Journal of Econometrics 156 (2010) 322–336

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Additive cubic spline regression with Dirichlet process mixture errors[☆]

Siddhartha Chib[a,*], Edward Greenberg[b]

[a] Olin Business School, Washington University in St. Louis, St. Louis MO 63130, United States

[b] Department of Economics, Washington University in St. Louis, St. Louis MO 63130, United States

ARTICLE INFO

Article history:
Received 5 May 2008
Received in revised form
23 October 2009
Accepted 7 November 2009
Available online 22 November 2009

Keywords:
Additive regression
Bayes factors
Cubic spline
Non-parametric regression
Dirichlet process
Dirichlet process mixture
Marginal likelihood
Markov chain Monte Carlo
Metropolis–Hastings
Model comparison
Ordinal data

ABSTRACT

The goal of this article is to develop a flexible Bayesian analysis of regression models for continuous and categorical outcomes. In the models we study, covariate (or regression) effects are modeled additively by cubic splines, and the error distribution (that of the latent outcomes in the case of categorical data) is modeled as a Dirichlet process mixture. We employ a relatively unexplored but attractive basis in which the spline coefficients are the unknown function ordinates at the knots. We exploit this feature to develop a proper prior distribution on the coefficients that involves the first and second differences of the ordinates, quantities about which one may have prior knowledge. We also discuss the problem of comparing models with different numbers of knots or different error distributions through marginal likelihoods and Bayes factors which are computed within the framework of Chib (1995) as extended to DPM models by Basu and Chib (2003). The techniques are illustrated with simulated and real data.

1. Introduction

Our objective in this article is to specify and estimate flexible Bayesian regression models for continuous and categorical outcomes. By flexible we mean models that are relatively free of assumptions about the functional forms through which covariates affect the response and of assumptions about the distribution of the unobserved error or, in the case of categorical outcomes, the distribution of the underlying latent data. Flexibility in the choice of such assumptions is especially desirable in fields such as biostatistics and the social sciences, where theory rarely provides guidance either about the form of the covariate effects, other than the presumption that the effects are smooth in the covariates, or about the error distribution.

The regression function has been modeled in several ways (see, for example, O'Hagan (1978); Angers and Delampady (1992); Müller et al. (1996); Chipman et al. (1997); Clyde et al. (1998); Vannucci and Corradi (1999)). In this article, we assume that the regression function is additive with each function of the covariates modeled as a cubic spline (for example, Härdle, 1990; Green and Silverman, 1994; Pagan and Ullah, 1999; Li and Racine, 2006). In this approach, it is necessary to specify a set of basis functions for the cubic spline. In the practice to date (for example, Congdon, 2007, chap. 4; Denison et al., 2002; Ruppert et al., 2003, chap. 16) attention has been restricted to the truncated power series basis and polynomial *B*-spline basis. In each case, the parameters of the basis functions have no easy interpretation. From the Bayesian viewpoint, the lack of interpretability of the coefficients is inconvenient because it hinders the construction of proper prior distributions that can be motivated by defensible a priori reasoning. The common strategy of specifying improper or default prior distributions is not satisfactory if the goal is to compare alternative non-parametric formulations through such formal means as marginal likelihoods and Bayes factors.

One innovation of this article is in the use of a relatively unexplored basis in which the spline coefficients have the attractive feature of being the unknown function ordinates at the knots. We exploit this feature to develop a proper prior distribution on the coefficients that involves the first and second differences of the ordinates, quantities about which one may be expected to have some prior knowledge. We also indicate how a simulation-based approach can be used to specify the hyperparameters of our

[☆] We thank the editor and four referees for their many useful comments on the article.

[*] Corresponding author. Tel.: +1 314 935 4657.
E-mail addresses: chib@wustl.edu (S. Chib), edg@artsci.wustl.edu (E. Greenberg).

