Statistics and Computing **11**: 313–322, 2001
© 2001 Kluwer Academic Publishers. *Manufactured in The Netherlands.*

Nonparametric regression using linear combinations of basis functions

ROBERT KOHN*, MICHAEL SMITH† and DAVID CHAN*

**Australian Graduate School of Management, University of New South Wales, Sydney 2052, Australia*
†Econometrics and Business Statistics Group, Faculty of Economics and Business, University of Sydney, NSW 2006, Australia

Received June 26, 2000 and accepted October 17, 2000

This paper discusses a Bayesian approach to nonparametric regression initially proposed by Smith and Kohn (1996. Journal of Econometrics 75: 317–344). In this approach the regression function is represented as a linear combination of basis terms. The basis terms can be univariate or multivariate functions and can include polynomials, natural splines and radial basis functions. A Bayesian hierarchical model is used such that the coefficient of each basis term can be zero with positive prior probability. The presence of basis terms in the model is determined by latent indicator variables. The posterior mean is estimated by Markov chain Monte Carlo simulation because it is computationally intractable to compute the posterior mean analytically unless a small number of basis terms is used. The present article updates the work of Smith and Kohn (1996. Journal of Econometrics 75: 317–344) to take account of work by us and others over the last three years. A careful discussion is given to all aspects of the model specification, function estimation and the use of sampling schemes. In particular, new sampling schemes are introduced to carry out the variable selection methodology.

Keywords: Bayesian estimation, hierarchical model, Markov chain Monte Carlo, Metropolis-Hastings, variable selection

1. Introduction

This article considers the estimation of a regression function of one or more independent variables when the regression function $f(x)$ is modeled as a linear combination of a family of basis functions $\{\phi_k(x), k = 1, \ldots, K\}$. That is,

$$f(x) = \sum_{k=1}^{K} \beta_k \phi_k(x), \tag{1}$$

where the β_k are the regression coefficients which are estimated from the data. Examples of basis functions are polynomials, polynomial splines, radial basis functions, or wavelet bases. To ensure that the function estimates are satisfactory for a wide variety of regression functions it is usually necessary to take a family with a large number of basis functions, i.e. K is large. For example, K may be 20 or 30 or may even be equal to, or greater than, the number of data points.

Although including a large number of basis functions in the regression (1) produces very flexible estimates, it can also result in over-fitting, and in the extreme case the estimate of the regression function will interpolate the data. To avoid overfitting the regression function we allow a positive prior probability that each regression coefficient is exactly zero. That is, with positive prior probability each basis term can be omitted from (1).

This article uses a Bayesian approach with the unknown regression function estimated by its posterior mean. We note that the Bayesian methods discussed below review and integrate previous work by us and others, as well as suggesting some extensions. The Bayesian approach has four aspects. First, a hierarchical model is developed that employs latent binary variables to indicate whether each coefficient is zero or non-zero. Second, it is necessary to specify proper priors for those regression coefficients that have a positive prior probability of being zero. If improper priors are placed on such coefficients, then, as noted in the Bayesian variable selection literature, these coefficients will be estimated as 0; for example, see Mitchell and Beauchamp (1988) and George and McCulloch (1993). Third, an efficient method of estimating the posterior mean is necessary because it is computationally intractable to evaluate the posterior mean

0960-3174 © 2001 Kluwer Academic Publishers

directly. Fourth, the estimate of the posterior mean of the regression function is constructed from the MCMC iterates.

In our article we use a proper Gaussian conjugate prior for the regression coefficients and show that using such a prior makes the posterior distribution of the regression function invariant to the scale of the independent and dependent variables. Markov chain Monte Carlo (MCMC) simulation is used for the computation and an important contribution of the paper is the introduction of several computationally efficient sampling schemes and a discussion of their statistical properties.

Estimating the regression function as a linear combination of basis terms has been used by a number of authors. One approach to avoid overfitting is regularization in which a penalty term is imposed on a functional of the regression function. For example, in the one dimensional case, the penalty is often taken as the integral of the square of the second derivative. Regularization usually results in the solution of a penalized regression problem with the penalty placed on the regression coefficients. For example, see Wahba (1990, Chapters 1 and 2) and Hastie and Tibshirani (1990, Chapters 1 and 2). A second approach to avoid overfitting is to use a data-driven method to select significant basis terms. This becomes a variable selection problem and has a number of solutions; for example, see Friedman and Silverman (1989), Friedman (1991) and Stone *et al.* (1997). The current paper uses the Bayesian solution proposed by Smith and Kohn (1996, 1997a), and developed further by Dennison, Mallick and Smith (1998) and Holmes and Mallick (1998).

The paper is organized as follows. Section 2 discusses some basis functions that are useful in nonparametric regression. Section 3 outlines the statistical model and the prior specifications. Section 4 gives details of the Bayesian analysis of the nonparametric model, including a discussion of efficient sampling schemes. Section 5 presents a simple empirical example.

2. Basis function representation

There are a number of basis functions that can be used to represent an unknown regression function. We first consider the one dimensional case and then generalize to multiple dimensions.

2.1. *Univariate bases*

A simple univariate basis is the polynomial basis $\{1, x, x^2, \ldots, x^d\}$ of degree d, where $K = d + 1$. A more flexible basis is obtained by using polynomial splines, also called regression splines. For example, a first order polynomial spline basis with knots at $\zeta_1, \ldots, \zeta_l$ is given by $\{1, x, (x - \zeta_1)_+, \ldots, (x - \zeta_l)_+\}$, where $(z)_+ = 0$ if $z \leq 0$ and $(z)_+ = z$ for $z \geq 0$. There are $K = l + 2$ terms in this basis and the knots are just representative points in the domain of x. Polynomial bases and polynomial spline bases are discussed by Hastie and Tibshirani (1990, Chapters 1 and 2).

Radial bases are popular in the neural net literature and generalize to several dimensions. A radial basis typically contains basis functions of the form $\phi_k(x) = \phi(|x - \zeta_k|)$, as well as some low order polynomial terms, where the ζ_k are knots as for the polynomial spline basis; see Powell (1987). The following are some examples of radial bases:

- One dimensional thin plate spline: $\{1, x, |x-\zeta_1|, \ldots, |x-\zeta_l|\}$.
- Quadratic radial basis: $\{1, x, |x - \zeta_1|^2, \ldots, |x - \zeta_l|^2\}$.
- Quasi-logarithmic basis: $\{1, x, |x - \zeta_1| \log|x - \zeta_1|, \ldots, |x - \zeta_l| \log|x - \zeta_l|\}$.

The knots in the polynomial spline bases and the radial bases can be chosen as percentiles of the independent variable vector as in Smith and Kohn (1996), or set to the observed x values as in Dennison, Mallick and Smith (1998). If all the x values are chosen as knots then $K > n$ so that either shrinkage or variable selection is necessary to avoid over-fitting.

2.2. *Multidimensional bases*

There are a number of ways of choosing a basis when the independent variables are multivariate. The choice of basis can impose some structures on the regression function, for example that the regression function is additive, or continuously differentiable up to a given degree. To illustrate, consider the bivariate case with the vector of independent variables equal to $x = (w, z)$. If the regression function is additive then $f(x) = \beta_0 + f_1(w) + f_2(z)$, with f_1 and f_2 univariate functions. To identify f_1 and f_2 we assume that $f_1(0) = 0 = f_2(0)$. It is straightforward to write both f_1 and f_2 as linear combinations of univariate basis functions $\mathcal{B}_1 = \{\phi_{1k}(w), k = 1, \ldots, K_1\}$ and $\mathcal{B}_2 = \{\phi_{2k}(z), k = 1, \ldots, K_2\}$. The resulting additive basis for the bivariate function f is $\{1, \mathcal{B}_1, \mathcal{B}_2\}$ and $K = K_1 + K_2 + 1$.

