the adjustment of the non-constant variance through transformations is valid only when the conditional variance is a function of the conditional mean.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

intrinsic interest by itself to understand how the variability changes with predictors or plays an important role in approximating other quantities.Therefore, an estimator of the variance function is required for predicting and estimating efficiently the mean function.

Different Bayesian and frequentist approaches have been implemented to simultaneously estimate the mean and variance functions where the intractability of both the likelihood function and the posterior distribution is the main issue of inference. Even though analytical approximations are faster than numerical approximation alternatives, there have been extensive research studies that have designed effective Markov chain Monte carlo, *MCMC*, algorithms to handle the generalized responses. However, there are two main drawbacks to *MCMC* fitting when using standard software. First, major difficulties are associated with assessing the convergence, which can be slow when samples of the dataset to be analyzed are large or the model to be fitted is complex. Early work regarding an estimation of the dispersion function in a fully nonparametric fashion dealing with the extended double exponential family and the use of *P-spline* approach was proposed by [5]. Likewise, [6] proposed a method for the mean and variance estimation where the responses were modeled using the double exponential family of the distributions and the mean and dispersion functions were specified as additive functions of the predictors. On the other hand, a frequentist iterative approach to heteroscedastic errors with the penalized spline methodology was developed by [7]. A fully Bayesian approach based on the *MCMC* algorithm, which provides the joint posterior of all parameters, was implemented by [8]. A flexible mean and dispersion function estimation in generalized additive models in the context of semiparametric models was proposed by [6]. [9] proposed a hybrid algorithm based on a combination of the Metropolis-Hasting algorithm and a Gibbs sampler for semiparametric joint mean and variance models on the basis of a *B-spline* approximation of nonparametric components. In large datasets, the problem of jointly estimating the mean and variance functions is handled by a neural network methodology and variational Bayesian approach, which avoids the simulation task of the *MCMC* algorithms and approximates the posterior distribution with a low bound. This approach has been recently implemented for mixed models and for much more complex models by [10]. On the other hand, an approximation of the posterior distribution by building transition distributions based on working variables, which follows from the relationship between the maximum likelihood estimation by using Fisher scoring and a weighted least squares method in the biparametric exponential family for the mean and variance function estimation for linear models, was implemented by [11] in a regression context.

Given the connection between linear mixed models, splines and Bayesian modeling, the goal of this paper is to extend this statistical partnership to jointly infer the mean and dispersion of response variables originated from biparametric exponential families. Specifically, the Gaussian model, which belongs to this family, is studied. The inference is focused on augmenting the vector of the regression coefficient and the design matrix to include random effects. Although the dimensionality of the parameter space increases dramatically with large datasets, shrinking by penalizing splines could render this approach practical.

In this paper, we contribute to this work by extending the advantage to the link between smoothing and the biparametric exponential family to jointly estimate mean and dispersion functions from a Bayesian perspective based on working variables. The benefits of this heteroskedastic semiparametric strategy lies in its connection with graphical model representations. Moreover, the systematic part of the mean and dispersion functions could be extended to take into account non-linearity in the parameters, which are more appropriate for some real applications in statistics and computing science. Furthermore, the modularity of the approach allows easy extensions to models with increasing complexity.

Recently, there has been growing interest in semiparametric models when the measurement error in a predictor distorts the relation to the response, which has led to the presence of intractable integrals in the estimation process. The Bayesian approach described in this paper is important for implementation. see [12] and [13]. Furthermore, other extensions that could be also applicable are related to accommodating more complex errors structures such as those arising in additive models, longitudinal studies or multilevel models [14]. However, new variants of GLM models where the link function relates the conditional mean of the response variable to a transformation of predictors using neural nets, which is a powerful tool for functional approximation, is a direction for future research [15]. Another development will be concerning new challenges to MCMC with better and large databases. For example, in the divide-and-conquer approach, the whole data set is partitioned into batches and run separately

with MCMC algorithms independently for each data batch. This methodology combines the simulated parameters to approximate the original posterior distribution. See [16] for a survey in accelerating MCMC algorithms.

Therefore, sub-models for the mean and variance functions depend on some covariates parametrically and others nonparametrically. A nonparametric function approximation is reached by the linear combination of the basis functions keeping the nonparametric regression part relatively simple by using low-rank penalized splines. Even though the *scatterplot smoothing* context is utilized in this paper, the methodology could be extended to allow a more efficient handling of standard and non-standard data such as longitudinal data and spatial correlation. Additionally, the analyst has to potentially deal with non-Gaussian models in real-world applications. For example, if the response variable is related to a proportion, then it could have categorical responses, outliers, data sparsity and missingness, among other issues.

There are two areas of data analytic research that could be used to handle the presence of non-linear relations that avoid the restrictions posed by parametric models. The first is in the statistics literature and is referred to as penalized splines, which have become a popular non-parametric tool because they use a low-rank basis and can be seen as mixed models. See [7] and [17] for an extensive review. it is worth to notice that under this method, the spline basis is chosen on some sufficiently large set of knots and the unnecessary structure is penalized. The second is in the computing science and refers to kernel machines as an important tool for classification and regression problems supported by the theory of reproducing a kernel Hilbert space (*RKHS*), which can also be formulated as fits in mixed model representation and the solution come from a minimization problem in a functional space. The theoretical issues can be found in [18], [19] and [20]. However, a friendly reference for implementing this approach is in [21]. Moreover, there is a connection between penalized splines and smoothing splines detailed in [22]

The link between penalized spline smoothing and linear mixed models has been studied extensively where an unknown smoothing function is estimated by replacing the function by a linear combination of the basis functions and there exists a mixed model representation of a penalized spline that could be implemented for the different models. Furthermore, generalized mixed models have been a vehicle not only for analyzing data handling grouping structures but also for using regression models that contain at least one function being modeled nonparametrically. Many applications of these models handle a range of applications, for example, to account for within-subject correlation, multilevel models, fixed and random components and smoothing. *GLMM*s can synthesize a likelihood-based approach for a variety of outcomes, accommodate the overdispersion, and model the dependence among outcome variables that are inherent in longitudinal studies or repeated measures designs. In addition, *GLMM* analyze complex datasets, and smoothing is derived from the connection between nonparametric models and mixed models. Nevertheless, a Bayesian modeling approach to semiparametric regression has advantages due to the attractiveness of the hierarchical Bayesian models for quantifying multiple sources of variability, dealing with missing data and measurement. Moreover, this approach treats parameters at random and benefits from the flexibility of nonparametric models and the exact inference provided by the Bayesian approach. The good mixing properties of the *MCMC* chains could be generated by using low-rank thin plate splines, which have good numerical properties because the correlation among parameters is much more smaller than that using another basis function. Fitting and testing could be conducted through the paradigm of the likelihood.

This paper contains six sections apart from this introduction and proceeds as follows. In Section 2, we describe the double stochastic generalized linear model with splines. In Section 3, we summarize the Bayesian strategy and describe the main steps involved in the *MCMC* algorithm to draw the inference from the proposed model. Furthermore, in Section 4, we present a simulation analysis to study the performance of the Bayesian methodology compared with those implemented in standard software. In Section 5, we present the results with a real dataset. Finally, the conclusions are presented in Section 6.

