Bayesian Analysis of Generalized Nonlinear Models with Gaussian Process Priors

Hideo Kozumi *
Graduate School of Economics
Hokkaido University

Abstract

This paper considers a Bayesian semiparametric approach to generalized nonlinear models, where response variables have an exponential family density and predictors are assumed to be an unknown function of covariates. Gaussian process priors are assumed for the unknown function. To estimate the model, we apply MCMC methods based on a genetic adaptive Metropolis and a block samplers. Our approach is illustrated by both simulated and real data.

Keywords: Block sampler; Gaussian process; Generalized linear models; Genetic adaptive Metropolis sampler; Markov chain Monte Carlo.

* *Address for correspondence*: Graduate School of Economics, Hokkaido University, Kita 9 Nishi 7, Kita-ku, Sapporo 060–0809, Japan. E-mail: kozumi@econ.hokudai.ac.jp

1 Introduction

Suppose that we observe the response y_i and the associated covariates x_i for the i–th individual and are interested in exploring the relationship between the responses and the covariates. One of the most popular and useful tools for this purpose is the linear regression model. In classical linear models it is assumed that the responses y_i are normally distributed and their means are a linear function of x_i, that is,

$$E(y_i|x_i) = x_i'\beta, \tag{1}$$

where β is a vector of unknown parameters. As an extension of linear regression models, Nelder and Wedderburn (1972) proposed generalized linear models (see also McCullagh and Nelder, 1989; Fahrmeir and Tutz, 1994).

There are certain distributional and structural assumptions associated with generalized linear models. The responses y_i are assumed to have an exponential family density

$$p(y_i|\theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right\}, \tag{2}$$

where θ_i are the canonical parameters, ϕ is the dispersion parameter, and $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are some specific functions. It is well known that important special cases of (2) include a binomial, a Poisson, a normal, a gamma and an inverse Gaussian distributions. The usual structural assumption is that the expectations $\mu_i = E(y_i|x_i)$ are related to the linear predictors $\eta_i = x_i'\beta$ through a link function $g(\cdot)$ such that

$$g(\mu_i) = \eta_i = x_i'\beta. \tag{3}$$

An important particular link, called the canonical link, is obtained when $g(\mu_i)$ is chosen so that $\eta_i = \theta_i$. From these assumptions, many useful models fall into the class of generalized linear models, including the normal, logit and Poisson regression models.

Classical inference for generalized linear models relies on maximum likelihood estimation of the parameters and the associated asymptotic distributional properties of the estimates. On the other hand, Bayesian inference puts prior distributions on the unknown regression coefficients β and employs Markov chain Monte Carlo (MCMC) methods to carry out the posterior analysis. See Del-

laportas and Smith (1993) and Dey et al. (2000) for an overview of Bayesian generalized linear models.

Though generalized linear models assume that the effect of the covariates is linear, this assumption may be often too restrictive in applications and needs to be extended by a nonlinear predictor. In order to relax the linear assumption, many researchers have proposed semiparametric approaches, where a predictor is treated as an unknown function of the covariates. Using a smoothing spline approach, Green and Yandell (1985), O'Sullivan et al. (1986) and Gu (1990) extended generalized linear models. A local likelihood approach was considered by Fan et al. (1995) and Carroll et al. (1997). Hastie and Tibshirani (1990) proposed generalized additive models, which were analyzed by Denison et al. (1998) in a Bayesian semiparametric framework.

A Bayesian semiparametric analysis must be based on a prior distribution over an unknown function which is an infinite dimensional parameter. In the Bayesian literature (see, e.g., Dey et al., 1998), Dirichlet process priors introduced by Ferguson (1973) were employed by Müller et al. (1996) and West et al. (1994) for Gaussian measurement data. Wood and Kohn (1998) applied integrated Wiener process priors to binary response data. Furthermore, Gaussian process priors were considered in density estimation (Leonard, 1978), regression models (O'Hagan, 1978, Neal, 1996), and binary response models (Hsu and Leonard, 1997, De Oliveira, 2000).

Diggle et al. (1998) and Gutierrez-Peòa and Smith (1998) also considered a Bayesian semiparametric approach to extend generalized linear models by using Gaussian process priors. Since their nonlinear generalized models are analytically intractable, efficient and computationally straightforward MCMC methods are desired for posterior inference. This paper considers the generalized nonlinear model and attempts to develop an efficient simulation algorithm for sampling the posterior distribution of the parameters. Specifically, we employ a genetic adaptive Metropolis sampler proposed by Holmes and Mallick (1998b) and a block sampling technique in order to improve over existing methods for sampling the posterior distribution.

The rest of the paper is organized as follows. In Section 2 we explain the model with Gaussian process priors. Section 3 discusses computational strat-

egy of MCMC methods. In Section 4 our approach is illustrated using both simulated and real data. Finally, brief conclusions are given in Section 5.

2 Model description

Let us assume that the responses y_i $(i = 1, \ldots, n)$ are independent with the probability distribution given by (2), and that the corresponding predictors η_i are an unknown function of $x_i = (x_{i1}, \ldots, x_{ip})'$ expressed as

$$\eta_i = f(x_i). \tag{4}$$

As in generalized linear models, the mean responses $\mu_i = E(y_i|x_i)$ are modeled as

$$g(\mu_i) = f(x_i), \tag{5}$$

for some known link function $g(\cdot)$. Since f is an unknown function, the choice of the link function may not have much effect on posterior inference. Thus, we simply suppose the canonical link function which is commonly used in practice, that is, $\theta_i = f(x_i)$. It should be mentioned that Gutierrez-Peòa and Smith (1998) considered more general distributions for the responses, which contain the exponential family density as a particular case. However, we are concerned with semiparametric estimation of the mean responses and confine ourself to the exponential family density for the responses.

Following the previous work, it is assumed that $f(x_i)$ follow a Gaussian process in the prior assessment. A Gaussian process is a stochastic process which has a joint multivariate Gaussian distribution for any finite set of points, and can be fully specified by its mean function

$$m(x_i) = E\left[f(x_i)\right], \tag{6}$$

and its covariance function

$$C(x_i, x_j) = E\left[\big(f(x_i) - m(x_i)\big)\big(f(x_j) - m(x_j)\big)\right]. \tag{7}$$

For the mean function, we assume that

$$m(x_i) = m_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = z_i'\beta, \tag{8}$$

where $z_i = (1, x_i')'$ and $\beta = (\beta_0, \ldots, \beta_p)'$. Although only the linear mean function is considered in this paper, more flexible functions such as polynomials can be used. Since a Gaussian process converges to its mean function as its variance goes to zero, the use of the linear mean function makes it possible to check whether a linear predictor is appropriate or not through estimates of the variance.

Though we parameterize them in hierarchical form, the predictors η_i can be equivalently written as

$$\eta_i = z_i'\beta + f_0(x_i), \tag{9}$$

where $f_0(x_i)$ follow a Gaussian process with mean zero. This parameterization is employed in Diggle et al. (1998) and Gutierrez-Peña and Smith (1998), but not to be recommended in the context of MCMC methods because of the poor mixing property (see Gelfand et al., 1996). In addition, as shown in the next section, our parameterization has an immediate advantage of obtaining the conditional posterior distribution for β.