A more general approach to bivariate surface estimation is to write $f(x)$ as a linear combination of bivariate basis functions $\{\phi_1(x), \ldots, \phi_K(x)\}$. For example, the radial bases generalize to the bivariate case by replacing the univariate norm by the vector norm, where the Euclidean norm is typically used. Therefore, the bivariate equivalents of the three radial bases given above are

- $\{1, w, z, \|x - \zeta_1\|, \ldots, \|x - \zeta_l\|\}$
- $\{1, w, z, \|x - \zeta_1\|^2, \ldots, \|x - \zeta_l\|^2\}$
- $\{1, w, z, \|x - \zeta_1\| \log(\|x - \zeta_1\|), \ldots, \|x - \zeta_l\|) \log(\|x - \zeta_l\|)\}$

The last basis is the thin plate basis in the bivariate case; see Wahba (1990, page 31). Choosing the knots is more difficult in the bivariate case because the set of independent variables $\{x_i, i = 1, \ldots, n\}$ can no longer be ordered. A suggestion for choosing bivariate knots is given by Smith and Kohn (1997). Holmes and Mallick (1998) choose all the x_i as knots, but for n large it may be unnecessary to use such a large number of knots. For regression functions that do not vary radically over the domain of x a promising approach is to cluster the x_i into a moderate number of groups and to use the cluster centers as knots.

It is straightforward to extend either additive bivariate bases or more general bivariate radial bases to higher dimensions. The knots can consist of all the x_i or some smaller number obtained by clustering.

3. Statistical model and prior specification

3.1. *Model description*

Suppose the data consist of the n observations on the dependent variable $y = (y_1, \ldots, y_n)'$, and corresponding observations on the independent variable $(x_1, \ldots, x_n)$. We assume that

$$y_i = f(x_i) + e_i, \qquad i = 1 \ldots, n,$$

where f is the regression function and the errors e_i are independent $N(0, \sigma^2)$. The regression function f is unknown, but we assume that it can be expressed as a linear combination of the basis functions ϕ_j as in (1). Then

$$y = X\beta + e, \tag{2}$$

where X is an $n \times K$ matrix with jth column $(\phi_j(x_1), \ldots, \phi_j(x_n))'$ and $\beta = (\beta_1, \ldots, \beta_K)'$. As in Smith and Kohn (1996), the regression coefficients β_j have a positive prior probability of being 0. This is accomplished by introducing the vector of binary variables $\gamma = (\gamma_1, \ldots, \gamma_K)$ such that $\gamma_i = 1$ if $\phi_i(x)$ is included and $\gamma_i = 0$ if it is not. For a given γ, let $q_\gamma = \sum_i \gamma_i$ be the number of basis terms included in the model, let X_γ be the $n \times q_\gamma$ submatrix of X consisting of those columns of X for which $\gamma_i = 1$, and let β_γ be the corresponding subvector of β. From (2)

$$y = X_\gamma \beta_\gamma + e.$$

A related framework for variable selection was developed and discussed by George and McCulloch (1993, 1997).

3.2. *Prior specification*

We follow the framework set out by Smith and Kohn (1996) for priors for γ, β_γ and σ^2, but refine it in the light of later work by Dennison, Mallick and Smith (1998) and Holmes and Mallick (1998), and ourselves.

The Bayesian model is hierarchical, with the prior for β_γ given conditional on γ and σ^2. The priors for γ and σ^2 are assumed independent *a priori*. The prior for β_γ is

$$p(\beta_\gamma \mid \gamma, \sigma^2) \sim N(\hat{\beta}_\gamma,\ c\sigma^2(X_\gamma' X_\gamma)^{-1}), \tag{3}$$

where $\hat{\beta}_\gamma = (X_\gamma' X_\gamma)^{-1} X_\gamma' y$ is the least squares estimate of β_γ and $c > 1$. To motivate this prior consider the likelihood for β_γ, conditional on γ and σ^2. The likelihood is

$$p(y \mid \gamma, \beta, \sigma^2) \propto \exp\left(-(\beta_\gamma - \hat{\beta}_\gamma)' X_\gamma' X_\gamma (\beta_\gamma - \hat{\beta}_\gamma)/2\sigma^2\right),$$

That is, for given σ^2, the likelihood is Gaussian $N(\hat{\beta}_\gamma,\ \sigma^2(X_\gamma' X_\gamma)^{-1})$ in β_γ. For given σ^2, the prior (3) is a dispersed version of the likelihood and is related to the fractional prior proposed by O'Hagan (1995). We take $c = n$ in which case $c(X_\gamma' X_\gamma)^{-1}$ is likely to remain reasonably constant as n increases.

The prior (3) has the following attractive properties. If the columns of X are rescaled, and hence β is rescaled, then the prior for β is rescaled accordingly. That is, if $X \to XD$ with D a diagonal matrix, then $\beta \to D^{-1}\beta$ and the prior for β_γ is rescaled appropriately. The prior (3) rescales automatically if y is rescaled because of the presence of σ^2 in (3). Furthermore, if the basis term $(1, \ldots, 1)'$ is always included in X_γ, then the prior for β_γ is invariant to location changes in the columns of X and in y.

The prior (3) is data-based because the mean of β_γ depends on y. Using the data-based prior allows proper centering of β and gives excellent results. However, it is straightforward to modify this prior to make it not reliant on the data by centering it at 0, that is,

$$p(\beta_\gamma \mid \gamma, \sigma^2) \sim N(0, c\sigma^2(X_\gamma' X_\gamma)^{-1}), \tag{4}$$

The prior (4) was used by Smith and Kohn (1996) and is related to the g-prior in Zellner (1986).

As in Smith and Kohn (1996), the prior for γ is

$$p(\gamma \mid \pi) = \prod_{i=1}^{K} p(\gamma_i \mid \pi_i),$$

where $\pi = (\pi_1, \ldots, \pi_K)$ is a vector of probabilities which we treat as parameters so that for each i, $p(\gamma_i = 1 \mid \pi_i) = \pi_i$. Henceforth, we take all the π_i to be equal, with $\pi_i = \pi$. In this case $E(q_\gamma \mid \pi) = K\pi$ and $\mathrm{var}(q_\gamma \mid \pi) = K\pi(1 - \pi)$. By choosing a specific value of π we can control the number of terms q_γ that enter the regression. Thus, if $\pi = 0.5$, as in Smith and Kohn (1996), then $E(q_\gamma \mid \pi) = 0.5K$ and $\mathrm{var}(q_\gamma \mid \pi) = 0.25K$. If $K = 25$, then $E(q_\gamma \mid \pi) = 12.5$ and $\mathrm{sd}\,(q_\gamma) = 2.5$, so that q_γ lies in the range 5 to 20 with prior probability close to 1.