2. Spline Double Generalized Linear Models

Consider the two-parameter exponential family distribution discussed by [23] and [24], which takes into account the exponential family of two parameters. This family has the following general density function:

$$f\left(y|\theta,\tau\right)=b\left(y\right)e^{\left[\theta y+\tau T(y)-\rho(\theta,\tau)\right]} \tag{1}$$

where f is a density from the p.d.f. of the parametric family, P_t, with respect to the finite appropriate finite measure. The function $\rho(\cdot)$ is the *cumulant function*. On the other hand, these authors stated that if the sufficient statistic $T(y)$, is convex, then for a common mean, $Var(y)$ increases in τ. A particular case occurs when $\tau=0$, which belongs to the one parametric exponential family of distributions. Distributions that belong to this family include Gauss, Poisson, and Gamma.

With the regularity conditions of Cramer Rao, the following properties are fulfilled:

$$\frac{\delta\rho}{\delta\theta}=E\left(y|\theta,\tau\right)=\mu$$

$$\frac{\delta^2\rho}{\delta\theta^2}=Var\left(y|\theta,\tau\right)$$

by reparametrizing the density function with respect to the mean, the likelihood function is established in terms of μ and τ.

The class of doubly semiparametric stochastic generalized linear models with splines as random effects can be stated as

$$y_i\sim DE(\mu_i,\tau_i) \tag{2}$$

$$h(\mu_i)=\mathbf{x_i'}\beta+f(x_i) \tag{3}$$

and

$$g(\tau_i)=\mathbf{z_i'}\gamma+l(z_i) \tag{4}$$

where $DE(\mu_i,\tau_i)$ denote the double exponential distribution, f and l are nonlinear functions. A particular case correspond to the Gauss distribution. Variance function estimation which allows the variance to be a function of the predictors and consequently treating the variance as if it were a regression function. The approach is penalized by assuming that the nonlinear coefficients are random effects. Stating the model as a double mixed model

$$E(y/x)=\beta_0+\beta_1x+\sum u_kz_k$$

$$g(\tau)=EXP\{\gamma_o+\gamma_1x+\sum v_kz_k\}$$

The entire model

$$y/u,v\sim DE\left(X\beta+Zu,diag\{exp(X\gamma+Zv\}\right)$$

with the random effect being doubled as well

$$\begin{bmatrix} u \\ v \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_u^2 I & 0 \\ 0 & \sigma_v^2 I \end{bmatrix}\right)$$

The Spline Double Generalized Linear Model, *SDGLM*, are defined by the following three components:

1. Random component: Let $Y_1, ..., Y_N$ be independent random observations, where Y_i conditional to the random effects comes from the biparametric exponential family distribution with the density function given in (1) with $E(Y_i) = \mu_i$ and $V(Y_i) = \sigma_i$
2. The systematic component: Assuming the penalized spline formulation, as adopted in [7], the systematic component $\eta_i = (\eta_{1i}, \eta_{2i})$ is given by:

$$\eta_{1i} = x_i'\boldsymbol{\beta} + \sum_{k=1}^{K_u} u_k z_k^u(x) \qquad u_k \overset{iid}{\sim} N(0, \sigma_u^2)$$

$$\eta_{2i} = z_i'\boldsymbol{\gamma} + \sum_{k=1}^{K_v} v_k z_k^v(x) \qquad v_k \overset{iid}{\sim} N(0, \sigma_v^2)$$

   where $\boldsymbol{\beta} = (\beta_1, \beta_2, ...\beta_p)'$ and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, ..., \gamma_r)'$ are unknown parameter vectors and $\{z_k^u : 1 \leq k \leq k_u\}$ and $\{z_k^v : 1 \leq k \leq k_v\}$ are set of fixed known knots of size k_u and k_v, respectively. The truncated polynomial spline basis is chosen for convenience. Other common basis functions include B-splines, RK basis, depending on the application.
3. The link functions: The link functions provide the relationship between the random and systematic components for the mean and dispersion of the bi-parametric family[‡]:

$$\mu_i = h^{-1}(\eta_{1i})$$

$$\tau_i = g^{-1}(\eta_{2i})$$

   where h and g are monotonic twice differentiable functions.

3. A Bayesian Estimation of the Mean and Dispersion Functions

To fit the submodels under the Bayesian paradigm, we rely on *MCMC* methods to simulate samples for the joint posteriors of interest. In this paper, the mean and dispersion regression structures are given by:

$$h(\mu_i) = x_{1i}'\beta + \sum_{k=1}^{K_u} u_k (x_i - k_k)_+^2 \qquad u \sim N(0, \sigma_c^2) \qquad k = 1, ..., K_u$$

$$g(\sigma_i^2) = z_{1i}'\delta + \sum_{k=1}^{K_v} v_k (z_i - k_k^v)_+^2 \qquad v \sim N(0, \sigma_v^2) \qquad k = 1, ..., K_v$$

[‡] The dispersion function is modeled as a linear mixed model. The term including the nonlinear term must be penalized to ensure a stable estimation, assuming that these coefficients are the random effects.

where β and δ are vectors for the fixed effects for the mean and dispersion functions respectively. *h* and *g* being appropriate real functions and $u(\cdot)$, $v(\cdot)$ represent the truncated spline line basis. Furthermore $\left\{K_k^u\right\}_{k=1}^{K_u}$ and $\left\{K_k^v\right\}_{k=1}^{K_v}$ are the knots.

According to [25] and [26], from the Bayesian perspective, the submodels can be placed in the mixed form by augmenting the vector of the regression coefficients and the design matrix with *h* and *g* as appropriate real functions.

$$\beta^{(aug)} = \begin{bmatrix} \beta \\ u_1 \\ \vdots \\ u_n \end{bmatrix} \quad \delta^{(aug)} = \begin{bmatrix} \delta \\ v_1 \\ \vdots \\ v_n \end{bmatrix}$$

$$x^{(aug)} = \begin{bmatrix} 1 & x_1 & (x_1 - k_1)_+ & \dots & (x_1 - k_k)_+ \\ \vdots & & & \ddots & \vdots \\ 1 & x_n & (x_n - k_1)_+ & \cdots & (x_n - k_k)_+ \end{bmatrix}$$

$$z^{(aug)} = \begin{bmatrix} 1 & z_1 & (z_1 - k_1^v)_+ & \dots & (z_1 - k_k^v)_+ \\ \vdots & & & \ddots & \vdots \\ 1 & z_n & (z_n - k_1^v)_+ & \cdots & (z_n - k_k^v)_+ \end{bmatrix}$$

The systematic components of the mean and variance sub-models can be written as:

$$\boldsymbol{\eta}_1 = \boldsymbol{x}^{(aug)} \boldsymbol{\beta}^{(aug)}$$
$$\boldsymbol{\eta}_2 = \boldsymbol{z}^{(aug)} \boldsymbol{\delta}^{(aug)}$$

To estimate the parameters of these sub-models by using a Bayesian approach, independent normal priors are assumed for the mean and variance parameters:

$$\boldsymbol{\beta}^{(aug)} \sim N(\boldsymbol{b}, \boldsymbol{B})$$
$$\boldsymbol{\beta}^{(aug)} \sim N((a', 0', \dots, 0')', (\boldsymbol{R_1}, \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_1))$$
$$\boldsymbol{\delta}^{(aug)} \sim N(g, \boldsymbol{G})$$

$$\boldsymbol{\delta}^{(aug)} \sim N((d', 0', \dots, 0')', (\boldsymbol{R_2}, \boldsymbol{\Sigma}_2, \dots, \boldsymbol{\Sigma}_2))$$

