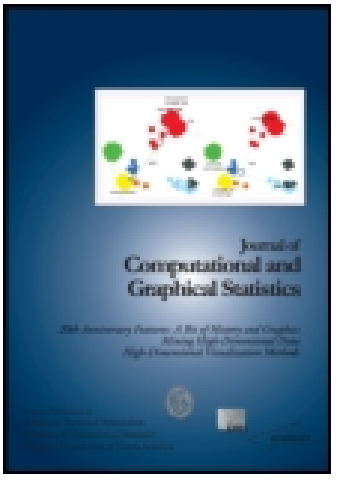

Journal of Computational and Graphical Statistics

Publication details, including instructions for authors and subscription information:
http://amstat.tandfonline.com/loi/ucgs20

Bayesian Variable Selection and Model Averaging in High-Dimensional Multinomial Nonparametric Regression

Paul Yau[a], Robert Kohn[a] & Sally Wood[a]

[a] Paul Yau is a Ph.D. Student, Robert Kohn is Professor, and Sally Wood is Senior Lecturer, Australian Graduate School of Management, University of New South Wales, UNSW Sydney NSW 2052, Australia ; and .

Published online: 01 Jan 2012.

To cite this article: Paul Yau, Robert Kohn & Sally Wood (2003) Bayesian Variable Selection and Model Averaging in High-Dimensional Multinomial Nonparametric Regression, Journal of Computational and Graphical Statistics, 12:1, 23-54, DOI: 10.1198/1061860031301

To link to this article: http://dx.doi.org/10.1198/1061860031301

Bayesian Variable Selection and Model Averaging in High-Dimensional Multinomial Nonparametric Regression

Paul YAU, Robert KOHN, and Sally WOOD

This article presents a Bayesian method for estimating nonparametrically a high-dimensional multinomial regression model. The regression functions are expressed as sums of main effects and interactions and our approach is able to select the significant components entering the model. Each of the main effects and interactions is written as a linear combination of basis terms with a variance components type prior on the regression coefficients. The conditional class probabilities are estimated using both variable selection and model averaging. Our approach can also be used for classification and gives results that are comparable to modern classification methods, but at the same time the results are highly interpretable to the practitioner. All computation is carried out using Markov chain Monte Carlo simulation.

Key Words: Classification; Markov chain Monte Carlo; Radial basis functions.

1. INTRODUCTION

This article considers the problem of nonparametric binary and multinomial regression. That is, we are interested in exploring the dependence of a variable w on the predictor variables $x_1, \ldots, x_p$, when w is categorical and takes on the values $1, \ldots, C$. In particular, we are interested in estimating the conditional probability of w being k given $x_1, \ldots, x_p$, that is, $\Pr(w = k|x_1, \ldots, x_p)$, for $k = 1, \ldots, C$. When $C = 2$ the problem reduces to binary nonparametric regression. An important application of multinomial regression is the multiple classification problem where the goal is to find $\arg\max_k \Pr(w = k|x_1, \ldots, x_p)$ for given values of $x_1, \ldots, x_p$. See Kooperberg, Bose, and Stone (1997) for a discussion.

Frequently, especially at the start of a study, a large number of potential predictor variables may be of interest, some of which may have a significant effect on the conditional class

Paul Yau is a Ph.D. Student, Robert Kohn is Professor, and Sally Wood is Senior Lecturer, Australian Graduate School of Management, University of New South Wales, UNSW Sydney NSW 2052, Australia (E-mail: pauly@agsm.edu.au; R.Kohn@unsw.edu.au; *and* sallyw@agsm.edu.au).

Journal of Computational and Graphical Statistics, Volume 12, Number 1, Pages 23–54
DOI: 10.1198/1061860031301

probabilities, while others may have little or no effect. Moreover, some potential predictor variables may jointly affect the response through their interaction. This may produce a large and complex regression model in which variable selection decisions need to be made and a large number of smoothing parameters need to be estimated. For example, in a binary regression with 10 potential predictor variables and with only main effects and two-way interactions included, the effective number of components is 55. In a multinomial problem with 5 categories and the same 10 predictor variables, the effective number of components is $55 \times (5 - 1) = 220$. In many problems, both the number of variables and the number of categories is much larger.

This article proposes a Bayesian variable selection and model averaging approach to multinomial nonparametric regression which can handle a large number of variables and their interactions. A different approach for solving this problem is to first choose a smaller subset of variables using some parametric variable selection technique and then to apply nonparametric multinomial regression to this smaller subset of variables; see, for example, Wahba, Gu, Wang, and Chappell (1995). However, selecting variables by assuming a parametric model may distort the true model and miss some important predictor variables. In contrast, our approach performs variable selection and nonparametric estimation simultaneously as illustrated with the Pima Indian diabetes dataset in Section 5.2.

We use a multinomial probit regression model with data augmentation (Albert and Chib 1993) to turn the multinomial regression problem into a sequence of smoothing problems with Gaussian errors. A functional ANOVA decomposition of the regression functions is used with main effects and two-way interactions. Each functional component of the regression functions is approximated as a linear combination of radial basis functions. To avoid overfitting we place a proper prior on the regression coefficients. This is also called a penalized likelihood approach in statistics; see, for example, Green and Silverman (1994, p. 50).

Nonparametric multinomial regression was studied by Lin (1998) and Kooperberg, Bose, and Stone (1997). Lin (1998) used a penalized smoothing spline approach and multiple logistic regression to model the regression functions. Shrinkage estimators based on roughness consideration are employed to smooth the regression function. Lin employed a reduction scheme similar to Xiang and Wahba (1998) to handle large datasets, but does not consider variable selection. Kooperberg, Bose, and Stone (1997) suggested using polynomial splines and their tensor products and multiple logistic regression techniques to model the regression functions. Instead of performing shrinkage, they carried out knot and variable selection to determine the significant spline terms and variables. They used stepwise addition and stepwise deletion to do the selection, but can handle relatively few basis terms because their method requires variable selection for a non-Gaussian likelihood. Marx and Eilers (1998) used a penalized likelihood approach with a polynomial spline basis to estimate nonparametrically an additive binary regression model. Their approach can be generalized to nonparametric multinomial regression, but they did not consider variable selection. Shively, Kohn, and Wood (1999) considered variable selection in binary additive nonparametric regression using smoothing splines to estimate the additive main effects.

Our article generalizes the variable selection approach of Shively, Kohn, and Wood (1999) in a number of ways. First, we allow for interaction effects. Second, we consider multinomial as well as binary regression. Third, our method can handle a large number of main effects and interactions because we write each component of the regression function as a linear combination of basis terms instead of representing it as a smoothing spline. Fourth, our method of computation is much more efficient than the method of Shively, Kohn, and Wood (1999). Our method of representing the unknown regression function is similar to the method of Eilers and Marx (1996), but we use the Bayesian machinery to estimate the smoothing parameters and carry out component selection. It is difficult computationally to estimate many functions simultaneously using the frequentist framework when the smoothing parameters are estimated by cross-validation or maximum likelihood because it is necessary to use a grid search. Similarly, it is computationally difficult to carry out component selection in the frequentist framework using component selection criteria such as AIC or BIC when the number of components m is large and each component can be in or out of the model because 2^m model comparisons are necessary.

The article is organized as follows. Section 2 describes the binary regression model, the way we select the basis functions, and sets up the Bayesian framework to do the estimation without variable selection. Section 3 extends the methodology to include variable and component selection and model averaging. Practical implementation of the estimation procedure for complex regression model is discussed. Section 4 uses data augmentation to convert the multinomial regression problem into a multivariate Gaussian problem. Section 5 applies our methodology to a number of examples. Two simulated datasets are used to illustrate the frequentist properties of our nonparametric variable selection approach. The Pima Indian diabetes dataset is used to illustrate how nonparametric variable selection can be used to model the dataset and recover a significant variable which was omitted in previous analyses when variable selection and nonparametric regression were performed separately. Two other datasets commonly used in the classification literature are used to compare our approach to some other classification methods. Section 6 discusses the application of our ideas to more general priors on the regression coefficients and to estimators that are more spatially adaptive. Section 7 summarizes the results in the article.

2. BASIC BINARY MODEL

Suppose a binary response variable w takes the values 1 or 2 and depends on p explanatory variables $x_1, \ldots, x_p$. Without loss of generality, we assume $x_1, \ldots, x_p$ are in the range $[0, 1]$. We model the dependence of w on $x_1, \ldots, x_p$ by the probit regression model

$$\Pr(w = 1|x_1, \ldots, x_p) = \Phi(f(x_1, \ldots, x_p)), \tag{2.1}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function and f is an unknown regression function in $x_1, \ldots, x_p$. To allow the probabilities to be estimated flexibly, we assume that f is a smooth function, but do not prescribe its functional form.

Following Albert and Chib (1993), we augment w with $y = f(x_1, \ldots, x_p) + e$, where $e \sim N(0, 1)$, so that $y > 0$ if $w = 1$ and $y < 0$ if $w = 2$. This device for data augmentation turns the binary regression problem into a Gaussian regression problem.

Throughout this article, we decompose f into additive functions of main effects, two-way interactions and so on. Such a functional decomposition into main effects and interactions is commonly used in the statistics literature. Our approach can be applied to models with general multi-way interactions, although in this article we do not go beyond models with two-way interactions.

2.1 BASIS FUNCTION APPROXIMATION

In this article we approximate each component of f by a linear combination of thin-plate basis functions as well as linear polynomial terms. Thin-plate basis function are defined as (Wahba 1990, p. 31),

$$\begin{aligned} \phi_k(x) &= \|x - \tilde{x}_k\|^{(2m-d)} \log(\|x - \tilde{x}_k\|), \quad 2m - d \text{ even} \\ \phi_k(x) &= \|x - \tilde{x}_k\|^{(2m-d)}, \quad 2m - d \text{ odd} \end{aligned}$$

We take $m = 2$ and d is the dimension of the abscissae x. For example, the basis terms for the main effect for x_i are

$$S_i = \{x_i, \|x_i - \tilde{x}_{i1}\|^3, \ldots, \|x_i - \tilde{x}_{iK}\|^3\}$$

and the basis terms for the interaction effect for x_i and x_j are

$$S_{ij} = \{\|(x_i, x_j) - (\tilde{x}_i, \tilde{x}_j)_1\|^2 \log(\|(x_i, x_j) - (\tilde{x}_i, \tilde{x}_j)_1\|), \ldots, \|(x_i, x_j) - (\tilde{x}_i, \tilde{x}_j)_K\|^2 \log(\|(x_i, x_j) - (\tilde{x}_i, \tilde{x}_j)_K\|)\}.$$

In the expressions above the $\tilde{x}_{ik}$ and the $(\tilde{x}_i, \tilde{x}_j)_k$ are the basis locations and K is the number of basis terms. We allow K to be different for each functional component.

Thin-plate splines are an example of a class of functions called radial basis functions that are used widely for approximation, especially in neural networks, for example, Bishop (1995, p. 169), and were used by Holmes and Mallick (1998) in their nonparametric smoother. The value of a radial basis function $\phi_k(x)$ depends only on the distance of the abscissae x from the basis location $\tilde{x}_k$ (and possibly some constants). The attraction of using radial basis functions in smoothing is that they are well defined in any dimension and maintain their simplicity in high dimensions. Although we use thin plate splines in our work, other bases can be used instead.

In nonparametric smoothing it is usually necessary to choose a large number of basis terms. For example, the traditional smoothing spline approach in Wahba (1990, chap. 2) includes all data points as basis terms and uses a roughness penalty to recover the smoothness of the regression function. Some parsimonious bases were proposed recently—for example, pseudosplines (Hastie 1996) and penalized regression splines (Eilers and Marx 1996). The studies by Xiang (1996) and Ruppert and Carroll (2000) show that for smoothing splines and

penalized regression splines respectively, the number of basis terms required can be, much smaller than the number of data points without significantly degrading the performance of the smoother. Xiang showed that in order to achieve the same convergence rate as the penalized smoothing spline, the number of basis terms required increases much more slowly than the number of data points. We follow their ideas and allow the number of basis terms to be large but far smaller than the number of data points. Unlike Xiang (1996) and Ruppert and Carroll (2000), we do not include a data-driven strategy to determine the number of basis terms because we focus on variable and component selection and on model averaging in high dimensional multinomial regression. However, in Section 6 we show how our approach can be used to obtain smoothing that is more spatially adaptive.