A more flexible approach is to place a hyperprior on π. We use a beta prior for convenience, so that

$$p(\pi) = \frac{\pi^{a_\pi - 1}(1 - \pi)^{b_\pi - 1}}{B(a_\pi, b_\pi)}, \tag{5}$$

with $a_\pi > 0$, $b_\pi > 0$, where $B(\cdot, \cdot)$ is the beta function. To choose a_π and b_π, we set the required values of $E(q_\gamma)$ and $\mathrm{var}(q_\gamma)$ and solve for a_π and b_π as shown below. First, note that

$$\begin{aligned} E(q_\gamma) &= KE(\pi), \\ \mathrm{var}(q_\gamma) &= \mathrm{var}(E(q_\gamma \mid \pi)) + E(\mathrm{var}(q_\gamma \mid \pi)), \\ &= K(K-1)E(\pi^2) + KE(\pi)(1 - KE(\pi)), \end{aligned}$$

so that $E(\pi) = E(q_\gamma)/K$. Now,

$$E(\pi) = \frac{a_\pi}{(a_\pi + b_\pi)}, \quad E(\pi^2) = E(\pi)\frac{a_\pi + 1}{(a_\pi + b_\pi + 1)},$$

and the parameters a_π and b_π are obtained by solving the two linear simultaneous equations

$$\begin{aligned} \frac{a_\pi}{(a_\pi + b_\pi)} &= E(\pi), \\ \frac{a_\pi + 1}{(a_\pi + b_\pi + 1)} &= \frac{\mathrm{var}(q_\gamma) - KE(\pi)(1 - KE(\pi))}{K(K-1)E(\pi)}, \end{aligned}$$

For example, suppose we set $E(q_\gamma)=5$ and $\mathrm{sd}(q_\gamma)=5$ so that q_γ is likely range from 0 to 20. For $K=50$, $a_\pi=0.975$ and $b_\pi=8.8$, and for $K=200$, $a_\pi=1.18$ and $b_\pi=46$. Specifying the standard deviation of q_γ as small means that the hyperprior for π is relatively informative, whereas if we make the standard deviation of q_γ large then the hyperprior is relatively uninformative about likely values of π. If $a_\pi=1$ and $b_\pi=1$ then the prior on π is uniformly distributed on (0, 1) and hence completely uninformative about π. Dennison *et al.* (1998) and Holmes and Mallick (1998) use a Poisson prior for the number of terms in the model with a very tight specification on the mean of the Poisson distribution.

We noted above that if a column of 1's is always included in X_γ for all γ, then the prior for β_γ is invariant to location shifts in the columns of X and in y. We can allow for adaptation to such shifts in location by including the basis term $(1,\ldots,1)'$ as the first column of X and setting $\pi_1=1$, with $\pi_i=\pi$ for $i>1$. The prior for π is specified as before except that $q_\gamma \geq 1$ for all γ.

Finally, we have generally used the improper prior $p(\sigma^2)\propto 1/\sigma^2$ for the error variance σ^2. It is unnecessary to have a proper prior for σ^2 because no variable selection is carried out on this parameter. However, it may be more satisfactory to use a proper prior for σ^2, but one that is relatively uninformative. We do so now using the inverse gamma prior

$$p(\sigma^2)=\frac{(\sigma^2)^{-(1+a_\sigma)}\exp(-b_\sigma/\sigma^2)b_\sigma^{a_\sigma}}{\Gamma(a_\sigma)}.$$

The mode of this distribution is at $b_\sigma/(1+a_\sigma)$ and a relatively uninformative prior is obtained by taking a_σ and b_σ close to 0, for example, $a_\sigma=10^{-10}$ and $b_\sigma=0.001$.

4. Bayesian analysis

A Bayesian analysis of the nonparametric regression problem computes the posterior distribution of the regression function f, or at least some of its posterior moments. Our discussion concentrates on estimating the posterior mean of f, but other moments are estimated similarly. The posterior mean of f is

$$E(f\mid y)=\sum_\gamma E(f\mid y,\gamma)p(\gamma\mid y); \tag{6}$$

the sum in (6) is over all permissible values of γ, i.e., over all permissible subsets of the basis functions. If $\pi_i<1$ for all i, then there are 2^K subsets. If K is moderate to large then it is infeasible to evaluate the sum in (6) directly and it is necessary to estimate $E(f\mid y)$ by simulation as outlined below.

Suppose $\gamma^{[l]}$, $l=1,\ldots,L$ are iterates from $p(\gamma\mid y)$ using one of the sampling schemes discussed in Sections 4.2 and 4.3. Then,

$$\hat{f}=\frac{1}{L}\sum_{l=1}^{L}E\big(f\mid y,\gamma^{[l]}\big)$$

is a consistent estimator of $E(f\mid y)$, as $L\to\infty$. For any γ and abscissae x.

$$E(f(x)\mid y,\gamma)=\sum_i \phi_i(x)E(\beta_i\mid y,\gamma)$$

with $E(\beta_i\mid y,\gamma)=0$ if $\gamma_i=0$. The posterior mean $E(\beta_\gamma\mid y,\gamma)$ of the nonzero elements of β is easy to calculate because β_γ is conditionally distributed as a multivariate t distribution; see Section 4.4 for further details.

4.1. *Calculating posterior probabilities*

The posterior probability of the subset γ is

$$p(\gamma\mid y)\propto p(y\mid\gamma)p(\gamma). \tag{7}$$

In our work it is important to have explicit expression for both the marginal likelihood $p(y\mid\gamma)$ and for $p(\gamma)$ with π integrated out. Following Smith and Kohn (1996),

$$\begin{aligned}p(y\mid\gamma)&=\int p(y\mid\gamma,\beta_\gamma,\sigma^2)p(\beta_\gamma\mid\gamma,\sigma^2)p(\sigma^2)d\beta_\gamma\,d\sigma^2\\&\propto(1+c)^{-q_\gamma/2}\,(S(\gamma)+2b_\sigma)^{-(\frac{n}{2}+a_\sigma)}\end{aligned} \tag{8}$$

For the prior (3), the term $S(\gamma)$ equals the error sum of squares $ESS(\gamma)=y'y-y'X_\gamma(X_\gamma'X_\gamma)^{-1}X_\gamma'y$, while for the prior (4), $S(\gamma)=ESS(\gamma)+RSS(\gamma)/(c+1)$ where $RSS(\gamma)=y'X_\gamma(X_\gamma'X_\gamma)^{-1}X_\gamma'y$ is the regression sum of squares. In both cases $S(\gamma)$ is positive, so that $S(\gamma)+2b_\sigma$ is strictly positive for $b_\sigma>0$, which helps numerical problems associated with small values of $S(\gamma)$.

To understand the effect of γ on the marginal likelihood $p(y\mid\gamma)$, consider the following partial ordering on γ. We say that $\gamma^{(1)}\leq\gamma^{(2)}$ if $\gamma_i^{(1)}\leq\gamma_i^{(2)}$ for all i. Then the term $(1+c)^{-q_\gamma/2}$ decreases in this partial ordering, while the term $(S(\gamma)+2b_\sigma)^{-(\frac{n}{2}+a_\sigma)}$ increases. In particular, if the addition of a new basis term results in only a slight or no increase in the regression sum of squares then the marginal likelihood will decrease. The probability $p(\gamma)$ in (7) is evaluated as

$$\begin{aligned}p(\gamma)&=\int p(\gamma\mid\pi)p(\pi)\,d\pi\\&=\frac{1}{B(a_\pi,b_\pi)}\int\pi^{q_\gamma+a_\pi-1}(1-\pi)^{K-q_\gamma+b_\pi-1}d\pi\\&=\frac{B(q_\gamma+a_\pi,K-q_\gamma+b_\pi)}{B(a_\pi,b_\pi)}.\end{aligned} \tag{9}$$

4.2. *One at a time sampling schemes*

The first sampling scheme we consider is the Gibbs sampler used by Smith and Kohn (1996). In this scheme the γ_i are generated one at a time conditional on $\gamma_{j\neq i}$, with β, σ^2 and π integrated out. This makes the sampling scheme more efficient than previous approaches which generated β and σ^2 as well.