The specification in a vector form is given by:

$$\theta = \begin{pmatrix} \beta^{(aug)} \\ \delta^{(aug)} \end{pmatrix} \sim N\left(\theta_0 = \begin{bmatrix} b_0 \\ g_0 \end{bmatrix}, \Sigma_0 = \begin{bmatrix} \beta_0 & c \\ c' & G_0 \end{bmatrix}\right)$$

From the Bayes theorem, the joint posterior distribution function is given by:

$$\pi(\boldsymbol{\beta}^{(aug)}, \boldsymbol{\delta}^{(aug)}|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}) \sim L(\boldsymbol{\beta}^{(aug)}, \boldsymbol{\delta}^{(aug)}|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z})P(\boldsymbol{\beta}^{(aug)}, \boldsymbol{\delta}^{(aug)})$$

$$\pi\left(\beta^{(aug)}, \delta^{(aug)}\right) \sim |\Sigma|^{-1/2} exp\left\{-\frac{1}{2}\left(Y - X'\beta^{(aug)}\right)^1 \Sigma^{-1}\left(Y - X'\beta^{(aug)}\right) - \frac{1}{2}\left(\theta - (\theta - \theta_0)^1 \Sigma_0^{-1}(\theta - \theta_0)\right)\right\}$$

where $L(\cdot)$ denotes the likelihood function and $P(\cdot)$ the joint prior distribution.

Given that $\pi(\beta^{(aug)}, \delta^{(aug)})$ is analytically intractable and difficult to obtain samples from it, it use conditional posterior distributions. We propose sampling $(\beta^{(aug)}, \delta^{(aug)}/Y, X)$ through an iterative process. Sampling $\beta^{(aug)}$ and $\delta^{(aug)}$ from the conditional distributions, which are also intractable. Therefore, in order to get samples from theses conditionals distributions we build transition kernels which allow us to simulate samples that will be part of the posterior distribution GLM working variables.

$$\pi(\boldsymbol{\beta}^{(aug)}|\boldsymbol{\delta}^{(aug)}, \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z})$$

$$\pi(\delta^{(aug)}|\beta^{(aug)}, Y, X, Z)$$

To get samples from $\pi(\beta^{(aug)}/\delta^{(aug)}, X, Y)$ we follow the methodology proposed by [25] and [27] by defining working observational variables as: $\tilde{y}_i, i = 1, 2, ..., n$ by:

$$\tilde{y}_i = h(\mu_i^{(c)}) + h'(\mu_i^{(c)})(y_i - \mu_i^{(c)})$$

This variable has $E(\tilde{y}_i) = h\left(\mu_i^{(c)}\right)$ with $h\left(\mu_i^{(c)}\right)$ following the linear structure assumed in the semiparametric model, which allows us to apply the first-order Taylor approximation of the function h around the current value of $\mu_i^{(c)}$. A key advantage of this method is that assuming normal prior distribution for the regression parameters and for the working variable, a normal transition kernel can be obtained by the combination of the working model and the prior regression parameter distributions. Additionally, given that the link functions are monotonically differentiable such as logarithmic and logistic function, a first-order Taylor approximation of h is a good approximation of h around the current values of θ.
In consequence, if $\boldsymbol{\beta}^{(aug)(c)}$ is the current value of $\boldsymbol{\beta}^{(aug)}$, the appropriate working observation variables to sample $\boldsymbol{\beta}^{(aug)}$ are:

$$\tilde{y}_i = x_i^{'(aug)}\boldsymbol{\beta}^{(aug)(c)} + h'(\mu_i^{(c)})(y_i - \mu_i^{(c)})\, i = 1, 2, \ldots, n,$$

for which

$$E(\tilde{y}_i) = x_i^{'(aug)}\boldsymbol{\beta}^{(aug)(c)}$$

and at the same time a simple and general expression is obtained for the variance:

$$V(\tilde{y}_i) = [h^{'}(\mu_i^{(c)})]^2 Var(y_i)$$

Therefore, the kernel transition function obtained from the combination of prior distribution with the working observational model, is given by:

$$q_1(\boldsymbol{\beta}^{(aug)}|\boldsymbol{\beta}^{(aug)(c)}, \boldsymbol{\delta}^{(aug)(c)}) \sim N(\boldsymbol{b}^*, \boldsymbol{B}^*)$$

where

$$\boldsymbol{b}^* = \boldsymbol{B}^*(\boldsymbol{B}^{-1}b + \boldsymbol{X'\Sigma}^{-1}\widetilde{Y}),\ \boldsymbol{B}^* = (\boldsymbol{B}^{-1} + \boldsymbol{X'\Sigma X})^{-1},\ \boldsymbol{\Sigma} = diag(\sigma_u^2)$$

On the other hand, to get samples of the variance parameters, the working variable is given by:

$$\hat{y}_i = g(\sigma_i^{2(c)}) + g'(\sigma_i^{2(c)})(t_i - \sigma_i^{2(c)})$$

where $E(t_i) = \sigma_i^2$ and $g(\sigma_i^{2(c)}) = z_i'\gamma$

$$\hat{y}_i = Z_i^{'(aug)}\boldsymbol{\delta}^{(aug)(c)} + g'(\sigma_i^{2(c)})(t_i - \sigma_i^{2(c)})\ i = 1, 2, \ldots, n$$

This random variable has mean and working observational variance given by

$$E(\hat{y}_i) = Z_i^{'(aug)}\boldsymbol{\delta}^{(aug)(c)}$$
$$V(\hat{y}_i) = [g'(\sigma_i^{2(c)})]^2 v(t_i)$$

Thus, this transition Kernel is given by:

$$q_2(\boldsymbol{\delta}^{(aug)}|\boldsymbol{\delta}^{(aug)(c)}, \boldsymbol{\beta}^{(aug)(c)}) \sim N(\boldsymbol{g}^*, \boldsymbol{G}^*)$$

where $\boldsymbol{g}^*$ and $\boldsymbol{G}^*$ are $\boldsymbol{g}^* = \boldsymbol{G}^*(\boldsymbol{G}^{-1}g + \boldsymbol{Z'}\psi^{-1}\widetilde{Y})$ and $\boldsymbol{G}^* = (\boldsymbol{G}^{-1} + \boldsymbol{Z'\psi^{-1}Z})^{-1}$, respectively. On the other hand, the values of $\boldsymbol{g}$ and $\boldsymbol{G}$ are given by the prior distribution $\boldsymbol{\delta}^{(aug)}|\boldsymbol{\beta}^{(aug)} \sim N(\boldsymbol{g}, \boldsymbol{G})$, and $\boldsymbol{\psi} = V(\hat{Y}_i)$

3.1. MCMC algorithm

Using the proposal densities q_1 and q_2 described, the algorithm consists of updating the mean and dispersion parameters associated with each $q - density$ until the posterior is approximated. During each step of the algorithm, the most recent parameters are used. With the sampling proposals derived, we will now provide the steps to implement an *MCMC* algorithm based on [27] that considers the joint modeling of the mean and dispersion parameters in the semiparametric model when the response variable follows the Gaussian distribution. The implementation for the other distributions of the exponential family, for example, the Poisson and Beta distributions, is similar. For the purposes of the mean comparison, the root mean squared errors were calculated.