We adopt the following strategy to select basis terms. In one dimension we select the basis terms to be equally spaced sample quantiles as in Smith and Kohn (1996) and Ruppert and Carroll (2000). If K basis terms are to be selected, then the basis locations are the $i/(K+1)$th sample quantiles of the data points, $i = 1, \ldots, K$. In higher dimensions, for example in two dimensions, we use a clustering algorithm to select the basis locations. Xiang (1996) used the FASTCLUS procedure from the SAS library to cluster the data points into distinct groups and then chose one data point randomly from each group as the basis location. A discussion of this method is given in Xiang and Wahba (1998). The basic idea is that clustering finds a fixed number of groups of data points that cover the data region and have maximum separation. We modify their method by using the clustering algorithm "clara" described by Kaufman and Rousseeuw (1990) and Struyf, Hubert, and Rousseeuw (1996), instead of FASTCLUS. The program "clara" is available from the Web site http://www.stat.ucla.edu/journals/jss. The program is written in Fortran and has an S-plus front end. As an illustration, Figure 1 displays the basis locations selected by "clara" for one of the interaction surfaces in the Pima Indian diabetes example presented in Section 5.2. It can be seen that the basis locations are spread over the domain of the interaction surface.

2.2 BAYESIAN ESTIMATION

Let $\mathbf{w} = (w_1, \ldots, w_n)', \mathbf{x}_1 = (x_{11}, \ldots, x_{1n})', \ldots, \mathbf{x}_p = (x_{p1}, \ldots, x_{pn})'$ be the observations and $\mathbf{y} = (y_1, \ldots, y_n)'$ the corresponding augmented variable. Writing the basis terms for the ith functional component of the regression function as $\phi_{ik}(\mathbf{x}_1, \ldots, \mathbf{x}_p), k = 1, \ldots, K$, the probit regression model in (2.1) can be rewritten as

$$\Pr(w = 1 | x_1, \ldots, x_p) = \Phi\left(\gamma + \sum_i \sum_k \beta_{ik}\phi_{ik}(\mathbf{x}_1, \ldots, \mathbf{x}_p)\right). \tag{2.2}$$

For conciseness, we only take two functional components from this point onward. Then (2.2) becomes

$$\Pr(w = 1 | x_1, \ldots, x_p) = \Phi\left(\gamma + \sum_{k=1}^{K_1} \beta_{1k}\phi_{1k}(\mathbf{x}_1, \ldots, \mathbf{x}_p) + \sum_{k=1}^{K_2} \beta_{2k}\phi_{2k}(\mathbf{x}_1, \ldots, \mathbf{x}_p)\right).$$

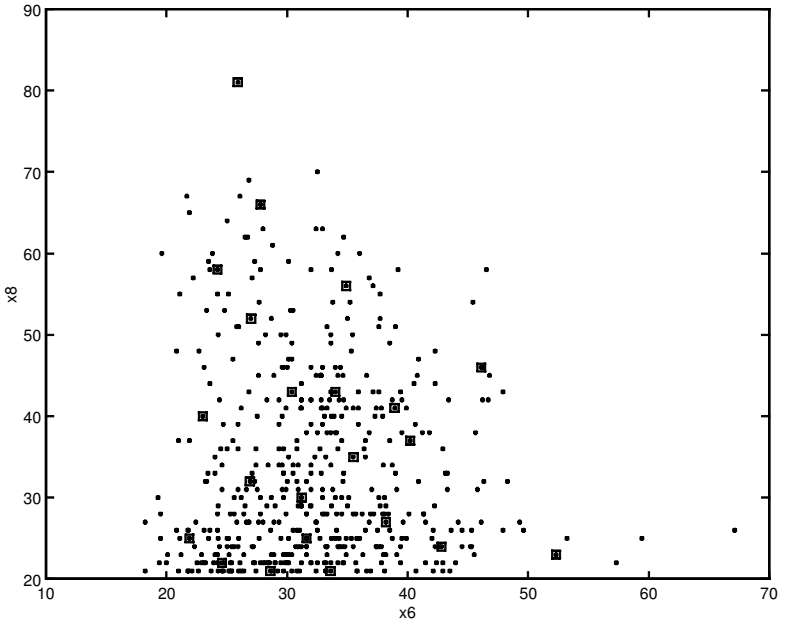

Figure 1. Basis locations selected by "clara" for the interaction surface between variables 6 and 8 of the of the Pima Indian diabetes dataset decribed in section 5.1. The dots are the observations and the squares are the basis locations.

The corresponding augmented equation is

$$y = \gamma + \sum_{k=1}^{K_1} \beta_{1k}\phi_{1k}(\mathbf{x}_1, \ldots, \mathbf{x}_p) + \sum_{k=1}^{K_2} \beta_{2k}\phi_{2k}(\mathbf{x}_1, \ldots, \mathbf{x}_p) + e \ , \tag{2.3}$$

where $e \sim N(0, 1)$. It is more convenient to write (2.3) in matrix form as

$$\mathbf{y} = \mathbf{Z}\gamma + \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{e} \ ,$$

where $\mathbf{Z}$ is a column vector of ones corresponding to the global intercept term γ.

To carry out a Bayesian analysis we place the following priors on γ, $\boldsymbol{\beta}_1$, and $\boldsymbol{\beta}_2$. Let $U(a, b)$ denote the uniform distribution on the interval (a, b). The prior on γ is $U(-10^{15}, 10^{15})$ which is proper, but very diffuse, so there is virtually no shrinkage of γ. The coefficient vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are assumed to be independent apriori, with $\boldsymbol{\beta}_i \sim N(0, c_i\mathbf{I}_{K_i})$. The hyperparameters c_i control the amount of shrinkage of the vectors $\boldsymbol{\beta}_i$. In this section we use proper but at priors $c_i \sim U(0, 10^{15})$. In the next section it is necessary to use data-based priors for the c_i to carry out variable and component selection. For given values of c_1 and c_2 the mode of the posterior density of $\gamma, \boldsymbol{\beta}_1$, and $\boldsymbol{\beta}_2$ is the same as the solution to the penalized likelihood problem

$$\arg \min_{\gamma, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2} \left\{ -\log\{p(\mathbf{w}|\gamma, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2)\} + \frac{1}{c_1}\boldsymbol{\beta}_1'\boldsymbol{\beta}_1 + \frac{1}{c_2}\boldsymbol{\beta}_2'\boldsymbol{\beta}_2 \right\} \ . \tag{2.4}$$

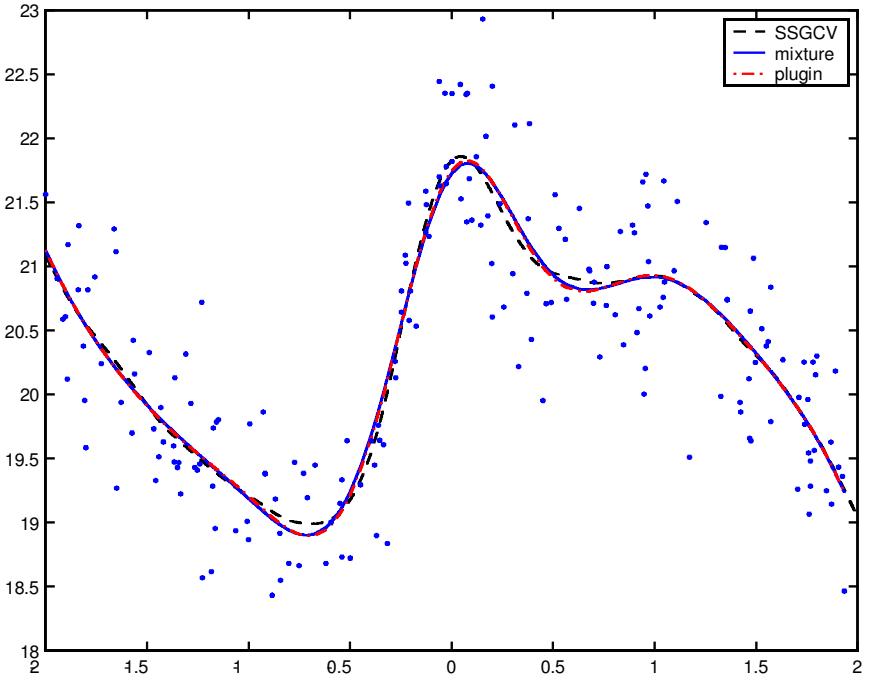

Figure 2. Nonparametric regression estimates of the function 20 + sin2x + 2 exp(−16 x²) using smoothing splines and generalized cross-validation (SSGCV), the posterior mean estimate with all parameters integrated out (mixture) and the Bayesian estimator with the posterior mean estimate of c_1 plugged in (plugin).

That is, the prior penalizes the size of the coefficient vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. Such an approach is used by Girosi, Jones, and Poggio (1995) in their application to radial basis neural networks and by Ruppert and Carroll (2000). These priors are different to the roughness penalty priors implicit in spline smoothing (Wahba 1990, chap. 2) and also used by Eilers and Marx (1996) and Marx and Eilers (1998). We use these priors because of their simplicity in multiple dimensions, but it is straightforward to extend our analysis to priors that impose a smoothness penalty. Such an extension is outlined in Section 6.

The following example illustrates that our prior for β acts similarly to a roughness penalty on the unknown function. We generated 200 observations from the model

$$y = 20 + \sin 2x + 2\exp(-16x^2) + e\,,$$

with x uniformly distributed on $(-2, 2)$ and $e \sim N(0, 0.5^2)$. The range of x was standardized to $[0, 1]$ and we used 40 basis terms to approximate the regression function. Figure 2 shows the posterior mean of the regression function when the smoothness parameter c_1 is set to its posterior mean, the posterior mean of the regression function with all unknown parameters integrated out and the smoothing spline estimate of the regression function with the smoothing parameter estimated by generalized cross-validation. The smoothing spline estimate was obtained using the Rkpack program applying the method of Gu, Bates, and Wahba (1989). The figure shows that the function estimates obtained using our prior for β

are smooth and very similar to the smoothing splines estimate, which is based on a roughness penalty.

We are interested in estimating the joint posterior distribution of $p(c_1, c_2, \gamma, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \mathbf{y}|\mathbf{w})$, and specifically the posterior means of $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, and γ. From these we estimate the regression function f and $\Pr(w = 1|x_1, \ldots, x_p)$. The joint distribution of $\{\mathbf{w}, \mathbf{y}, \gamma, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, c_1, c_2\}$ can be expressed in hierarchical form as

$$p(\mathbf{w}, \mathbf{y}, \gamma, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, c_1, c_2) = p(\mathbf{w}|\mathbf{y})p(\mathbf{y}|\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \gamma)p(\boldsymbol{\beta}_1|c_1)p(\boldsymbol{\beta}_2|c_2)p(\gamma)p(c_1)p(c_2).$$

To simplify the calculations of the posterior conditional probabilities in the MCMC sampling scheme, we make the following transformations. First, we make $\mathbf{X}_i$ orthogonal to $\mathbf{Z}$ by the transformation

$$\mathbf{X}_i \longrightarrow \mathbf{X}_i - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}_i.$$

If $\mathbf{X}_i$ is a two-way interaction component, we also make $\mathbf{X}_i$ orthogonal to its corresponding main effects components by a similar transformation. Second, we transform $\mathbf{X}_i$ so that $\mathbf{X}_i'\mathbf{X}_i$ becomes an orthonormal matrix. This is done by finding the eigenvalue decomposition of $\mathbf{X}_i'\mathbf{X}_i$. Suppose $\mathbf{X}_i'\mathbf{X}_i = \mathbf{Q}_i\mathbf{D}_i\mathbf{Q}_i'$ such that $\mathbf{Q}_i'\mathbf{Q}_i = \mathbf{I}$ and $\mathbf{D}_i$ is diagonal. We transform $\mathbf{X}_i$ by

$$\mathbf{X}_i \longrightarrow \mathbf{X}_i\mathbf{Q}_i\mathbf{D}_i^{-\frac{1}{2}}.$$

The prior for $\boldsymbol{\beta}_i$ is changed to $N(0, c_i\mathbf{D}_i)$ accordingly.

2.3 MCMC SAMPLING SCHEME

We use a MCMC sampling scheme to explore the posterior distribution of the unknown parameters and, in particular, to estimate functionals of these parameters, such as the class probabilities. The sampling scheme is run for a warmup period at the end of which it is assumed that the sampling scheme generates iterates from the posterior distribution. The sampling scheme is then run for a further period which is used for inference and is called the sampling period.