0304-4076/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2009.11.002

prior distribution. The basis we employ is described in Lancaster and Šalkauskas (1986, secs. 3.7 and 4.2). We, henceforth, refer to it as the LS basis. This basis is mentioned briefly by Wood (2006, Sec. 4.1.2) in relation to cubic regression splines, but it is not used in the computations. Other bases parameterized in terms of ordinate values appear in Poirier (1973) and Green and Silverman (1994), but without any connection to Bayesian problems or to the issues that concern us in this article.

In the spline literature, the error distribution is usually assumed to be parametric, especially when the outcome is not continuous. In ordinal models, which we use as the running example for categorical outcomes, the model is almost always specified with logit or probit links. One of our goals is to show that it is not difficult to move beyond such parametric assumptions. In our development, we assume that the distribution of the error is an unknown parameter that we model as a Dirichlet process mixture (DPM). The DPM is a general family of prior distributions on probability measures that was introduced by Ferguson (1973) and Antoniak (1974) and has found many applications in statistics and econometrics. Some of the early uses of this specification in economics are Tiwari et al. (1988), Hirano (2002), and Chib and Hamilton (2002). In modeling the error distribution for continuous outcomes, we assume that, conditioned on an unknown location and variance, the error distribution is normal. We then assume that the location and variance of this normal distribution have an unknown distribution that is modeled as a Dirichlet process, leading to an error distribution that is an arbitrary location–variance mixture of normal distributions. A key property of this DPM specification, established by Ferguson (1983), is that it is capable of approximating any unknown distribution. An alternative approach, which we do not pursue, could involve the weighted mixtures of Dirichlet processes (Dunson et al., 2007). For ordinal outcomes, where the ordered category probabilities are defined in terms of an increasing sequence of cut-points, we invoke a version of the DPM model in which mixing occurs only over the variance parameter because in that case the unknown location is confounded with the cut-points.

An important characteristic of our models is that they are easy to understand and estimate. The models can be expressed in the form of a linear regression for observed outcomes in the case of continuous outcomes and for latent outcomes in the case of ordinal outcomes. The predictors in these regressions are derived from the data on the underlying covariate and our basis functions for the cubic spline. The coefficients are the unknown function ordinates at the various knots. The derivative of each function with respect to its covariate is easily calculated. As far as estimation is concerned, the posterior distribution of the model parameters and other unknowns can be summarized by relatively straightforward Markov chain Monte Carlo (MCMC) methods. For continuous outcomes, conditioned on the parameters of the DPM process, the set-up is similar to a Gaussian heteroskedastic regression model, which simplifies several sampling steps. Similarly, conditioned on all the other unknowns and the data, the sampling of the DPM parameters is done according to the methods of Escobar and West (1995) and MacEachern and Müller (1998). The fitting is completed by steps in which the unknown smoothness parameters are sampled. The MCMC sampling algorithm for the ordinal model is similar in the latent variable framework of Albert and Chib (1993). It differs from the continuous model because the posterior distribution includes the cut-points and the latent variables that are introduced to model the ordinal outcomes.

A major focus of our work is on the comparison of different versions of our models (defined, for example, by alternative covariates, fewer or additional knots, or parametric assumptions about the error distribution). For this purpose, we discuss the computation of marginal likelihoods and Bayes factors. We provide algorithms for computing the marginal likelihood for both the continuous and ordinal models within the framework of Chib (1995) as extended to DPM models by Basu and Chib (2003). These algorithms are not complicated and require virtually the same code that is used in the fitting of the models.

Our article can be viewed as a contribution to an emerging literature on flexible Bayesian regression models. For instance, Leslie et al. (2007) pursue similar objectives but in the context of regression splines and with a different basis than ours. The article does not consider the question of the prior on the spline coefficients or the computation of the marginal likelihood. Griffin and Steel (2007) analyze a new Dirichlet process regression smoother in which the functional form for the covariate structure is centered over a class of regression models rather than taking the form of a spline. Finally, Geweke and Keane (2007) and Villani et al. (2007) consider a Bayesian regression model in which the error distribution is modeled by a discrete mixture of normal variables. The mean function in the former is modeled by general quadratic, cubic, and quartic polynomials in two covariates, while splines are used in the latter. These two articles primarily focus on time series problems and do not tackle the question of Bayes factors for model comparisons. None of these four contributions extend their methods to models of ordinal outcomes.