There are many choices of covariance functions (see, for example, Cressie, 1993). Diggle et al. (1998) used an isotropic covariance function which depends only on the distance between x_i and x_j. However, as argued in Holmes and Mallick (1998a) and Williams (1998), the use of an isotropic covariance function may not be appropriate in many applications. Therefore we choose the following covariance function

$$C(x_i, x_j) = C_{ij} = \tau_0 \exp\left\{-\frac{1}{2}\sum_{l=1}^{p}\tau_l(x_{il} - x_{jl})^2\right\}, \tag{10}$$

where τ_i (> 0) are unknown parameters (see Neal, 1996). This covariance function has the length scale parameters τ_l $(l > 0)$ corresponding to each covariate which characterize the distance in that particular direction over which f is expected to vary significantly. For irrelevant covariates, the corresponding τ_l will become small, and f is expected to be essentially a constant function of that covariate. This is closely related to the automatic relevance determination idea of MacKay (1994) and Neal (1996). The variance τ_0 gives the overall scale of the Gaussian process. Note that our model reduces to a special case of generalized linear mixed effects models considered in Breslow and Clayton (1993) if $C_{ij} = 0$ for $i \neq j$. Some features of Gaussian processes are illustrated in Figure

1, showing some random samples drawn from Gaussian processes with mean zero. From Figure 1, we can verify the flexibility of Gaussian processes.

To complete the prior specification, we assign a normal distribution to β,

$$\beta \sim N(\beta_0, V_0). \tag{11}$$

Following Neal (1996), it is assumed that the τ_i's are independent and

$$\tau_i \sim IG(m_{0i}/2, \omega_{0i}/2), \tag{12}$$

where $IG(a, b)$ denotes an inverse gamma distribution with a density proportional to $x^{-(a+1)}e^{-b/x}$. In the case of a normal distribution, the dispersion parameter $\phi = \sigma^2$ is assumed to have an inverse gamma prior $IG(n_0/2, s_0/2)$. Finally we mention that $F = (f(x_1), \ldots, f(x_n))'$ has a probability density

$$\pi\left(F|\beta, \tau\right) \propto |C|^{-1/2} \exp\left\{-\frac{1}{2}(F-m)'C^{-1}(F-m)\right\}, \tag{13}$$

where $\tau = (\tau_0, \tau_1, \ldots, \tau_p)$, $m = Z\beta$, $Z = (z_1, \ldots, z_n)'$ and $C = \{C_{ij}\}$ is an $n \times n$ covariance matrix.

3 Posterior simulation

From the distributional assumption of $Y = (y_1, \ldots, y_n)'$ and the prior distributions for the parameters specified above, the likelihood function can be obtained as

$$p(Y|\beta, \tau, \phi) = \int \prod_{i=1}^{n} p(y_i|f_i, \phi)\pi(F|\beta, \tau)dF, \tag{14}$$

where $p(y_i|f_i, \phi)$ is the exponential family density given in (2) and $f_i = f(x_i)$. In the distribution of y_i the canonical parameter θ_i is replaced with f_i since the canonical link function is considered in this paper. The likelihood is complicated and intractable because this multiple integral cannot be in general solved in closed form (an exception is a normal regression model). Therefore we resort to MCMC sampling techniques to estimate the model. For a review of MCMC methods, see Gelfand and Smith (1990), Tierney (1994), Gamerman (1997) and the references therein.

To develop an operational MCMC scheme for simulating the posterior distribution, it is necessary to include F in the simulation. Consequently, our MCMC

algorithm consists of sampling β, τ and F from their conditional posterior distributions recursively. When the response variables have a normal distribution, the sampling of σ^2 is also included in a cycle of our MCMC algorithm.

It is easily derived that the conditional posterior distribution of β is

$$\pi(\beta|F,\tau,\phi,Y) = N(\hat{\beta},\hat{B}), \tag{15}$$

where $\hat{B}^{-1} = Z'C^{-1}Z + V_0^{-1}$ and $\hat{\beta} = \hat{B}\left(Z'C^{-1}F + V_0^{-1}\beta_0\right)$. The conditional posterior distribution of σ^2 is given as

$$\pi(\sigma^2|F,\beta,\tau,Y) = IG\left(\frac{n_0+n}{2}, \frac{s_0+(Y-F)'(Y-F)}{2}\right). \tag{16}$$

However, the computational problem arises in sampling F and τ since their conditional posterior distributions can not be sampled by standard methods. Therefore, we adopt the Metropolis–Hastings (MH) algorithm (Metropolis et al., 1953 and Hastings, 1970) to overcome the problem.

3.1 Sampling τ

The conditional posterior distribution of τ is written as

$$\begin{aligned}\pi(\tau|F,\beta,\phi,Y) \quad &\propto \quad |C|^{-1/2}\exp\left\{-\frac{1}{2}(F-m)'C^{-1}(F-m)\right\} \\ &\times \prod_{i=0}^{k}\tau_i^{-(m_{i0}/2+1)}e^{-\omega_{0i}/2\tau_i},\end{aligned} \tag{17}$$

for which the construction of good proposals is not trivial. Neal (1996) proposed the hybrid MCMC algorithm based on continuous time processes. Since it requires discretization of systems and partial derivatives of the target density during simulation, the hybrid MCMC algorithm is difficult to implement and time consuming. Gutierrez-Peña and Smith (1998) approximated the conditional posterior distribution by discretizing the parameter space of τ.

Alternatively, we advocate the use of a genetic adaptive Metropolis (GAM) sampler proposed by Holmes and Mallick (1998b). The GAM sampler is a modification of the genetic algorithm (GA) (see, e.g., Goldberg, 1989) to fulfill the requirements of a MCMC sampler, and augments the parameter space to accommodate multiple chains in parallel. The idea behind the GAM sampler is related to the snooker algorithm proposed by Gilks et al. (1994) and Roberts

and Gilks (1994). An advantage of the GAM sampler is that gradient information is implicitly encoded within the distribution of the parameters across the parallel chains, which improves mixing property of the chain. In addition the algorithm is very simple to implement requiring only a few extra lines of code to a conventional MH sampler.

Since the τ_i's are restricted to be positive, the GAM algorithm is not directly applicable for the sampling of τ. Thus we transform τ to $\gamma = \log \tau$ and consider to sample γ instead. The GAM sampler generates a proposal value γ^* as follows: First augment γ to form a population $\Gamma = \{\gamma^1, \ldots, \gamma^M\}$ where M is a population size. At each iteration select one member of the population, say γ^a, for updating. Then apply the mutation or the crossover operator, which are explained below, to the population with probabilities M_R and $1 - M_R$, respectively (M_R is called a mutation rate).

Mutation operator: The mutation operator generates a proposal value γ^* from the random walk sampler,

$$\gamma^* = \gamma^a + u, \;\; u \sim N(0, \delta^2), \tag{18}$$

which is accepted with probability

$$\min\left\{1, \frac{\pi(\gamma^*|F, \beta, \phi, Y)}{\pi(\gamma|F, \beta, \phi, Y)}\right\}. \tag{19}$$

The acceptance probability is simply the ratio of the conditional posterior distributions of γ evaluated at the proposed and the current values.

Crossover operator: The crossover operator plays a central role in the GAM sampler. In the cross operator two parent states γ^i and γ^j are selected from the population such that $i \neq j \neq a$. Then an offspring state γ^o is created by using a GA crossover scheme and $\{\gamma^i, \gamma^j\}$. Holmes and Mallick (1998b) suggest that each parameter in γ^o is taken from either γ^i or γ^j with probability 1/2. After creating the offspring γ^o, a proposal value is sampled by

$$\gamma^* = \begin{cases} \gamma^a + 2(\gamma^o - \gamma^a), & \text{with probability } F_R, \\ \gamma^a + r(\gamma^o - \gamma^a)/||\gamma^o - \gamma^a||, & \text{otherwise,} \end{cases} \tag{20}$$

where F_R is a flip rate, $||\cdot||$ indicates the Euclidean norm, and $r \sim N(0, \lambda^2)$ with $\lambda = ||\gamma^i - \gamma^j||$. The fist move type is a reflection of γ^a about γ^o, and the second one is a random sampler along the direction $\gamma^o - \gamma^a$. It should be

noted that the proposal density is symmetric in all of the above move types. Therefore, the proposed value γ^* is accepted with probability given by (19).