MCMC Algorithm

The components are updated in these steps:

1. Set the iteration counter chain to $j = 1$ and give initial values β_0^{aug}, δ_0^{aug} and $\sigma_{v_o}^2$
2. Propose a new value ξ generated from the proposed density $q_1(\beta^{aug(j-1)}|.)$
3. Compute the acceptance probability of the movement $\alpha(\beta^{aug(j-1)}, \xi)$. If it is accepted, then $\beta^{aug(j)} = \xi$; otherwise $\beta^{aug(j)} = \beta^{aug(j-1)}$
4. Propose a new value ξ, generated from the proposed density $q_2(\delta^{aug(j-1)}|.)$
5. Compute the acceptance probability of movement $\alpha(\delta^{aug(j-1)}, \xi)$. If the movement is accepted then $\delta^{aug(j)} = \xi$; otherwise $\delta^{aug(j)} = \delta^{aug(j-1)}$
6. Finally, update the counter from j to $j + 1$ and return to step 2.

4. Simulation Study

To confirm the effectiveness of the proposed Bayesian algorithm for general heteroscedastic semiparametric models where the true model is known, we conducted a comprehensive simulation study. As seen from Table 1, a set of four mean and standard deviation functions with different patterns was generated according to model (1). There were *n=500* replications for each simulation, and the predictor $x \sim U(0, 1)$. In the four experiments advocated by [10], the functions chosen for mean and dispersion provide different patterns of evolution. In setting A the mean function follow a smooth sinusoidal pattern and the standard deviation take a double u form. Next, the setting B generate a mean with rapid rise, followed by a steep fall at the end of the simulation. The standard deviation function decline smoothly with some jumps. Next, the mean from setting C presents one section that remained steady and the other that exhibit a declining pattern. Finally, the mean function from setting D fluctuated widely and the standard deviation present two clusters. In other words, these mean and standard deviation functions display different non-linearities that can be estimated in a semiparametric fashion. On the other side, following standard assumptions about the hyper-parameters. [8] The fixed effects parameters for the mean function are assumed apriori independent with a very large variance. That is, $\beta_i \sim N(0, 10^6)$. Moreover, for the fixed effects used in the dispersion model, we also employed independent $N(0, 10^6)$.

We compared the mean function for our method with those that include spline and the *RK* basis in the customized software *PROC MIXED* from *SAS*. The results presented here are based on *P-splines* but the checks show that smoothing splines provide a similar result when 200 simulated samples were generated.

Setting	$f(x)$	$\log g(x)$
A	$sin(3\pi x^2)$	$0.1 + cos(4\pi x)$
B	$-1.02x + 0.018x^2 + 0.4\phi(x; 0.38, 0.08)$ $+0.08\phi(x; 0.75, 0.03)$	$-0.5 - \Phi(x; 0.2, 0.1) + 0.3x^2$
C	$0.35\phi(x; 0.01, 0.08) + 1.9\phi(x; 0.45, 0.23)$ $+1.8\{1 - \phi(x; 0.7, 0.14)\}$	$0.3\phi(x; 0, 0.2) + 0.4\phi(x; 1, 0.1)$
D	$sin(3\pi x^2) - 1.02x + 0.018x^2$ $+0.4\phi(x; 0.38, 0.08)$	$cos(4\pi x) - 0.4 + 0.3x^2$ $-\Phi(x; 0.2, 0.1)$

Table 1: data were simulated according to model (1) with n=500 and $x \sim U(0, 1)$, $\phi(\cdot, \mu, \sigma^2)$ and $\Phi(\cdot, \mu, \sigma^2)$ represent the density and distribution functions of the Normal distribution
Source: [10].

In each panel of Figure 1, all curve estimates in the interior are visually similar . However, the *P-splines* have a tendency to deviate from the underlying curve in settings (b) and (d). Moreover, a potentially serious problem with *smoothing splines* is a lack of spatial adaptivity, which is the ability to impose less smoothing where the regression

function exhibits a sharp curvature. An attractive advantage of penalized splines compared to smoothing splines is the ease at which *MCMC* schemes fit semiparametric models with reduced basis functions.
We will compare the frequentist properties of the proposed Bayesian algorithm with two classical models that incorporate the spline and *RK* basis in the architecture of the mixed models customized software. The mixed model software provides an automatic smoothing parameter choice via the restricted maximum likelihood of variance components. We now examine how well the underlying curve is recovered for each simulation setting with the three smoothing models described. For all the simulations settings, we evaluated the quality of the obtained fits via the mean squared error, *MSE*. That is, between the true curve and the fits from the proposed *MCMC* algorithm and the *P-spline* and *RK* mixed model respectively. The results are summarized in each panel of Figure 2. The *box plots* shows the distributions, over 500 samples, of differences in *MSE* between each method. We found that in the simulation settings *A,B,C,D*, our *MCMC* algorithm has lower or equal *MSE* than the competing methods. However, *RK* has a substantial tail of poor fits. Note also the skew in the comparisons: this seems that *MCMC* and *P-spline* share similar results. Summarizing, the simulation evidence supports that *MCMC* may have practical advantages over *P-spline* and *RK* for estimating the mean function.

According to the proposed Bayesian statistical inference procedure, we generated 10,000 samples of the fixed parameters for both the mean and dispersion functions. We discard the first 1,000. Figures 3 and 4 give the sample autocorrelation functions, the trajectories of the sample, and the posterior densities for the fixed parameters β and γ from model (1). After disposing of the training period, it is observed that the trajectories are stable, which indicates that the chains produced non-correlated samples for the fixed parameters. We computed other diagnosis of convergence such as the *CD* statistics and the inefficiency factor, which are available upon request. Based on this evidence, we concluded that the posterior sampling is efficient and consequently the fitted obtained by the *MCMC* proposed is accurate and credible.

5. The *LIDAR* Monitoring of Air Pollutants

In this section, we illustrate the flexible estimation method based on a typical real dataset. The *LIDAR* data was obtained from atmospheric monitoring of pollutants. This technique refers to the *Light Detection and Ranging technique* used for monitoring the distribution of meteorological parameters and several atmospheric species of importance. This application is discussed in [28]. The *LIDAR* equation deterministically describes the received signal power $P(\lambda, x)$ as function of range x and wavelength λ; it relates the concentration of mercury at range x. A typical *LIDAR* dataset is shown in panel 1a) of Figure 5. This result is an example where a log or power transformation will not stabilize the variance since the variance does not depend on the mean but rather on the range variable. In addition, the panel 1b) of Figure 5 provides evidence that the linear model does not help to remove the heteroscedasticity according to the residual plot.

The fits from the *MCMC* algorithm and the *P-spline* and *RK* mixed model strategy are depicted in figure 6. The *MCMC* algorithm provided reliable results and is reasonably fast for any sample size. The difference in run times is important in MCMC algorithms that require a large number of iterations for complex simulations. In this paper we rely on SAS/IML software and its power lies in faster run time for these repetitions. Although run times for experiments with less than 10,000 samples are similar, for simulations beyond 10,000, SAS/IML is the unique software that remain stable. However, more in-depth research into run times would need to take further conclusions. see [29] for the figures of the comparisons.

We implemented our algorithm from scratch using the software *SAS-IML*. With the proposals stated, the acceptance rates were reported between 30% and 50%.

6. Conclusions

In this paper, we have proposed a flexible Bayesian framework for modeling the mean and variance functions for heteroscedastic semiparametric models where the response comes from the biparametric exponential family.