Sampling Scheme 1

1. Choose initial values $\mathbf{y}^{[0]}, \gamma^{[0]}, \boldsymbol{\beta}_1^{[0]}, \boldsymbol{\beta}_2^{[0]}, {c_1}^{[0]}, {c_2}^{[0]}$. These may be generated from some distribution or fixed.
2. Generate $\mathbf{y}$ from $p(\mathbf{y}|\mathbf{w}, \gamma, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, c_1, c_2)$, which is a constrained normal distribution with mean $\mathbf{Z}\gamma + \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2$ and variance $\mathbf{I}_n$. If $w_i = 1$, then y_i is constrained to be positive and if $w_i = 2$, then y_i is constrained to be negative.
3. Generate c_1 and $\boldsymbol{\beta}_1$ as a block from $p(c_1, \boldsymbol{\beta}_1|\mathbf{y}, c_2, \boldsymbol{\beta}_2)$ by first generating c_1 from $p(c_1|\mathbf{y}, c_2, \boldsymbol{\beta}_2)$ and then generating $\boldsymbol{\beta}_1$ from $p(\boldsymbol{\beta}_1|\mathbf{y}, c_2, \boldsymbol{\beta}_2, c_1)$ conditional on the value generated for c_1.
   (a) To generate c_1, we note that by construction $\mathbf{X}_1$ and $\mathbf{Z}$ are orthogonal and $\mathbf{X}_1'\mathbf{X}_1 = \mathbf{I}$. From this we deduce that

$$
\begin{aligned}
p(c_1|\mathbf{y}, c_2, \boldsymbol{\beta}_2) &\propto p(\mathbf{y}|c_1, c_2, \boldsymbol{\beta}_2)p(c_1) \\
&\propto |I + c_1\mathbf{D}_1|^{-\frac{1}{2}} \\
&\quad \times \exp\left(\frac{1}{2}(\mathbf{y} - \mathbf{X}_2\boldsymbol{\beta}_2)'\mathbf{X}_1\left(I + \frac{1}{c_1}\mathbf{D}_1^{-1}\right)^{-1}\right. \\
&\quad \left. \times\, \mathbf{X}_1'(\mathbf{y} - \mathbf{X}_2\boldsymbol{\beta}_2)\right) p(c_1). \qquad (2.5)
\end{aligned}
$$

The log density $\log\{p(c_1|\mathbf{y}, c_2, \boldsymbol{\beta}_2)\}$ is evaluated on a grid of points that are equidistant on the log scale and in the range 10^{-15} to 10^{15}. The computation is fast because

$$
\log\left(p(c_1|\mathbf{y}, c_2, \boldsymbol{\beta}_2)\right) = -\frac{1}{2}\sum_{i=1}^{K_1}(1 + c_1 d_{1_i}) + \frac{1}{2}\sum_{i=1}^{K_1}\frac{v_i^2}{1 + \frac{1}{c_1 d_{1_i}}} + \log p(c_1)\ ,
$$

where $\mathbf{D}_1 = \text{diag}(d_{1_1}, \ldots, d_{1_{K_1}})$ and $\mathbf{v} = (v_1, \ldots, v_{K1})' = \mathbf{X}_1'(\mathbf{y} - \mathbf{X}_2\boldsymbol{\beta}_2)$. The vector $\mathbf{v}$ is calculated once and the remaining calculations involve only $O(K_1)$ operations. The log density $\log\{p(c_1|\mathbf{y}, c_2, \boldsymbol{\beta}_2)\}$ is approximated on both sides of its maximum by piecewise linear functions and the approximation is normalized to be a density function. The parameter c_1 is generated from this approximation. Because an approximation is used to generate c_1, the whole step of generating c_1 and β_1 is modified by the Metropolis–Hastings method to ensure that the sampling scheme converges to the correct posterior distribution. In our examples, the acceptance rate of the Metropolis–Hastings step is very high because the approximation is made very accurate. For example, the acceptance rate is over 99% when 100 grid points are used.

(b) Once c_1 is generated, the coefficient vector $\boldsymbol{\beta}_1$ is generated from a normal distribution with mean $(I + \frac{1}{c_1}\mathbf{D}_1^{-1})^{-1}\mathbf{X}_1'(\mathbf{y} - \mathbf{X}_2\boldsymbol{\beta}_2)$ and variance $(I + \frac{1}{c_1}\mathbf{D}_1^{-1})^{-1}$.

4. Generate c_2 and $\boldsymbol{\beta}_2$ as a block from $p(c_2, \boldsymbol{\beta}_2|\mathbf{y}, c_1, \boldsymbol{\beta}_1)$ similarly to the way c_1 and $\boldsymbol{\beta}_1$ are generated.
5. Generate γ from $p(\gamma|\mathbf{y})$, which is a normal density with mean $(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$ and variance $(\mathbf{Z}'\mathbf{Z})^{-1}$. This conditional density does not depend on the values of c_1, c_2, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ because $\mathbf{Z}$ is orthogonal to $\mathbf{X}_1$ and $\mathbf{X}_2$.

The output of the sampling scheme can be used to estimate the unknown parameters and class probabilities as follows. The posterior means of $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and γ can be expressed as the integral

$$
\text{E}(\gamma, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2|\mathbf{w}) = \int \text{E}(\gamma, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2|\mathbf{w}, \mathbf{y}, c_1, c_2)p(\mathbf{y}, c_1, c_2|\mathbf{w})dc_1 dc_2 d\mathbf{y}. \qquad (2.6)
$$

Suppose the sequence ${c_1}^{[j]}, {c_2}^{[j]}, {\boldsymbol{\beta}_1}^{[j]}, {\boldsymbol{\beta}_2}^{[j]}, \mathbf{y}^{[j]}, j = 1, \ldots, M$, is generated in the sampling period from sampling scheme 1. Then, from (2.6),

$$\hat{\boldsymbol{\beta}}_1 = \frac{1}{M}\sum_{j=1}^{M} \mathrm{E}(\boldsymbol{\beta}_1|\mathbf{w}, \mathbf{y}^{[j]}, {c_1}^{[j]}, {c_2}^{[j-1]}, {\boldsymbol{\beta}_2}^{[j-1]}),$$

$$\hat{\boldsymbol{\beta}}_2 = \frac{1}{M}\sum_{j=1}^{M} \mathrm{E}(\boldsymbol{\beta}_2|\mathbf{w}, \mathbf{y}^{[j]}, {c_2}^{[j]}, {c_1}^{[j]}, {\boldsymbol{\beta}_1}^{[j]}),$$

$$\hat{\gamma} = \frac{1}{M}\sum_{j=1}^{M} \mathrm{E}(\gamma|\mathbf{w}, \mathbf{y}^{[j]}, {c_1}^{[j]}, {c_2}^{[j]}, {\boldsymbol{\beta}_1}^{[j]}, {\boldsymbol{\beta}_2}^{[j]}),$$

are estimates of the posterior means of $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and γ. The regression function f is estimated by

$$\hat{f}(\mathbf{x}_1, \ldots, \mathbf{x}_p) = \widehat{\gamma} + \sum_{k=1}^{K_1} \widehat{\beta}_{1k}\phi_{1k}(\mathbf{x}_1, \ldots, \mathbf{x}_p) + \sum_{k=1}^{K_2} \widehat{\beta}_{2k}\phi_{2k}(\mathbf{x}_1, \ldots, \mathbf{x}_p)$$

and $\Pr(\mathbf{w} = 1|\mathbf{x}_1, \ldots, \mathbf{x}_p)$ is estimated by $\Phi\left(\hat{f}(x)\right)$. An unbiased, but more expensive, estimate of the posterior mean of $\Pr(w = 1|x_1, \ldots, x_p)$ is

$$\widehat{\Pr}(w = 1|x_1, \ldots, x_p) = \frac{1}{M}\sum_{j=1}^{M} \Phi(f^{[j]}(x)),$$

where

$$f^{[j]}(x) = \gamma^{[j]} + \sum_{k=1}^{K_1} \beta_{1k}^{[j]}\phi_{1k}(\mathbf{x}_1, \ldots, \mathbf{x}_p) + \sum_{k=1}^{K_2} \beta_{2k}^{[j]}\phi_{2k}(\mathbf{x}_1, \ldots, \mathbf{x}_p)\,.$$

3. MODEL SELECTION AND MODEL AVERAGING

3.1 INTRODUCTION

It is of interest to know which main effects and two-way interactions enter the model. That is, which components enter the model. It is also of interest to know which variables enter the model. Model parsimony is important in order to improve our understanding of which variables and interactions affect the outcome, and it may also lead to improved prediction of class probabilities and classification into one of several alternatives.

To allow components of the regression to be in or out of the model, we introduce the binary indicator variables J_i such that the regression vector $\boldsymbol{\beta}_i$ (and hence c_i) is identically 0 if $J_i = 0$ and $\boldsymbol{\beta}_i \sim N(0, c_i\mathbf{I})$ if $J_i = 1$. We assume that the J_i are independent of each other apriori, with $\Pr(J_i = 1|\pi_{\text{main}}, \pi_{\text{inter}}) = \pi_{\text{main}}$ if J_i corresponds to a main effect, and $\Pr(J_i = 1|\pi_{\text{main}}, \pi_{\text{inter}}) = \pi_{\text{inter}}$ if J_i corresponds to a second order interaction effect. That

is, π_{main} is the prior probability that a main effect is in and π_{inter} is the prior probability that a bivariate interaction is in. Because the indicators are independent,

$$\Pr(J_1, J_2, \ldots | \pi_{\text{main}}, \pi_{\text{inter}}) = \prod_{i \in \text{main}} \Pr(J_i | \pi_{\text{main}}) \prod_{i \in \text{inter}} \Pr(J_i | \pi_{\text{inter}}).$$

We assume that the probabilities π_{main} and π_{inter} are independent and uniformly distributed on $[0, 1]$. The prior of the indicator variables J_i is uniform in the number of included components. To understand the implication of the assumption on the J_i, suppose there are m_{main} main effect components. Let q_{main} be the number of J_i which corresponds to a main effect and takes the value 1, that is, there are q_{main} effective components in the regression model. Then it is straightforward to show that, with π_{main} integrated out, q_{main} has a discrete uniform distribution on $[0, m_{\text{main}}]$, i.e., $\Pr(q_{\text{main}} = k) = 1/(m_{\text{main}} + 1)$ for $k = 0, \ldots, m_{\text{main}}$. A similar interpretation applies to interaction effects components.

We define a model by the main effects and interactions that it includes. Thus, when there are five variables, and all possible models with main effects and two-way interactions are allowed, then there are 15 components and 2^{15} possible models. We say that a variable is included in a model if either its main effect or any one of its second-order interactions with other variables is included in the model. Otherwise, the variable is excluded from the model. We note that when the class of models we consider contains just main effects, then component selection is the same as variable selection.

To choose a model by component selection we estimate the posterior probabilities $\Pr(J_i = 1|w)$ and use a rule to determine which components to retain. For example, we could retain those components that have a posterior probability greater than some cutoff κ_c. We often take $\kappa_c = 0.5$, but it is useful to consider other values such as 0.3 and 0.1.

To choose a model by variable selection we proceed similarly. Let $V_j = 1$ if the jth variable is in the model and let $V_j = 0$ if it is not, and use the posterior probability $\Pr(V_j = 1|w)$ to determine if the jth variable is retained. A cutoff value κ_v for these probabilities is necessary, and a natural choice is $\kappa_v = 0.5$, but it is again useful to consider other values. In choosing a model by variable selection, if a variable is retained then its main effect is retained, and a decision needs to be made on which of its interactions to also include in the model.

An alternative way of choosing a model is to calculate the posterior probability of each model that is under consideration and select that model(s) having the highest posterior probability. Let $\{M_l, l = 1, 2, \ldots\}$ be a set of binary variables such that $M_l = 1$ if the lth model is selected. Then $\Pr(M_l = 1|w)$ is the posterior probability of the lth model.

Once a model M is selected by one of the above approaches the probability that a new observation w is equal to 1, given the explanatory variables $x_1, \ldots, x_p$, is $\Pr(w = 1|x_1, \ldots, x_p, M)$. An alternative to using a single model to predict the probability of a future observation is to take a weighted average of the predictions made by several models. The Bayesian approach to do such forecast averaging is called model averaging where the weights are the model posterior probabilities; see Raftery, Madigan, and Hoeting (1997). In model averaging we can consider all possible models, or a subset of models, and the probability that $w = 1$ for a new observation is calculated as a weighted sum of the probabilities for each

model, with the weights being the posterior model probabilities, that is,

$$\Pr(w = 1|x_1, \ldots, x_p) = \sum_l \Pr(w = 1|x_1, \ldots, x_p, M_l) \Pr(M_l = 1|w).$$

3.2 DATA-BASED PRIORS

It is well known in ordinary variable selection that placing diffuse priors on individual regression coefficients results in the simplest model always being selected. That is, any variable that is allowed to be in or out is always selected to be out. This is known as Lindley's paradox. Our case is somewhat more complex because we are doing variable selection on variance components. However, the same problem occurs in this context as pointed out in Shively, Kohn, and Wood (1999). Using the prior for the c_i given in Section 2, that is $c_i \sim U(0, 10^{15})$, results in poor variable selection with most variables omitted. It is therefore necessary to use more informative priors for the c_i. We do so now by refining the data-based approach to setting the prior proposed by Shively, Kohn, and Wood (1999) who used smoothing splines. To obtain the data-based prior for the c_i, we first run Sampling Scheme 1 as in section 2 (using a at prior for the c_i) and collect the iterates $c_i^{[j]}, j = 1, \ldots, M$, generated in the sampling period. We use a lognormal distribution for the data-based prior $p_{\mathrm{DB}}(c_i)$ for c_i, such that

$$p_{\mathrm{DB}}(c_i) = \frac{1}{c_i\sqrt{2\pi n\sigma_i^2}} \exp\left(-\frac{1}{2n\sigma_i^2}(\log(c_i) - \mu_i)^2\right),$$

with μ_i the median of the $\log(c_i^{[j]})$ and σ_i^2 the sample variance of the $\log(c_i^{[j]})$. The data-based prior in ates the posterior variance of the $\log(c_i^{[j]})$ by the sample size n. The justification for using this data-based prior was given by Shively, Kohn, and Wood (1999) who related it to BIC. We take μ_i as the median of the $\log(c_i^{[j]})$, whereas Shively, Kohn, and Wood (1999) used the mean. In our work we found the median to be more satisfactory than the mean because the median of $p_{\mathrm{DB}}(c_i)$ is then $\exp(\mu_i)$.