If the algorithm starts with the dispersed population, the value of λ is likely to be large at the beginning of the simulation. This results in large update proposals in the crossover operator, and the GAM sampler explores the state space more widely. As the simulation proceeds, the population settles down in a high density region. Thus the value of λ will be reduced and smaller steps will be taken.

Recently Liang and Wong (2000, 2001) proposed a very similar algorithm which they called the evolutionary Monte Carlo method. In their algorithm the same mutation operator is used, but the crossover operator is implemented by the snooker algorithm. They also devised the exchange operator of parallel tempering (Geyer, 1991) to take account of multimodal distributions.

It should be mentioned that the conditional posterior distribution of τ_0 is an inverse gamma distribution and its sampling is straightforward. Since, however, correlations between the τ_i's were found from preliminary experiments, we consider to sample the τ_i's jointly for a better mixing property.

3.2 Sampling F

We now discuss to sample F from the conditional posterior distribution given by

$$\pi(F|\beta,\tau,\phi,Y) \propto \prod_{i=1}^{n} p(y_i|f_i,\phi) \times \exp\left\{-\frac{1}{2}(F-m)'C^{-1}(F-m)\right\}. \qquad (21)$$

The most common strategy for sampling F may be to update f_i one at a time, as in Diggle et al. (1998) and Gutierrez-Peña and Smith (1998). However, this single move sampler would exhibit poor mixing (see Liu et al., 1994). Instead, it can be considered to sample all the f_i's simultaneously. Since the dimension of F can be large, it is hard to find proposal densities which approximate well the conditional posterior distribution given in (21). Consequently, this approach suffers from slow mixing due to low acceptance probabilities. As a compromise of these two approaches, we divide F into B blocks and sample each block in turn.

It is easily derived that the conditional prior distribution of $f_{[i:j]} = (f_i, \ldots, f_j)'$

given the rest is normal with mean

$$m_{(i,j)} = \begin{cases} m_{[i:j]} - K_{[i:j]}^{-1} K_{[j+1:n]} \left(f_{[j+1:n]} - m_{[j+1:n]} \right), & \text{if } i = 1, \\ m_{[i:j]} - K_{[i:j]}^{-1} K_{[1:i-1]} \left(f_{[1:i-1]} - m_{[1:i-1]} \right), & \text{if } j = n, \\ m_{[i:j]} - K_{[i:j]}^{-1} \left\{ K_{[1:i-1]} (f_{[1:i-1]} - m_{[1:i-1]}) \right. & \\ \quad \left. + K_{[j+1:n]} (f_{[j+1:n]} - m_{[j+1:n]}) \right\}, & \text{otherwise,} \end{cases} \tag{22}$$

and covariance matrix

$$\Sigma_{(i,j)} = K_{[i:j]}^{-1}, \tag{23}$$

where $m_{[i:j]} = (m_i, \ldots, m_j)'$ and $K_{[i:j]}$ denotes the submatrix of $K = C^{-1}$ given by the rows and columns numbered i to j. In addition $K_{[1:i-1]}$ and $K_{[j+1:n]}$ are the matrices left and right of $K_{[i:j]}$, respectively.

To define a suitable proposal density, we consider the following adjusted linear model (see McCullagh and Nelder, 1989):

$$\tilde{y}_i = f_i + \epsilon_i, \quad \epsilon_i \sim N(0, v_i), \tag{24}$$

where $\tilde{y}_i = g(\mu_i) + (y_i - \mu_i) g'(\mu_i)$ and $v_i = b''(\theta) g'(\mu_i)^2$. It should be noted that $\tilde{y}_i$, μ_i and v_i are evaluated at the current values of the parameters. Then, combining the adjusted linear model (24) with the conditional prior distribution derived above, we have

$$f_{[i:j]} \sim N \left(\hat{m}_{(i,j)}, \hat{\Sigma}_{(i,j)} \right), \tag{25}$$

where

$$\begin{aligned} \hat{\Sigma}_{(i,j)}^{-1} &= V_{(i,j)}^{-1} + \Sigma_{(i,j)}^{-1}, \\ \hat{m}_{(i,j)} &= \hat{\Sigma}_{(i,j)} \left(V_{(i,j)}^{-1} \tilde{y}_{[i:j]} + \Sigma_{(i,j)}^{-1} m_{(i,j)} \right), \\ V_{(i,j)} &= \mathrm{diag}(v_i, \ldots, v_j), \end{aligned}$$

and $\tilde{y}_{[i:j]} = (\tilde{y}_i, \ldots, \tilde{y}_j)'$. Although this distribution can be a good approximation to the conditional posterior distribution of $f_{[i:j]}$ which is the product of the contributions of y_l $(l = i, \ldots, j)$ and the conditional prior, the ratio of the target and the proposal densities can be unbounded. Thus, as in Chib et al. (1998), we adopt the multivariate t distribution as the proposal density, that is,

$$q \left(f_{[i:j]}^* | f_{[i:j]} \right) = MVt \left(\hat{m}_{(i,j)}, \hat{\Sigma}_{(i,j)}, \nu \right),$$

where ν denotes degrees of freedom. It should be noted that if a normal distribution is assumed for responses, the conditional posterior distribution of F is normal with mean $(I/\sigma^2 + C^{-1})^{-1}(Y/\sigma^2 + C^{-1}m)$ and covariance matrix $(I/\sigma^2 + C^{-1})^{-1}$. In this case we do not have to rely on the MH algorithm to update F.

To implement the block sampling, we must select the block sizes k_i $(i = 1, \ldots, B)$. Following Shephard and Pitt (1997), we select k_i randomly with U_i being independent uniforms and

$$k_i = \text{int}\left[n \times \frac{i + U_i}{B + 2}\right], \quad i = 1, \ldots, B, \tag{26}$$

and treat B as a tuning parameter. This allows the points of conditioning to change over the iterations and ensure that the method does not become stuck by an excessive amount of rejections.

3.3 Estimation of $\mu(x)$

Suppose that we are concerned with estimation of $\mu(x) = E(y|x)$ for a given value of x. Since $\mu(x)$ is related to $f(x)$ through $f(x) = g(\mu(x))$, all we require is the posterior distribution of $f(x)$ written as

$$\pi(f(x)|Y) = \int \pi(f(x)|F, \beta, \tau)\pi(F, \beta, \tau|Y)dFd\beta d\tau, \tag{27}$$

where $\pi(f(x)|F, \beta, \tau)$ is the conditional distribution of $f(x)$ given F. It follows from the properties of the Gaussian process that $\pi(f(x)|F, \beta, \tau)$ is normal with

$$E(f(x)|F, \beta, \tau) = m(x) + \kappa(x)'C^{-1}(F - m), \tag{28}$$

and

$$\text{Var}\,(f(x)|F, \beta, \tau) = \tau_0 - \kappa(x)'C^{-1}\kappa(x), \tag{29}$$

where $\kappa(x) = (C(x, x_1), \ldots, C(x, x_n))'$. Thus, given a sample from $\pi(F, \beta, \tau|Y)$, it is straightforward to draw $f(x)$ from the posterior distribution. After obtaining a simulated sample of $f^{(t)}(x)$ $(t = 1, \ldots, T)$, we can estimate $\mu(x)$ by

$$\hat{\mu}(x) = \frac{1}{T}\sum_{t=1}^{T} g^{-1}(f^{(t)}(x)), \tag{30}$$

and other posterior summaries are easily calculated.