Gibbs Sampling Scheme
For $i = 1$ to K

Generate γ_i from $p(\gamma_i \mid y,\ \gamma_{j\neq i})$

End

The sampling scheme gives one complete scan of the Gibbs sampler, that is, all the γ_i are generated once. Note that instead of generating the γ_i in some specific order, they can be generated in a random order. To implement the Gibbs sampler it is necessary to evaluate the probability $p(\gamma_i = 1 \mid y,\ \gamma_{j\neq i})$. This probability is obtained as follows. Let $\gamma^{(1)} = (\gamma_1, \ldots, \gamma_{i-1},\ \gamma_i = 1,\ \gamma_{i+1}, \ldots, \gamma_K)$ and $\gamma^{(0)} = (\gamma_1, \ldots, \gamma_{i-1},\ \gamma_i = 0, \gamma_{i+1}, \ldots, \gamma_K)$, let $L_1 \propto p(y \mid \gamma^{(1)})$ and $L_0 \propto p(y \mid \gamma^{(0)})$ be evaluated as in (8), and let $\theta_1 = p(\gamma_i = 1 \mid \gamma_{j\neq i})$, $\theta_0 = 1 - \theta_1$, and $D = \theta_1 L_1 + \theta_0 L_0$. It is straightforward to show using (7) to (9) that

$$p(\gamma_i = 1 \mid y,\ \gamma_{j\neq i}) = \frac{\theta_1 L_1}{D}, \quad \text{where} \tag{10}$$

$$\theta_1 = p(\gamma_i = 1 \mid \gamma_{j\neq i}) = \frac{q_{\gamma^1} + a_\pi - 1}{K + a_\pi + b_\pi - 1}. \tag{11}$$

Expressions for L_1 and L_0 are given in (8) and it is only necessary to compute one of L_1 or L_0 for each i because the value of the other is available from the previous step of the sampling scheme.

In our experience the Gibbs sampler works reliably even when the number of knots is very large. However, using the Gibbs sampler means that it is necessary to compute $S(\gamma)$ once for each i and this computation requires a Cholesky factorization of the matrix $X'_\gamma X_\gamma$, or at least its update, as in Smith and Kohn (1996). Such repeated computation may be wasteful because q_γ is usually much smaller than K, so that θ_1 is small from (11), and if the current value of $\gamma_i = 0$ then γ_i is likely to be regenerated as 0.

We consider three Metropolis-Hastings schemes (Hastings 1970) to generate γ that offer significant potential savings over the Gibbs sampler. The following notation is used in all three schemes. Let $Q(0 \to 1) = Q(\gamma_i = 0 \to \gamma_i = 1)$ be the proposal probability to generate $\gamma_i = 1$ when it is currently 0, and define the proposal probability $Q(1 \to 0)$ of going from 1 to 0 similarly. The Metropolis-Hastings acceptance probability of going from 0 to 1 is (Hastings 1970)

$$\alpha(0 \to 1) = \min\left(1,\ \frac{\theta_1 L_1}{\theta_0 L_0} \times \frac{Q(1 \to 0)}{Q(0 \to 1)}\right),$$

and $\alpha(1 \to 0)$ is defined similarly.

The first sampling scheme uses a proposal that attempts to make the Gibbs sampler more efficient. The second sampling scheme uses the prior for γ as a proposal. The proposal in the third sampling scheme is accepted with probability one. It is not clear a priori which sampling scheme is more efficient. We prove that sampling schemes 1 and 2 are equivalent when generating the γ_i one at a time, and are more efficient than sampling scheme 3. We subscript the proposal probabilities Q and the acceptance probabilities α with the sampling scheme number.

Sampling Scheme 1
For $i = 1$ to K
Let $Q_1(0 \to 1) = \theta_1 \min(1, \frac{L_1}{D})$ and $Q_1(1 \to 0) = \theta_0 \min(1, \frac{L_0}{D})$.
The acceptance probabilities are

$$\alpha_1(0 \to 1) = \begin{cases} 1 & \text{if } L_1/L_0 \geq 1 \\ D/L_0 & \text{if } L_1/L_0 < 1 \end{cases}$$

$$\alpha_1(1 \to 0) = \begin{cases} 1 & \text{if } L_0/L_1 \geq 1 \\ D/L_1 & \text{if } L_0/L_1 < 1 \end{cases}$$

End

The acceptance probabilities for the sampling scheme above follow from the following lemma. Its proof is straightforward and is omitted.

Lemma 1. *The following three conditions are equivalent.* (*a*) $L_1/L_0 \geq 1$; (*b*) $L_1/D \geq 1$; *and* (*c*) $L_0/D < 1$.

To understand why the first sampling scheme may be more efficient than the Gibbs sampler in the nonparametric regression problem, we note that γ_i is zero for most i. That is, there are usually many superfluous basis terms at the end of any given scan because $q_\gamma \ll K$. Suppose that a given γ_i is currently 0. To generate γ_i using the Gibbs sampler, we compute L_1 and generate a uniform U. If $U \leq \theta_1 L_1/D$ then γ_i is set to 1, otherwise it remains at 0. Thus, in the Gibbs sampler it is always necessary to evaluate L_1 when γ_i is currently 0. In sampling scheme 1 we generate a uniform U. If $U > \theta_1$, then $U > Q_1(0 \to 1)$ and γ_i stays at 0. Only if $U \leq \theta_1$ is it necessary to evaluate L_1 and determine whether $U \leq Q_1(0 \to 1)$. However, the event that $U > \theta_1$ happens most of the time because θ_1 is small and it is then unnecessary to evaluate L_1. Similar computations are necessary if γ_i is currently 1. Therefore, the big computational advantage that sampling scheme 1 has over the Gibbs sampler in the nonparametric regression problem is that for most i it is unnecessary to evaluate L_0 or L_1. Instead, the prior probability θ_1 is calculated as given in equation (11), which is computationally simple. For example, Smith and Kohn (1997b) apply sampling scheme 1 to estimate nonparametric seemingly unrelated regression models with many binary indicator variables.

The argument just given that sampling scheme 1 is more efficient than the Gibbs sampler is suggestive rather than formal and is based on the number of floating point operations per iteration required by both samplers. However, statistical efficiency per floating point operation is a better measure of the efficiency of a sampling scheme. Statistical efficiency can be measured by the variance or the root-mean-squared error of an estimator, when a given sampling scheme is run. We found empirically that

sampling scheme 1 is more efficient than the Gibbs sampler per floating point operation.

The second sampling scheme uses the prior as a proposal and is described as follows. We show in Lemma 2 that sampling schemes 1 and 2 are equivalent. Thus, sampling scheme 1 helps us to understand intuitively why sampling scheme 2 will be statistically more efficient than the Gibbs sampler.

