

# ESTIMATION AND MODEL CHOICE IN NONPARAMETRIC ADDITIVE REGRESSION\*

SIDDHARTHA CHIB<sup>†</sup>

Washington University in St. Louis

IVAN JELIAZKOV<sup>‡</sup>

University of California, Irvine

November 1, 2005

## Abstract

This article revisits the Bayesian inferential problem for the important class of nonparametric additive models. We propose a new identification restriction on the unknown covariate functions under which the model can be estimated efficiently by Markov chain Monte Carlo simulation techniques. In contrast to previous discussions in the literature, our estimation procedure is based on proper smoothness priors on the unknown functions. This opens the way for model comparisons on the basis of marginal likelihoods and Bayes factors, implemented via the approach of Chib (1995). A simulation study is used to illustrate the performance of the proposed techniques. The entire methodology is conveniently adapted to generalized additive models for both non-clustered and clustered data.

*Keywords:* Additive models; Bayes factor; Bayesian model comparison; Marginal likelihood; Markov chain Monte Carlo; Markov process priors; Gibbs sampling; Sherman-Morrison formula.

## 1 Introduction

This article considers three aspects of the inferential problem—specification, estimation, and model selection—for the class of nonparametric additive models (Hastie and Tibshirani 1990). For the first of these aspects, we focus on identification and the use of proper smoothness priors for the unknown covariate functions. We approach the problem from a Bayesian perspective by modeling the unknown covariate functions through appropriate smoothness priors (Whittaker 1923, Whittaker and Robinson 1924, Wahba 1978, 1990, Shiller 1984, Silverman 1985, Besag *et al.* 1995, Fahrmeir and Tutz 1997, Fahrmeir and Lang 2001, Chib and Jeliazkov 2005, Koop and Poirier 2004). We analyze the two most common identification restrictions used when the unknown covariate functions enter the model additively, and propose a modification of one of these restrictions which yields considerable computational benefits.

---

\*PRELIMINARY AND INCOMPLETE DRAFT – COMMENTS WELCOME.

<sup>†</sup>*Address for correspondence:* John M. Olin School of Business, Washington University, Campus Box 1133, 1 Brookings Drive, St. Louis, MO 63130. E-mail: chib@wustl.edu.

<sup>‡</sup>*Address for correspondence:* Department of Economics, University of California, Irvine, 3151 Social Science Plaza, Irvine, CA 92697-5100. E-mail: ivan@uci.edu.

A key challenge in the estimation of nonparametric models is computational efficiency. This is because estimation of the nonparametric functions involves simulation of quantities whose dimension can easily exceed the sample size. Efficient computational algorithms are of particular importance in large cross-sectional or in longitudinal data, where the sample size can easily run into the thousands or tens of thousands. As discussed in detail below, the algorithms in this paper build on, and extend, previous algorithms exploiting the specific (banded) structure implied by the priors on the unknown functions (Silverman 1985, Fahrmeir and Lang 2001, Koop and Poirier 2004). Hence, the methods reduce the computational costs and make feasible the analysis of large-dimensional applied problems.

The recent interest in Bayesian nonparametric estimation, aided by the computational advances in Markov chain Monte Carlo (MCMC) simulation methods, has been quite strong (Besag *et al.* 1995, Wood and Kohn 1998, Denison *et al.* 1998, Shively *et al.* 1999, Hastie and Tibshirani 2000, Fahrmeir and Lang 2001, DiMatteo *et al.* 2001, Wood *et al.* 2002, Koop and Poirier 2004). Nonparametric functional modeling has appealing frequentist and Bayesian properties, but an advantage of the simulation-based approach is that, in conjunction with the latent variable augmentation technique of Albert and Chib (1993), it allows for the straightforward analysis of nonparametric models for binary and ordinal response data (Kohn and Wood 1998, Shively *et al.* 1999). The entire methodology can also be adapted to nonparametric modeling for both clustered and non-clustered data (Chib and Jeliazkov 2005).

Two aspects of this analysis distinguish it from most of the discussion in the current Bayesian nonparametric literature. One is tied to the specification of the model and the other to the problem of model comparison. For the former, we specify a fully-Bayesian model which combines proper smoothness priors on the unknown functions with a likelihood that involves identification restrictions on the unknown functions (since the functions can only be identified up to an additive constant). In contrast, much of the literature involves partially improper smoothness priors, which prevents formal Bayesian model selection through marginal likelihoods and Bayes factors. Moreover, many models are often defined without clearly resolving the identification problem and providing model-based interpretation; instead, identification is often addressed in the course of sampling, e.g by “centering on the fly” (Besag *et al.* 1995, and Hastie and Tibshirani 2000, Fahrmeir

and Lang 2001). This scheme lacks a model-based interpretation, especially when the analysis is conducted with improper priors on the functions, as in the aforementioned papers. It should be noted that if proper priors are used with the recentering scheme of Hastie and Tibshirani (2000), the posterior updates no longer involve banded matrices, and the computational burden becomes excessive.

We address the problem of model comparison because of its importance since it enables the comparison of nonparametric models to other competing specifications, including parametric and semiparametric alternatives. The related literature to date has been quite sparse and, as far as we can tell, the formal Bayesian approach (based on marginal likelihoods and Bayes factors) has not been examined in this context. The little literature that does exist on this problem relies on the AIC and BIC information measures (Shively *et al.* 1999, Wood *et al.* 2002, DiMatteo *et al.* 2001, Hansen and Kooperberg 2002). However, as discussed by Wood *et al.* (2002), the use of these measures in the additive case raises some theoretical difficulties. In addition, the computation of the maximum likelihood values (the inputs into the AIC and BIC) is demanding, especially with large data sets. To deal with the latter problem, Wood *et al.* (2002) employ some computational shortcuts (discussed in the sequel) whose validity may be questioned. We contribute to this literature by presenting a method for computing marginal likelihoods and Bayes factors, based on the framework of Chib (1995).

The model comparison analysis yields surprising results. Previous research has avoided the computation of marginal likelihoods because they were thought to be infeasible high-dimensional integration problems. Instead, researchers had focused on the computation of AIC and BIC. Interestingly, however, here we show that the marginal likelihood (from the approach we develop) is more easily computed than the AIC and BIC measures.

The article is organized as follows. In Section 3, we briefly review a nonparametric model with a single unknown function, and present an efficient fitting algorithm. In Section 2, we present the full additive model, discuss a new identification scheme, and generalize the fitting algorithm. Section 4 deals with the problem of computing the marginal likelihood. In Section 5, we provide a simulation study showing the performance of our MCMC fitting algorithm and of the model comparison criterion. In Section 6, we outline extensions of the method to additive models for

non-clustered and clustered binary data. Section 7 considers a multiple-regression example and Section 8 concludes.

## 2 Additive Models

Additive models provide a natural way of extending univariate models to the multiple regression setting. Relative to linear parametric models, nonparametric models maintain additivity but do not require that the estimated functions lie in a particular class of functions. Here the response  $y_i$  is assumed to depend on the vector of covariates  $(s_{i1}, \dots, s_{ip})$  in the form

$$y_i = c + g_1(s_{i1}) + \dots + g_p(s_{ip}) + \varepsilon_i, \quad (i = 1, \dots, n), \quad (1)$$

where  $\varepsilon_i \sim N(0, \sigma^2)$ , and  $\{g_j(\cdot)\}_{j=1}^p$  are unknown smooth functions that are to be estimated non-parametrically. For an introduction to additive models, with applications, see Hastie and Tibshirani (1990).

### 2.1 Prior Distributions

For each of the  $j = 1, \dots, p$  functions in (1), let the  $n$  observations in the covariate vectors  $\mathbf{s}_j = (s_{j1}, \dots, s_{jn})'$  determine the corresponding  $m_j \times 1$  *design point vectors*  $\mathbf{v}_j = (v_{j1}, \dots, v_{jm_j})'$  with entries equal to the *unique ordered* values of  $\mathbf{s}_j$ , that is  $v_{j1} < \dots < v_{jm_j}$ ,  $m_j \leq n$ , and with corresponding function evaluation vectors  $\mathbf{g}_j = (g(v_{j1}), \dots, g(v_{jm_j}))'$ . The idea is to model the function evaluations as a stochastic process which controls the degree of local variation between neighboring states.

To avoid cumbersome notation, in the remainder of this section we drop the function index  $j$  (we will return to it later), but note that this should not confuse the reader because every function is treated similarly. We place a second-order Markov process prior on  $\mathbf{g} = (g(v_1), \dots, g(v_m))' = (g_1, \dots, g_m)'$ . Specifically, defining  $h_t \equiv v_t - v_{t-1}$ , our (second-order) Markov process prior is given by

$$g_t = \left(1 + \frac{h_t}{h_{t-1}}\right) g_{t-1} - \frac{h_t}{h_{t-1}} g_{t-2} + u_t, \quad u_t \sim N(0, \tau^2 h_t), \quad (2)$$

where  $\tau^2$  is a smoothness parameter, such that small values of  $\tau^2$  produce smoother functions, while larger values allow the function to be more flexible and interpolate the data more closely. There is always some degree of flexibility in the choice of any such prior, and there is a variety



and

$$\boldsymbol{\Sigma} = \begin{pmatrix} \mathbf{G}_0 & & & \\ & h_3 & & \\ & & \ddots & \\ & & & h_m \end{pmatrix},$$

the global smoothness representation of the second order Markov process prior equivalent to (2) and (3) becomes

$$\mathbf{g}|\tau^2 \sim N(\mathbf{g}_0, \tau^2 \mathbf{K}^{-1}), \quad (4)$$

where  $\mathbf{g}_0 = \mathbf{H}^{-1}\tilde{\mathbf{g}}$ , with  $\tilde{\mathbf{g}} = (g_{10}, g_{20}, 0, \dots, 0)'$ , and the *penalty matrix*  $\mathbf{K}$  is given by  $\mathbf{K} = \mathbf{H}'\boldsymbol{\Sigma}^{-1}\mathbf{H}$ . In the preceding,  $\mathbf{g}_0$  can equivalently be derived by taking recursive expectations of (2) starting with the mean in (3), so as to avoid the inversion of  $\mathbf{H}$ . A key feature of the prior in (4) is that it is proper. This offers an important refinement on much of the literature on smoothness priors for nonparametric function estimation where, in contrast, partially improper priors and reduced rank penalty matrices  $\mathbf{K}$  are used. This refinement removes an important impediment to formal Bayesian model selection.