3.3 SAMPLING SCHEMES

As before, the posterior distributions of all the parameters of interest are obtained by a Markov chain Monte Carlo sampling scheme, which is a modification of Sampling Scheme 1. With component selection, for the ith functional component we associate the indicator variable J_i to the smoothing parameter c_i and the regression coefficients $\boldsymbol{\beta}_i$. The next paragraph shows how to generate J_i, c_i, and $\boldsymbol{\beta}_i$ as a block conditional on the parameters of the other components, $J_{j\neq i}, c_{j\neq i}$ and $\boldsymbol{\beta}_{j\neq i}$, and the augmented vector $\mathbf{y}$. The augmented vector $\mathbf{y}$ and regression coefficient γ are generated as in Sampling Scheme 1.

The indicator J_i is generated first, with c_i and $\boldsymbol{\beta}_i$ integrated out, from its conditional distribution

$$\begin{aligned}\Pr(J_i|\mathbf{y}, J_{j\neq i}, c_{j\neq i}, \boldsymbol{\beta}_{j\neq i}) &\propto p(\mathbf{y}|J_i, J_{j\neq i}, c_{j\neq i}, \boldsymbol{\beta}_{j\neq i}) \Pr(J_i|J_{j\neq i}) \\ &\propto \Pr(J_i|J_{j\neq i}) \int p(\mathbf{y}|J_i, c_i, \boldsymbol{\beta}_i, J_{j\neq i}, c_{j\neq i}, \boldsymbol{\beta}_{j\neq i}) p(c_i|J_i) \\ &\quad \times p(\boldsymbol{\beta}_i|c_i, J_i) dc_i d\boldsymbol{\beta}_i .\end{aligned}$$

The integral is evaluated by first integrating out $\boldsymbol{\beta}_i$ analytically as in (2.5), but it is necessary to numerically integrate out c_i by using the same approximation to $p(c_i|\mathbf{y}, c_{j\neq i}, \boldsymbol{\beta}_{j\neq i})$ as in part (3a) of Sampling Scheme 1. The value of $\Pr(J_i|J_{j\neq i})$ can be evaluated analytically with π_{main} and π_{inter} integrated out. Suppose there are m_{main} main effect components and m_{inter} interaction effect components. Let $q_{\text{main}} = \#\{J_j = 1; j \neq i, j \in \text{main}\}$ be the number of indicator variables $J_{j\neq i}$ corresponding to main effect components and having value 1. Analogously let $q_{\text{inter}} = \#\{J_j = 1; j \neq i, j \in \text{inter}\}$ be the number of indicator variables $J_{j\neq i}$ corresponding to interaction effect components and having value 1. It is straightforward to check that

$$\Pr(J_i = 1|J_{j\neq i}) = (q_{\text{main}} + 1)/(m_{\text{main}} + 1) ,$$

if J_i corresponds to a main effect, and

$$\Pr(J_i = 1|J_{j\neq i}) = (q_{\text{inter}} + 1)/(m_{\text{inter}} + 1) ,$$

if J_i corresponds to an interaction effect. If J_i is generated as 1 then c_i and $\boldsymbol{\beta}_i$ are generated as in Sampling Scheme 1. If J_i is generated as 0 then c_i and $\boldsymbol{\beta}_i$ are set to 0. The block generation of J_i, c_i and $\boldsymbol{\beta}_i$ is adjusted by a Metropolis-Hastings step to correct for approximating the conditional densities of J_i and c_i.

The sampling scheme we use is summarized as follows. For simplicity, we demonstrate with only two functional components.

Sampling Scheme 2

1. Choose initial values $\gamma^{[0]}, \boldsymbol{\beta}_1^{[0]}, \boldsymbol{\beta}_2^{[0]}, c_1{}^{[0]}, c_2{}^{[0]}, J_1{}^{[0]}, J_2{}^{[0]}$. These may be generated from some distribution or fixed.
2. Generate $\mathbf{y}$ from $p(\mathbf{y}|\mathbf{w}, \gamma, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, c_1, c_2, J_1, J_2)$ as in Step 2 of Sampling Scheme 1.
3. Generate J_1, c_1 and $\boldsymbol{\beta}_1$ as a block from $p(J_1, c_1, \boldsymbol{\beta}_1|\mathbf{y}, J_2, c_2, \boldsymbol{\beta}_2)$. The indicator J_1 is generated first from the $\Pr(J_1|\mathbf{y}, c_2, \boldsymbol{\beta}_2)$, with c_1 and $\boldsymbol{\beta}_1$ integrated out. Then c_1 and $\boldsymbol{\beta}_1$ are generated as in Sampling Scheme 1 conditional on the generated value of J_1. A Metropolis–Hastings step is necessary because J_1 and c_1 are generated from approximations.
4. Generate J_2, c_2, and $\boldsymbol{\beta}_2$ as a block from $p(J_2, c_2, \boldsymbol{\beta}_2|\mathbf{y}, J_1, c_1, \boldsymbol{\beta}_1)$ in a similar manner as the generation of J_1, c_1, and $\boldsymbol{\beta}_1$.
5. Generate γ from $p(\gamma|\mathbf{y})$ as in Step 5 of Sampling Scheme 1.

3.4 ESTIMATION

Sampling Scheme 2 can be used to carry out model selection and model averaging. Let $J_i^{[k]}, V_j^{[k]}$, and $M_l^{[k]}, k = 1, \ldots, K$, be the kth iterates of J_i, V_j, and M_l during the sampling period. Then the posterior probabilities $\Pr(J_i = 1|w), \Pr(V_j = 1|w)$, and $\Pr(M_l = 1|w)$, are estimated by

$$\widehat{\Pr}(J_i = 1|w) \quad = \quad \frac{1}{K}\sum_{k=1}^{K} J_i^{[k]},$$

$$\widehat{\Pr}(V_j = 1|w) = \frac{1}{K}\sum_{k=1}^{K} V_j^{[k]}, \quad \text{and}$$

$$\widehat{\Pr}(M_l = 1|w) = \frac{1}{K}\sum_{k=1}^{K} M_l^{[k]}.$$

Once a model is determined, parameters and probabilities are estimated as in Section 2.

The model averaged estimate of the regression function $f(x)$ is

$$E\{f(x)|w\} = \frac{1}{K}\sum_{k=1}^{K} E\left\{f(x)|w, y^{[k]}, \mathbf{c}^{[k]}, \boldsymbol{\beta}^{[k]}, \mathbf{J}^{[k]}\right\},$$

where $y^{[k]}$, $\mathbf{c}^{[k]}$, $\boldsymbol{\beta}^{[k]}$, and $\mathbf{J}^{[k]}$ are the vector of latent y's, the vector of c's, the vector of β's and the vector of J's, all at the kth iterate. Similarly, the model averaging estimate of the probability $\Pr(w = 1|x_1, \ldots, x_p)$ is

$$\widehat{\Pr}(w = 1|x_1, \ldots, x_p) = \frac{1}{K}\sum_{k=1}^{K} \Phi\left\{f^{[k]}(x)\right\},$$

where $f^{[k]}(x)$ is the value of the regression function at the kth iteration.

3.5 PRACTICAL IMPLEMENTATION

Because the number of components in the regression function grows quadratically with the number of variables, it is important to know how reliable and efficient the sampling schemes are for a given number of components and a given sample size. To gain some understanding of these two issues, we experimented with a large number of simulated datasets. In each case, we assessed the efficiency of the simulation by examining the autocorrelation of the iterates of elements of the latent vector y and the regression function $f(x)$ at a number of abscissae. We assessed the reliability of the simulation by comparing estimates of the regression function $f(x)$ from independent runs at a number of abcissae. We define a sampling scheme to be reliable if the estimates from independent runs of the sampling scheme are consistent and efficient if the autocorrelations of the latent vector y and the regression function $f(x)$ at a number of abscissae between successive draws die out after 50 iterates. In our experiments we found that for a given sample size, Sampling Scheme 1 performs reliably and efficiently for models not exceeding a certain complexity. For example, for 500 observations, Sampling Scheme 1 can handle up to 30 components, but its performance degrades rapidly after that. In general, the sampling scheme can handle more components when the number of observations increases. Given data-based priors, Sampling Scheme 2 can in general handle regression models with more components than Sampling Scheme 1 since the number of active components at any one iteration, that is, those components with indicator variables $J_i = 1$, is usually much smaller than the total number of components in the model.

The conclusion from our experiments is that for models with a large number of components, the estimation method must be carefully designed to ensure that the sampling

schemes are reliable in the sense described above. This applies in particular to Sampling Scheme 1 because it is used to construct the data-based priors.

In practice, we adopt the following estimation strategies. First, Sampling Scheme 1 is applied to the full regression model, including interactions. The simulation output is examined to see if it is reliable using the criterion outlined above. If the simulation is determined to be reliable, then data-based priors for smoothing parameters c_i are obtained using the simulation output and Sampling Scheme 2 is then run using these data-based priors. The output from Sampling Scheme 2 is then used to select the model using component or variable selection, or alternatively to obtain the model-averaged estimates of the regression function. In general this strategy works well for main effects models with up to 30 variables and interaction models with up to 7 variables for binary data with 500 observations. The analyses in Section 5.1 and part of 5.2 used this strategy.

For larger regression models, the output from Sampling Scheme 1 may not be reliable when the complete model is estimated. In this case, we then estimate a main effects model using Sampling Scheme 1 and determine from its output if the sampling scheme is reliable. Based on our experience on both real and simulated datasets, the output from a main effects model is highly likely to be reliable because in all our work this turned out to be the case. Sampling Scheme 2 is then run on the main effects model using the output from Sampling Scheme 1. We now discuss the three strategies that we employed in our experiments. The first is to stay with a main effects model. In that case the output from Sampling Scheme 2 can be used to obtain a model-averaged model, or we can use variable selection to determine a more parsimonious model.

The second strategy is to add interaction effects in stages. In each stage, 10 to 15 second-order interactions are added to the current regression model. Sampling Scheme 1 is used to obtain data-based priors for all the included components and then Sampling Scheme 2 is run to select those components whose posterior probabilities exceed the cutoff. The reliability of both sampling schemes is examined, and if either one is found to be unreliable then we restart with fewer interactions terms added to the model determined at the last stage. At the end of each stage, the current working model is re-estimated to include the significant interaction effects, that is, those interaction effects whose posterior probabilities exceed the cutoff. This staged process is continued until all interaction effects have had a chance to be included in the model. The order in which interaction effects are added is based on the posterior probabilities of the main effects components in the main effects only model, with interaction effects for the highest posterior main effects going in first. The final model estimates are those obtained from the last stage. We found that this strategy works well in practice, and in most cases only a few interaction effects are added at each stage to the current model. This keeps the size of the regression model at each stage sufficiently small to produce reliable estimates.

The third strategy is a more parsimonious version of the second strategy. Here we include only interactions for those variables whose posterior probabilities exceeded a threshold value in the main effects only estimation. If there are too many such second-order interactions, then the second-order effects can be added in stages as in the second strategy. The rationale for the third strategy is that if a main effect for a variable is not included then

in most cases it is less likely that the second-order interactions for that variable will be included. The third strategy is particularly useful for very large regression models, where it is likely to save a lot of time compared to the second strategy. In Section 5.3, we used strategy 2 for the binary datasets and strategy 3 for the multinomial datasets.

For all the examples in the article, we assessed convergence of the sampling schemes by using multiple runs with different random number seeds and different starting values. In all cases, we used sufficiently large numbers of iterations for both the warmup and sampling periods so that very similar results were obtained from independent runs. This leads us to believe that with our choice of warmup period, convergence was attained before the start of the sampling period. We found that with the complex models that we were estimating, using multiple runs was a more reliable indicator of convergence than convergence diagnostics based on a single run.

Because the applications in the article are highly intensive computationally, it was important that our algorithms take advantage of any mathematical structure in our problems. This meant that using a Bayesian package such as BUGS (Gilks, Thomas, and Spiegelhalter 1994) would make the programs too slow to be practical, because it does not take account of this special structure. All the MCMC sampling schemes described in this article were implemented in Matlab and run on an SGI Unix machine.