The rest of the article is organized as follows. In Section 2, we present the basic models. Section 3 contains the cubic spline basis for modeling the unknown covariate functions and introduces the identifying restrictions. We specify the prior distribution for the parameters in Section 4 and, in Section 5, develop the prior-posterior analyses of the models and show how the posterior distribution of the unknowns can be summarized by MCMC methods. The computation of the marginal likelihood is considered in Section 6. Section 7 deals with some special cases. Examples with simulated and real data are contained in Section 8. Section 9 has our conclusions.

2. Models

2.1. Continuous outcomes

Assume that y_i is the ith observation in a sample of n observations $y = (y_1, \dots, y_n)$ and that the model generating y_i depends on a k_0-vector of covariates x_{i0}, consisting of an intercept and nominal variables, and q additional covariates $w_{i1}, \dots, w_{iq}$. Now, let

$$y_i = x'_{i0}\beta_0 + g_1(w_{i1}) + \cdots + g_q(w_{iq}) + \varepsilon_i, \quad i \le n, \tag{2.1}$$

where the $g_j(\cdot)(j \le q)$ are unknown functions, and the error ε_i is independent of the covariates. Thus, in this model, the covariates x_{i0} are assumed to have a parametric effect on the expected value of the response, and the w_{ij} are assumed to enter the model nonparametrically.

The distribution of the error is assumed to be a DPM. Although other non-parametric formulations of the error distributions are possible, the DPM specification has the strengths of being both parsimonious and tractable. Formally, conditioned on an unknown location μ_i and positive variance σ_i^2, we assume that the error distribution is normal $\mathsf{N}\left(\mu_i, \sigma_i^2\right)$. We then suppose that $\phi_i = (\mu_i, \sigma_i^2)$ has an unknown probability measure G over $((-\infty, \infty) \times (0, \infty), \mathcal{B} \times \mathcal{B}_+)$, where the prior on G is given by the Dirichlet process (Ferguson, 1973) with concentration parameter α and base distribution G_0. Marginalized over ϕ_i and G, it follows that the error distribution is an arbitrary location–variance mixture of normal distributions. In particular, we assume that

$$\varepsilon_i|\phi_i \sim \mathsf{N}\left(\mu_i, \sigma_i^2\right)$$
$$\phi_i|G \sim G$$
$$G \sim \mathrm{DP}(\alpha G_0),$$

where

$$G_0 = \mathrm{N}(\mu_i|0, g\sigma_i^2) \text{ inv gamma}\left(\sigma_i^2\middle|\frac{a}{2}, \frac{b}{2}\right)$$

for given values of the hyperparameters (g, a, b). Note that the distribution of μ_i under G_0 has a mean of zero because x_{i0} is assumed to contain an intercept. By writing $\varepsilon_i|\phi_i = \mu_i + \mathrm{N}\left(0, \sigma_i^2\right)$, it can be seen that μ_i in this model plays the same role as a random effect. From the literature on random effects models, it then follows that a zero prior mean of the random effect is sufficient to identify the intercept. Alternatively, one can drop the intercept from x_{i0} and let the mean of μ_i under G_0 be non-zero and unknown. Although the latter mean would not equal the intercept, the two parameterizations are equivalent.