Since the prior on  $\mathbf{g}$  is defined conditional of the hyperparameter  $\tau^2$ , in the next level of the modeling hierarchy we specify the prior distribution

$$\tau^2 \sim IG\left(\frac{\nu_0}{2}, \frac{\delta_0}{2}\right). \quad (5)$$

The prior distribution on the variance parameter  $\sigma^2$  in (8) is similarly taken to be

$$\sigma^2 \sim IG\left(\frac{s_0}{2}, \frac{d_0}{2}\right). \quad (6)$$

We conclude the discussion on Markov process priors by making two remarks. First, from an estimation point of view, it is important to note that the  $m \times m$  penalty matrix  $\mathbf{K}$  is banded. This fact is of considerable practical utility, as manipulations involving banded matrices take  $O(m)$  operations, rather than the usual  $O(m^3)$  for inversions or  $O(m^2)$  for multiplication by a vector. Given that  $m$  may be large (potentially as large as the total number of observations  $n$  in the data sample) this has important ramifications for the numerical efficiency of the estimation procedure. Second, Markov process priors are conceptually simple and easily adaptable to different orders, enabling them to match problem-specific tasks more closely (Besag *et al.* 1995, Fahrmeir and Lang 2001). For example, a simple first order Markov process prior  $g_t = g_{t-1} + u_t$  will penalize abrupt

jumps between successive states of the random walk process, while higher order priors embody more complex notions of “smoothness” related to the rates of change in the function; such priors share many similar features and are easily specified using the general ideas outlined above. However, before we can consider the estimation of the model, we must address the likelihood identification problem that emerges from the additive structure in (1).

## 2.2 Identification

As written, the model in (1) is likelihood unidentified because the functions  $\{g_j\}$  are unrestricted, so the likelihood will remain unchanged if one redefines simultaneously  $g_j^* = g_j + \alpha$  and  $g_i^* = g_i - \alpha$  for  $i \neq j$  and some constant  $\alpha$ . Even though it is well known that Bayesian models with proper priors do not suffer from identification problems even when the likelihood is not identified (Lindley 1971), likelihood identification is essential in providing well-behaved posteriors and well-behaved MCMC algorithms that will quickly and efficiently explore the posterior distribution. To achieve likelihood identification, the functions  $\{g_j\}$  have to be “anchored” by imposing restrictions that remove any free constants.

One approach to identification is suggested in the work of Shively *et al.* (1999), where the functions  $\{g_j\}_{j=1}^p$  are restricted to start at zero. The approach produces a properly identified likelihood function, but the identification crucially depends on the smoothness parameters. Identification can be weak when some of the  $\{\tau_j^2\}$  are large, implying a small penalty to vertical shifts of the functional values beyond the first one. In addition, the identification restrictions produce funnel-like error bands for the function estimates whose interpretation is not entirely transparent.

Another possibility for removing the free constant from an unknown function is to center that function using the restriction  $\sum_{i=1}^{m_j} g_j(v_i) = 0$ , or in vector form  $\mathbf{g}'_j \mathbf{1} = 0$  (Hastie and Tibshirani 1990). It will be sufficient to apply this centering to  $p - 1$  of the unknown functions, allowing the overall intercept to be absorbed into the remaining function, or alternatively, centering can be applied to all functions after the introduction of an additional intercept parameter in (1). Gelfand (2000) points out that this identification scheme has often been applied in ways that do not correspond to well-defined Bayesian models, because recentering is introduced “on the fly” after each iteration, merely as a step in the fitting algorithm for models that are otherwise built upon par-

tially improper smoothness priors and/or unidentified likelihood functions (e.g. Besag *et al.* 1995, Hastie and Tibshirani 2000). Adapting the notation in Lin and Zhang (1999) to the case where  $p-1$  of the unknown functions are centered, this identification scheme implies the following stacked representation of the additive model

$$\mathbf{y} = \mathbf{Q}_1 \mathbf{g}_1 + \mathbf{Q}_2 \mathbf{M}_{02} \mathbf{g}_2 + \dots + \mathbf{Q}_p \mathbf{M}_{0p} \mathbf{g}_p + \varepsilon,$$

where the  $\{\mathbf{Q}_j\}$  represent the  $n \times m_j$  incidence matrices corresponding to each of the covariates, and the  $\{\mathbf{M}_{0j}\}$  represent the  $m_j \times m_j$  symmetric and idempotent mean-differencing matrices

$$\mathbf{M}_{0j} = \left( \mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{m_j} \right), \quad j = 1, \dots, p.$$

Unfortunately, with proper priors on each of the unknown functions, as in this paper, this identification scheme becomes computationally demanding because the posterior updates no longer involve banded matrices, and the computational burden becomes excessive. We believe that it is because of this difficulty that no work has been done on additive models with proper priors on the unknown functions.

We solve this problem by proposing a different (but closely related) identification scheme that is computationally simple to execute, even though the posterior updates do not involve banded matrices. In stacked form our identification scheme for additive models can be represented as

$$\mathbf{y} = \mathbf{Q}_1 \mathbf{g}_1 + \mathbf{M}_0 \mathbf{Q}_2 \mathbf{g}_2 + \dots + \mathbf{M}_0 \mathbf{Q}_p \mathbf{g}_p + \varepsilon, \quad (7)$$

where the mean differencing matrix

$$\mathbf{M}_0 = \left( \mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{n} \right)$$

now premultiplies the incidence matrices  $\{\mathbf{Q}_j\}$  and centers the expanded vector of functional evaluations. In other words, we recenter the functions using the restrictions  $\sum_{i=1}^n g_j(s_i) = 0$ . Hence, in determining the centering constants, this is equivalent to weighted averaging of the covariate functions where the weights are proportional to the number of times each element of  $\mathbf{v}_j$  is represented in  $\mathbf{s}_j$ . Note that if there are no repeating elements in  $\mathbf{s}_j$ , our approach is equivalent to the scheme discussed above, but the two differ by a constant when there are repeating values in  $\mathbf{s}_j$ . The specific benefits from this identification method will be discussed shortly.

It should be clear from the representation in (7) that the model will be unidentified if any two incidence matrices  $\mathbf{Q}_h$  and  $\mathbf{Q}_k$  ( $h \neq k$ ) are identical. In other words, if for some  $h \neq k$ ,  $\mathbf{v}_h$  and  $\mathbf{v}_k$  are coincidental, so that their elements occur in the same exact manner in  $\mathbf{s}_h$  and  $\mathbf{s}_k$ , respectively, the data will not be informative about the individual effects of  $\mathbf{s}_h$  and  $\mathbf{s}_k$ . This multicollinearity is related to the problem of concurvity (Hastie and Tibshirani 1990).

### 3 Nonparametric Modeling

To motivate the general approach, we begin by considering the important special case of univariate regression. Given data  $\{y_i, s_i\}_{i=1}^n$ , the responses  $y_i$  are assumed to depend on the (scalar) covariate  $s_i$  in the form

$$y_i = g(s_i) + \varepsilon_i, \quad (i = 1, \dots, n), \quad (8)$$

where  $\varepsilon_i \sim N(0, \sigma^2)$ , and  $g(\cdot)$  is an unknown smooth function that is to be estimated nonparametrically.

#### 3.1 Efficient Estimation

The model in (8) can be written in stacked form as

$$\mathbf{y} = \mathbf{Q}\mathbf{g} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}), \quad (9)$$

where  $\mathbf{Q}$  is an *incidence matrix* of dimension  $n \times m$ , with entries  $\mathbf{Q}(i, j) = 1$  if  $s_i = v_j$  and 0 otherwise. In other words,  $\mathbf{Q}$  determines the correspondence between the covariate vector  $\mathbf{s}$  and the design point vector  $\mathbf{v}$  of unique and ordered elements of  $\mathbf{s}$ , so that  $\mathbf{s} = \mathbf{Q}\mathbf{v}$ . This definition of  $\mathbf{Q}$  also implies that the  $i$ th component of  $\mathbf{Q}\mathbf{g}$  is  $g(s_i)$ .

The algorithm described below is derived from (9) using standard Bayes updates, and has been used for sampling in many applications (Besag *et al.* 1995, Hastie and Tibshirani 2000, Fahrmeir and Lang 2001, Müller *et al.* 2001). Extensions of this algorithm to additive models will be discussed in the next section, while generalizations to semiparametric and binary data models will be discussed in Section 6.