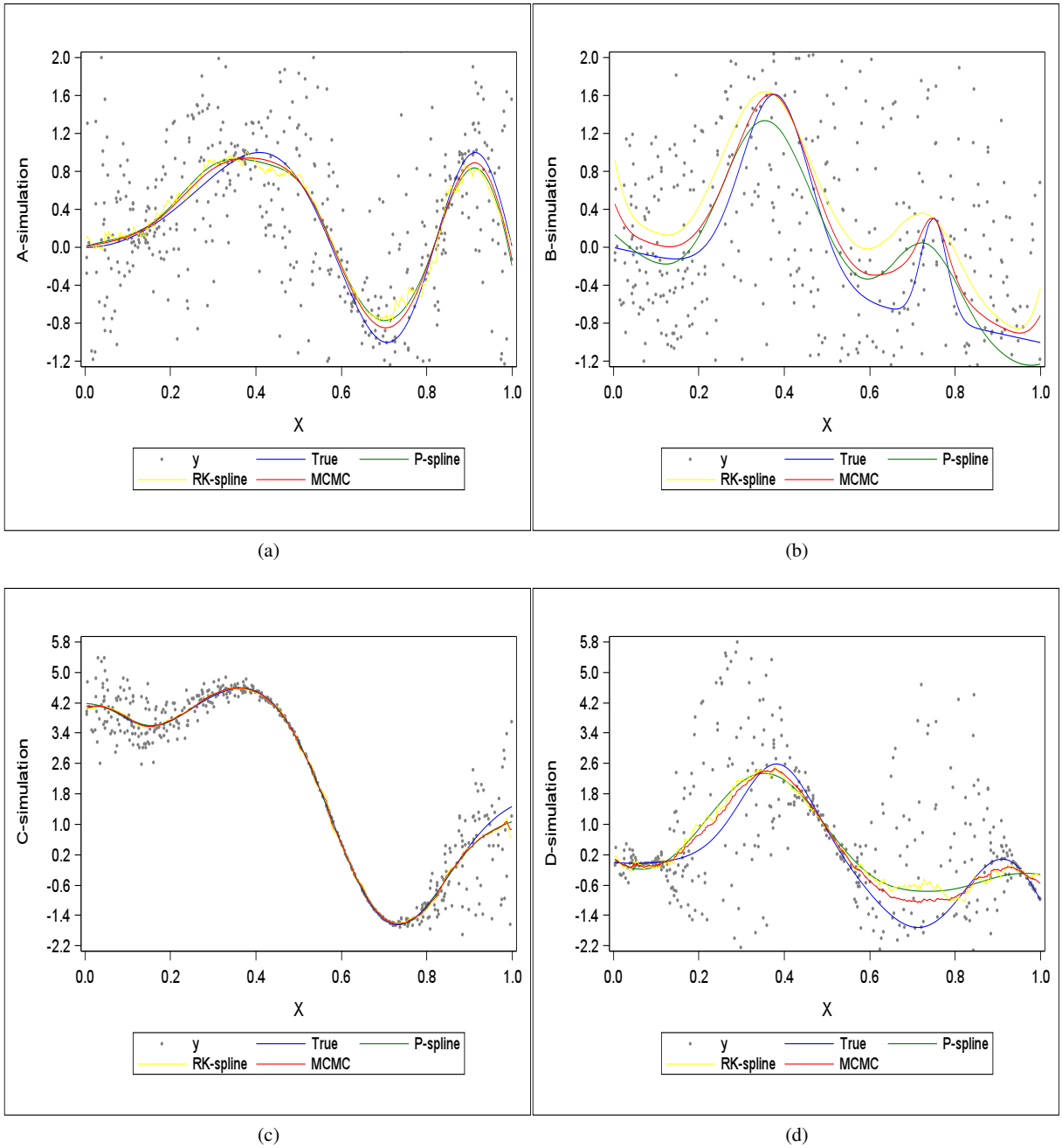


Figure 1. Comparison of MCMC and PROC MIXED with P-splines and RKHS basis in each of the four simulations settings

Our *MCMC* simulation algorithm provided reliable results and is reasonably fast for any sample size. The *MCMC* and mixing properties were assessed by both visual inspection and diagnostic tests of the chain histories of the parameters of interest. We compared the frequentist properties of the Bayesian methodology with two classical models that incorporate the spline and *RKHS* basis in the architecture of the mixed model's customized software (*PROC MIXED*). The results of this study indicate that our method is comparable with the mean function fit. Furthermore, the Bayesian approach could display the estimation of the variance function. Future research could

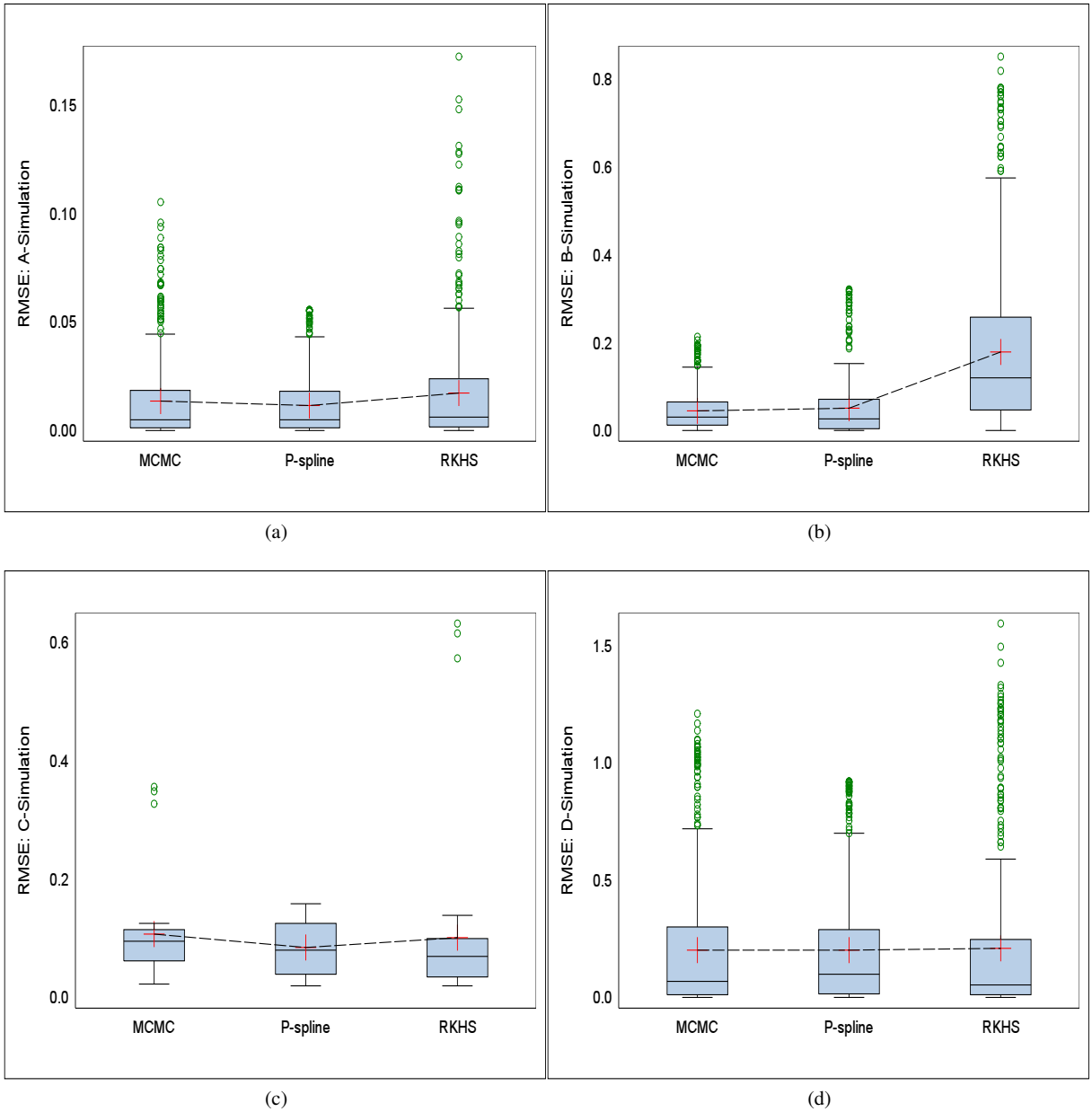


Figure 2. Box plots of the RMSE for the simulation setting

be extended to models with measurement error problems and models that allow complex error structures, including additive models, longitudinal studies and multilevel models.