4 Numerical examples

This section illustrates our approach using simulated and real data sets. In Section 4.1 we simulated four data sets from a normal, a Poisson and a binomial distributions. Section 4.2 briefly considers a data set studied by Brinkman (1981) and fits a Gaussian regression model. Finally Section 4.3 examines a count data set examined by Kennan (1985) and Jaggia (1991), and a Poisson regression model is applied.

To implement the MCMC algorithm explained in the previous section, we must choose several tuning parameters. There are four tuning parameters in the GAM algorithm, namely M, M_R, F_R and δ^2. Following Holmes and Mallick (1998b), we set $M = 10$ and $F_R = 0.1$. The values of M_R and δ^2 were obtained in short preliminary runs by examining the acceptance rates. The degrees of freedom for the multivariate t proposal density is fixed to $\nu = 10$. The priors for the estimation are defined by the hyper–parameters

$$\begin{aligned} \beta_0 &= 0, \quad & V_0 &= 1000 \times I, \\ m_{0i} &= 4, \quad & \omega_{i0} &= 0.02, \\ n_0 &= 4, \quad & s_0 &= 0.02, \end{aligned}$$

which reflect weak prior information. Unless otherwise stated, all of the results reported here are based on these parameter values.

4.1 Simulated data

Following Denison et al. (1998), response variables were generated from

$$\begin{aligned} y_i &\sim N(\mu_i, 0.2^2), \ \ i = 1, \ldots, 150, \\ \mu_i &= x_i + 2\exp(-16x_i^2), \end{aligned}$$

where x_i were equally spaced in the interval $(-2, 2)$. With the simulated data set, we fitted the Gaussian model by running the MCMC algorithm for 10,000 iterations following a burn–in phase of 5,000 iterations.

Figure 2 shows the true and the posterior estimates of the mean response $\mu(x)$ together with the 90% intervals. It can be seen from Figure 2 that the proposed method produces good estimates of $\mu(x)$. Table 1 reports the posterior estimates of β and γ. The posterior mean and standard deviation of β_1 are

0.989 and 0.155, respectively, providing an evidence that the mean function of the Gaussian process captures linear trend in the data.

To examine its convergence and mixing performances, we applied the GAM samplers with $M = 5$, 10 and 20 for updating γ. For comparison, the random walk sampler was also conducted. In this experiment, the starting values of f_i were set to y_i, and those of β and σ^2 were chosen from the ordinary least squares estimates under a linear regression model. The population of γ was initialized by a sample from a uniform distribution on $(-3, 3)^2$. The resulting acceptance rates for the GAM and the random walk samplers were about 50% and 40%, respectively. We tried several initial values of γ and found quick convergence of the GAM algorithm for any initial values of γ. However, the random walk sampler sometimes exhibited slow convergence as shown in Figure 3, which plots the sampling paths of the fist 2,000 iterations of γ_1 and the log values of the probability distribution of Y. It can be seen from the figure that the random walk sampler is trapped around up to 1,3000 iterations and then suddenly moves to a high density region. In contrast, the GAM sampler traverses the parameter space widely at the start of the chain and shows better convergence properties than the random walk sampler. We can also observe that the larger the value of M becomes, the longer time the GAM sampler needs to settle down in a high density region.

Figure 4 shows the Euclidean distance between $\gamma^{(t)}$ and $\gamma^{(t+1)}$ over the iterations, where $\gamma^{(t)}$ denotes the value of γ at the t-th iteration. It can be seen that the GAM algorithm attempts larger changes than the random walk sampler. This result suggests a good mixing property of the GAM sampler. Figure 5 shows the estimated autocorrelations for the GAM ($M = 10$) and the random walk samplers. The autocorrelations for the GAM sampler decay more quickly than those for the random walk sampler. It is interesting to note that the autocorrelations under the GAM sampler decrease very slowly after lag 5. Nevertheless, it can be concluded that the GAM sampler possesses good convergence and mixing properties from these findings.

To examine other distributions for response variables, we considered a Poisson and a binomial distributions. For the Poisson distribution, we considered

the following two means:

$$\mu_i = \begin{cases} \exp\left\{-200x_i(x_i-0.5)^2(x_i-1.0)\right\}, & \text{(polynomial)} \\ \exp(1.0+2.0x_i). & \text{(linear)} \end{cases}$$

As an example of a binomial distribution, we used a logit model given by

$$\begin{aligned} y_i &\sim Bi(\mu_i, 1), \;\; i = 1, \ldots, 150, \\ \text{logit}(\mu_i) &= 3\cos(2\pi x_i). \end{aligned}$$

In all of the models, x_i were equally spaced in the interval $(0, 1)$. We fitted the models based on the same MCMC design as in the normal case, and the results are shown in Figures 6 and 7. The posterior estimates of β and γ are also summarized in Table 1.

From the figures, we can see that our approach works well again and the flexibility of Gaussian processes can be verified. The posterior mean of γ_1 under the Poisson model with polynomial is the largest among all the cases, which reflects complexity of the test functions. In the polynomial case, the corresponding posterior mean and standard deviation of β_1 are -0.402 and 2.831 respectively, indicating that its posterior distribution is dispersed. Therefore, it can be concluded that the mean function of the Gaussian process is constant and a linear predictor is not appropriate. In the linear case of the Poisson model, the posterior estimates of β are very close to the true values and the mean function of the Gaussian process explains most part of the variation of the response variables. Consequently, the small posterior means of γ_0 and γ_1 are observed as discussed in Section 2.

Using the simulated data for the Poisson model with polynomial, we examined the effects of the number of blocks B on the convergence performances of the block sampler. For this purpose we run the short MCMC algorithm of 200 iterations for $B = 1$, 15 and 150 with $f(x_i)$ initialized to zero. It should be noted that the case of $B = 150$ corresponds the single move sampler. For comparison, we also employed the conditional prior distribution given in (22) and (23) as the proposal distribution for F. The use of the conditional prior proposal was suggested by Knorr–Held (1999) in the context of dynamic models.

Figure 8 shows the posterior means of $\mu(x)$ after 20, 100 and 200 iterations. The algorithm with $B = 1$ does not converge at all because of large rejection

rate. In fact all the proposal values of $f(x_i)$ were rejected during the MCMC updates in this case. On the other hand, we obtained the high average acceptance rates of $f(x_i)$ for $B = 15$ and 150, which were 79.9% and 96.1% respectively. Consequently, $f(x_i)$ were frequently updated, and fast convergence was achieved in these cases as shown in Figure 8. This finding is different from that in Knorr–Held (1999). He carried out similar experiments using conditional prior proposals in dynamic models and found slow convergence of the single move sampler was found. When the conditional prior proposal was applied to our model by setting $B = 15$, we observed that the average acceptance rate was 42.5%, which was smaller than that of our proposal. Figure 8 also shows slow convergence of the conditional prior proposal. Thus the proposal distribution based on the adjusted linear model approximates the target density better than the conditional prior proposal, and its use results in quick convergence of the MCMC algorithm.

To compare the efficiency of the block ($B = 15$) and the single ($B = 150$) samplers, the autocorrelations were calculated from 10,000 draws beyond a burn-in period of 5,000 iterations. Figure 9 shows that the single move sampler shows quite large correlations compared with the block sampler with $B = 15$. Though Diggle et al. (1998) and Gutierrez-Peña and Smith (1998) used the single move sampler for updating the f_i's, this result suggests that the single move sampler is less efficient and the use of the block sampler increases the reliability of the MCMC algorithm.

4.2 Ethanol data

To illustrate our approach on a real data set, we consider the ethanol data examined by Brinkman (1981). The data set has 88 observations on NOx exhaust emissions from a single cylinder engine (NOx), the engine's equivalence ratio (E) and compression ratio. We modeled the data using a Gaussian nonlinear regression model where the response is NOx and the covariate is E. To estimate the model, we run our algorithm for 15,000 iterations following a burn–in phase of 5,000 iterations. For comparison, we fitted the model using the smoothing spline (Wahba, 1990) and the local polynomial regression (Cleveland et al., 1992) techniques. The results for these techniques were created by using the R

functions `smooth.spline` and `loess` with defaults parameter values.