**Algorithm 1** *Univariate Gaussian Nonparametric Model: MCMC Implementation*

1. Sample  $\mathbf{g}|\mathbf{y}, \tau^2, \sigma^2 \sim N(\hat{\mathbf{g}}, \mathbf{G})$ , where  $\mathbf{G}$  and  $\hat{\mathbf{g}}$  are the usual Bayes updates for linear regression, namely  $\mathbf{G} = (\mathbf{K}/\tau^2 + \mathbf{Q}'\mathbf{Q}/\sigma^2)^{-1}$  and  $\hat{\mathbf{g}} = \mathbf{G}(\mathbf{K}\mathbf{g}_0/\tau^2 + \mathbf{Q}'\mathbf{y}/\sigma^2)$ . Remark 1 below presents important notes on the sampling in this step.
2. Sample  $\tau^2|\mathbf{g} \sim IG\left(\frac{\nu_0+m}{2}, \frac{\delta_0+(\mathbf{g}-\mathbf{g}_0)'\mathbf{K}(\mathbf{g}-\mathbf{g}_0)}{2}\right)$ , where we note that conditional on  $\mathbf{g}$ ,  $\tau^2$  is independent of the remaining parameters and the data.
3. Sample  $\sigma^2|\mathbf{y}, \mathbf{g} \sim IG\left(\frac{s_0+n}{2}, \frac{d_0+(\mathbf{y}-\mathbf{Q}\mathbf{g})'(\mathbf{y}-\mathbf{Q}\mathbf{g})}{2}\right)$ .

**Remark 1 *Sampling of g***

In sampling  $\mathbf{g}$ , one should note that  $\mathbf{Q}'\mathbf{Q}$  is a diagonal matrix whose  $j$ th diagonal entry equals the number of values in  $\mathbf{s}$  corresponding to the design point  $v_j$ . Since  $\mathbf{K}$  and  $\mathbf{Q}'\mathbf{Q}$  are banded,  $\mathbf{G}^{-1}$  is banded as well. Thus sampling of  $\mathbf{g}$  need not include an inversion to obtain  $\mathbf{G}$  and  $\hat{\mathbf{g}}$ . The mean  $\hat{\mathbf{g}}$  can be found instead by solving  $\mathbf{G}^{-1}\hat{\mathbf{g}} = (\mathbf{K}\mathbf{g}_0/\tau^2 + \mathbf{Q}'\mathbf{y}/\sigma^2)$ , which is done in  $O(n)$  operations by back substitution. Also, let  $\mathbf{P}'\mathbf{P} = \mathbf{G}^{-1}$ , where  $\mathbf{P}$  is the Cholesky decomposition of  $\mathbf{G}^{-1}$  and is also banded. To obtain a random draw from  $N(\hat{\mathbf{g}}, \mathbf{G})$  efficiently, sample  $\mathbf{u} \sim N(\mathbf{0}, \mathbf{I})$ , and solve  $\mathbf{P}\mathbf{x} = \mathbf{u}$  for  $\mathbf{x}$  by back substitution. It follows that  $\mathbf{x} \sim N(\mathbf{0}, \mathbf{G})$ . Adding the mean  $\hat{\mathbf{g}}$  to  $\mathbf{x}$ , one obtains a draw  $\mathbf{g} \sim N(\hat{\mathbf{g}}, \mathbf{G})$ .

We note that the MCMC approach to estimating  $\tau^2$  in this hierarchical model offers an alternative to cross-validation. MCMC estimation accounts fully for parameter uncertainty, unlike plug-in approaches, which do not account for the variability due to estimating the smoothing parameters. Also, as will be discussed in Section 6, the MCMC approach can be easily extended to discrete data settings.

**3.2 Efficient Estimation**

The main advantage of the identification scheme proposed above is computational. Under this identification and the priors discussed in Section 2, the sampling from the posterior distribution is done in the following steps.

**Algorithm 2 *Gaussian Additive Model: MCMC Implementation***

1. Sample  $\mathbf{g}_1 | \mathbf{y}, \tau_1^2, \sigma^2, \{\mathbf{g}_i\}_{i=2}^p \sim N(\hat{\mathbf{g}}_1, \hat{\mathbf{G}}_1)$ , where,

$$\hat{\mathbf{G}}_1 = \left( \frac{1}{\tau_1^2} \mathbf{K}_1 + \frac{1}{\sigma^2} \mathbf{Q}'_1 \mathbf{Q}_1 \right)^{-1}$$

and

$$\hat{\mathbf{g}}_1 = \hat{\mathbf{G}}_1 \left( \frac{1}{\tau_1^2} \mathbf{K}_1 \mathbf{g}_{10} + \frac{1}{\sigma^2} \mathbf{Q}'_1 \left( \mathbf{y} - \sum_{i=2}^p \mathbf{M}_0 \mathbf{Q}_i \mathbf{g}_i \right) \right).$$

The sampling in this step is carried out in  $O(n)$  operations as discussed in Remark 1.

2. Sample  $\mathbf{g}_j | \mathbf{y}, \tau_j^2, \sigma^2, \{\mathbf{g}_i\}_{i \neq j} \sim N(\hat{\mathbf{g}}_j, \hat{\mathbf{G}}_j)$  for  $j = 2, \dots, p$ , where

$$\hat{\mathbf{G}}_j = \left( \frac{1}{\tau_j^2} \mathbf{K}_j + \frac{1}{\sigma^2} \mathbf{Q}'_j \mathbf{M}_0 \mathbf{Q}_j \right)^{-1}$$

and

$$\hat{\mathbf{g}}_j = \hat{\mathbf{G}}_j \left( \frac{1}{\tau_j^2} \mathbf{K}_j \mathbf{g}_{j0} + \frac{1}{\sigma^2} \mathbf{Q}'_j \mathbf{M}_0 \left( \mathbf{y} - \mathbf{Q}_1 \mathbf{g}_1 - \sum_{i \geq 2, i \neq j} \mathbf{M}_0 \mathbf{Q}_i \mathbf{g}_i \right) \right).$$

Remark 2 below shows how the sampling in this step can be carried out efficiently in  $O(n)$  operations, even though  $\hat{\mathbf{G}}_j$  is not banded.

3. Sample  $\tau_j^2$ ,  $j = 1, \dots, p$ , from

$$\tau_j^2 | \mathbf{g}_j \sim IG \left( \frac{\nu_{j0} + m_j}{2}, \frac{\delta_{j0} + (\mathbf{g}_j - \mathbf{g}_{j0})' \mathbf{K}_j (\mathbf{g}_j - \mathbf{g}_{j0})}{2} \right),$$

where conditional on  $\mathbf{g}_j$ ,  $\tau_j^2$  is independent of the remaining parameters, functions, and the data.

4. Sample  $\sigma^2 | \mathbf{y}, \{\mathbf{g}_j\} \sim IG \left( \frac{s_0 + n}{2}, \frac{d_0 + \|\mathbf{y} - \mathbf{Q}_1 \mathbf{g}_1 - \sum_{j=2}^p \mathbf{M}_0 \mathbf{Q}_j \mathbf{g}_j\|}{2} \right)$ .

While the above algorithm involves steps which are similar to those in Algorithm 1, Step 2 involves  $p - 1$  non-banded  $n \times n$  matrices and requires special care in order to sample efficiently. That efficient  $O(n)$  sampler is discussed in Remark 2 below.

**Remark 2 Efficient Sampling of  $\mathbf{g}_j | \mathbf{y}, \tau_j^2, \sigma^2, \{\mathbf{g}_i\}_{i \neq j}$  for  $j = 2, \dots, p$**

In drawing  $\mathbf{g}_j \sim N(\hat{\mathbf{g}}_j, \hat{\mathbf{G}}_j)$ ,  $j = 2, \dots, p$ , using the definition of  $\mathbf{M}_0$  write

$$\begin{aligned} \hat{\mathbf{G}}_j &= \left( \frac{1}{\tau_j^2} \mathbf{K}_j + \frac{1}{\sigma^2} \mathbf{Q}'_j \mathbf{M}_0 \mathbf{Q}_j \right)^{-1} \\ &= \left( \frac{1}{\tau_j^2} \mathbf{K}_j + \frac{1}{\sigma^2} \mathbf{Q}'_j \mathbf{Q}_j - \frac{\mathbf{c}_j \mathbf{c}'_j}{\sigma^2 n} \right)^{-1}, \end{aligned}$$

where  $\mathbf{c}_j = \mathbf{Q}'_j \mathbf{1}$  is an  $m_j$ -vector of cluster sizes induced by the correspondence between the elements of  $\mathbf{s}_j$  and  $\mathbf{v}_j$ , that is the  $i$ th entry of  $\mathbf{c}_j$  contains the number of times the  $i$ th entry of  $\mathbf{v}_j$  is repeated in  $\mathbf{s}_j$ . Let  $\mathbf{A}_j = \frac{1}{\tau_j^2} \mathbf{K}_j + \frac{1}{\sigma^2} \mathbf{Q}'_j \mathbf{Q}_j$  and  $\mathbf{u}_j = \frac{1}{\sqrt{\sigma^2 n}} \mathbf{c}_j$ . Then, by the Sherman-Morrison formula, one can write  $\hat{\mathbf{G}}_j$  in the above as

$$\begin{aligned} \hat{\mathbf{G}}_j &= (\mathbf{A}_j - \mathbf{u}_j \mathbf{u}'_j)^{-1} \\ &= \mathbf{A}_j^{-1} + \frac{\mathbf{A}_j^{-1} \mathbf{u}_j \mathbf{u}'_j \mathbf{A}_j^{-1}}{1 - \lambda_j} \end{aligned} \quad (10)$$