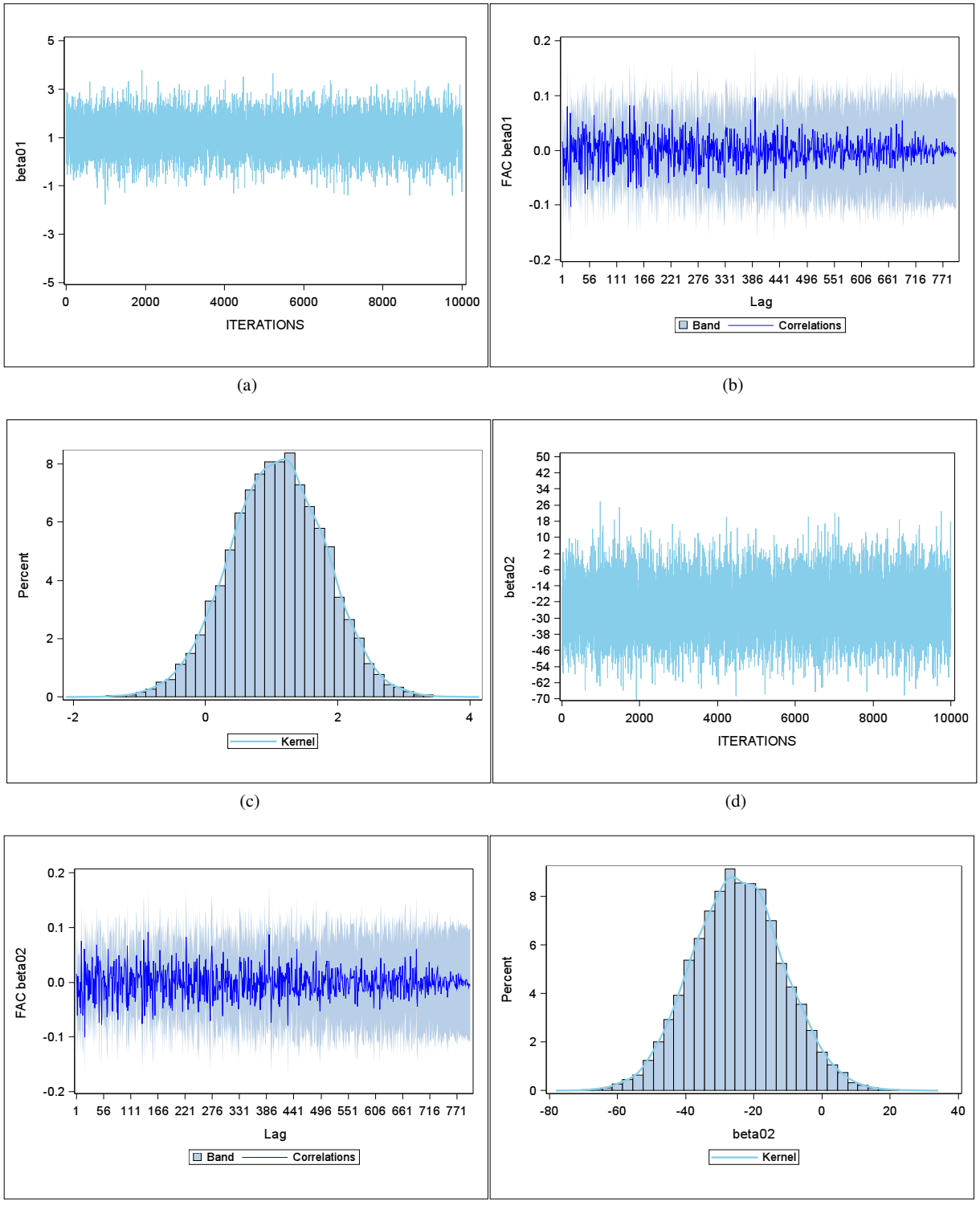


Figure 3. Trace plot, sample autocorrelation function and kernel estimates of the posterior density of β coefficients