4. MULTINOMIAL REGRESSION MODEL

This section extends the binary regression model to the multinomial case. Suppose the response variable $\mathbf{w} = (w_1, \ldots, w_n)'$ takes values $1, \ldots, C$. We are interested in estimating the conditional class probabilities $\Pr(w = i|\mathbf{x})$ for $i = 1, \ldots, C$. Similar to the binary case, we follow Albert and Chib (1993) and augment $\mathbf{w}$ with C latent variables to convert the multinomial nonparametric problem to a multivariate Gaussian problem with inequality constraints on the latent variables.

Define the C latent variables $y_1, \ldots, y_C$ as

$$y_j = f_j(\mathbf{x}) + e_j, \qquad j = 1, \ldots, C, \tag{4.1}$$

where $f_j, j = 1, \ldots, C$, are C regression functions and the errors $\mathbf{e} = (e_1, \ldots, e_C)' \sim N(0, \Sigma)$. For $t = 1, \ldots, n$, let $\mathbf{y}_{.t} = (y_{1t}, \ldots, y_{Ct})'$ be the vector of latent variables corresponding to the tth discrete observation w_t. The latent vector $\mathbf{y}_{.t}$ is connected to the discrete observation w_t by requiring that if $w_t = j$ then $y_{jt} > y_{it}$ for all $i \neq j$. In the marketing literature this is known as a discrete choice model (McFadden 1973). The model (4.1) is not identified without further restrictions because we only observe the w_t and not the $\mathbf{y}_{.t}$. Thus, if $w_t = j$, then we know that $y_{jt} - y_{it} > 0$ for $i \neq j$, but we do not know the actual latent variables. To achieve identification we assume without loss of generality that the regression function $f_C(\mathbf{x})$ is identically zero and that $\Sigma = I_C$. The second assumption is unnecessary for identification (see Bunch 1991; Keane 1992), but is computationally convenient when we estimate class probabilities. It is also not overly restrictive compared to current statistical practice as argued below. When the error $\mathbf{e}$ is Gaussian, the model (4.1)

is called a multinomial probit model. Although the model (4.1) is relatively unfamiliar to statisticians, McFadden (1973) showed that when the errors $e_j, j = 1, \ldots, C$, are independent and identically distributed with a Gumbel distribution, then (4.1) corresponds to the logistic regression model which is used extensively in statistics. Therefore, our assumption that the covariance matrix $\Sigma = I_C$ is not overly restrictive because it is well known that similar class probabilities are obtained for probit multinomial regression with $\mathbf{e} \sim N(0, I)$ as for logistic regression.

As in Section 2, we decompose the regression functions $f_j(\mathbf{x})$ into main effects and second order interactions and approximate each component by radial basis functions. For simplicity, we again assume that there are only two functional components. Let $\mathbf{y}_j = (y_{j1}, \ldots, y_{jn})'$ and $\mathbf{e}_j = (e_{j1}, \ldots, e_{jn})'$. Writing (4.1) in matrix form we have

$$\mathbf{y}_j = \mathbf{Z}\gamma_j + \mathbf{X}_1\boldsymbol{\beta}_{j1} + \mathbf{X}_2\boldsymbol{\beta}_{j2} + \mathbf{e}_j \,,$$

for $j = 1, \ldots, C$, with $\gamma_j, \boldsymbol{\beta}_{j1}$ and $\boldsymbol{\beta}_{j2}$ identically zero for $j = C$.

It is of interest to note that for the binary case ($C = 2$) the set of Equations (4.1) is equivalent to the single equation

$$y_1 - y_2 = f_1(\mathbf{x}) + e_1 - e_2 \,,$$

because we only know the sign of $y_1 - y_2$ from the data w. That is, the set of equations in (4.1) is equivalent to the single Equation (2.3).

4.1 SAMPLING SCHEME

There are three differences between estimating a multinomial regression with $C > 2$ and binary regression. First, the constraints between the latent vectors $\mathbf{y}_j, j = 1, \ldots, C$, are more complex which means that sampling $\mathbf{y}_j$ is a little more complicated. Second, it is necessary to estimate $C - 1$ regression functions instead of just one regression function. Third, it is necessary to use simulation to estimate the C conditional class probabilities.

We introduce smoothing parameters c_{j1} and c_{j2} and indicator variables J_{j1} and J_{j2} for $j = 1, \ldots, C - 1$, as a straightforward extension of the approach in Sections 2 and 3. We can sample the $C - 1$ sets of parameters independently conditional on the latent vectors $\mathbf{y}_j, j = 1, \ldots, C$, because the errors $\mathbf{e}_j$ in these regression functions are independent by assumption. For each observation, $w_t, t = 1, \ldots, n$, the latent vector variable $\mathbf{y}_{\cdot t}$ is sampled as follows. If $w_t = i$, then y_{it} is generated first. If $i \neq C$, then y_{it} is generated from a normal distribution with mean $\gamma_i + \mathbf{X}_{1t}\boldsymbol{\beta}_{i1} + \mathbf{X}_{2t}\boldsymbol{\beta}_{i2}$ and variance 1 subject to the constraints $y_{it} > y_{jt}$ for $j \neq i$. If $i = C$, then y_{it} is generated from a normal distribution with mean 0 and variance 1 subject to the same constraint. Next, for $j = 1, \ldots, C, j \neq i$, y_{jt} is generated from a normal distribution with mean $\gamma_j + \mathbf{X}_{1t}\boldsymbol{\beta}_{j1} + \mathbf{X}_{2t}\boldsymbol{\beta}_{j2}$ and variance 1, subject to the constraint that $y_{jt} < y_{it}$. In particular, if $j = C$, then y_{jt} is generated from a constrained normal distribution with mean 0 and variance 1.

This sampling scheme is summarized as follows.

Sampling Scheme 3

1. Choose initial values ${\gamma_j}^{[0]}, {\boldsymbol{\beta}_{j1}}^{[0]}, {\boldsymbol{\beta}_{j2}}^{[0]}, {c_{j1}}^{[0]}, {c_{j2}}^{[0]}, {J_{j1}}^{[0]}$, and ${J_{j2}}^{[0]}$, where $j = 1, \ldots, C$ for the $\mathbf{y}_j$ and $j = 1, \ldots, C-1$ for the other parameters.
2. Draw $\mathbf{y}_j, j = 1, \ldots, C$, as described above.
3. For $j = 1, \ldots, C-1$, generate $J_{j1}, c_{j1}, \boldsymbol{\beta}_{j1}, J_{j2}, c_{j2}, \boldsymbol{\beta}_{j2}$ and γ as in Step 3, 4, and 5 in Sampling Scheme 2.

4.2 ESTIMATION

The regression coefficients, the regression functions and the probabilities for the indicator variables are estimated as in Section 3.2. To estimate the conditional class probabilities $\Pr(w_t = i|\mathbf{x}_t)$ we use the following simplified version of the Geweke-Keane-Hajivassiliou (GHK) estimator (Geweke 1991; Keane 1994).

$$\begin{aligned}
\Pr(w_t = i|\mathbf{x}_t) &= \Pr\left(y_{it} > \max_{j \neq i} y_{jt} | \mathbf{x}_t\right) \\
&= \int \Pr(y_{jt} < y_{it}, j \neq i | y_{it}, \mathbf{x}_t)\; p(\mathbf{y}_{j\neq i,t}|y_{it}, \mathbf{x}_t) p(y_{it}|\mathbf{x}_t)\; d\mathbf{y}_{j\neq i,t} dy_{it} \\
&= \int \left(\prod_{j \neq i} \int \Pr(y_{jt} - y_{it} < 0 | y_{it}, f_i(\mathbf{x}_t), f_j(\mathbf{x}_t)) \right. \\
&\qquad \left. \times\, p(y_{jt}|\mathbf{x}_t) dy_{jt} \right) p(y_{it}|\mathbf{x}_t)\; dy_{it} \\
&= \int \left(\prod_{j \neq i} \int \Phi(y_{it} - f_j(\mathbf{x}_t)|\mathbf{y}_t, \mathbf{x}_t) df_j(\mathbf{x}_t) \right) p(y_{it}|\mathbf{x}_t)\; dy_{it}. \qquad (4.2)
\end{aligned}$$

Using the output from Sampling Scheme 3, (4.2) is estimated as

$$\widehat{\Pr}(w_t = i|\mathbf{x}_t) = \frac{1}{M} \sum_{k=1}^{M} \prod_{j \neq i} \Phi\left\{ {y_{it}}^{[k]} - f_j^{[k]}(\mathbf{x}_t) \right\}.$$

For variable or component selection we can again drop the jth component in equation i if $\widehat{\Pr}(J_{ij} = 1|\mathbf{w})$ is small. However, the conditional class probabilities do not depend on the jth component if and only if the jth component does not appear in any of the C equations. This is also the case for multinomial model using logistic regression. Empirically, we say that the jth component does not appear in all of the conditional class probabilities if $\widehat{\Pr}(J_{ij} = 1|\mathbf{w})$ is small for $i = 1, \ldots, C-1$.

5. EXAMPLES

Section 5.1 carries out two simulation experiments to study the frequentist properties of our methodology. Section 5.2 carries out variable and component selection on a real

Table 1. Component Selection Results for Simulation Experiments 1 and 2, Giving the Number of Times the Corresponding Component is Selected Out of 50 Replications

	Cutoff	Regression Function Component									
Experiment	Probability	x_1	x_2	x_3	x_4	(x_1, x_2)	(x_1, x_3)	(x_1, x_4)	(x_2, x_3)	(x_2, x_4)	(x_3, x_4)
1	0.25	50	2	3	50	0	1	0	46	0	1
	0.5	50	0	2	50	0	0	0	44	0	0
	0.75	50	0	0	50	0	0	0	41	0	0
2	0.25	50	50	13	50	0	0	0	0	0	0
	0.5	50	50	3	50	0	0	0	0	0	0
	0.75	50	50	0	50	0	0	0	0	0	0

dataset and compares it to an existing analysis of that data. Section 5.3 studies how well our approach classifies binary and multinomial data compared to modern classification methods.

5.1 SIMULATION EXPERIMENTS

Two simulation experiments were carried out to study the frequentist properties of our Bayesian methodology. Each experiment involved a binary outcome and consisted of 50 replications, using a sample size of 500 for each replication. There are four predictor variables for each experiment, with the predictor variables generated from a uniform distribution on (0,1), and they are independent of each other. In each experiment, the observations are generated from $\Pr(w = 1|x_1, x_2, x_3, x_4) = \Phi\left(f(x_1, x_2, x_3, x_4)\right)$, where we use the functions f_1 and f_2 given by

$$\begin{aligned} f_1(x_1, x_2, x_3, x_4) &= -1.5 + \exp(1.1x_1^3) + 4(x_2 - 0.5)(x_3 - 0.5) + \sin(4\pi x_4)\ , \\ f_2(x_1, x_2, x_3, x_4) &= -0.5 + \exp(1.1x_1^3) - 2x_2 + \sin(4\pi x_4) \end{aligned}$$

for experiments 1 and 2 respectively. Similar components, except for the bivariate term, were used by Shively et al. (1999) in their simulations. In each experiment, we included all main effects and two-way interactions, giving 10 components. Sampling Scheme 1 was run first to obtain data-based priors, and Sampling Scheme 2 was then run for model selection and model averaging. We used 1,000 iterations for both warmup and sampling periods.

We first report the model selection results. Table 1 shows the frequency (out of 50) with which the components were selected when the cutoff used for the posterior probability of component inclusion took on three values, 0.25, 0.5, and 0.75. In addition, for experiment 1, the correct combination of functional components was selected 40, 42, and 41 times out of 50. The corresponding figures for experiment 2 were 37, 47, and 50. The results suggest that for the cases studied in the simulation, component selection works well and that using a cutoff of 0.5 is reasonable.