where  $\lambda_j = \mathbf{u}'_j \mathbf{A}_j^{-1} \mathbf{u}_j$ . The efficiency benefits from this representation are significant, because the application of (10) obviates the matrix inversion necessary to obtain  $\hat{\mathbf{G}}_j$  and  $\hat{\mathbf{g}}_j$  in Step 2 of Algorithm 2. Instead, the mean  $\hat{\mathbf{g}}_j$  can be obtained in  $O(n)$  operations from

$$\hat{\mathbf{g}}_j = \left( \mathbf{A}^{-1} + \frac{\mathbf{A}^{-1} \mathbf{u} \mathbf{u}' \mathbf{A}^{-1}}{1 - \lambda} \right) \left( \frac{1}{\tau_j^2} \mathbf{K}_j \mathbf{g}_{j0} + \frac{1}{\sigma^2} \mathbf{Q}'_j \mathbf{M}_0 \mathbf{d}_j \right),$$

where

$$\mathbf{d}_j = \left( \mathbf{y} - \mathbf{Q}_1 \mathbf{g}_1 - \sum_{i \geq 2, i \neq j} \mathbf{M}_0 \mathbf{Q}_i \mathbf{g}_i \right)$$

by working with  $\mathbf{A}$  without inverting to  $\mathbf{A}^{-1}$  as outlined in Remark 1. Furthermore, let

$$\mathbf{B}_j = \left( \mathbf{A}_j + \frac{\mathbf{u}_j \mathbf{u}'_j}{1 - \lambda_j} \right),$$

which implies that  $\hat{\mathbf{G}}_j$  in (10) can be written as  $\mathbf{A}_j^{-1} \mathbf{B}_j \mathbf{A}_j^{-1}$ . Thus, if  $\mathbf{x} \sim N(\mathbf{0}, \mathbf{B}_j)$ , it follows that  $\mathbf{z} = \mathbf{A}_j^{-1} \mathbf{x}$  has a distribution  $\mathbf{z} \sim N(\mathbf{0}, \hat{\mathbf{G}}_j)$ , and to obtain a draw for  $\mathbf{g}_j \sim N(\hat{\mathbf{g}}_j, \hat{\mathbf{G}}_j)$ , one simply forms the sum  $\mathbf{g}_j = \hat{\mathbf{g}}_j + \mathbf{z}$ . Now, to generate  $\mathbf{x} \sim N(\mathbf{0}, \mathbf{B}_j)$ , simply let  $\mathbf{x} = \mathbf{w}_1 + \tilde{\mathbf{u}}_j w_2$ , where  $\mathbf{w}_1 \sim N(\mathbf{0}, \mathbf{A}_j)$  and  $w_2 \sim N(0, 1)$  are independent, and  $\tilde{\mathbf{u}}_j = \mathbf{u}_j / \sqrt{1 - \lambda_j}$ .

Due to the shortcuts afforded by the Sherman-Morrison formula and the fact that  $\{\mathbf{A}_j\}$  are banded, all operations above are  $O(n)$ . Also, we emphasize that although the algorithm in this section is concerned with continuous cross-section data, the estimation methods are quite general and easily applicable to other related problems. Some such extensions will be outlined in Section 6. We next turn to the problem of model comparison.

## 4 Marginal Likelihood Estimation

Model comparison is a central issue in statistical data analysis, since the appropriate specification is generally subject to uncertainty. Given a collection of models  $\{\mathcal{M}_1, \dots, \mathcal{M}_L\}$ , the formal Bayesian

approach to model comparison is based on the posterior model probabilities, or their ratios, known as the posterior odds. For any two models  $\mathcal{M}_i$  and  $\mathcal{M}_j$  that attempt to explain the data  $\mathbf{y}$ , we have

$$\frac{\Pr(\mathcal{M}_i|\mathbf{y})}{\Pr(\mathcal{M}_j|\mathbf{y})} = \frac{\Pr(\mathcal{M}_i)}{\Pr(\mathcal{M}_j)} \times \frac{m(\mathbf{y}|\mathcal{M}_i)}{m(\mathbf{y}|\mathcal{M}_j)},$$

so that the posterior odds are equal to the prior odds times the ratio of the marginal likelihoods (the Bayes factor), where the marginal likelihood

$$m(\mathbf{y}|\mathcal{M}_l) = \int f(\mathbf{y}|\mathcal{M}_l, \theta_l) \pi_l(\theta_l|\mathcal{M}_l) d\theta_l \quad (11)$$

is the integral of the likelihood function  $f(\mathbf{y}|\mathcal{M}_l, \theta_l)$  with respect to the prior distribution on the model parameters  $\pi(\theta_l|\mathcal{M}_l)$ .

To our knowledge there is no existing work on the calculation of the marginal likelihood in the context of nonparametric additive models. It should be clear that direct analytic calculation of the marginal likelihood is infeasible given that the dimension of the integral over  $\theta = (\mathbf{g}_1, \dots, \mathbf{g}_p, \tau_1^2, \dots, \tau_p^2, \sigma^2)$  will typically be very large (possibly bigger than the sample size  $n$ ). We deal with this problem by utilizing the approach in Chib (1995) which has been widely applied to deal with other high-dimensional problems. In this approach, the infeasible integration of the likelihood over the prior is reduced to the more tractable problem of finding an estimate of the posterior, at a single point  $\theta^*$ . Specifically, the marginal likelihood is available from

$$m(\mathbf{y}) = \frac{f(\mathbf{y}|\theta^*)\pi(\theta^*)}{\pi(\theta^*|\mathbf{y})} \quad (12)$$

where we have suppressed the model index. Since the numerator terms are available by direct calculation, the marginal likelihood can be estimated by finding an estimate of the posterior ordinate  $\pi(\theta^*|\mathbf{y})$ .

When multi-block MCMC algorithms are used to sample  $\theta$ , as in the case of the models discussed in this paper, it is useful to break up the estimation of  $\pi(\theta^*|\mathbf{y})$  into several pieces. Let  $\theta = (\theta_1, \dots, \theta_B)$ , and denote by  $\psi_i = (\theta_1, \dots, \theta_i)$  the blocks up to  $i$  and by  $\psi^{i+1} = (\theta_{i+1}, \dots, \theta_B)$  the blocks beyond  $i$ , and write the posterior ordinate at  $\theta^*$  as

$$\pi(\theta_1^*, \dots, \theta_B^*|\mathbf{y}) = \prod_{i=1}^B \pi(\theta_i^*|\mathbf{y}, \theta_1^*, \dots, \theta_{i-1}^*) = \prod_{i=1}^B \pi(\theta_i^*|\mathbf{y}, \psi_{i-1}^*). \quad (13)$$

In the context of Gibbs sampling when the full-conditional densities, including their normalizing constants, are fully known, Chib (1995) proposed finding the ordinate  $\pi(\theta_i^*|\mathbf{y}, \psi_{i-1}^*)$  by Rao-Blackwellization

$$\begin{aligned}\pi(\theta_i^*|\mathbf{y}, \psi_{i-1}^*) &= \int \pi(\theta_i^*|y, \psi_{i-1}^*, \psi^{i+1}) \pi(\psi^i|\mathbf{y}, \psi_{i-1}^*) d\psi^i \\ &\approx T^{-1} \sum_{t=1}^T \pi(\theta_i^*|\mathbf{y}, \psi_{i-1}^*, \psi^{i+1,(t)}) ,\end{aligned}$$

where  $\psi^{i,(t)} \sim \pi(\psi^i|\mathbf{y}, \psi_{i-1}^*)$ ,  $t = 1, \dots, T$ , come from a reduced run for  $1 < i < B$ , and sampling is only over  $\psi^i$ , with the blocks  $\psi_{i-1}^*$  being held fixed. The ordinate  $\pi(\theta_1^*|\mathbf{y})$  for the first block of parameters  $\theta_1$  is estimated with draws  $\theta \sim \pi(\theta|\mathbf{y})$  from the main MCMC run, while the ordinate  $\pi(\theta_B^*|\mathbf{y}, \psi_{B-1}^*)$  is available directly.

The choice of a suitable decomposition in (13) is quite important, as it determines an appropriate balance between computational and statistical efficiency. To see this, consider the case when a large dimensional block (such as  $\{\mathbf{g}_j\}$ ) is placed towards the front of the decomposition in (13). Because this block is held fixed in subsequent reduced runs, the computational demands are lower. This, however, may increase the variability in the Rao-Blackwellization step, where the full-conditional density for this large block is averaged over a conditioning set which changes with every iteration. Alternatively, if the large dimensional block is placed towards the end in (13), the Rao-Blackwell average will be more stable as now more blocks in the conditioning set stay fixed; this strategy leads to higher statistical efficiency but comes at a higher computational cost, since a large dimensional block (rather than a different block of lower dimension) is simulated in all of the preceding reduced runs.

Before turning to our suggested procedure for finding the marginal likelihood we mention that in small samples, and under the assumptions of the model above, the marginal likelihood can be obtained by direct marginalization over the  $\{\mathbf{g}_j\}$ , so that the marginal likelihood can be determined after estimating only the low-dimensional posterior ordinates of the remaining model parameters  $\{\tau_j^2\}$  and  $\sigma^2$ . We call this the direct marginalization approach. We emphasize that it is only feasible for small sample sizes.