Figure 10 summarizes the results, and we can observe that the mean response is evidently nonlinear. It can be seen from Figure 10 that our Bayesian estimate fits the data well and is very similar to the smoothing spline estimate. Among the three approaches, the local polynomial regression exhibits the worst fit. It should be mentioned that, as discussed in MacKay (1998), Gaussian processes are closely related to smoothing splines. This may be a reason why we obtained the similar results for the Gaussian process and the smoothing spline approaches.

4.3 Strike data

In this section we consider the strike data analyzed by Kennan (1985) and Jaggia (1991). The data set consists of 108 strike frequencies recorded from January 1968 through December 1976. Though Kennan (1985) and Jaggia (1991) examined the data using duration models, we model the number of strikes based on a Poisson model with the canonical link function. The response variable (STRIKES) is the number of contract strikes in U.S. manufacturing beginning each month. The covariates include a time trend (TIME) and a measure of the cyclical departure of aggregate production from its trend level (OUTPUT).

Using the strike data described above, we estimated the two Poisson models. The first model (Model 1) includes only TIME and the second model (Model 2) has both TIME and OUTPUT in a set of the covariates. Table 2 reports the posterior estimates of the parameters which were based on 15,000 posterior draws following a burn–in of 5,000 iterations. Judging from the estimates of β, we can conclude that a linear predictor is not appropriate for this data set. Although the posterior means of β and γ_0 were almost the same in both models, the posterior mean of γ_1 became smaller by including OUTPUT. It is of interest to note that the posterior mean of γ_2 is very small compared with that of γ_1. This finding implies that OUTPUT explains only a small part of the variation of the response variables. Therefore the level of economic activity has no effect on strike frequency.

Figure 11 plots the posterior estimates of the mean responses along with the 95% intervals obtained from Model 1. It can be seen that the mean responses exhibit some periodical movement and have several peaks. In particular we can

observe the highest peak around in 1974.

5 Conclusions

This paper has considered a Bayesian semiparametric approach to generalized nonlinear models, where the response variables have an exponential family density and the predictors are assumed to be an unknown function of covariates. We have employed Gaussian process priors for the unknown function and developed MCMC methods based on the genetic adaptive Metropolis and the block samplers for posterior inference. Our approach has been illustrated by both simulated and real data.

The results for our experiments have shown that the Gaussian process priors produce smooth and adequate estimates for nonlinear functions. It has been also shown that both the GAM and the block samplers have good convergence and mixing properties. We believe that the use of these algorithms improves over the existing methods considered in, for example, Gutierrez-Peña and Smith (1998).

One drawback of the use of Gaussian process priors is that the inversion of the $n \times n$ covariance matrix C is required for sampling τ, where n is a sample size. It is well known that computational complexity of the matrix inversion scales as $O(n^3)$. Therefore our approach takes long time for a large dataset and further algorithmic improvements and/or approximations are necessary. Furthermore, we have applied only the crossover operator explained in Section 3.1 in conducting the GAM sampler. Since the crossover operator is a core part of the GAM sampler, further improvements might be possible by applying other algorithms such as the snooker algorithm. These topics need further investigations and are left for the future research.

Acknowledgment

The author is grateful to two anonymous referees for their useful comments, which improved an earlier version of the paper. This research was partially supported by the Japanese Ministry of Education, Culture, Sports, Science and Technology under Grant-in-Aid for Encouragement of Young Scientists

(#12730019) .

References

[1] Brinkman, N. D. (1981). "Ethanol fuel - a single-cylinder engine study of efficiency and exhaust emissions," *SAE transactions*, **90**, 1414–1424.

[2] Breslow, N. E. and Clayton, D. G. (1993). "Approximate inference in generalized linear mixed models," *Journal of the American Statistical Association*, **88**, 9–25.

[3] Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. (1997). "Generalized partially linear single–index models," *Journal of the American Statistical Association*, **92**, 477–489.

[4] Chib, S., Greenberg, E., and Winkelmann, R. (1998). "Posterior simulation and Bayes factors in panel count data models," *Journal of Econometrics*, **8**6, 33–54.

[5] Cleveland, W. S., Grosse, E., and Shyu, W. M. (1992). "Local regression models," in J. M. Chambers and T. J. Hastie (eds.), *Statistical Models in S*. Wadsworth & Brooks: California.

[6] Cressie, N. A. C. (1993). *Statistics for Spatial Data*, Revised edition. John Wiley: New York.

[7] Dellaportas, P. and Smith, A. F. M. (1993). "Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling," *Applied Statistics*, **42**, 443–460.

[8] Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1998). "Automatic Bayesian curve fitting," *Journal of the Royal Statistical Society Ser.B*, **60**, 333–350.

[9] De Oliveira, V. (2000). "Bayesian prediction of clipped Gaussian random fields," *Computational Statistics & Data Analysis*, **34**, 299–314.

[10] Dey, D. K., Ghosh, S. K., and Mallick, B. K. (2000). *Generalized Linear Models: A Bayesian Perspective*. Marcel Dekker: New York.

[11] Dey, D. K., Müller, P., and Sinha, D. (1998). *Practical Nonparametric and Semiparametric Bayesian Statistics*. Springer: New York.

[12] Diggle, P. J., Tawn, J.A., and Moyeed, R. A. (1998). "Model–based geostatistics (with discussion)," *Applied Statistics*, **47**, 299–326.

[13] Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer: New York.

[14] Fan, J., Heckman, N. E., and Wand, M. P. (1995). "Local polynomial kernel regression for generalized linear models and quasi-likelihood functions," *Journal of the American Statistical Association*, **90**, 141–150.

[15] Ferguson, T. S. (1973). "A Bayesian analysis of some nonparametric problems," *Annals of Statistics*, **1**, 209–230.

[16] Gamerman, D. (1997). *Markov Chain Monte Carlo*. Chapman and Hall: London.

[17] Gelfand, A. E. and Smith, A. F. M. (1990). "Sampling based approaches to calculating marginal densities," *Journal of the American Statistical Association*, **85**,398–409.

[18] Gelfand, A. E., Sahu, S. K. and Carlin, B. P. (1996). "Efficient parameterizations for generalized linear mixed models (with discussion)," in J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (eds.), *Bayesian Statistics 5*. Oxford University Press: Oxford.

[19] Geyer,C. J. (1991). "Markov chain Monte Carlo maximum likelihood," in E. M. Keramigas (ed.), *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Inference*. Interface Foundation: Fairfax.

[20] Gilks, W. R., Roberts, G. O. and George, E. I. (1994). "Adaptive direction sampling," *Statistician*, **43**, 179–189.

[21] Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison–Wesley: New York.

[22] Green, P. and Yandell, B. (1985). "Semi–parametric generalized linear models," in R. Gilchrist, B. Francis and J. Whittaker (eds.), *Generalized Linear Models*. Springer: Heidelberg.

[23] Gu, C. (1990). "Adaptive spline smoothing in non–gaussian regression models," *Journal of the American Statistical Association*, **85**, 801–807.

[24] Gutierrez-Peña, E. and Smith, A. F. M. (1998). "Aspects of smoothing and model inadequacy in generalized regression," *Journal of Statistical Planning and Inference*, **67**, 273–286.

[25] Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall: London.

[26] Hastings, W. K. (1970). "Monte Carlo sampling methods using Markov chain and their applications," *Biometrika*, **57**, 97–109.