2.2. Ordinal outcomes

As a variation of the preceding model, consider the situation where y_i takes one of the ordered values $\{0, 1, \ldots, J-1\}$. In that case, the model is specified in terms of the latent variables y_i^* and ordered category cut-points $c_{-1} < c_0 < c_1 < \cdots < c_{J-2} < c_{J-1}$, where c_{-1} is normalized to minus infinity, c_0 to zero, and c_{J-1} to plus infinity, as

$$y_i = j \quad \text{if } c_{j-1} < y_i^* \le c_j,$$

where

$$y_i^* = x_{i0}'\beta_0 + g_1(w_{i1}) + \cdots + g_q(w_{iq}) + \varepsilon_i, \quad i \le n.$$

In contrast to the preceding model, we do not start from the assumption that $\varepsilon_i|\phi_i \sim \mathrm{N}\left(\mu_i, \sigma_i^2\right)$, because of an ambiguity it causes between the μ_i and the c_j's. Instead, following Basu and Mukhopadhyay (2000), we assume that $\phi_i = \lambda_i$ and let

$$\begin{aligned}&\varepsilon_i|\phi_i \sim \mathrm{N}\left(0, \lambda_i^{-1}\right)\\&\lambda_i|G \sim G\\&G \sim \mathrm{DP}(\alpha G_0),\end{aligned} \tag{2.2}$$

where

$$G_0 = \text{gamma}\left(\lambda_i\middle|\frac{\nu}{2}, \frac{\nu}{2}\right)$$

for a given value of the hyperparameter ν. Note that despite appearances, it is not possible in this model to scale the y_i^* and the cut-points and leave the model unchanged. This is because scaling y_i^* involves a scaling of the λ_i, which changes the distribution of λ_i from the one given above, and hence the probabilities of the outcomes.

It is then readily seen that, under G_0, the probability that $y_i = j$ has the Student-t link function given by

$$\mathrm{T}_\nu(c_j - x_{i0}'\beta_0 - g_1(w_{i1}) - \cdots - g_q(w_{iq}), 1) - \mathrm{T}_\nu(c_{j-1} - x_{i0}'\beta_0 - g_1(w_{i1}) - \cdots - g_q(w_{iq}), 1),$$

where $\mathrm{T}_\nu(a, b)$ is the cdf of the Student-t distribution at the point a, with dispersion b and ν degrees of freedom. Under G, however, the link function is of unknown form. This model is, therefore, a DPM generalization of the Student-t link binary and ordinal models introduced in Albert and Chib (1993). It also generalizes the DPM binary response models considered in Basu and Mukhopadhyay (2000) and Basu and Chib (2003) to ordinal outcomes with additive spline-based covariate effects.

3. Modeling g_j

We represent each of the unknown g_j's in terms of a natural cubic spline. Cubic splines (for example, Green and Silverman, 1994; Wasserman, 2006, chaps. 5, 8, and 9) are piecewise cubic polynomials that have continuous second derivatives at and between knot points. The word "natural" refers to the assumption that the second derivatives of the approximating spline are set to zero at the smallest and largest knot points. This assumption or a similar one is necessary to determine a unique approximating function.

We let the knots be equally spaced and set the smallest and largest knots, respectively, to the minimum and maximum of each covariate. Ruppert (2002) finds that function approximations are not very sensitive to the number of knots beyond some minimum and that excessively many knots can worsen the mean squared error. Our work with simulated and real data supports this conclusion. A simple strategy that we find useful is to start with (say) five knots for each function and then to incrementally adjust the number of knots on the basis of the model marginal likelihoods. This exploration can be done quickly with a Gaussian error model because, in our studies, we find that the choice of the number of knots is not very sensitive to the distribution of the error term.

3.1. LS basis

Consider the jth function $g_j(w)$. Let $\tau_j = (\tau_{1j}, \ldots, \tau_{M_jj})$, where $\tau_{1j} = \min_i(w_{ij})$, $\tau_{M_jj} = \max_i(w_{ij})$, and $\tau_{mj} < \tau_{m+1,j}$ for $m = 1, \ldots, M_j - 1$, denote the M_j knot points. We require $M_j \ge 4$, because four is the minimum number of basis functions required to fit an unrestricted cubic polynomial. For this discussion, we assume that M_j is fixed and discuss later how it may be determined from a model choice perspective.