To understand the idea behind the direct marginalization method, note that the marginal

likelihood can be computed using

$$m(\mathbf{y}) = \frac{f(\mathbf{y} | \{\tau_i^{2*}\}, \sigma^{2*}) \pi(\tau_i^{2*}, \sigma^{2*})}{\pi(\{\tau_i^{2*}\}, \sigma^{2*} | \mathbf{y})},$$

where all quantities are marginalized over the high dimensional blocks  $\{\mathbf{g}_j\}$ . This marginalization is possible due to the Gaussian structure of the model, because conditional on  $\{\tau_i^{2*}, \sigma^{2*}\}$ , the density  $f(\mathbf{y} | \{\tau_i^{2*}, \sigma^{2*}\})$ , marginalized over  $\{\mathbf{g}_j\}$  with respect to the prior distributions, is directly available, and is given by

$$f(\mathbf{y} | \{\tau_i^{2*}, \sigma^{2*}\}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where

$$\boldsymbol{\mu} = \mathbf{Q}_1 \mathbf{g}_{10} + \mathbf{M}_0 \mathbf{Q}_2 \mathbf{g}_{20} + \dots + \mathbf{M}_0 \mathbf{Q}_p \mathbf{g}_{p0}$$

and

$$\boldsymbol{\Sigma} = \sigma^2 \mathbf{I} + \tau_1^{2*} \mathbf{Q}_1 \mathbf{K}_1^{-1} \mathbf{Q}_1' + \mathbf{M}_0 \left\{ \tau_2^{2*} \mathbf{Q}_2 \mathbf{K}_2^{-1} \mathbf{Q}_2' + \dots + \tau_p^{2*} \mathbf{Q}_p \mathbf{K}_p^{-1} \mathbf{Q}_p' \right\} \mathbf{M}_0.$$

Because of this analytical tractability,  $m(\mathbf{y})$  can then be found after the main run where one computes

$$\pi(\{\tau_i^{2*}\}, \sigma^{2*} | y) \approx T^{-1} \sum_{t=1}^T \left\{ IG(\sigma^{2*} | y, \{\mathbf{g}_j^{(t)}\}) \prod_{i=1}^p IG(\tau_i^{2*} | \mathbf{g}_i^{(t)}) \right\}$$

using draws  $\{\mathbf{g}_j^{(t)}\}$  from the main run, and the conditional independence of the densities in Steps 3 and 4 of Algorithm 2. This approach saves further reduced runs, and reduces the numerical standard error of the estimate, because it does not require knowledge of the reduced conditional ordinates for  $\{\mathbf{g}_j\}$ . However, the computational burden is  $O(n^3)$  since the covariance matrix of the density  $f(y | \{\tau_i^{2*}, \sigma^{2*}\})$  is  $n \times n$  and is not banded, which can become excessive with large data sets (large  $n$ ). Nonetheless, this decomposition is useful when  $n$  is small, and we will use it to evaluate the precision of the reduced run approach, which we present next.

We now turn to the more general way of finding the marginal likelihood, one that can be applied even when the sample size is large. This procedure relies on  $p - 1$  additional reduced runs after the main MCMC run and, for this reason, we refer to it as the reduced run approach. It is based on the following expression of the marginal likelihood

$$m(\mathbf{y}) = \frac{f(\mathbf{y} | \{\tau_i^{2*}\}, \sigma^{2*}, \{\mathbf{g}_j^*\}) \pi(\tau_i^{2*}, \sigma^{2*}, \{\mathbf{g}_j^*\})}{\pi(\{\tau_i^{2*}\}, \sigma^{2*}, \{\mathbf{g}_j^*\} | \mathbf{y})},$$

where the  $\{\mathbf{g}_j^*\}$  are not marginalized out analytically. In estimating  $\pi(\{\tau_i^{2*}\}, \sigma^{2*}, \{\mathbf{g}_j^*\} | \mathbf{y})$ , the reduced run approach relies on the decomposition

$$\pi(\{\tau_i^{2*}\}, \sigma^{2*} | \mathbf{y}) \prod_{i=1}^p \left\{ \pi(\mathbf{g}_i^* | \mathbf{y}, \{\tau^{2*}\}, \sigma^{2*}, \{\mathbf{g}_j^*\}_{j < i}) \right\}. \quad (14)$$

The  $\{\mathbf{g}_j\}$  are placed last in (14) to ensure that the estimate is stable, because each  $\mathbf{g}_j$  may potentially be of dimension up to the total number of observations  $n$  in the sample. We note, however, that because the simulation algorithm for  $\{\mathbf{g}_j\}$  is  $O(n)$  as discussed in Section 2, this particular choice comes at a small increase in computational cost, while the statistical efficiency benefits may be substantial, especially for large-dimensional problems. We have presented results for an alternative decomposition in Section 5 below, where the  $\{\mathbf{g}_j\}$  are placed at the beginning of the posterior ordinate decomposition, but in line with the arguments above, we found that (14) reduced variability in the marginal likelihood estimate, and worked well for both small- and large-dimensional  $\{\mathbf{g}_j\}$ .

An important special case is the univariate nonparametric model (8), where the marginal likelihood estimation by the method of Chib (1995) will not require any reduced runs, because in the decomposition

$$\pi(\tau^{2*}, \sigma^{2*} | \mathbf{y}) \pi(\mathbf{g}^* | \mathbf{y}, \tau^{2*}, \sigma^{2*})$$

of posterior ordinate in the denominator of (12) the first term can be estimated with draws from the main run (that quantity is the average of the product of two independent inverse gammas with shape and scale parameters given in Steps 2 and 3 of Algorithm 1), while the second term is available immediately (it is the density ordinate of a multivariate normal density with a banded precision matrix).

It is interesting to note the following feature of the marginal likelihood for additive models. Suppose that we have formulated the priors for the various functions  $\{\mathbf{g}_j\}$ . After deciding which  $p - 1$  functions will be centered, the centering constraints introduced in Section 2.2 will imply that the prior on the first function will have to be adjusted to absorb the sum of the constants of the remaining functions. But does that imply that we must also change the priors on the last  $p - 1$  functions? The interesting result is that those priors can be kept the same, and doing so will not influence the marginal likelihood. This can be easily seen from the fact that if the likelihood does not depend on a given parameter that is present in a proper prior (the constant term in the centered

functions in our case), then the marginal likelihood will not depend on the particular prior for that parameter. In general, if the likelihood is given by  $f(\mathbf{y}|\theta_1)$  and we have the prior  $\pi(\theta_1, \theta_2)$ , then the marginal likelihood

$$\begin{aligned} m(y) &= \int f(\mathbf{y}|\theta_1) \pi(\theta_1, \theta_2) d\theta_1 d\theta_2 \\ &= \int f(\mathbf{y}|\theta_1) \pi(\theta_1) \int \pi(\theta_2|\theta_1) d\theta_2 d\theta_1 \\ &= \int f(\mathbf{y}|\theta_1) \pi(\theta_1) d\theta_1, \end{aligned}$$

so it will not be influenced by the (proper) prior on  $\theta_2$ . In practice this is important for modeling, because it implies that once the priors on the functions  $\{\mathbf{g}_j\}$  are determined, the particular choice of which functions to center in the identification scheme of Section 2.2 will require a revision of only one of these priors, and the remaining priors can stay as initially specified, without impacting model comparison.

We now turn attention to an important computational aspect used in the estimation of the marginal likelihood. While banded matrix algorithms are applicable to the calculation of the exponents of the full conditional densities for the functions  $\mathbf{g}_j$ ,  $j = 2, \dots, p$ , calculating the determinants of the posterior precision matrices may be more computationally intensive because those matrices are not banded but rather have the form  $(\mathbf{A} - \mathbf{u}\mathbf{u}')$ , as in Remark 2. However, because of the special structure of the covariance matrices, we can find the determinant in  $O(n)$  operations. This is possible because

$$\begin{aligned} \det(\mathbf{A} - \mathbf{u}\mathbf{u}') &= \det\{\mathbf{A}(\mathbf{I} - \mathbf{A}^{-1}\mathbf{u}\mathbf{u}')\} \\ &= \det(\mathbf{A}) \det(1 - \mathbf{u}'\mathbf{A}^{-1}\mathbf{u}), \end{aligned}$$

where the second line follows from the identity  $\det(\mathbf{I} + \mathbf{A}\mathbf{B}) = \det(\mathbf{I} + \mathbf{B}\mathbf{A})$ , and where  $(1 - \mathbf{u}'\mathbf{A}^{-1}\mathbf{u})$  is a scalar (we used the definition  $\lambda = \mathbf{u}'\mathbf{A}^{-1}\mathbf{u}$  in Remark 2). Because of the computational shortcuts that are afforded by the identification scheme in Section 2.2, the calculation of the posterior ordinates only takes  $O(n)$  operations.

In closing, we mention that the numerical standard error of the marginal likelihood estimate can be derived by following the method given by Chib (1995). In this context, the numerical standard error provides the variation that can be expected in the marginal likelihood estimate if

the simulation were to be repeated. To check the accuracy of the numerical standard error estimates obtained by Chib's (1995) approach in the presence of high dimensional blocks such as the  $\{\mathbf{g}_j\}$ , we used 50 simulations to find a second estimate of the variability of the marginal likelihood estimate. The two estimates were found to be virtually identical.

## 5 Simulation Study

The key aspect of our implementation is that it relies on a fully Bayesian, finite sample methodology for the analysis of the additive model in Section 2. This is enabled by our use of proper priors for the parameters and the unknown function  $g(\cdot)$  as discussed in Section 3, under the identification scheme introduced in Section 2.2, and may be contrasted with previous studies (Silverman (1985), Wood and Kohn (1998), Hastie and Tibshirani (2000), Fahrmeir and Lang (2001)) where partially improper priors were used. In our simulation study we demonstrate the performance of the estimation technique proposed in Section 3.2. A second goal is to evaluate and compare the performance of the marginal likelihood estimation procedures of Section 4. Finally, we contrast the results from the marginal likelihood approach to those from AIC and BIC.