We next report the performance of three methods to estimate the binary class probabilities. The first method uses Sampling Scheme 1 to estimate the binary class probabilities with all components included. That is, uninformative priors are used and component selection is not carried out. The second method estimates the probabilities using model averaging. The probabilities are obtained by running Sampling Scheme 1 followed by Sampling Scheme

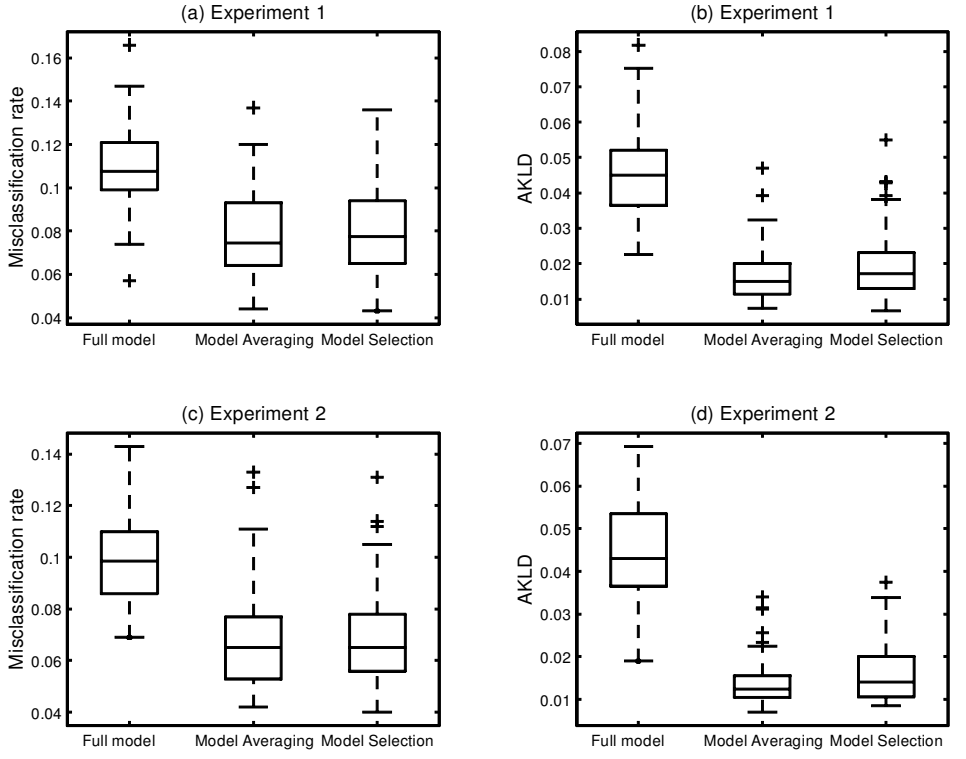

Figure 3. Boxplots of the misclassification rates and AKLD from experiments 1 and 2. In panel (a) to (d), the three boxplots from left to right, correspond to the estimates obtained by unconstrained estimation of the full regression model, model averaging and model selection via component selection.

2. The third method estimates the probabilities by using component selection to select the model. The components are selected using the output of Sampling Scheme 2 with a cutoff of 0.5 for the posterior probabilities. The estimation in method 3 is completed by running Sampling Scheme 1 on the selected components.

The three estimation methods were compared using two measures of performance. The first is the average Kullback–Leibler distance (AKLD), which is defined as

$$\begin{aligned}\text{AKLD} \;=\; & \frac{1}{N}\sum_{i=1}^{N}\Big(\Pr(w_i=1|x_1,x_2,x_3,x_4) \\ & \times \log\{\Pr(w_i=1|x_1,x_2,x_3,x_4)/\widehat{\Pr}(w_i=1|x_1,x_2,x_3,x_4)\} \\ & + \Pr(w_i=0|x_1,x_2,x_3,x_4) \\ & \times \log\{\Pr(w_i=0|x_1,x_2,x_3,x_4)/\widehat{\Pr}(w_i=0|x_1,x_2,x_3,x_4)\}\Big)\,, \quad (5.1)\end{aligned}$$

where $\Pr(.)$ and $\widehat{\Pr}(.)$ are the true and estimated probabilities applied to a test dataset, respectively. The second measure is the misclassification rate. The misclassification rate is the usual measure of performance for binary and multinomial estimators, but it is a rather coarse measure of the quality of the estimator because it does not distinguish between a

probability of 0.6 and a probability of 0.9 when classifying a binary outcome as 1. The AKLD measures more directly the quality of the probability estimators and was used by Shively et al. (1999). As mentioned by Shively et al. (1999), the AKLD is always non-negative and the closer it is to 0, the better the estimator; if the estimated and true probabilities are identical, then AKLD $= 0$.

Both of the measures were calculated using 1,000 new observations that were randomly generated in the same way as the 500 observations that were used for the estimation. Figure 3 presents boxplots of the AKLD and the misclassification rate for experiments 1 and 2. It is clear from the boxplots that the model selection and model averaging estimators outperform the estimator obtained from the unconstrained regression model. The model averaging and model selection estimators perform similarly in terms of misclassification rate, but the model averaging estimator is marginally better than the model selection estimator for the AKLD.

5.2 PIMA INDIAN DIABETES DATASET

The dataset was obtained from the UCI Repository of Machine Learning Databases (Web address http://www.ics.uci.edu/mlearn/MLRepository.html), and analyzed by Smith et al. (1988) and Wahba et al. (1995). The dataset was selected from a larger dataset held by the National Institutes of Diabetes and Digestive and Kidney Diseases with the selection done by several criteria. In particular, all patients represented in this dataset are Pima-Indian women at least 21 years old and living near Phoenix, Arizona, USA. There are eight attribute variables and the response class variable. The attribute variables are :

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)

The class variable y takes on value "0" or "1", where "1" means a positive test for diabetes and "0" is a negative test for diabetes. There are 268 cases in class "1" and 500 cases in class "0". It is reported that there are no missing values, but as pointed out by Wahba et al. (1995), there are cases with zero body mass index and zero plasma glucose concentration. We also found cases with zero diastolic blood pressure, which is impossible. We deleted these aberrant cases, leaving 724 remaining cases, of which 249 are of class "1" sand 475 are of class "0". Wahba et al. (1995) used 752 cases.

Wahba et al. (1995) divided the dataset into two parts: 500 records were selected randomly for the training set and the remaining 252 records were used as the evaluation set. The evaluation set was used to choose between models. Since their penalized log-likelihood smoothing spline approach was very intensive computationally, they did not attempt to estimate all possible models. They used the GLM function in S-plus (Chambers

and Hastie, 1992) to fit several linear models to the data in order to select a few of the most in uential predictors. They identified variables 1, 2, 6, and 7 as significant predictors. Among the models built from these variables, they identified the following four models as candidates for further investigation.

$$\begin{aligned} f(\mathbf{x}) &= \mu + f_2(x_2) + f_6(x_6) \\ f(\mathbf{x}) &= \mu + f_2(x_2) + f_6(x_6) + f_{2,6}(x_2, x_6) \\ f(\mathbf{x}) &= \mu + f_1(x_1) + f_2(x_2) + f_6(x_6) + f_{2,6}(x_2, x_6) \quad \text{and} \\ f(\mathbf{x}) &= \mu + f_1(x_1) + f_2(x_2) + f_6(x_6) + f_7(x_7), \end{aligned}$$

where $f(\mathbf{x})$ is the logit transformation. In these four models, Wahba et al. (1995) treated variable x_1 (the number of times pregnant) as a categorical variable with the four categories corresponding to $x_1 = 0$, $x_1 = \{1, 2\}$, $x_1 = \{3, 4, 5\}$ and $x_1 > 5$. They fitted the models using a penalized log-likelihood smoothing spline approach and used an unbiased estimate of the average Kullback–Leibler distance applied to the evaluation set to measure the goodness of fit of an estimated model. As in (5.1), the average Kullback–Leibler distance on the evaluation set is

$$\text{AKLD} = \frac{1}{252} \sum_{i=1}^{252} \left(\Pr(y_i = 1) \log \left\{ \frac{\Pr(y_i = 1)}{\widehat{\Pr}(y_i = 1)} \right\} + \Pr(y_i = 0) \log \left\{ \frac{\Pr(y_i = 0)}{\widehat{\Pr}(y_i = 0)} \right\} \right) , \tag{5.2}$$

where $\Pr(y_i)$ and $\widehat{\Pr}(y_i)$ are the true and estimated probabilities respectively. Since $\Pr(y_i)$ is unknown, Wahba et al. (1995) used the following unbiased estimator of (5.2),

$$\widehat{\text{AKLD}} = -\frac{1}{252} \sum_{i=1}^{252} \left(y_i \log \left\{ \widehat{\Pr}(y_i = 1) \right\} + (1 - y_i) \log \left\{ \widehat{\Pr}(y_i = 0) \right\} \right) . \tag{5.3}$$

For logistic regression, (5.3) can be further simplified to

$$\widehat{\text{AKLD}} = -\frac{1}{252} \sum_{i=1}^{252} \left(y_i \hat{f}(\mathbf{x}_i) - \log(1 + e^{\hat{f}(\mathbf{x}_i)}) \right) ,$$

where $\hat{f}$ is the estimated function. The Kullback–Leibler distance is always non-negative and the closer $\widehat{\text{AKLD}}$ is to zero the better the fit of the model. The third model was best when using this criterion.

To compare our approach with that of Wahba et al. (1995), we also divided the dataset into two parts and selected 500 cases randomly as the training set. The remaining 224 cases constituted the evaluation set. The number of cases in the evaluation set was reduced from 252 to 224 because we deleted more cases than Wahba et al. (1995). We note, however, that our method does not require an evaluation set for model comparison and selection.

We performed variable selection and nonparametric regression simultaneously under the Bayesian framework, putting all the 8 predictors into the model and including all main effects and the 28 two-way interaction effects. For each of the main effects we used 11 basis terms and we used 22 basis terms for each of the two-way interactions. We kept the number

Table 2. Posterior Probabilities of the Components for the Pima Indian Diabetes Dataset.

Component	*Posterior probability*
x_1	0.10
x_2	1.00
x_3	0.07
x_4	0.04
x_5	0.06
x_6	0.99
x_7	0.07
x_8	0.95

of basis terms relatively small because with such a relatively small amount of data we do not expect to detect any fine structure in the underlying model. For this dataset, simultaneously estimating all main effects and interaction effects significantly downgraded the reliability and efficiency of the sampling schemes. Hence, the second estimation strategy stated in Section 3.5 was used and the estimation proceeded as follows. A main effects model was first estimated, with all 8 variables included. Sampling Scheme 1 was run and followed by Sampling Scheme 2. The posterior probabilities of inclusion of the main effects were ranked. Then the 28 interaction effects were added to the regression model in two stages as described in Section 3.5. The order in which the interaction effects were added was based on the posterior probabilities of inclusion of the variables in the main effects model, with the interaction effects for the main effects with highest posterior probabilities entering first. In each of the two stages, component selection was performed on the interaction effects with a cutoff probability of 0.1. Components of the interaction effects were retained in the regression model whenever their posterior probabilities of inclusion were larger than the cutoff probability. After the two stages, a final run of Sampling Scheme 1 followed by Sampling Scheme 2 was performed with all main effects and the selected interaction effects in the regression model. Model selection and model averaging estimates from this run were used for the final estimation. In all the MCMC runs we used 2,000 iterations for both the warmup and sampling periods. Table 2 shows the component selection results and Table 3 shows the posterior probabilities of the individual models. From Table 2, the main effects for variables 2, 6, and 8 were chosen more than 95% of time, while the main effects for other variables were chosen 4% to 10% of time. No interaction terms were in the final model because none were chosen in the earlier stages of the estimation. Figure 4 displays

Table 3. Posterior Probabilities of the Models for the Pima Indian Diabetes Dataset. Only models with posterior probabilities greater than 0.01 are listed.

Model	*Posterior probability*
x_2,x_6,x_8	0.698
x_2,x_3,x_6,x_8	0.047
x_2,x_6,x_7,x_8	0.046
x_2,x_5,x_6,x_8	0.040
x_1,x_2,x_6,x_8	0.038
x_1,x_2,x_6	0.036
x_2,x_4,x_6,x_8	0.029
x_2,x_8	0.012

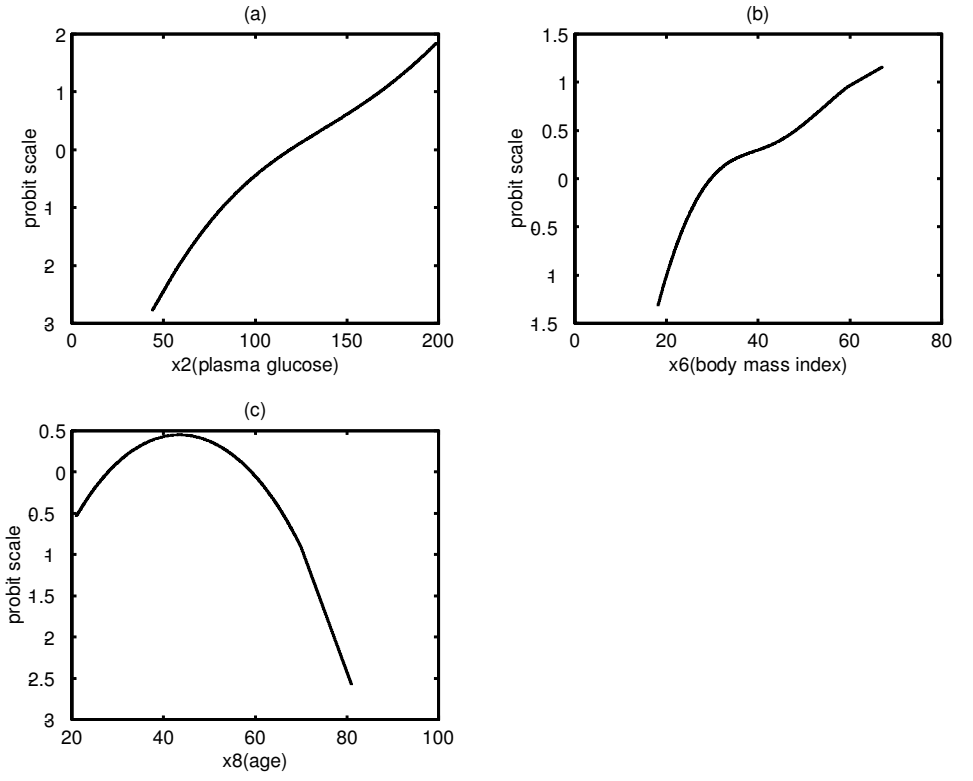

Figure 4. Estimates for the Pima Indian diabetes dataset. Panel (a), (b), and (c) are main effects for variables 2, 6, and 8, respectively. All plots are in the probit scale.

the estimates of the regression function for main effects of variable 2, 6, and 8, with the estimates of the other main effects and interaction effects being negligible. We note that the same results were obtained with a cutoff probability of 0.5.