[27] Holmes, C. C. and Mallick, B. K. (1998a). Discussion of "Model–based geostatistics," *Applied Statistics*, **72**.

[28] Holmes, C. C. and Mallick, B. K. (1998b). "Parallel Markov chain Monte Carlo sampling: An evolutionary based approach," Technical Report, Imperial College.

[29] Hsu, J. S. and Leonard, T. (1997). "Hierarchical Bayesian semiparametric procedures for logistic regression," *Biometrika*, **84**, 85–93.

[30] Jaggia, S. (1991). "Specification tests based on the generalized gamma model of duration – with an application to Kennan strike data," *Journal of Applied Econometrics*, **6**, 169–180.

[31] Kennan, J. (1985). "The duration of contract strikes in U.S. manufacturing," *Journal of Econometrics*, **28**, 5–28.

[32] Knorr–Held, L. (1999). "Conditional Prior Proposals in Dynamic Models," *Scandinavian Journal of Statistics*, **26**, 129-144.

[33] Leonard, T. (1978). "Density estimation, stochastic processes, and prior information," *Journal of the Royal Statistical Society Ser.B*, **40**, 113–146.

[34] Liang, F., and Wong, W. H. (2000). "Evolutionary Monte Carlo sampling: Applications to C_p model sampling and change–point problem," *Statistica Sinica*, **10**, 317–342.

[35] Liang, F., and Wong, W. H. (2001). "Real–parameter evolutionary Monte Carlo with applications to Bayesian mixture models," *Journal of the American Statistical Association*, **96**, 653–666.

[36] Liu, J. S., Wong, W. H. and Kong, A. (1994). "Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes," *Biometrika*, **81**, 27–40.

[37] MacKay, D. J. C. (1994). "Bayesian methods for backpropagation networks," in E. Domany, J.L. van Hemmen and K. Schulten (eds.), *Models of Neural Networks III*. Springer: New York.

[38] MacKay, D. J. C. (1998). "Gaussian processes – a replacement for supervised neural networks?," Technical Report, Department of Physics, Cambridge University.

[39] McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*, 2nd edition. Chapman and Hall: London.

[40] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A.H. and Teller, E. (1953). "Equations of state calculations by fast computing machines," *Journal of Chemical Physics*, **21**, 1087–1091.

[41] Müller, P., Erkanli, A., and West, M. (1996). "Bayesian curve fitting using multivariate normal mixtures," *Biometrika*, **83**, 67–79.

[42] Neal, R. M. (1996). *Bayesian Learning for Neural Networks*. Springer: New York.

[43] Nelder, J. A. and Wedderburn, R. W. M. (1972). "Generalized linear models," *Journal of the Royal Statistical Society, Ser.A*, **135**, 370–384.

[44] O'Hagan, A. (1978). "On curve fitting and optimal design for regression," *Journal of the Royal Statistical Society Ser.B*, **40**, 1–42.

[45] O'Sullivan, F., Yandell, B. and Raynor, W. (1986). "Automatic smoothing of regression functions in generalized linear models," *Journal of the American Statistical Association*, **81**, 96–103.

[46] Roberts, G. D. and Gilks, W. R. (1994). "Convergence of adaptive direction sampling," *Journal of Multivariate Analysis*, **49**, 287–298.

[47] Shephard, N. and Pitt, M. K. (1997). "Likelihood analysis of non–Gaussian measurement time series," *Biometrika*, **84**, 653–667.

[48] Tierney, L. (1994). "Markov chains for exploring posterior distributions (with discussion)," *Annals of Statistics*, **22**, 1701–1762.

[49] Wahba, G. (1990). *Spline Models for Observational Data*. SIAM: Philadelphia.

[50] West, M., Müller, P. and Escobar, M. D. (1994). "Hierarchical priors and mixture models, with application in regression and density estimation," in A. F. M. Smith and P. Freeman (eds.), *Aspects of Uncertainty: A Tribute to D. V. Lindley*. John Wiley: New York.

[51] Williams, C. K. I. (1998). Discussion of "Model–based geostatistics," *Applied Statistics*, **72**.

[52] Wood, S. and Kohn, R. (1998). "A Bayesian approach to robust binary nonparametric regression," *Journal of the American Statistical Association*, **93**, 203–213.

Table 1: Simulated data: Posterior means and standard deviations are shown.

	Normal		Binomial	
	mean	std dev	mean	std dev
β_0	0.200	0.205	0.911	4.010
β_1	0.989	0.155	0.022	5.594
γ_0	-1.344	0.399	2.398	0.820
γ_1	2.689	0.195	2.665	0.364
	Poisson			
	polynomial		linear	
	mean	std dev	mean	std dev
β_0	0.868	1.818	1.003	0.118
β_1	-0.402	2.831	1.933	0.118
γ_0	1.275	0.613	-5.091	0.641
γ_1	3.613	0.255	-5.078	0.686

Table 2: Strike data: Posterior means and standard deviations are shown.

	Model 1		Model 2	
	mean	std dev	mean	std dev
β_0	1.487	0.161	1.485	0.157
β_1 (TIME)	-0.188	0.152	-0.179	0.148
β_2 (OUTPUT)			0.094	0.120
γ_0	-1.418	0.370	-1.429	0.328
γ_1 (TIME)	4.163	0.330	4.208	0.256
γ_2 (OUTPUT)			-5.033	0.661

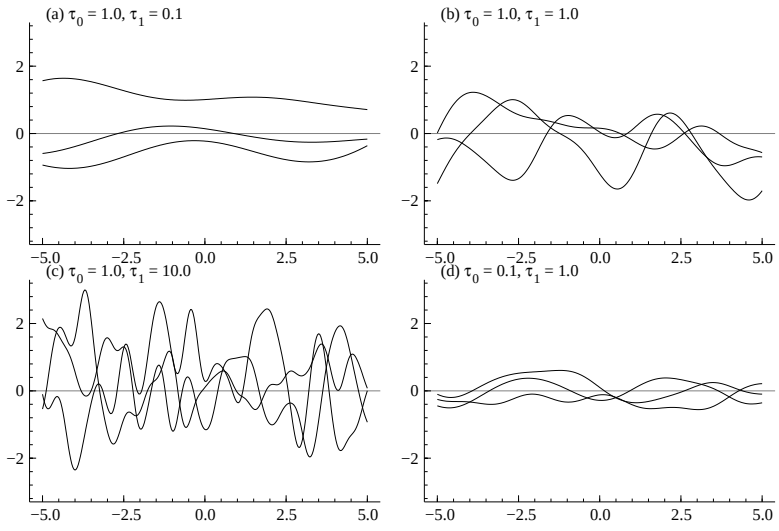

Figure 1: Samples drawn from Gaussian processes: The figure plots three samples drawn from Gaussian processes with parameters (a) $\tau_0 = 1.0$ and $\tau_1 = 0.1$; (b) $\tau_0 = 1.0$ and $\tau_1 = 1.0$; (c) $\tau_0 = 1.0$ and $\tau_1 = 10.0$; (d) $\tau_0 = 0.1$ and $\tau_1 = 1.0$.