### 5.1 Estimation

The posterior mean estimates  $\hat{\mathbf{g}}_j = E\{\mathbf{g}_j|\mathbf{y}\}$ ,  $j = 1, \dots, p$ , are found from MCMC runs of length 10000 following burn-ins of 1000 draws. We calculate mean squared errors for the estimates of the unknown function, which are reported for several designs. We also demonstrate the performance of the MCMC algorithm by reporting the autocorrelations and the inefficiency factors for the sampled parameters under alternative model specifications and sample sizes. We find that the MCMC algorithm performs very well and that its performance improves with larger sample sizes. Data are simulated from an additive model in (1) with the following three functional specifications:

1.  $g_1(s) = \sin(2\pi s)$ , for  $s \in [0.6, 1.4]$ ;
2.  $g_2(s) = -1 + s + 1.6s^2 + \sin(5s)$ , for  $s \in [0, 1.1]$ ;
3.  $g_3(s) = -0.8 + s + \exp\{-30(s - 0.5)^2\}$ , for  $s \in [0, 1]$ .

The three functions are plotted in Figure 1. Each of them is evaluated on a regular grid of

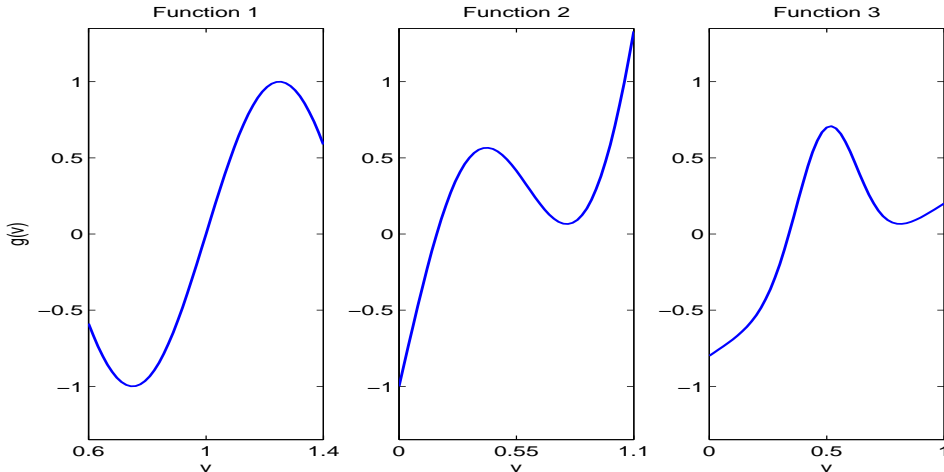


Figure 1: The three true functions in the simulation study.

$m = 51$  points. We have chosen these functions to capture a range of specifications used in the literature – for example, the first function achieves its extrema in the interior of its domain, while the second does so at the endpoints of the domain; the third function has a minimum at the end, and a maximum in the interior, of its domain. In addition, the first function is symmetric, while the other two are asymmetric. Data are generated from the additive model with  $\sigma = 0.25$ , and some of the resulting descriptive statistics for the three functions presented in Table 1.

Generated Functions			
	$g_1$	$g_2$	$g_3$
$SD(g_i)/\sigma$	3.10	1.81	1.85
$Range(g_i)/\sigma$	7.99	9.32	6.03

Table 1: Descriptive statistics for the functions in the simulation study.

As already discussed, due to the centering constraints in additive models, the estimated functions will generally be vertical translations of the true functions – the first function will absorb the overall constant, while the remaining functions will be centered. We therefore gauge the performance of the method in fitting the above functions using mean squared error applied to the appropriately translated true functions  $\{\tilde{g}_j(\cdot)\}$ .

$$MSE_i = \frac{1}{m} \sum_{j=1}^m \{\hat{g}_i(v_j) - \tilde{g}_i(v_j)\}^2.$$

The average  $MSE_i$ , together with the standard errors based on 15 data samples, are reported in Table 2 for several sample sizes. The results in Table 2 illustrate that as the sample size grows, the

Average Mean Squared Errors			
Observations	$g_1$	$g_2$	$g_3$
$n = 250$	0.00260 (0.00157)	0.00514 (0.00437)	0.00837 (0.00848)
$n = 500$	0.00088 (0.00052)	0.00292 (0.00486)	0.00302 (0.00268)
$n = 1000$	0.00055 (0.00043)	0.00235 (0.00187)	0.00245 (0.00301)

Table 2: Average mean squared errors based on 15 samples, with estimated standard errors in parentheses.

functions are estimated more and more precisely, as expected. As an illustration of the technique, in Figure 2 we show several regressions for the three sample sizes simulated above. In Figure 2 the original functions are now vertically translated relative to those in Figure 1 because of the centering constraints.

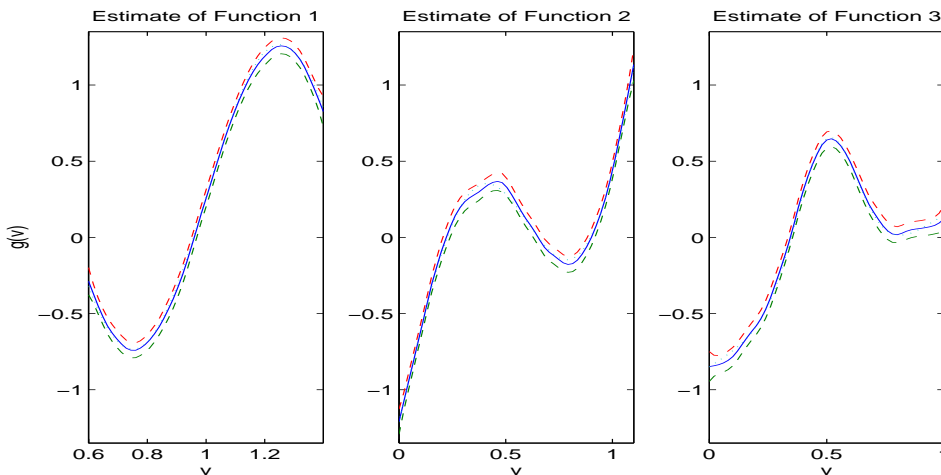


Figure 2: Simulation Study. Three examples of estimated functions (solid lines), true functions (dotted lines), and confidence bands at two standard deviations (dashed lines).

An example of the performance of the MCMC sampler for the problem with  $n = 500$  is illustrated in Figure 3. Table 3 shows the inefficiency factors corresponding to the parameters for the same model. The parameters appear to be estimated very well.

In summary, the results suggest that the MCMC algorithm performs well, and that the estimation method recovers the parameters and functions used to generate the data. The performance of the method in recovering the nonparametric function  $g(\cdot)$  and the model parameters improves with the sample size, as expected.

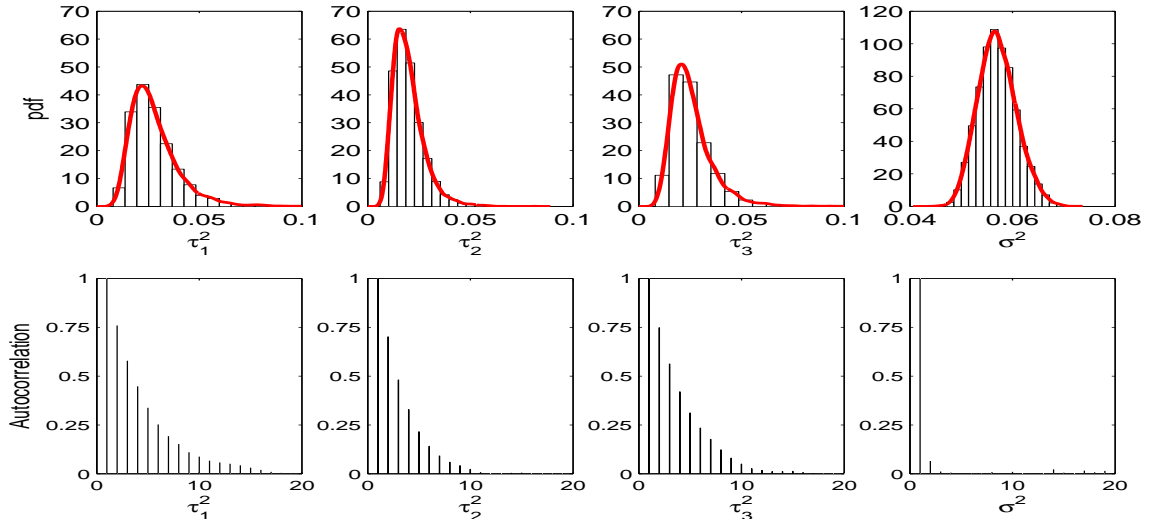


Figure 3: Posterior samples and autocorrelations for the parameters of a nonparametric additive model with three unknown functions.

	Inefficiency Factors			
	$\tau_1^2$	$\tau_2^2$	$\tau_3^2$	$\sigma^2$
$n = 250$	8.885	6.261	8.286	1.220
$n = 500$	6.684	4.872	5.488	1.068
$n = 1000$	5.758	5.689	4.966	1.000

Table 3: Examples of estimated inefficiency factors (autocorrelation times) for the parameters of the additive model for various sample sizes.