Sampling Scheme 2
For $i = 1$ to K

Let $Q_2(0 \to 1) = \theta_1$ and $Q_2(1 \to 0) = \theta_0$. The acceptance probabilities are

$$\alpha_2(0 \to 1) = \begin{cases} 1 & \text{if } L_1/L_0 \geq 1 \\ L_1/L_0 & \text{if } L_1/L_0 < 1 \end{cases}$$

$$\alpha_2(1 \to 0) = \begin{cases} 1 & \text{if } L_0/L_1 \geq 1 \\ L_0/L_1 & \text{if } L_0/L_1 < 1 \end{cases}$$

End

The third sampling scheme is described as follows.

Sampling Scheme 3
For $i = 1$ to K
Let

$$Q_3(0 \to 1) = \theta_1 \frac{L_1}{L_1 + L_0} \quad \text{and}$$

$$Q_3(1 \to 0) = \theta_0 \frac{L_0}{L_1 + L_0}$$

The acceptance probabilities $\alpha_3(0 \to 1)$ and $\alpha_3(1 \to 0)$ are always 1.
End

The following lemma compares sampling schemes 1 to 3 and the Gibbs sampler and shows that the transition probabilities of the first two sampling are equal and are at least as large as the transition probabilities of the third sampling scheme. It follows from Peskun (1973) that the first two sampling schemes are statistically more efficient than the third sampler. We conclude that sampling schemes 1 and 2 are equivalent and are prefered to sampling scheme 3. The Gibbs sampler is more efficient per scan than the first two sampling schemes, but it is also computationally more demanding. We show empirically in Section 5, for a given example, that the first sampling scheme is statistically more efficient per floating point operation than the Gibbs sampler. In Lemma 2, let $Q_G(0 \to 1)$ be the transition probability defined at equation (10) for the Gibbs sampler and define $Q_G(1 \to 0)$ similarly. The acceptance probabilities for the Gibbs sampler are of course 1, when it is viewed as a special case of the Metropolis-Hastings method.

Lemma 2. *The following results hold for sampling schemes* 1 *to* 3 *and follow from Lemma* 1.

$$\begin{aligned} Q_G(0 \to 1) &> Q_1(0 \to 1)\alpha_1(0 \to 1) \\ &= Q_2(0 \to 1)\alpha_2(0 \to 1) > Q_3(0 \to 1)\alpha_3(0 \to 1) \end{aligned}$$

$$\begin{aligned} Q_G(1 \to 0) &> Q_1(1 \to 0)\alpha_1(1 \to 0) \\ &= Q_2(1 \to 0)\alpha_2(1 \to 0) > Q_3(1 \to 0)\alpha_3(1 \to 0) \end{aligned}$$

4.3. *Block sampling schemes*

It is straightforward to generalize to block sampling the ideas of one at a time sampling, that is generating the γ_i one at a time, as in sampling schemes 1 to 3 and the Gibbs sampler. Suppose that γ_B is a block of the γ_i; for example, suppose the γ_i are ordered in some way and are generated in blocks of 2 so that $\gamma_B = (\gamma_{2i-1},\ \gamma_{2i})$. Alternatively, we can pick 2 of the γ_i at random, out of the K possible. Let γ^C be the current value of γ, that is the value of γ at the end of the last generation, and let $\gamma_{\setminus B}$ be the elements of γ not in γ_B. The block γ_B becomes the basic unit for generation, and $\gamma_{\setminus B} = \gamma^C_{\setminus B}$.

The conditional prior probability $p(\gamma_B \mid \gamma_{\setminus B})$ is obtained from (9) as follows. Let q_{γ_B}, $q_{\gamma_{B^C}}$ and $q_{\gamma_{\setminus B}}$ be the number of nonzero elements in γ_B, γ_B^C and $\gamma_{\setminus B}$ respectively. Then,

$$\frac{p(\gamma_B \mid \gamma_{\setminus B})}{p(\gamma_B^C \mid \gamma_{\setminus B})} = \frac{B(q_{\gamma_B} + q_{\gamma_{\setminus B}} + a_\pi,\ K - q_{\gamma_B} - q_{\gamma_{\setminus B}} + b_\pi)}{B(q_{\gamma_B^C} + q_{\gamma_{\setminus B}} + a_\pi,\ K - q_{\gamma_B^C} - q_{\gamma_{\setminus B}} + b_\pi)}$$

Hence,

$$\begin{aligned} \frac{1}{p(\gamma_B^C \mid \gamma_{\setminus B})} &= \sum_{\gamma_B} \frac{p(\gamma_B \mid \gamma_{\setminus B})}{p(\gamma_B^C \mid \gamma_{\setminus B})} \\ &= \sum_{\gamma_B} \frac{B(q_{\gamma_B} + q_{\gamma_{\setminus B}} + a_\pi,\ K - q_{\gamma_B} - q_{\gamma_{\setminus B}} + b_\pi)}{B(q_{\gamma_B^C} + q_{\gamma_{\setminus B}} + a_\pi, K - q_{\gamma_B^C} - q_{\gamma_{\setminus B}} + b_\pi)} \end{aligned}$$

from which it is possible to solve for $p(\gamma_B^C \mid \gamma_{\setminus B})$ and hence obtain $p(\gamma_B \mid \gamma_{\setminus B})$ for all γ_B.

We now show how sampling schemes 1 to 3 generalize to the block case. It is convenient to introduce the following notation. Let $\theta_B = p(\gamma_B \mid \gamma_{\setminus B})$, $\theta_B^C = p(\gamma_B^C \mid \gamma_{\setminus B})$, $L(\gamma) \propto p(y \mid \gamma)$ as in (8), $L_B = L(\gamma_B,\ \gamma_{\setminus B})$, and $L_B^C = L(\gamma_B^C,\ \gamma_{\setminus B})$. Let

$$D = \sum_{\gamma_B} \theta_B L_B \quad \text{and} \quad D_L = \sum_{\gamma_B} L_B,$$

where the sums are over all possible values of γ_B. Finally, for the ith sampling scheme let $Q_i(\gamma_B^C \to \gamma_B)$ be the proposal probability of going from γ_B^C to γ_B and $\alpha_i(\gamma_B^C \to \gamma_B)$ be the acceptance probability. Then, the three sampling schemes generalize with proposals for $\gamma_B \neq \gamma_B^C$,

Sampling Scheme 1	$Q_1(\gamma_B^C \to \gamma_B) = \theta_B \min(1,\ \frac{L_B}{D})$
Sampling Scheme 2	$Q_2(\gamma_B^C \to \gamma_B) = \theta_B$
Sampling Scheme 3	$Q_3(\gamma_B^C \to \gamma_B) = \theta_B L_B/D_L$
Gibbs	$Q_G(\gamma_B^C \to \gamma_B) = \theta_B L_B/D$

The acceptance probability for sampling scheme i is

$$\alpha_i(\gamma_B^C \to \gamma_B) = \min\left(1, \frac{\theta_B\, L_B\, Q_i(\gamma_B \to \gamma_B^C)}{\theta_B^C\, L_B^C\, Q_i(\gamma_B^C \to \gamma_B)}\right).$$

It is straightforward to show as in Section 4.2 that the acceptance probability $\alpha_2(\gamma_B^C \to \gamma_B) = \min(1,\ L_B/L_B^C)$, and α_3 is identically 1. Lemma 3 generalizes Lemma 2 to the block sampling case.