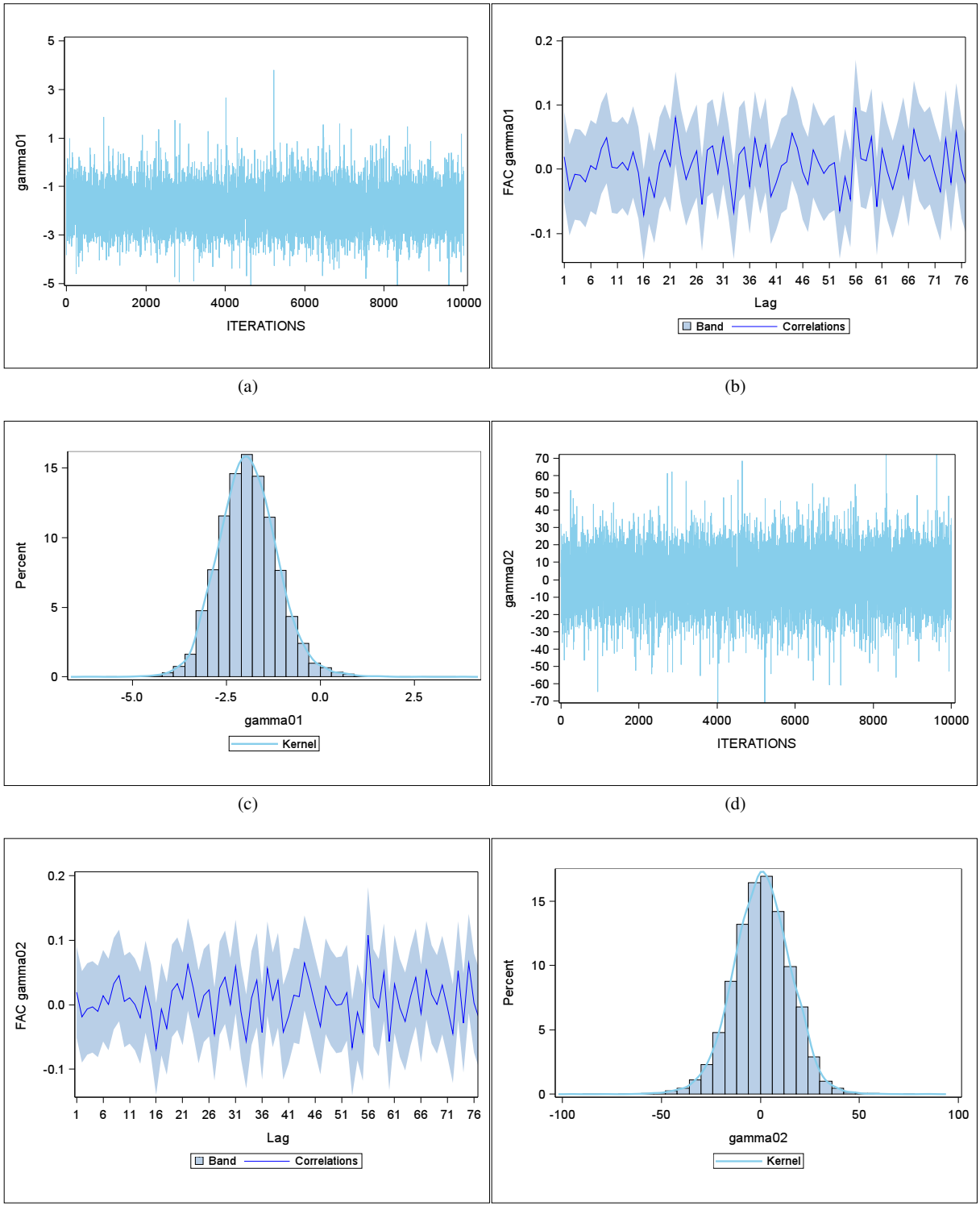
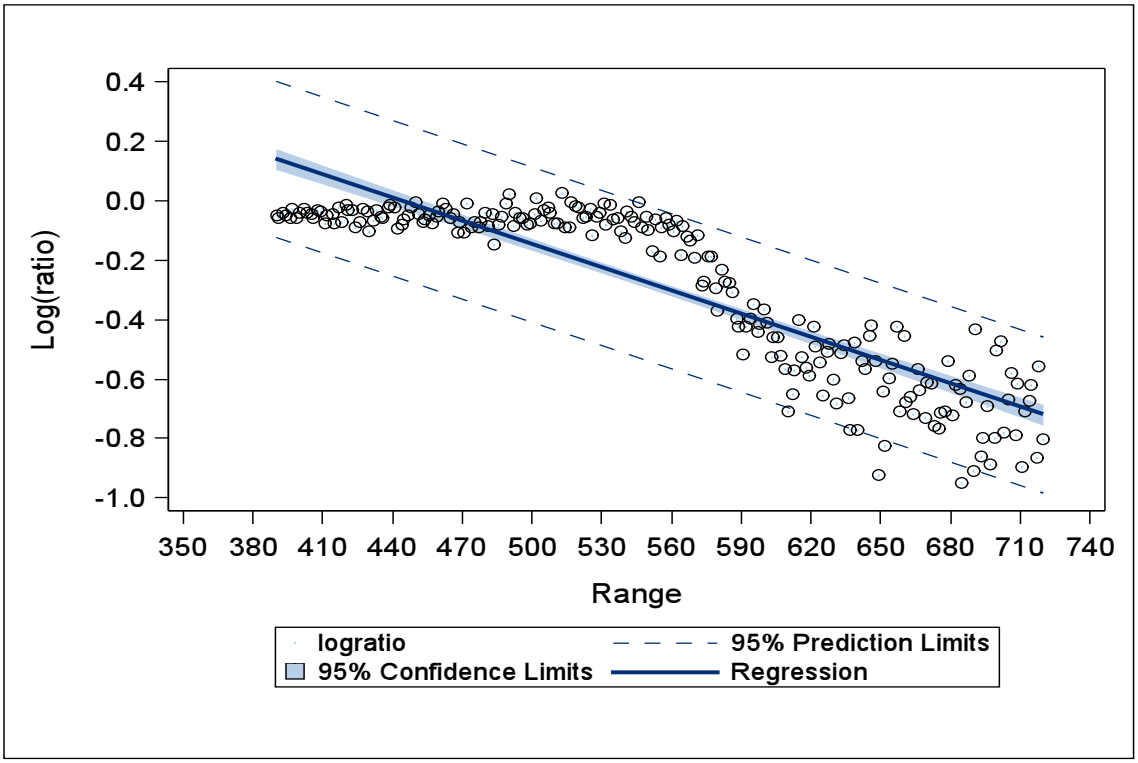
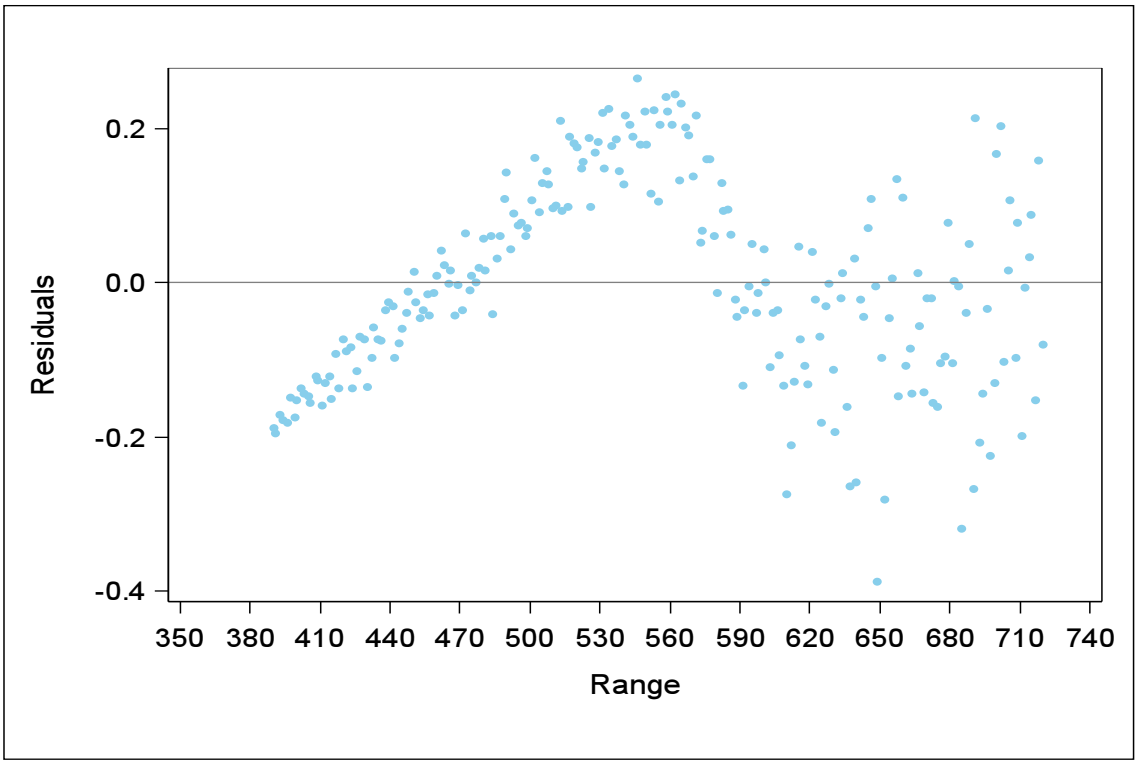


Figure 4. Trace plot, sample autocorrelation function and kernel estimates of the posterior density of γ coefficients



(1a)



(1b)

Figure 5. LIDAR regression and residuals

Source: [28]