## 5.2 Performance of the Model Choice Method

As discussed in Section 4, some care is required in applying the identity (12) because the dimension of the vectors of functional evaluations  $\{\mathbf{g}_j\}$  may be very large, with each  $\mathbf{g}_j$  potentially as large as the sample size  $n$ . For this reason, we recommended the direct marginalization method which marginalizes out the  $\{\mathbf{g}_j\}$  analytically, and the reduced run method in which the calculation of the posterior ordinates for  $\{\mathbf{g}_j\}$  is done only after all remaining parameters in the posterior decomposition have been fixed. To illustrate that these methods outperform an alternative decomposition which places the posterior ordinates of  $\{\mathbf{g}_j\}$  at the beginning of the decomposition (13), we applied this alternative approach to a simulated data set. The alternative decomposition produced highly variable results, even for small  $m$ . Specific results for  $m = 25$  are presented in Table 4 for the three methods. The table shows that the direct marginalization method and the reduced run method perform best, and their numerical standard errors are approximately ten times lower than those

Numerical Standard Error of the Marginal Likelihood Estimate			
	Reduced Run	Direct Marginalization	Alternative Decomposition
$n = 250$	0.013	0.013	0.150
$n = 500$	0.009	0.009	0.116
$n = 1000$	0.008	0.008	0.083

Table 4: Numerical standard error of the marginal likelihood estimates for three marginal likelihood decompositions for various sample sizes.

from the alternative decomposition where the ordinates for  $\{\mathbf{g}_j\}$  were placed at the beginning of (13). What is more, with larger  $m$ , the alternative decomposition required much longer MCMC runs to produce accurate estimates. Because these results provide a clear illustration of the advantages of the reduced run and the direct marginalization methods, we only focus on these methods in this paper.

The direct marginalization method is preferable for smaller sample sizes, because the evaluation of the ordinate  $f(y|\sigma^{2*}, \{\tau^{2*}\})$  requires  $O(n^3)$  operations and  $O(n^2)$  storage spaces. The reduced run method is applicable for larger sample sizes, even though it requires  $p-1$  reduced runs, because only univariate quantities are stored during the reduced runs, and each of the runs requires  $O(nT)$  operations, where  $T$  is the MCMC simulation size. Although there will be considerable variation in computing times due to differences in computer platforms and software, in the context of this example, our experiments have shown that the single evaluation of  $f(y|\sigma^{2*}, \{\tau^{2*}\})$  may take longer than doing the two reduced runs when  $n$  is larger than approximately 2000.

To evaluate the statistical properties of the direct marginalization and the reduced run methods, we next consider the numerical standard errors of the marginal likelihood estimates obtained by these two methods. Table 5 presents results for  $n = 2500$ . The table illustrates that the numerical

Numerical Standard Error of the Marginal Likelihood Estimate		
	Reduced Run	Direct Marginalization
$m = 501$	0.049	0.049
$m = 1001$	0.061	0.059
$m = 1501$	0.086	0.065

Table 5: Numerical standard error of the marginal likelihood estimates for three  $m$ .

standard error of the reduced run method is not larger than that of the direct marginalization method, even though several additional quantities are estimated in reduced runs. Because the

dimensions of the  $\{\mathbf{g}_j\}$  may vary across data samples, this table also presents results on the effect of  $m$  on the precision of the marginal likelihood estimate. We see that the precision of the direct method is very high, and that even though there are high-dimensional blocks in the posterior ordinate in the reduced run method, it is nonetheless very precise.

To assess the performance of the model selection technique in practice, we show the results on comparing three nonparametric models. Realizations were generated from an additive model using two covariates and the first two regression functions in Figure 1. The first fitted model used only the first covariate, omitting the second relevant covariate. The second model used both relevant covariates, while the third model included an additional covariate that was not used in the generation of the data. We performed 50 data replications for three possible sample sizes, where for each data sample we estimated all three models and their marginal likelihoods. The model with the highest marginal likelihood was selected. Table 6 illustrates the results of this Monte Carlo study. The results in Table 6 are quite encouraging, and illustrate that marginal likelihoods can

Sample Size	Selection Frequency		
	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$
$n = 250$	0.00	1.00	0.00
$n = 500$	0.00	1.00	0.00
$n = 1000$	0.00	1.00	0.00

Table 6: Proportion of times each model was selected using the marginal likelihood estimates. The model used to generate the data was  $\mathcal{M}_2$ . The models  $\mathcal{M}_1$ ,  $\mathcal{M}_2$ , and  $\mathcal{M}_3$  include one, two, or three covariates, respectively.

guard against both underparameterization and overparameterization.

### 5.2.1 Comparison with AIC and BIC

Until now, marginal likelihoods and Bayes factors have not been calculated for nonparametric additive models, and instead comparisons have made use of the AIC and BIC measures. In the context of the nonparametric additive model, the BIC approach of Schwarz (1978) approximates the logarithm of the marginal likelihood  $\log m(y)$  by  $\log f\left(\mathbf{y} \mid \left\{\hat{\tau}_j^2\right\}, \hat{\sigma}^2\right) - (q/2) \log(n)$ , where  $q = p + 1$  is the number of parameters in the model, and  $\left\{\hat{\tau}_j^2\right\}_{j=1}^p$  and  $\hat{\sigma}^2$  are the values that maximize  $f\left(\mathbf{y} \mid \left\{\tau_j^2\right\}, \sigma^2\right)$ . In this expression, the term  $-(q/2) \log(n)$  is the penalty on complexity. The related AIC criterion (Akaike 1973) is given by  $\log f\left(\mathbf{y} \mid \left\{\hat{\tau}_j^2\right\}, \hat{\sigma}^2\right) - q$ .

The use of these measures is not without its disadvantages. First, the calculation of the AIC or BIC can be demanding for large samples because

$$f(\mathbf{y} | \{\tau_j^2\}, \sigma^2) = \int f(\mathbf{y} | \{\tau_j^2\}, \{\mathbf{g}_j\}, \sigma^2) \pi(\{\mathbf{g}_j\} | \{\tau_j^2\}) d\{\mathbf{g}_j\}$$

is a high dimensional integral, and each evaluation requires  $O(n^3)$  operations, and of course we need multiple evaluations for optimization. Because of the high dimensionality and computational costs, optimization will rarely be a viable option. For this reason Wood, Kohn, Shively, and Jiang (2002) have used a simulation-based optimization method. According to that method, given the sample of MCMC draws  $\left\{ \left\{ \tau_j^2 \right\}^{(t)}, \left\{ \mathbf{g}_j \right\}^{(t)}, \sigma^{2,(t)} \right\}$ , one simply takes the draw which maximizes  $\pi\left(\left\{ \tau_j^2 \right\}, \left\{ \mathbf{g}_j \right\}, \sigma^2 | \mathbf{y}\right)$ , and using the values for  $\left\{ \tau_j^2 \right\}$  and  $\sigma^2$  from that draw, one computes  $f\left(\mathbf{y} | \left\{ \tau_j^2 \right\}, \sigma^2\right)$ . The problem with this method is that the mode of the joint density over  $\left(\left\{ \tau_j^2 \right\}, \left\{ \mathbf{g}_j \right\}, \sigma^2\right)$ , need not be the same as the mode of the marginal density over  $\left(\left\{ \tau_j^2 \right\}, \sigma^2\right)$ . In fact, our experiments have shown that frequently it is the case that  $f\left(\mathbf{y} | \left\{ \tau_j^2 \right\}, \sigma^2\right)$  will have a higher value evaluated at the mean, rather than at the mode of the joint distribution. The problem of obtaining reliable MLE estimates still remains open.

From a methodological perspective, the AIC and BIC measures are based on asymptotic considerations and they do not, in a formal sense, produce posterior model probabilities for the given set of models. Moreover, as discussed in Wood *et al.* (2002), one may have to be concerned with boundary complexity problems when comparing alternative models.

## 6 Model Extensions

The estimation techniques presented in this paper are quite general and readily applicable in many settings. Here we briefly outline some of them. One important extension of the techniques is to partially linear (semiparametric) models, where the effect of a given covariate  $s$  is modeled nonparametrically as in Section 3, but the model also includes linear effects of the type  $\mathbf{X}\beta$  which are added to the mean function on the right hand side of (9). Estimation of such a partially linear is a straightforward extension of Algorithm 1, and proceeds by recursively forming partial residuals  $\mathbf{y} - \mathbf{X}\beta$  when estimating  $\mathbf{g}$ , and  $\mathbf{y} - \mathbf{Q}\mathbf{g}$  when estimating  $\beta$  (and the quantities are sampled from the standard posterior updates). The partially linear model can also easily be extended to the semiparametric additive case (with several nonlinear functions in addition to the linear effects) by

pursuing a similar strategy in the context of Algorithm 2. Of course, if  $\mathbf{X}$  contains an intercept, for identification reasons all of the unknown functions will have to be centered and sampling will proceed as in Step 2 of Algorithm 2. If  $\mathbf{X}$  does not contain an intercept, all but one of the functions should be centered.

The methods in this paper, including the above extensions, can also be applied to the analysis of binary and polychotomous regression using the latent variable augmentation framework of Albert and Chib (1993). In that framework, the MCMC sampler explicitly includes the (continuous) latent variables underlying the (discrete) observed responses. The main advantage of this approach is that conditionally on the latent data, estimation of the parameters and the unknown functions closely mirrors the methods for continuous data. Another benefit of the approach is that it allows for analysis under various link function specifications, such as probit and Student- $t$  links (Albert and Chib 1993), as well as under any mixture-of-normals link function, including the logit link function (Wood and Kohn 1998).

Finally, the approach provided in this paper is also applicable to the class of continuous and binary data additive mixed models (Lin and Zhang 1999). For example, Chib and Jeliazkov (2005) discuss the specification and estimation of a semiparametric partially linear model for dynamic binary panel data. The method is based on the latent variable augmentation of Albert and Chib (1993) and on the efficient sampling algorithms for mixed effect models provided in Chib and Carlin (1999). The estimation algorithm proposed by Chib and Jeliazkov (2005) can easily be modified to account for a semiparametric or nonparametric additive structure simply by adapting the methods from Section 2 above.