Lemma 3. *The following results hold for the block sampling schemes* 1 *to* 3 *and the Gibbs sampler.*

(a) $\alpha_2(\gamma_B^C \to \gamma_B)Q_2(\gamma_B^C \to \gamma_B) \geq \alpha_1(\gamma_B^C \to \gamma_B)Q_1(\gamma_B^C \to \gamma_B)$.
(b) $\alpha_2(\gamma_B^C \to \gamma_B)Q_2(\gamma_B^C \to \gamma_B) > \alpha_3(\gamma_B^C \to \gamma_B)Q_3(\gamma_B^C \to \gamma_B)$.
(c) *For a block size of* 2 *or greater, sampling scheme* 2 *neither dominates*, *nor is dominated*, *by the Gibbs sampler.*

Proof: For convenience, we write α_i and Q_i for $\alpha_i(\gamma_B^C \to \gamma_B)$ and $Q_i(\gamma_B^C \to \gamma_B)$, respectively. To prove the lemma, we substitute in the proposal Q_i and transition probability α_i for each of the sampling schemes and consider the results in the following four different cases.

Case (i). $L_B/D > 1$ and $L_B^C/D > 1$. For this case we can check that $Q_1 = \theta_B$ and $\alpha_1 = \min(1,\ L_B/L_B^C)$. Hence $\alpha_G Q_G > \alpha_1 Q_1 = \alpha_2 Q_2 > Q_3$.
Case (ii). $L_B/D > 1$ and $L_B^C/D < 1$. For this case $Q_1 = \theta_B = Q_2$ and $\alpha_1 = 1 = \alpha_2$. Furthermore, $Q_G > Q_2 > Q_3$.
Case (iii). $L_B/D < 1$and $L_B^C/D > 1$. For this case $Q_1 = \theta_B L_B/D$ and $\alpha_1 = D/L_C^B$. Hence, $\alpha_G Q_G > \alpha_1 Q_1 = \theta_B L_B/L_B^C = \alpha_2 Q_2 > \alpha_3 Q_3$.
Case (iv). $L_B/D < 1$ and $L_B^C/D < 1$. For this case $Q_1 = \theta_B$ $L_B/D = Q_G$ and $\alpha_1 = 1$. $\alpha_2 Q_2 = \theta_B \min(1,\ L_B/L_B^C) \geq \alpha_1 Q_1$. Furthermore, it is straightforward to check that $\alpha_2 Q_2 > \alpha_3 Q_3$ because $D_L \geq L_C^B$. □

An alternative way of traversing the space of γ is to use the reversible jump approach of Green (1995) as in Dennison, Mallick and Smith (1998) and Holmes and Mallick (1998). The reversible jump method is a Metropolis-Hastings scheme with 3 basic proposals: (a) a birth proposal in which a γ_i that is currently a 0 is proposed to be 1; (b) a death proposal in which a γ_i that is currently 1 is proposed to be a 0; and (c) a swap proposal in which a γ_i that is currently 0 and a γ_j that is currently 1 are proposed to exchange values. The three steps are done according to a probabilistic method described in Green (1995).

The reversible jump Metropolis-Hastings method allows a limited block move whereas the one at a time sampling schemes do not. However, the block sampling schemes shows how to carry out block moves with blocks of size 2 or greater in a way that may be more flexible than the reversible jump Metropolis-Hastings method. We do not compare these two different approaches formally in the present article.

4.4. *Posterior mean estimation*

This section gives some extra details of how to estimate the posterior mean of $f(x)$. For given γ and abscissae x, let $\phi_\gamma(x)$ be the vector of $\phi_i(x)$ for which $\gamma_i = 1$. Then

$$E(f(x) \mid y,\ \gamma) = E(\beta_\gamma \mid y,\ \gamma)'\phi_\gamma(x)$$

which is readily computed for each γ. The posterior distribution of $p(\beta_\gamma \mid y,\ \gamma)$ is obtained as in Smith and Kohn (1996) with the derivation sketched out below.

$$\begin{aligned} p(\beta_\gamma \mid y, \gamma) &= \int p(\beta_\gamma, \sigma^2 \mid y, \gamma) d\sigma^2 \\ &\propto \int p(y \mid \beta_\gamma,\ \gamma,\ \sigma^2) p(\beta_\gamma \mid \gamma,\ \sigma^2) p(\sigma^2) d\sigma^2 \\ &\propto \Bigg(S(\gamma) + 2b_\sigma + \frac{1+c}{c}(\beta_\gamma - \tilde{\beta}_\gamma)' \\ &\quad \times X_\gamma' X_\gamma (\beta_\gamma - \tilde{\beta}_\gamma) \Bigg)^{-(\frac{n+q_\gamma}{2} + a_\sigma)} \end{aligned}$$

where $\tilde{\beta}_\gamma = \hat{\beta}_\gamma$ if the prior is (3) and $\tilde{\beta}_\gamma = \frac{c}{c+1}\hat{\beta}_\gamma$ if the prior is (4), where $\hat{\beta}_\gamma$ is defined in Section 3.2. Therefore,

$$\sqrt{\frac{c+1}{c}}(X_\gamma' X_\gamma)^{-\frac{1}{2}}(\beta_\gamma - \tilde{\beta}_\gamma) = \frac{\zeta}{\sqrt{v/m}}$$

with $\zeta \sim N(0,\ I)$, v is independent of ζ and has a gamma density with parameters $(n + 2a_\sigma)/2$ and $1/2$, and $m = S(\gamma) + 2b_\sigma$ and $E(\beta_\gamma \mid y,\ \gamma) = \tilde{\beta}_\gamma$.

5. Example

This section illustrates the methods discussed in the article for estimating a regression function nonparametrically. We compare the statistical efficiency of the Gibbs sampling scheme and the second sampling scheme with block sizes 1, 2 and 4 for a bivariate example by comparing the variance of the function estimates obtained using each of the sampling schemes as well as the substantive number of floating point operations required for each scheme for a given number of iterations.

5.1. *Inefficiency factor*

To compare the statistical efficiency of the Gibbs sampler and sampling scheme 2, consider first the estimate $\hat{f}_G(x)$ of the regression function f, obtained using the Gibbs sampler, at the abscissae x. From Section 4,

$$\hat{f}_G(x) = \frac{1}{L}\sum_{l=1}^{L} E\big(f(x) \mid y,\ \gamma^{[l]}\big)$$

where $\gamma^{[l]}$, $l = 1, \ldots, L$ are the iterates of γ after the sampler has converged. Let $Z_l = E(f(x) \mid y,\ \gamma^{[l]})$. Then $Z_1, \ldots, Z_L$ form a stationary sequence and $\hat{f}_G(x) = \bar{Z}$, the sample mean of the Z_l. As in Hastings (1970),

$$\operatorname{var}(\hat{f}_G(x)) = \operatorname{var}(\bar{Z}) = \frac{\sigma_G^2}{L}\left(1 + 2\sum_{l=1}^{L-1}\left(1 - \frac{l}{L}\right)\rho_{G,l}\right), \tag{12}$$

Table 1. *Inefficiencies for the Gibbs sampler and sampling scheme 2 (SS2) with block sizes of 1, 2 and 4*