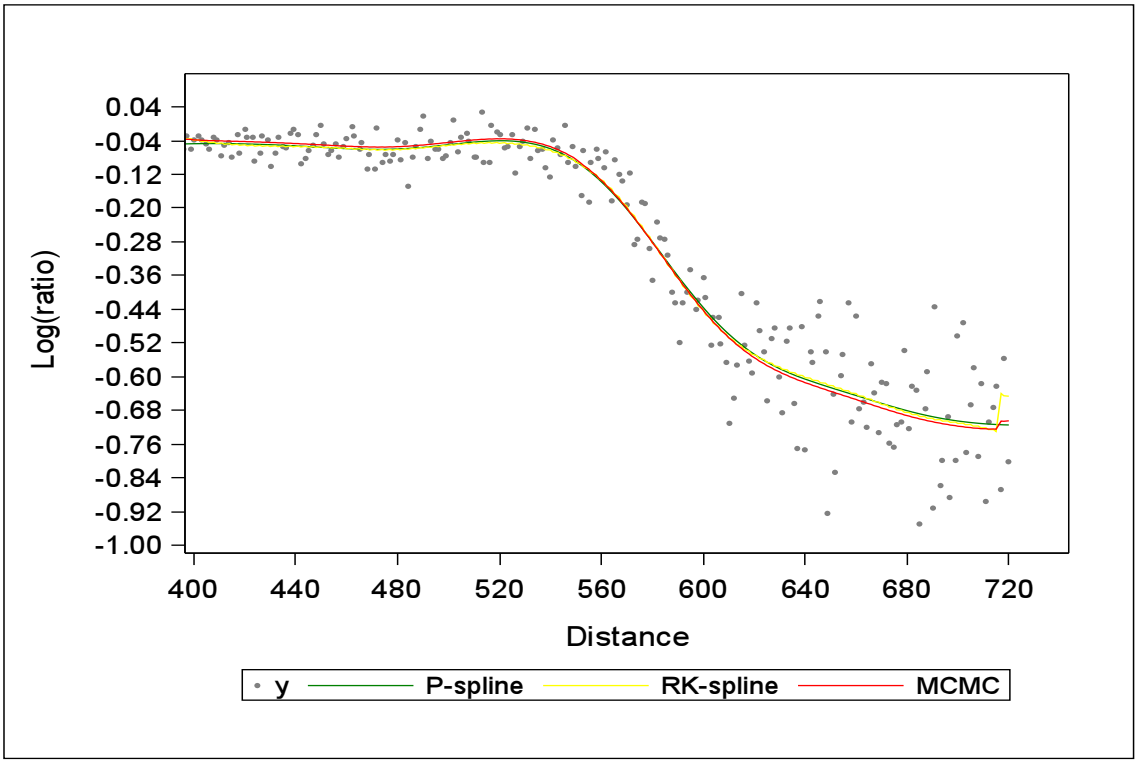


Figure 6. Comparison among MCMC, P-splines and RK-splines

REFERENCES

1. R. Littell and O. Schabenberger, *SAS for Mixed Models*. No. 2, 2006.
2. J. Pinheiro and D. Bates, *Mixed-Effects Models in S and S-Plus*. Springer Verlag, 2009.
3. D. Spiegelhalter and N. Best, "Bayesian approaches to multiple sources of evidence in complex cost-effectiveness modelling," *Statistics in Medicine*, no. 23, pp. 3687 – 3709, 2003.
4. M. B. Denison, D and F. Smith, "A bayesian cart algorithm," *Biometrika*, no. 2, pp. 363 – 367, 1998.
5. D. Nott, "Semiparametric estimation of mean and variance functions for non-gaussian data," *Computational Statistics*, no. 3-4, pp. 603–620, 2006.
6. I. Gijbels and I. Prosdocimi, "Flexible mean and dispersion function estimation in extended generalized additive models," *Communications in statistics - Theory and Methods*, no. 41, pp. 3259 – 3277, 2012.
7. D. Ruppert, M. Wand, and R. J. Carroll, "Semiparametric regression during 2003-2007," *Electronic Journal of Statistics*, vol. 3, pp. 1193–1256, 2009.
8. C. Crainiceanu, "Spatially adaptative bayesian penalized splines with heteroscedastic errors," *Journal of Computational and Graphical Statistics*, no. 2, pp. 265–288, 2007.
9. D. Xu and Z. Zhang, "A semiparametric bayesian approach to joint mean and variance models," *Statistics & Probability Letters*, vol. 83, no. 7, pp. 1624 – 1631, 2013.
10. M. Mencitas and M. Wand, "Variational inference for heteroscedastic semiparametric regression," *School of mathematical sciences, University of Technology. Sydney, Australia*, 2014.
11. E. Cepeda and D. Gamerman, "Bayesian modeling of variance heterogeneity in normal regression models," *J. Prob.Stat*, vol. 14, pp. 207–221, 2001.
12. Y. Ma and C. R. J., "Locally efficient estimators for semiparametric models with measurement error," *Journal of the American Statistical Association*, no. 101, p. 14651474, 2006.
13. C. R. R. D. Berry, S., "Bayesian smoothing and regression splines for measurement error problems," *Journal of the American Statistical Association*, no. 457, pp. 160–169, 2011.
14. M. B. Eilers, P. and M. Durbn, "Twenty years of p-splines," *SORT (Statistics and Operations Research Transactions)*, no. 39, 2014.
15. N. N. N. D. Tran, M. and R. Kohn, "Bayesian deep net glm and glmm," *SORT (arXiv:1805.10157v1 [stat.CO])*, 2018.
16. E. V. T. N. Robert, C. P. and W. C., "Accelerating mcmc algorithms," *Journal of the American Statistical Association*, 2018.
17. B. M. Currie, I., "Flexible smoothing with p-splines : a unified approach," *Statistical Modelling*, no. 4, pp. 333–349, 2002.
18. C. Gu, *Smoothing Spline ANOVA Models*. Springer, West Lafayette,USA, 2002.
19. G. Wahba, *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, 1990.

20. P. Green and B. Silverman, *Nonparametric Regression and Generalized Linear Models: A roughness penalty approach*. Chapman and Hall, London, 1994.
21. S. C. T. C. C. R. Nosedal-Sanchez, A., "Reproducing kernel hilbert spaces for penalized regression : A tutorial," *The American Statistician*, no. 66, pp. 50–60, 2012.
22. W. D. Pierce, N., "Penalized splines and reproducible kernel methods," *American Statistical Association*, no. 3, pp. 233–240, 2006.
23. D. K. Dey, A. E. Gelfand, and F. Peng, "Overdispersed generalized linear models," *Journal of statistical planning and inference,*, vol. 64, no. 64, pp. 93–108, 1997.
24. E. Cepeda, *Variability modeling in Generalized Linear models*. PhD thesis, Unpublished Ph.D thesis, Matematics Institute Universidade Federal do Rio de Janeiro, 2001.
25. D. Gamerman, "Sampling from the posterior distribution in generalized linear mixed models," *Instituto de matemtica, Universidade Federal do Rio de Janeiro*, pp. 59 – 68, 1997.
26. E. Cepeda, J. A. Achcar, and L. G. Lopera, "Bivariate beta regression models: joint modeling of the mean, dispersion and association parameters," *Journal of Applied statistics*, vol. 41, pp. 677–687, Marzo 2014.
27. E. Cepeda and D. Gamerman, "Bayesian methodology for modeling parameters in the two parametric exponential family," *Estadstica*, vol. 57, pp. 93–105, 2005.
28. D. Ruppert, M. Wand, U. Holst, and O. Hssjer, "Local polynomial variance-function estimation," *Technometrics*, no. 39, pp. 262–273, 1997.
29. L. Chelsea, "Mcmc in sas: From scratch or by proc," *Wetern users of SAS software 2016*, vol. 1, no. 1, pp. 1 – 19, 2016.