## 7 Application To Exam Score Data

As an illustrative example we apply the methodology to the final exam data in Wooldridge (2002). The data include 680 observations on the standardized score ( $y$ ) on the final exam in microeconomic principles. To obtain the standardized outcome  $y$ , the score on a student's final exam is represented in terms of the number of standard deviations away from the class mean. This transformation makes the performance measure interpretable in relation to the rest of the class, and helps to ensure that the regression assumptions are not violated. In the analysis we use the following independent

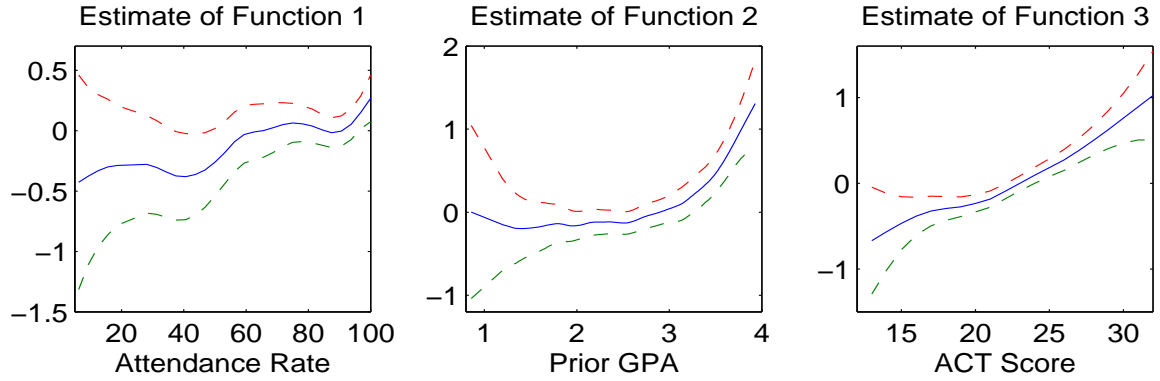


Figure 4: The three estimated functions functions for the exam data.

variables: class attendance rate ( $s_1$ ), cumulative college GPA prior to taking the class ( $s_2$ ), and ACT exam score ( $s_3$ ) (the ACT is an exam that is widely used in college admissions). The data are modeled as the additive nonparametric regression

$$y_i = g_1(s_{i1}) + g_2(s_{i2}) + g_3(s_{i3}) + \varepsilon_i, \quad (i = 1, \dots, 680),$$

where each of the functions  $g_1(s_{i1})$ ,  $g_2(s_{i2})$ ,  $g_3(s_{i3})$  is *a priori* modeled as a second order Markov process, and is appropriately restricted for likelihood identifiability as outlined in Section 2.2. Estimation of the model was done after a warm-up period of 1000 iterations followed by a main simulation run of 10000 MCMC draws. Estimates of the model parameters are presented in Table 7, and the three functions are plotted in Figure 4.

Parameter	Mean	SD	Median	Lower	Upper	Ineff
$\tau_1^2$	0.002	0.001	0.002	0.001	0.005	5.580
$\tau_2^2$	0.002	0.001	0.002	0.001	0.004	22.750
$\tau_3^2$	0.003	0.002	0.003	0.001	0.010	5.420
$\sigma^2$	0.755	0.042	0.753	0.677	0.841	1.000

Table 7: Parameter estimates and summaries for the exam data model. The table also reports 95% confidence intervals and sampling inefficiency factors.

From Table 7 we see that the parameters are estimated well and have generally low autocorrelations. The marginal likelihood estimates are also well behaved – the marginal likelihood of the model was estimated to be  $-907.031$  with numerical standard error of 0.040. From Figure 4 we see that the nonparametric estimates of the three functions depict a relationship that is consistent with commonly held notions of the determinants of good academic performance. For example, students with demonstrated scholastic abilities, as demonstrated by a cumulative GPA above 3.0 or high

ACT scores, appear much more likely to score above average on the final exam. The general effect of attendance is also positive, but there appear to be interesting nonlinearities in the data around 42 and 85 percent.

In addition to this model, we were interested in the hypothesis that class attendance might have a different effect for students who have performed differently in the past (as measured either by prior cumulative GPA or by ACT scores). For this reason we estimated additive models including the function  $g_4(s_1s_2)$  or  $g_4(s_1s_3)$ , however these models did not perform competitively with the model above on the basis of their marginal likelihoods. This indicates that additivity is a reasonable restriction in this setting and the structure of the model above is sufficient to capture the effects of attendance on exam performance.

## 8 Concluding Remarks

This article has examined the specification, estimation, and comparison of nonparametric additive models based on proper smoothness priors. A new scheme is used to identify the unknown covariate functions and an efficient Markov chain Monte Carlo sampling procedure is developed for estimating the model. We discuss methods for finding the marginal likelihood based on the framework of Chib (1995), thus paving the way for a comparison of additive models on the basis of Bayes factors. We provide a simulation study to demonstrate the applicability of the estimation and model choice methods. The techniques are quite general and are applicable to semiparametric, mixed effect, and discrete data models.

## References

- Albert, J. and S. Chib (1993), “Bayesian Analysis of Binary and Polychotomous Response Data,” *Journal of the American Statistical Association*, 88, 669–679.
- Besag, J., P. Green, D. Higdon, and K. Mengersen (1995), “Bayesian Computation and Stochastic Systems” (with discussion), *Statistical Science*, 10, 3-66.
- Chib, S. (1995), “Marginal Likelihood from the Gibbs Output,” *Journal of the American Statistical Association*, 90, 1313–21.
- Chib, S. and B. Carlin (1999): “On MCMC Sampling in Hierarchical Longitudinal Models,” *Statistics and Computing*, 9, 17-26.
- Chib, S. and I. Jeliazkov (2005), “Inference in Semiparametric Dynamic Models for Binary Longitudinal Data,” *Journal of the American Statistical Association*, forthcoming.

- Denison, D. G. T., B. K. Mallick, and A. F. M. Smith (1998), “Automatic Bayesian Curve Fitting,” *Journal of the Royal Statistical Society, B*, 60, 333-350.
- DiMatteo, I., C. R. Genovese, and R. E. Kass (2001), “Bayesian Curve-Fitting with Free-Knot Splines,” *Biometrika*, 88, 1055-1071.
- Fahrmeir, L. and G. Tutz (1997), “Multivariate Statistical Modelling Based on Generalized Linear Models.” New York: Springer-Verlag.
- Fahrmeir, L. and S. Lang (2001), “Bayesian Inference for Generalized Additive Mixed Models Based on Markov Random Field Priors,” *Journal of the Royal Statistical Society, C*, 50, 201-220.
- Gersovitz, M. and J. MacKinnon (1978): “Seasonality in Regression: An Application of Smoothness Priors,” *Journal of the American Statistical Association*, 73, 264–273.
- Hansen, M. H., and C. Kooperberg (2002), “Spline Adaptation in Extended Linear Models” (with discussion), *Statistical Science*, 17, 2-51.
- Hastie, T. and R. Tibshirani (1990), *Generalized Additive Models*. New York: Chapman & Hall.
- Hastie, T. and R. Tibshirani (2000), “Bayesian Backfitting” (with discussion), *Statistical Science*, 15, 196-223.
- Lin, X., and D. Zhang (1999), “Inference in Generalized Additive Mixed Models by Using Smoothing Splines,” *Journal of the Royal Statistical Society, B*, 61, 381-400.
- Müller, P., G. Rosner, L. Inoue, and M. Dewhurst (2001): “A Bayesian Model for Detecting Acute Change in Nonlinear Profiles,” *Journal of the American Statistical Association*, 96, 1215–1222.
- Shiller, R. (1973): “A Distributed Lag Estimator Derived From Smoothness Priors,” *Econometrica*, 41, 775-788.
- Shiller, R. (1984): “Smoothness Priors and Nonlinear Regression,” *Journal of the American Statistical Association*, 79, 609–615.
- Shively, T. S., R. Kohn, and S. Wood (1999), “Variable Selection and Function Estimation in Additive Nonparametric Regression Using a Data-Based Prior” (with discussion), *Journal of the American Statistical Association*, 94, 777-806.
- Silverman, B. (1985): “Some Aspects of the Spline Smoothing Approach to Non-parametric Regression Curve Fitting” (with discussion), *Journal of the Royal Statistical Society, B*, 47, 1-52.
- Wahba, G. (1978), “Improper Priors, Spline Smoothing and the Problem of Guarding Against Model Errors in Regression,” *Journal of the Royal Statistical Society, B*, 40, 364-372.
- Wahba, G. (1990), *Spline Models for Observational Data*. Philadelphia: SIAM.
- Wood, S. and R. Kohn (1998): “A Bayesian Approach to Robust Binary Nonparametric Regression,” *Journal of the American Statistical Association*, 93, 203–213.
- Wood, S., R. Kohn, T. Shively, and W. Jiang (2002), “Model Selection in Spline Nonparametric Regression,” *Journal of the Royal Statistical Society, B*, 119-139.

- Whittaker, E. (1923): "On a New Method of Graduation," *Proceedings of the Edinburgh Mathematical Society*, 41, 63-75.
- Whittaker, E., and C. Robinson (1924), *The Calculus of Observations*. Glasgow: Blackie & Son.
- Wooldridge, J. M. (2002), *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.