Point	1	2	3	4	5	6	7	8	9	10
Gibbs	139.5	153.0	116.8	56.5	88.9	21.1	114.3	123.5	53.2	77.1
SS2(1)	204.4	210.7	137.6	191.7	94.0	47.7	117.3	129.8	67.3	94.8
SS2(2)	208.1	210.1	140.8	213.5	115.0	51.4	141.5	141.4	68.9	118.7
SS2(4)	207.5	237.1	171.5	115.1	137.2	72.9	181.4	196.9	115.8	129.0

where $\sigma_G^2 = \text{var}(Z_l)$ and $\rho_{G,l}$ is the lth autocorrelation of the Z_l sequence. If $\rho_{G,l}$ is negligible for some $M \ll L$ then (12) simplifies to

$$\text{var}(\bar{Z}) = \frac{\sigma_G^2}{L}\left(1 + 2\sum_{l=1}^{M}\left(1 - \frac{l}{L}\right)\rho_{G,l}\right)$$

which is estimated by

$$\widehat{\text{var}}(\hat{f}_G(x)) = \frac{\hat{\sigma}_G^2}{L}\left(1 + 2\sum_{l=1}^{M}\left(1 - \frac{l}{L}\right)\hat{\rho}_{G,l}\right) \tag{13}$$

where $\hat{\sigma}_G^2$ is the sample estimate of σ_G^2 and $\hat{\rho}_{G,l}$ is the sample estimate of $\rho_{G,l}$. We use the estimators

$$\hat{\sigma}_G^2 = \sum_{j=1}^{L}(Z_j - \bar{Z})^2 \Big/ L \quad \text{and}$$

$$\hat{\rho}_{G,l} = \sum_{j=1}^{L}(Z_j - \bar{Z})(Z_{j+l} - \bar{Z}) \Big/ (L\hat{\sigma}_G^2).$$

The term

$$\widehat{\text{INEFF}}_G = \left(1 + 2\sum_{l=1}^{M}\left(1 - \frac{l}{L}\right)\hat{\rho}_{G,l}\right)$$

is called in the literature an estimate of the inefficiency factor of the Gibbs sampler because it measures the factor by which the number of iterates L needs to increase in order that the estimate based on the Gibbs sampler has the same variance as an independent sample of size L.

The inefficiency factor for sampling scheme 2 is estimated similarly. We call the estimate $\widehat{\text{INEFF}}_{SS2}$. From Lemma 2 in Section 4 we expect that $\widehat{\text{INEFF}}_{SS2}/\widehat{\text{INEFF}}_{G>1}$. However, in comparing the two sampling schemes it is also necessary to take into account the number of floating point operations per iteration required by both samplers. The number required by sampling scheme 2 will be much smaller than for the Gibbs sampler. This is illustrated using the bivariate example in the next section.

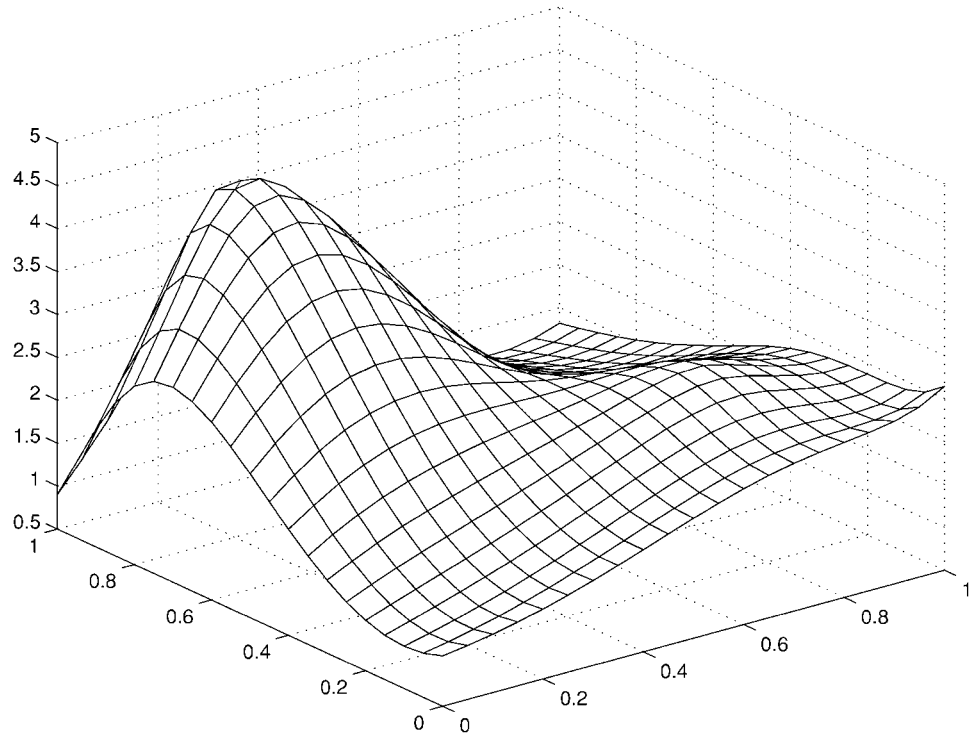

Fig. 1. *Plots of bivariate regression function estimates using the Gibbs sampler and SS2(1). The estimates are so similar that it is hard to distinguish the two different estimates*

Table 2. *Some summary statistics for the Gibbs sampler and sampling scheme 2 (SS2) with block sizes of 1, 2 and 4*

	Gibbs	SS2(1)	SS2(2)	SS2(4)
Integrated squared error	0.0683	0.0679	0.0696	0.0669
Cholesky updates per iteration	103.00	13.81	11.93	10.63
Expected number of terms	7.62	7.61	7.38	7.64
Standard deviation of number of terms	1.13	1.14	1.06	1.10
Acceptance rate requiring Cholesky	—	9.77%	8.63%	7.80
Overall acceptance rate	—	87.90%	89.42%	90.49

5.2. *Bivariate smoothing*

This section considers the bivariate regression function, based on a mixture of two normals.

$$f(x) = 1 + \mathrm{N}(\mu_1,\ \Sigma_1,\ x) + \mathrm{N}(\mu_2,\ \Sigma_2,\ x)$$

where $N(\mu,\ \Sigma,\ x)$ is a bivariate normal density evaluated at the abcissa x, with mean μ and covariance matrix Σ. In the expression above, we have $(\mu_1 = (0.25,\ 0.75)'$, $\Sigma_1 = (\Sigma_{1,ij})$ with $\Sigma_{1,11} = \Sigma_{1,22} = 0.05$ and $\Sigma_{1,12} = 0.01$ and $\mu_2 = (0.75,\ 0.25)'$, $\Sigma_2 = (\Sigma_{2,ij})$ with $\Sigma_{2,11} = \Sigma_{2,22} = 0.1$ and $\Sigma_{2,12} = 0.01$.

The abscissae x are generated uniformly over the unit square. The sample size is $n = 600$ and the error standard deviation is $\sigma = 0.5$. We use the bivariate thin plate basis with $n/6 = 100$ knots so the number of terms in the basis is $K = 103$. The knots are chosen by a clustering algorithm. The warmup period was 5,000 iteration and the sampling period 50,000 iterations. The function was estimated on a 21×21 equally spaced grid on the unit square and the inefficiency factors for both sampling scheme 2 and the Gibbs sampler were calculated at 10 points chosen from this grid of 441 points. Each of the inefficiency factors was estimated based on 150 autocorrelation estimates. The parameters a_π and b_π in the prior for π were chosen so that $E(q_\gamma) = 5$ and $\mathrm{sd}(q_\gamma) = 5$. However, we obtained similar results to those reported below by taking both a_π and b_π equal to 1.

Table 1 presents the inefficiencies at each of the 10 points for the different sampling schemes and Table 2 presents some summary statistics.