For any point $w \in R$ and the set of knots τ_j, the basis functions are the collections of cubic splines $\left\{\Phi_{mj}(w)\right\}_{m=1}^{M_j}$ and $\{\Psi_{mj}(w)\}_{m=1}^{M_j}$, where

$$\Phi_{mj}(w) = \begin{cases}0, \quad w < \tau_{m-1,j},\\ -(2/h_{mj}^3)(w - \tau_{m-1,j})^2(w - \tau_{mj} - 0.5h_{mj}),\\ \quad \tau_{m-1,j} \le w < \tau_{mj},\\ (2/h_{m+1,j}^3)(w - \tau_{m+1,j})^2(w - \tau_{mj} + 0.5h_{m+1,j}),\\ \quad \tau_{mj} \le w < \tau_{m+1,j},\\ 0, \quad w \ge \tau_{m+1,j},\end{cases} \tag{3.1}$$

$$\Psi_{mj}(w) = \begin{cases}0, \quad w < \tau_{m-1,j},\\ (1/h_{mj}^2)(w - \tau_{m-1,j})^2(w - \tau_{mj}),\\ \quad \tau_{m-1,j} \le w < \tau_{mj},\\ (1/h_{m+1,j}^2)(w - \tau_{m+1,j})^2(w - \tau_{mj}),\\ \quad \tau_{mj} \le w < \tau_{m+1,j},\\ 0, \quad w \ge \tau_{m+1,j},\end{cases} \tag{3.2}$$

and $h_{mj} = \tau_{mj} - \tau_{m-1,j}$ is the spacing between the $(m-1)$st and mth knots. Note that Φ_{1j} and Ψ_{1j} are defined by the last two lines of Eqs. (3.1) and (3.2), respectively, and that Φ_{M_jj} and Ψ_{M_jj} are defined by only the first two lines. In both cases, the strong inequality at the upper limit should be replaced by a weak inequality. We plot these basis functions in Fig. 1 at the first knot $(m = 1)$, an intermediate knot m, and the last knot $(m = M_j)$.

For any point $w_{lj} \in R$ in the support of the covariate w_j, our representation of $g_j(w_{lj})$ as a natural cubic spline is given by

$$g_j(w_{lj}) = \sum_{m=1}^{M_j}\left(\Phi_{mj}(w_{lj})f_{mj} + \Psi_{mj}(w_{lj})s_{mj}\right), \tag{3.3}$$

where

$$f_j = (f_{1j}, \ldots, f_{M_jj})' \quad \text{and} \quad s_j = (s_{1j}, \ldots, s_{M_jj})'$$

are the coefficients of this cubic spline. These coefficients have the convenient interpretation of being, respectively, the ordinate and slope of $g_j(w_{ji})$ at the mth knot. Specifically,

$$g_j(\tau_{mj}) = f_{mj} \quad \text{and} \quad g_j'(\tau_{mj}) = s_{mj}.$$

That $f_{mj} = g_j(\tau_{mj})$ can be seen by evaluating $g_j(w_{lj})$ at τ_{mj} and applying the facts, which can be checked by substitution, that

$$\begin{aligned}&\Phi_{mj}(\tau_{mj}) = 1,\\&\Phi_{mj}(\tau_{lj}) = 0, \quad l \ne m,\\&\Psi_{lj}(\tau_{lj}) = 0, \quad l = 1, \ldots, M_j.\end{aligned}$$