The main difference between our results and those in Wahba et al. (1995) is that we chose variable 8 along with variables 2 and 6, while they chose variable 1 along with variables 2 and 6. We argue that the difference is due to the different methods used in variable selection. We performed variable selection on a nonparametric regression model, while Wahba et al. (1995) performed variable selection in a linear logistic regression model in the first stage of their analysis. The first issue to be resolved is whether variable 8 should be included in the regression model and, if it should be included, whether it should enter linearly or nonlinearly. The regression function estimate in Figure 4(c) looks like a quadratic function with a maximum around age 45, suggesting that a linear function will not estimate this function well. To investigate this, we repeated the analysis done by Wahba et al. (1995) on our dataset. We fitted a linear model for all the 8 variables using the GLM function in S-plus. Variables 1, 2, and 6 were significant, variable 7 was marginally insignificant while variable 8 was not significant. This agreed with the analysis in Wahba et al. (1995). The analysis suggested that variable 8 was excluded as a significant

Table 4. Misclassication Rates and Estimated AKLD for the Pima Indian Diabetes Dataset. The gures for ADAP, PSA, and GLIM are extracted from Wahba et al. (1995). ADAP is the learning algorithm in Smith et al. (1988). PSA is the penalized smoothing spline method in Wahba et al. (1995). GLIM is the generalized linear model obtained by glm function in Splus. BAYES is the Bayesian estimator proposed in this article.

Smoother	*Misclassication rate on evaluation set*	*Estimate AKLD on evaluation set*
BAYES with variables 2 and 6	24.1%	0.4903
BAYES with variables 1, 2, and 6	23.2%	0.4718
BAYES with variables 2, 6, and 8	21.8%	0.4611
BAYES with all variables	21.8%	0.4595
ADAP (Smith et al.,1988)	24%	N/A
PSA (Wahba et al.,1995)	26%	N/A
GLIM (Wahba et al.,1995)	28%	N/A

predictor when a linear model was used to identify significant predictors. However, when a nonparametric model was used in the variable selection procedure, variable 8 was identified as a significant predictor variable. We concluded that variable 8 entered the regression model nonparametrically and was selected as a significant predictor variable only when variable selection was performed on the nonparametric regression model. The second issue is whether variable 1 is significant. In Wahba et al. (1995), variable 1 was selected in the absence of variable 8 in their nonparametric estimation process, while variable 8 was eliminated as a significant predictor variable early in their analysis. However, in our analysis, when both variable 1 and 8 were available our nonparametric variable selection estimator preferred variable 8 to variable 1, and only one of them was required to model the response variable. This is supported by Table 3, which shows that the model containing exactly variable 2, 6, and 8 was selected almost 70% of time. Models containing variable 2, 6, and 8 but not variable 1 were selected more than 86% of time. Models containing variables 1, 2, and 6 but not variable 8 were selected less than 5% of time. Models containing both variable 1 and 8 were selected less than 5% of time.

We can also assess the performance of individual models using measures such as the AKLD and the misclassification rate. We estimated the following models individually and compared their estimated AKLD and misclassification rates on the evaluation set.

(a) The model with variables 1, 2, and 6.
(b) The model with variables 2, 6, and 8.
(c) The full model, that is, the model with all 8 variables.
(d) The model with variables 2 and 6.

The aims were to determine whether model (b) was better than model (a), and whether models (a) or (b) were adequate in modeling the response. Model (d) was included as a reference point.

The results are given in Table 4, which shows that model (b) is almost as good as model (c), which is the full variable selection model. Model (a) was worse than model (b), both in terms of the estimated AKLD and the misclassification rate. Model (d) performed worst. Comparisons of models (a), (b), and (d) show that, in addition to variables 2 and 6, variable 8 or variable 1 improved the estimation significantly, and variable 8 gave the most

Table 5. Misclassication Rates for the Pima Indian Diabete Dataset With the Same Setting as in Statlog. Logdisc is Logistic Discrimination. Quadisc is Quadratic Discrimination. NaiveBay is Naive Bayes and *k-NN* is *k*-nearest Neighbor. BAYES is the Bayesian estimator proposed in this article.

Smoother	*Misclassication rate by 12-fold cross-validation*
best of Statlog (Logdisc)	22.3%
median of Statlog (NaiveBay and Quadisc)	26.2%
worst of Statlog (*k*-NN)	32.4%
BAYES with main effect only	22.0%
BAYES with interactions	22.2%

improvement. A comparison of models (b) and (c) also shows that the model with variables 2, 6, and 8 adequately models the responses. The results agree with the Bayesian model selection estimation, which selected variables 2, 6, and 8.

In addition to the misclassification rates of the estimated models (a) to (d), Table 4 also shows the misclassification rates obtained by Wahba et al. (1995) and Smith et al. (1988). The misclassification rate for the full Bayesian variable selection model is the lowest, but these misclassification rates must be interpreted cautiously because we used a different subset of the data to the other authors in determining the misclassification rates.

5.3 SOME CLASSIFICATION DATASETS

To compare the predictive performance of the Bayesian nonparametric estimator with other classification methods and nonparametric regression methods, we analyzed several datasets used previously. For each dataset, we tried two regression models—a model with main effects only and a model with main effects and two-way interaction effects. In all cases, the estimates used to obtain the final results are the model averaging estimates for the regression models and a cutoff probability of 0.5 was used in component selection.

The first example is the Pima Indian diabetes dataset which was analyzed in Section 5.2. The Statlog project (see Michie, Spiegelhalter, and Taylor 1994) performed a comparative study of more than 20 classification methods applied to this dataset, including classical and modern statistical methods, machine learning approaches and neural networks. There are 768 observations, 8 predictor variables, and 2 classes in the dataset. A 12-fold cross-validation procedure was used to determine the classification rates. The best misclassification rate was 22.3% by the "Logdisc" method. The median misclassification rate was 26.2% by the "NaiveBay" and "Quaddisc" methods. We tested our Bayesian variable selection estimator using the same criteria. We first estimated a main effects model. The misclassification rate for this model was 22.0%. We then estimated a model with two-way interaction effects. For this interaction effects model, we started with the main effects components in the regression model and added the 28 two-way interaction effects components to the regression model in two stages. The estimation procedure was described in Section 3.5. The misclassification rate for this model was 22.2%. Table 5 shows that the Bayesian approach compares favorably with the other approaches.

The second example is the Wisconsin breast cancer dataset which is available at the

Table 6. Misclassication Rates for the Wisconsin Breast Cancer Dataset. LogitBoost, Real AdaBoost, Gentle AdaBoost and Discrete AdaBoost are methods used in Friedman et al. (2000). Results for the boosting methods are extracted from Friedman et al. (2000). BAYES is the Bayesian estimator proposed in this article.

Smoother	*Misclassication rate by ve-fold cross-validation*
LogitBoost(Stumps)	2.9%
Real AdaBoost(Stumps)	4.0%
Gentle AdaBoost(Stumps)	4.1%
Discrete AdaBoost(Stumps)	4.0%
LogitBoost(8 Terminal Nodes)	3.8%
Real AdaBoost(8 Terminal Nodes)	3.4%
Gentle AdaBoost(8 Terminal Nodes)	3.1%
Discrete AdaBoost(8 Terminal Nodes)	3.7%
BAYES with main effect only	3.2%
BAYES with interactions	3.6%

UCI Repository of Machine Learning Databases and was originally studied by Wolberg and Mangasarian (1990). There are 699 cases, 16 of which contained missing values. The dependent variable is binary, with the tumor being benign (458 cases) or malignant (241 cases), and there are 9 explanatory variables. Friedman, Hastie, and Tibshirani (2000) applied their boosting methods to this dataset using a five-fold cross-validation to determine the classification rates. We followed their setting and tested our Bayesian variable selection approach with the dataset. We estimated a main effects model as well as a model with two-way interaction effects. For the interaction effects model, we started with the main effects components in the regression model and subsequently added the 36 interaction effects in five stages. The estimation procedure proceeded as described in Section 3.5. Table 6 shows that our results are comparable to other modern classification methods.

The third example is the Vowel dataset which is also available at the UCI Repository of Machine Learning Databases. There are 990 cases, 10 predictor variables, and 11 classes. The training set consisted of 528 cases and the evaluation set consisted of the remaining 462 cases. Friedman et al. (2000) analyzed this dataset, and we reanalyzed it using our Bayesian approach. Table 7 summarizes the results and shows that our Bayesian approach is comparable with modern classification methods.

In addition, we analyzed the dataset using a 8-fold cross-validation procedure. The misclassification rate was 43.4% for the main effects only model and 46.8% for the model with interactions. The misclassification rates decreased because more observations were used in the training set.

For the interaction effects model, we started with main effects components and subsequently added interaction effects to the regression model in stages. To save estimation time, we considered only interaction effects between significant main effects components. Hence the estimation procedure started with a variable selection procedure on the main effects model. Only interaction effects between variables selected in this main effects model were subsequently added to the regression model in the later stages of the estimation. The estimation was completed when all eligible interaction effects had been added to the regression model.

Table 7. Misclassication Rates for the Vowel Dataset. LogitBoost, Real AdaBoost, Gentle AdaBoost, and Discrete AdaBoost are methods used in Friedman et al. (2000). Results for the boosting methods are extracted from Friedman et al. (2000). BAYES is the Bayesian estimator proposed in this article.

Smoother	*Misclassication rate on testing dataset*
LogitBoost(Stumps)	51.1%
Real AdaBoost(Stumps)	54.8%
Gentle AdaBoost(Stumps)	58.4%
Discrete AdaBoost(Stumps)	56.3%
LogitBoost(8 Terminal Nodes)	51.7%
Real AdaBoost(8 Terminal Nodes)	49.6%
Gentle AdaBoost(8 Terminal Nodes)	49.6%
Discrete AdaBoost(8 Terminal Nodes)	50.0%
BAYES with main effect only	53.7%
BAYES with interactions	50.2%

Overall, the Bayesian classification results are comparable to those obtained by modern classification methods such as boosting. However, they are far easier to interpret than tree-based boosting methods because they are model based rather than being an average of a large number of trees. Moreover, the boosting approach may not perform well when the training data is noisy, that is when the training data contains wrongly classified observations, because the method puts increasing weight on such cases. An additional advantage of our methods is that they produce genuine estimates of conditional class probabilities.

6. EXTENSIONS

6.1 DIFFERENT PENALTY FUNCTIONS

It is straightforward to apply our computational framework to different priors for the regression coefficients, or equivalently to different penalty functions. In particular, it is straightforward to use our methods with roughness penalties corresponding to smoothing splines. It is sufficient to consider the univariate nonparametric regression problem with Gaussian errors, which we write as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}.$$

In our article we take the prior for $\boldsymbol{\beta}$ as $N(0, cI)$, which corresponds to solving the penalized least squares problem

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \frac{1}{c}\boldsymbol{\beta}'\boldsymbol{\beta}. \tag{6.1}$$

This is different from the roughness penalty,

$$\int \left(\hat{f}^{(2)}(x)\right)^2 dx\,,$$

which is used in smoothing splines, where $\hat{f}(x)$ is the regression function. As in Green and Silverman (1994, p. 142), this penalty can be written as $\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}$ for some symmetric positive-semidefinite matrix $\boldsymbol{\Sigma}$, which is equivalent to placing the prior $\beta \sim N(0, \boldsymbol{\Sigma}^-)$ on $\boldsymbol{\beta}$, where $\boldsymbol{\Sigma}^-$ is the generalized inverse of $\boldsymbol{\Sigma}$, and results in the penalized least squares problem

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \frac{1}{c}\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}\,. \tag{6.2}$$

It is straightforward to express the problem (6.2) in the form (6.1) by using the transformations $\tilde{\boldsymbol{\beta}} = \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\beta}$ and $\tilde{\mathbf{X}} = \mathbf{X}(\boldsymbol{\Sigma}^{1/2})^-$ resulting in $\mathbf{X}\boldsymbol{\beta} = \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}}$ and $\mathrm{Cov}(\tilde{\boldsymbol{\beta}}) = c\mathbf{I}$.