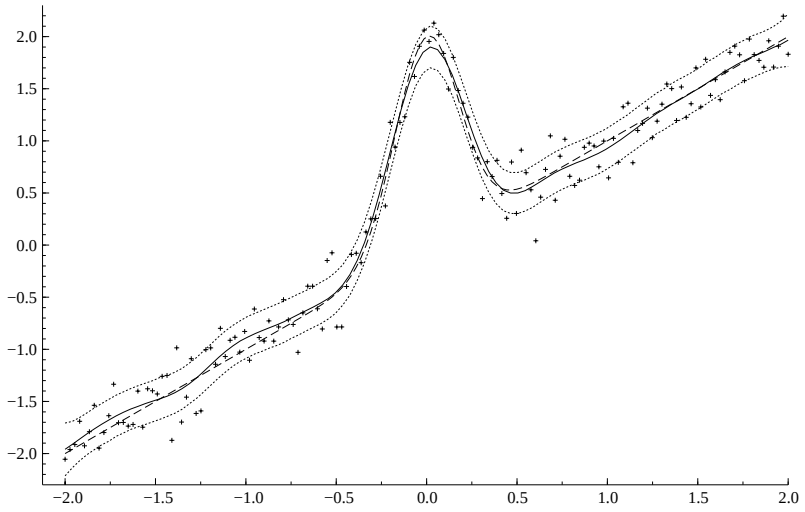

Figure 2: Simulated data (Normal): The true (dashed line) and the posterior means of $\mu(x)$ (solid line) with 90% intervals (dotted line) are shown together with data points (plus).

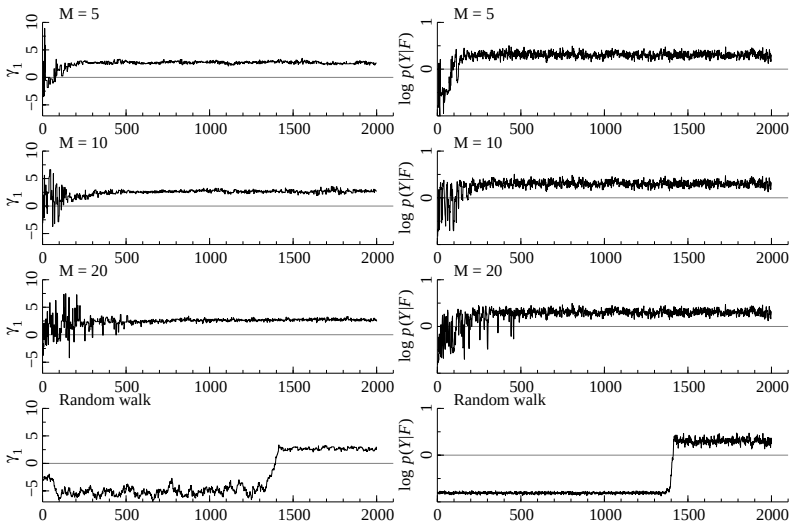

Figure 3: Simulated data (Normal): The figure shows the time series plots of γ_1 (left panels) and the log values of the probability distribution of Y (right panels). From top to bottom, the panels show the results for the GAM samplers with $M = 5$, 10, 20 and the random walk sampler, respectively.

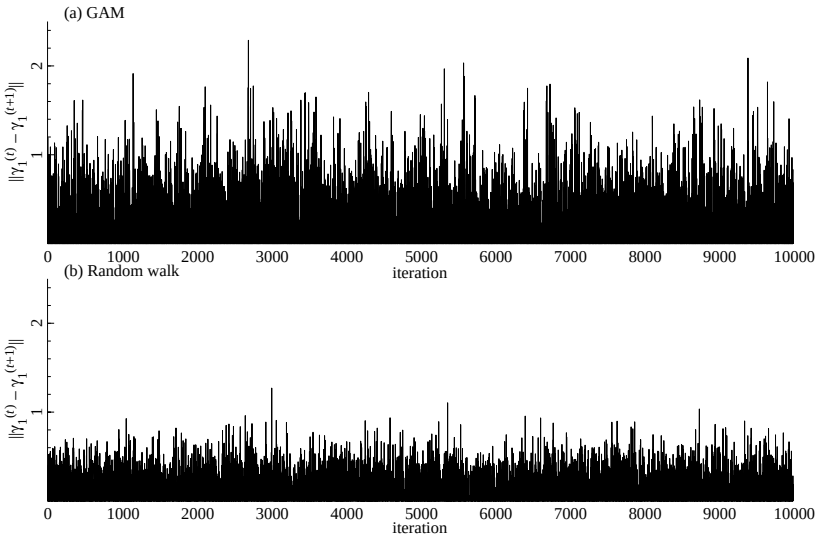

Figure 4: Simulated data (Normal): The figure shows the time series plots of Euclidean norm of the successive values of γ_1 for (a) the GAM sampler ($M = 10$) and (b) the random walk sampler, respectively.

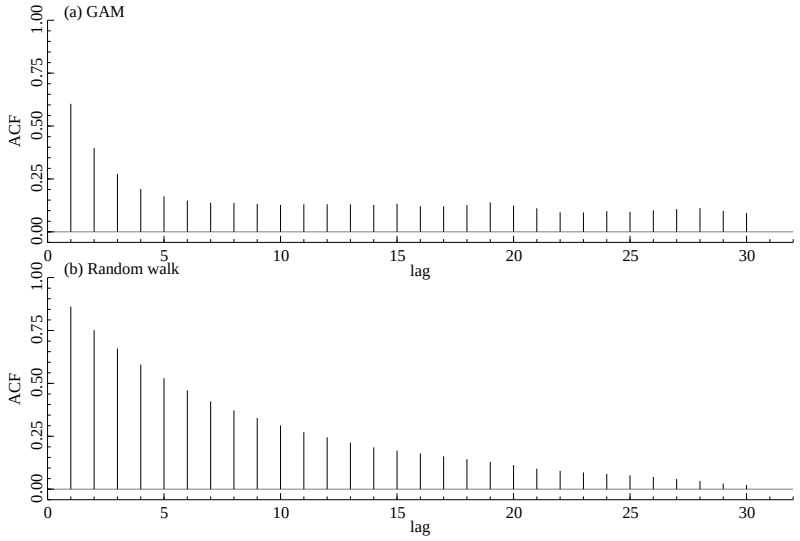

Figure 5: Simulated data (Normal): The figure shows the estimated autocorrelations of γ_1 for (a) the GAM sampler ($M = 10$) and (b) the random walk sampler, respectively.

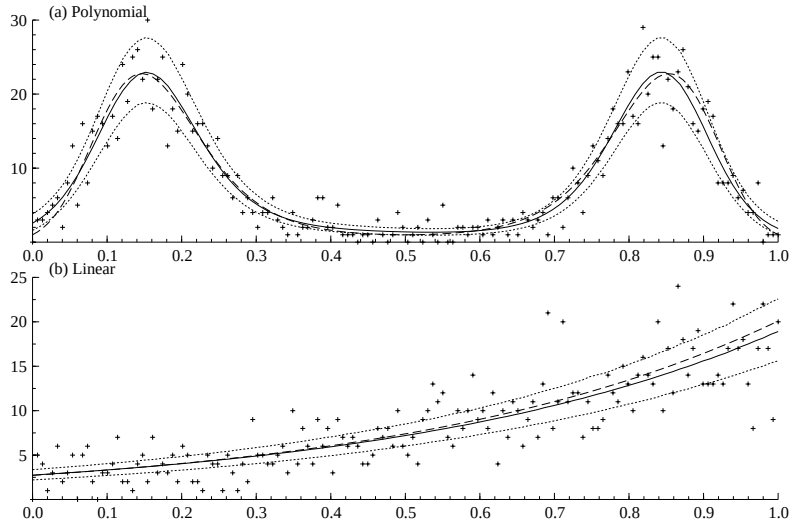

Figure 6: Simulated data (Poisson): The true (dashed line) and the posterior means of $\mu(x)$ (solid line) with 90% intervals (dotted line) are shown together with data points (plus). The panel (a) shows the results for the polynomial case and the panel (b) for the linear case.