The ratio of the inefficiencies of sampling scheme 2 to Gibbs is at most 3.45, using a block size of 4, which means that for the 10 function estimates considered, it takes at most about 3.45 times as many iterations to obtain the same accuracy from sampling scheme 2 as it does for the Gibbs sampler. However, the number of Cholesky updates per iteration for the Gibbs sampler is 103 and it is 10.6 for SS2(4). That is, the Gibbs sampler requires about 9.7 times as many Cholesky updates as SS2(4), which

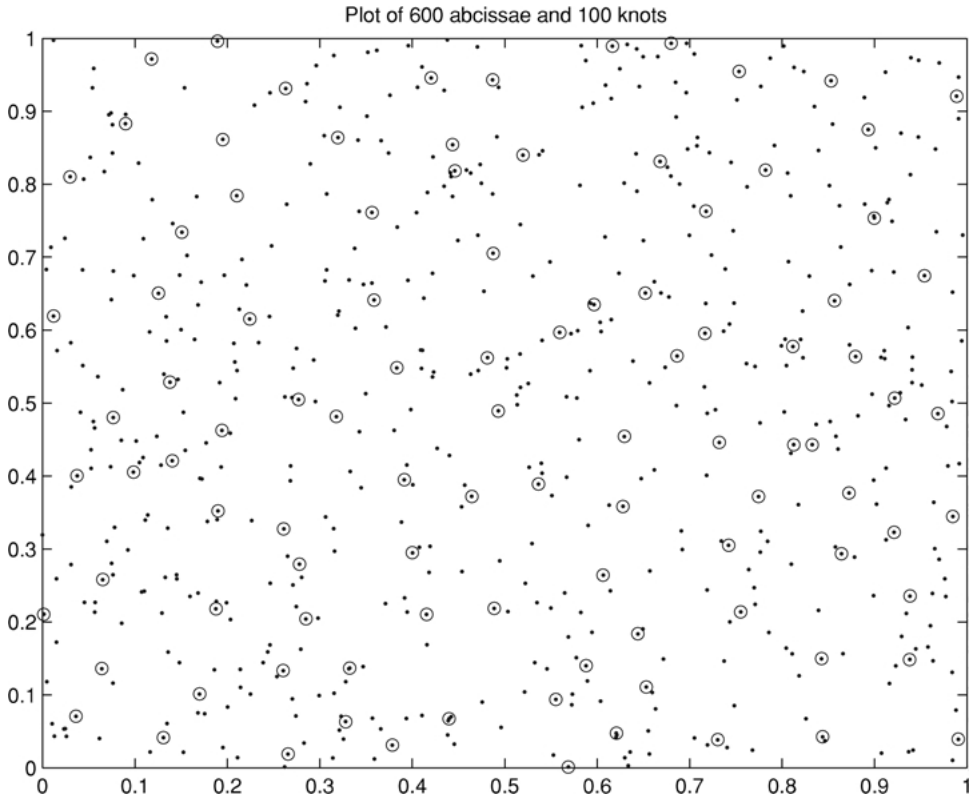

Fig. 2. *Plot of the 600 abscissae and the 100 knots chosen by a clustering algorithm*

means that SS2(4) is at least 2.8 times as efficient as the Gibbs sampler per floating point operation.

We see that on average, increasing the block size increases the inefficiency factor, but this is offset by a decrease in the number of Cholesky updates required per iteration. All schemes have roughly the same expected number of terms and the integrated squared error for all 4 schemes are almost identical.

The integrated squared errors, the expected number of terms per iteration and the standard deviation of the number of terms per iteration reported in Table 2 suggest that the Gibbs sampler and sampling scheme 2 with block sizes of 1, 2 and 4 converge to the same posterior distribution.

The last two lines in Table 2 show the average acceptance rates for the Metropolis-Hastings method used in SS2. The second last line gives the acceptance rates for SS2 when the transition requires performing a Cholesky decomposition. That is, when $\gamma_B^C \neq \gamma_B^N$. This acceptance rate decreases as block size increases because the correct density and the proposed density are more likely to differ for larger blocks. The final line in Table 2 has the overall acceptance rates for SS2 including transitions when $\gamma_B^C = \gamma_B^N$, where the acceptance probability is always 1. The overall acceptance rates increases as block size increases because we have more transitions of the type $\gamma_B^C \to \gamma_B^N$ where the block of γ's are all zeros.

Figure 1 plots the surface estimates using both the Gibbs sampler and sampling scheme 2 with block size 1; the figure shows the two estimates are almost identical.

Figure 2 plots the 600 abscissae and the 100 knots chosen by the clustering algorithm. It is clear from the plot that for these abcissae, the 100 knots easily cover the 600 abscissae and in fact a smaller number of knots will be adequate.

Acknowledgment

We thank an anonymous referee for suggestions that improved the presentation of the paper. Robert Kohn and Michael Smith were partially supported by Australian Research Council grants.

References

Dennison D.G.T., Mallick B.K., and Smith A.F.M. 1998. Automatic Bayesian curve fitting. Journal of the Royal Statistical Society B 60: 333–350.

Friedman J.H. 1991. Multivariate adaptive regression splines. The Annals of Statistics 19: 1–141.

Friedman J.H. and Silverman B.W. 1989. Flexible parsimonious smoothing and additive modeling. Technometrics 31: 3–39.

George E.I. and McCulloch R.E. 1993. Variable selection via Gibbs sampling. Journal of the American Statistical Association 88: 881–889.

George E.I. and McCulloch R.E. 1997. Approaches for Bayesian variable selection. Statistica Sinica 7: 339–373.

Green P.J. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82: 711–732.

Hastie T.J. and Tibshirani R.J. 1990. Generalized Additive Models. Chapman Hall, New York.

Hastings W.K. 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57: 97–109.

Holmes C.C. and Mallick B.K. 1998. Radial basis functions of variable dimension. Neural Computation 10: 1217–1233.

Mitchell T.J. and Beauchamp J.J. 1988. Bayesian variable selection in linear regression. Journal of the American Statistical Association 83: 1023–1036.

O'Hagan A. 1995. Fractional Bayes factors for model comparison (with discussion). Journal of the Royal Statistical Society B 57: 99–138.

Peskun P.H. 1973. Optimum Monte Carlo sampling using Markov chains. Biometrika 60: 607–612.

Powell M.J.D. 1987. Radial basis functions for multivariate interpolation. In: Mason J.M. and Cox M. (Eds.), Algorithms for Approximation. Oxford University Press, Oxford, pp. 143–167.

Smith M. and Kohn R. 1996. Nonparametric regression via Bayesian variable selection. Journal of Econometrics. 75: 317–344.

Smith M. and Kohn R. 1997a. A Bayesian approach to nonparametric bivariate regression. Journal of the American Statistical Association 92: 1522–1535.

Smith M. and Kohn R. 2000. Nonparametric seemingly unrelated regression. Journal of Econometrics 98: 257–281.

Stone C., Hansen M., Kooperberg C., and Truong Y. 1996. Polynomial splines and their tensor products in extended linear modeling. Annals of Statistics 25: 1371–1470.

Wahba G. 1990. Spline Models for Observational Data. CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 59. SIAM, Philadelphia.

Zellner A. 1986. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In: P.K. Goel and A. Zellner (Eds.), Bayesian Inference and Decision Techniques—Essays in Honor of Bruno de Finetti. North Holland, Amsterdam, 11:233–243.