This single component example using a roughness penalty prior is readily extended to roughness penalties for additive main effects and two-way interactions. That is, our computational approach can handle such roughness penalties for the full model considered in the article.

6.2 SPATIAL ADAPTATION

In our article, each of the main effects and interactions used one smoothing parameter. This is usually adequate to capture the shape of the regression function, especially in the multinomial case where the observations gives us far less information about the regression function than in the Gaussian case. Some regression functions, however, display a great deal of spatial adaptation and it may be useful to have a more spatially adaptive estimator for such functions. Ruppert and Carroll (2000) considered a spatially adaptive penalty in their penalized regression spline with Gaussian errors. An obvious way to proceed is to extend the approach proposed by Smith and Kohn (1996) and Denison, Mallick, and Smith (1998). It is straightforward to do so using data augmentation, but the results were disappointing when we tried because the estimates displayed too much spurious variability.

We now show how the Bayesian framework proposed in this article can also allow for spatial variability in the regression function. It is sufficient to consider the univariate binary case with the independent variable x in the interval $(0, 1)$. Suppose that $\Pr(w = 1|x) = \Phi(f(x))$, with

$$f(x) = \exp(-0.5((x - 0.15)/0.05)^2)/0.3) + \exp(-0.5((x - 0.6)/0.2)^2)/1.2) - 1.0.$$

The function $f(x)$ is plotted in Figure 5(a) and $\Pr(w = 1|x)$ in Figure 5(c). We generated 400 binary observations from this probability. An estimate of f based on 20 knots and a single smoothing parameter is plotted in Figure 5(b) and the corresponding estimate of the probability is in Figure 5(d). This estimate shows the basic shape of the function and adequately estimates the probability, but displays some spurious local variability because it fails to adapt adequately to the regression function. To obtain a more locally adaptive estimator we divided the interval $(0, 1)$ into the three overlapping subintervals, $(0, 0.4)$, $(0.3, 0.7)$, and $(0.6, 1.0)$. In each subinterval we selected 20 knots as in the single smoothing parameter case and approximated the true regression function $f(x)$ additively by three components, each with its own smoothing parameter. We also estimated the regression function using five components based on the five overlapping subintervals (0,0.2), (0.1,0.3), (0.2,0.4), (0.3,0.7),

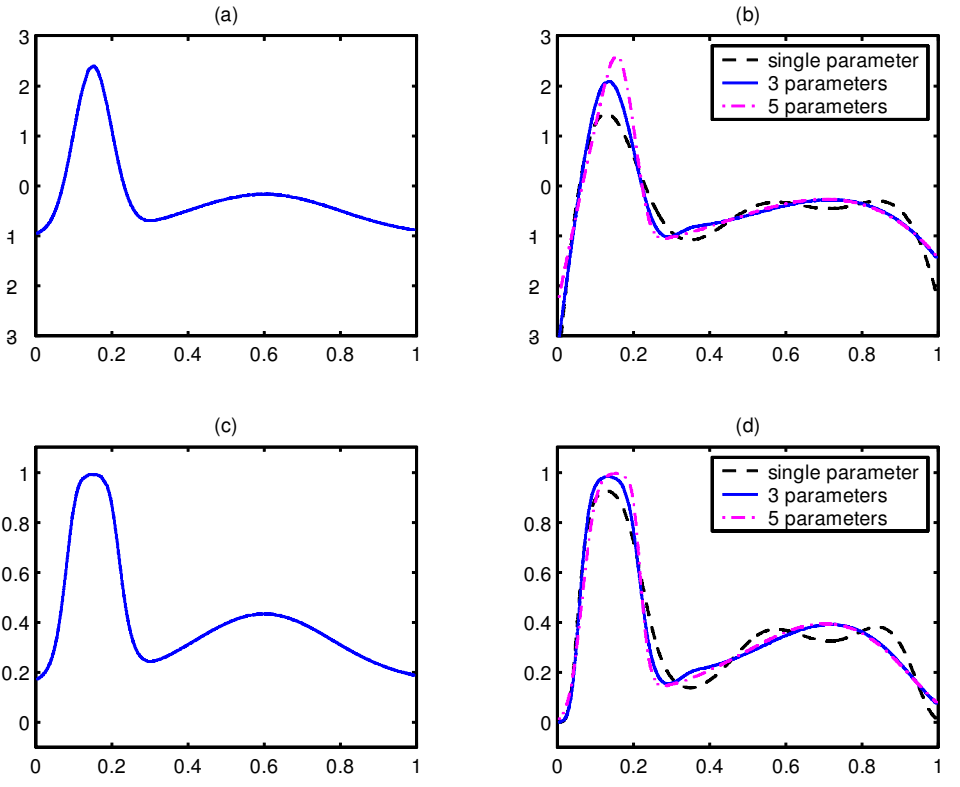

Figure 5. Estimation of spatially adaptive regression function and the corresponding probability. Panel (a) plots the function and panel (c) the corresponding probability. Panel (b) plots three estimates of the function in panel (a) based on 1, 3, and 5 smoothing parameters. Panel (d) presents the corresponding estimates of the probability.

(0.6,1.0). The model averaging estimates obtained using Sampling Scheme 2 are plotted in Figure 5(b) with the probabilities plotted in Figure 5(d). The two estimators based on multiple smoothing parameters are similar to each other and are more spatially adaptive than the estimator based on a single smoothing parameter. Interestingly, however, the estimator based on a single smoothing parameter is still adequate for estimating the conditional probabilities.

The discussion above suggests how our approach can be extended to make univariate function estimation more adaptive spatially. The extension to models with additive univariate components is straightforward, while similar ideas can be used to model bivariate and higher order interactions. These ideas will be explored in future work.

7. CONCLUSION

The article presents a unified Bayesian approach to nonparametric regression for binary and multinomial response data which allows for variable selection and model averaging. The approach can handle a large number of observations, multiple categories, and multiple

independent variables. It is easy to interpret the regression results because we write the regression function additively as a sum of main effects and interactions. We show that our approach can be used for classification, especially for binary classification, and the classification results are comparable to the best methods currently available, such as boosting. However, our results are model based and therefore much easier to interpret than tree-based boosting methods which average over many trees. Moreover, the boosting approach may also not perform well when the training data is noisy, that is when the training set contains observations that are wrongly classified, because it puts increasing weight on such cases. We showed in Section 6.1 that our approach can be adapted to handle a variety of priors on the smoothing parameters including priors that correspond to penalties on the second derivative of the regression function. Our approach can also be made more spatially adaptive in the univariate case by representing the function as a sum of components on overlapping subintervals. The extension of these ideas to additive models and models with interactions will be explored in future work.

ACKNOWLEDGMENTS

Robert Kohn was partially supported by a large ARC grant. We thank the associate editor and three referees for suggestions that improved the presentation of the article.

[Received January 2000. Revised September 2001.]

REFERENCES

Albert, J., and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669–679.

Bishop, C. (1995) *Neural Networks for Pattern Recognition*, Oxford: Clarendon Press.

Bunch, D. S. (1991),"Estimability in the Multinomial Probit Mode," *Transportation Research*, B, 25B, 1–12.

Chambers, J. M., and Hastie, T. J. (eds.) (1992), *Statistical Models in S*, Pacific Grove, CA: Wadsworth and Brooks/Cole.

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.

Eilers, P. H. C., and Marx, B. D. (1996), "Flexible Smoothing with B-splines and Penalties," *Statistical Science*, 11, 89–121.

Friedman, J., Hastie, T., and Tibshirani, R. (2000), "Additive Logistic Regression: A Statistical View of Boosting," *The Annals Statistics*, 28, 337–407.

Geweke, J. (1991), "Efficient Simulation from the Multivariate Normal and Student-t Distributions Subject to Linear Constraints and the Evaluation of Constraints Probabilities," in *Proceedings of the Twenty-Third Symposium on the Interface Between Computing Science and Statistics*, ed. Elaine Kermamidas, 571–578.

Gilks, W. R., Thomas, A., and Spiegelhalter, D. J. (1994), "A Language and Program for Complex Bayesian Modelling," *The Statistician*, 43, 169–178.

Girosi, F., Jones, M., and Poggio, T. (1995), "Regularization Theory and Neural Networks," *Neural Computation*, 7, 219–269.

Green, P. J., and Silverman, B. W. (1994), *Nonparametric Regression and Generalized Linear Models*, London: Chapman and Hall.

Gu, C., Bates, D., and Wahba, G. (1989), "The Computation of GCV Functions Through Householder Tridiagonalization With Application to the Fitting of Interaction Spline Models," *SIAM Journal of Matrix Analysis and Applications*, 10, 457–480.

Hastie, T. (1996), "Pseudosplines," *Journal of the Royal Statistical Society*, Series B, 58, 379–396.

Holmes, C. C., and Mallick, B. K. (1998), "Bayesian Radial Basis Functions of Variable Dimension," *Neural Computation*, 10, 1217–1233.

Kaufman, L., and Rousseeuw, P. J. (1990), *Finding Groups in Data*, New York: Wiley.

Keane M. P. (1992) "A Note on Identification in the Multinomial Probit Model," *Journal of Business and Economic Statistics*, 10, 193–200.

Keane, M. P. (1994),"A Computationally Practical Simulation Estimator for Panel Data," *Econometrica*, 62, 95–116.

Kooperberg, C., Bose, S., and Stone, C. J. (1997), "Polychotomous Regression," *Journal of the American Statistical Association*, 92, 117–127.

Lin, X. (1998), "Smoothing Spline Analysis of Variance for Polychotomous Response Data," Unpublished doctorial thesis, University of Wisconsin-Madison.

McFadden, D. (1973), "Conditional Logit Analysis of Qualitative Choice Behavior," in *Frontiers in Econometrics*, ed. Paul Zarembka, New York, Academic Press, pp. 105–142.

Marx, B. D., and Eilers, P. H. C. (1998), "Direct Generalized Additive Modelling With Penalized Likelihood," *Computational Statistics and Data Analysis*, 28, 193–209.

Michie, D., Spiegelhalter, D. J., and Taylor, C. C. (eds.) (1994), *Machine Learning, Neural and Statistical Classification*, Chichester: Ellis Horwood.

Raftery, A., Madigan, D., and Hoeting, J. (1997), "Bayesian Model Averaging for Linear Regression Models," *Journal of the American Statistical Association*, 92, 179–191.

Ruppert, D., and Carroll, R. J. (2000), "Spatially-Adaptive Penalties for Spline Fitting," *Australian and New Zealand Journal of Statistics*, 45, 205–223.

Shively, T., Kohn, R., and Wood, S. (1999), "Variable Selection and Function Estimation in Additive Nonparametric Regression Models Using a Data-Based Prior" (with discussion), *Journal of the American Statistical Association*, 94, 777–806.

Smith, M., and Kohn, R. (1996), "Nonparametric Regression Using Bayesian Variable Selection," *Journal of Econometrics*, 75, 317–344.

Smith, J., Everhart, J., Dickson, W., Knowler, W., and Johannes, R. (1988), "Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus," in *Proceedings of the Symposium on Computer Applications and Medical Care*, IEEE Computer Society Press, pp. 261–265.

Struyf, A., Hubert, M., and Rousseeuw, P. J. (1996), "Clustering in an Object-Oriented Environment," *Journal of Statistical Software*, 1, 1–30.

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

Wahba, G., Gu, C., Wang, Y., and Chappell, R. (1995), "Soft Classification, aka Risk Estimation, via Penalized Log Likelihood and Smoothing Spline Analysis of Variance," in *The Mathematics of Generalization, Santa Fe Institute Studies in the Science of Complexity, Proc. Vol. XX*, ed. D. Wolpert, Reading, MA: Addison-Wesley, pp. 329–360.

Wolberg, W. H., and Mangasarian, O. L. (1990), "Multisurface Method of Pattern Separation for Medical Diagnosis Applied to Breast Cytology," in *Proceedings of the National Academy of Sciences*, 87, pp. 9193–9196.

Xiang, D. (1996), "Model Fitting and Testing for Non-Gaussian Data with Large Data Sets," Unpublished doctorial thesis, University of Wisconsin-Madison.

Xiang, D., and Wahba, G. (1998), "Approximate Smoothing Spline Methods for Large Datasets in the Binary Case," in *Proceedings of the 1997 ASA Joint Statistical Meetings, Biometrics Section*, pp. 94–98.