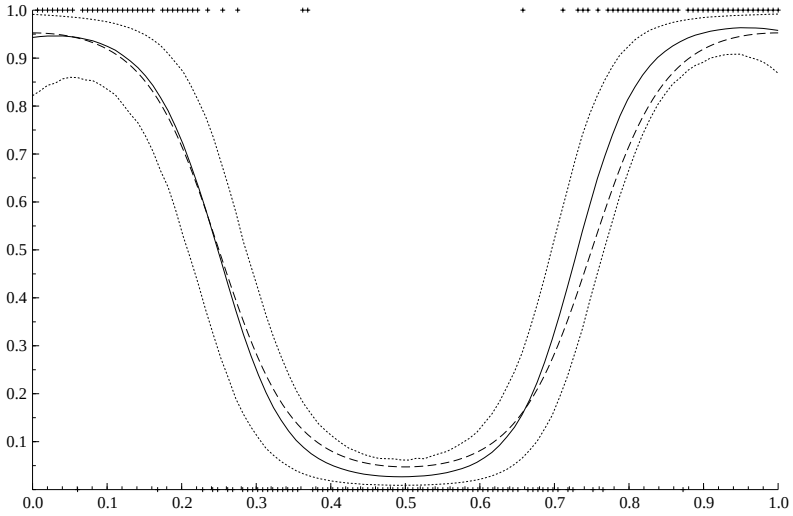

Figure 7: Simulated data (Binomial): The true (dashed line) and the posterior means of $\mu(x)$ (solid line) with 90% intervals (dotted line) are shown together with data points (plus).

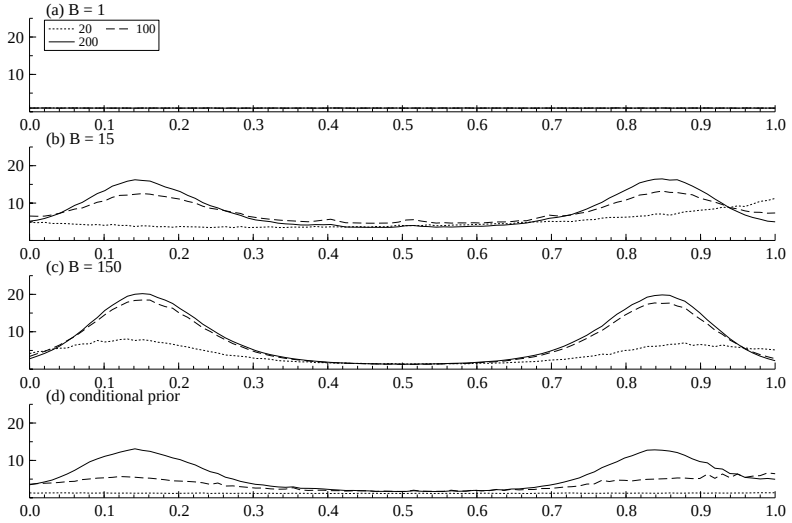

Figure 8: Simulated data (Poisson): The figure shows the posterior means of $\mu(x)$ after 20 (dotted line), 100 (dashed line) and 200 (solid line) iterations for the number of blocks (a) $B = 1$; (b) $B = 15$; (c) $B = 150$, and for (d) the conditional prior proposal.

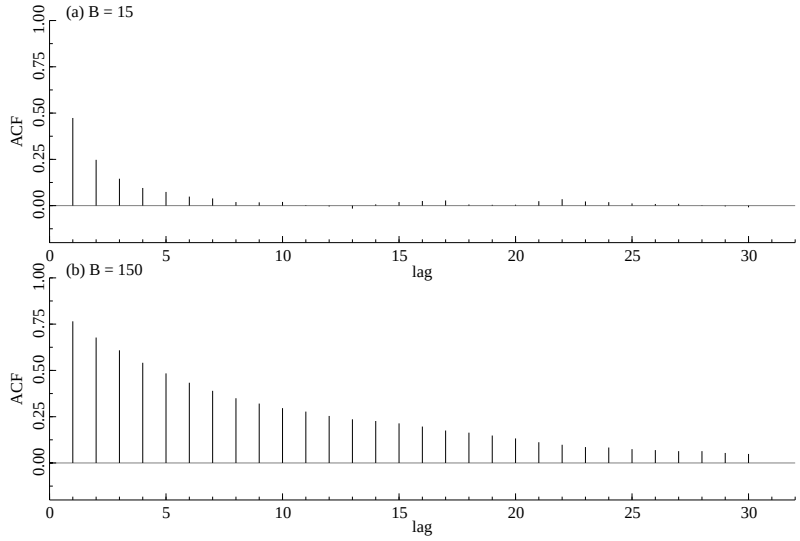

Figure 9: Simulated data (Normal): The figure shows the estimated autocorrelations for (a) the block sampler ($B = 15$) and (b) the single move sampler ($B = 150$), respectively.

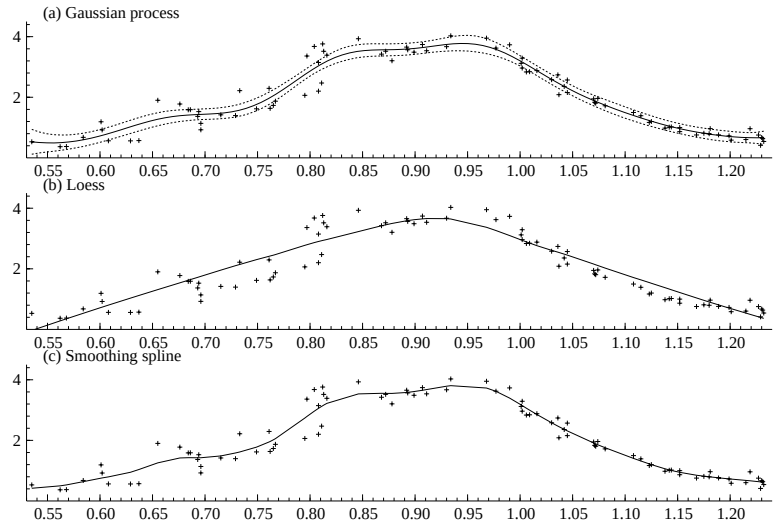

Figure 10: Ethanol data: The panel (a) shows the posterior means of $\mu(x)$ (solid line) with 90% intervals (dotted line) and data points (plus). The panels (b) and (c) show the estimated mean responses (solid line) obtained from the local polynomial regression and the smoothing spline methods respectively.

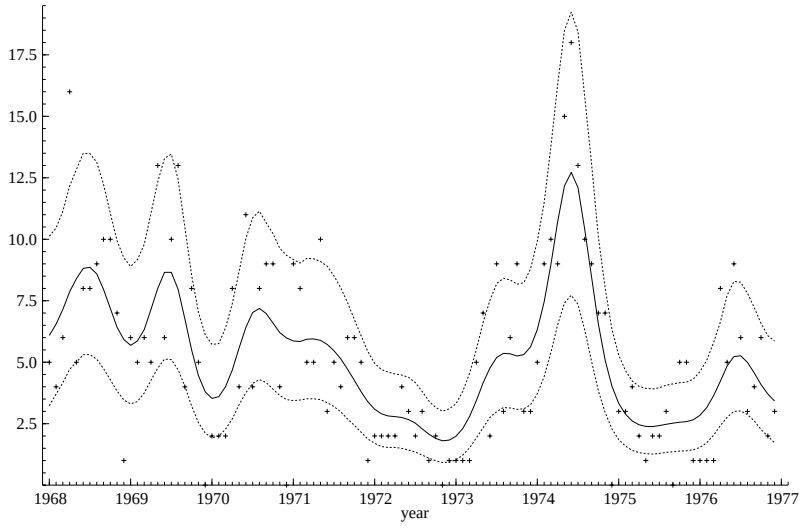

Figure 11: Strike data: The figure shows the posterior means of $\mu(x)$ (solid line) with 90% intervals (dotted line) and data points (plus).