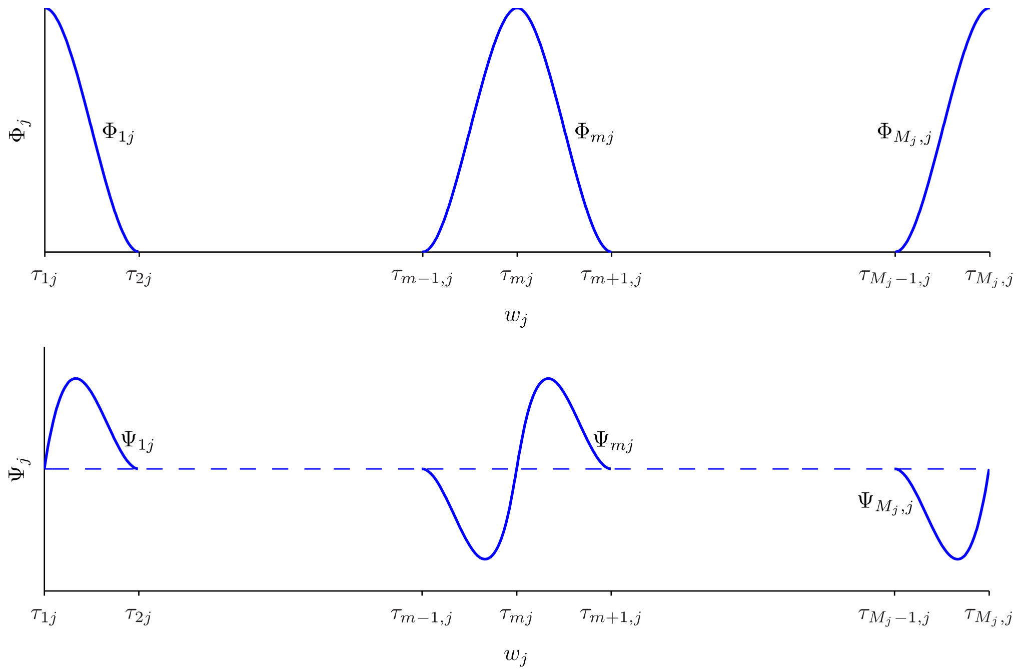

Fig. 1. Φ_j and Ψ_j.

Confirmation that $g_j'(\tau_{mj}) = s_{mj}$ requires the derivatives of the basis functions. These derivatives are the quadratic splines

$$\Phi_{mj}'(w) = \begin{cases} 0, \quad w < \tau_{m-1,j}, \\ -\dfrac{2}{h_{mj}^3}[(w-\tau_{m-1,j})^2 + 2(w-\tau_{mj} \\ \quad -0.5h_{mj})(w-\tau_{m-1,j})], \quad \tau_{m-1,j} \le w < \tau_{mj}, \\ \dfrac{2}{h_{m+1,j}^3}[(w-\tau_{m+1,j})^2 + 2(w-\tau_{mj} \\ \quad +0.5h_{m+1,j})(w-\tau_{m+1,j})], \quad \tau_{mj} \le w < \tau_{m+1,j}, \\ 0, \quad w \ge \tau_{m+1,j}, \end{cases} \tag{3.4}$$

$$\Psi_{mj}'(w) = \begin{cases} 0, \quad w < \tau_{m-1,j}, \\ \dfrac{1}{h_{mj}^2}[(w-\tau_{m-1,j})^2 + 2(w-\tau_{m-1,j})(w-\tau_{mj})], \\ \quad \tau_{m-1,j} \le w < \tau_{mj}, \\ \dfrac{1}{h_{m+1,j}^2}[(w-\tau_{m+1,j})^2 + 2(w-\tau_{mj})(w-\tau_{m+1,j})], \\ \quad \tau_{mj} \le w < \tau_{m+1,j}, \\ 0, \quad w \ge \tau_{m+1,j}, \end{cases} \tag{3.5}$$

respectively, where Φ_{1j}' and Ψ_{1j}' are computed from the last two lines of (3.4) and (3.5), and Φ_{M_jj}' and Ψ_{M_jj}' from the first two lines. On evaluating the derivative of $g_j(w_{lj})$ at τ_{mj}, we find

$$\Psi_{mj}'(\tau_{mj}) = 1,$$
$$\Psi_{mj}'(\tau_{lj}) = 0, \quad l \ne m,$$
$$\Phi_{mj}'(\tau_{lj}) = 0, \quad l = 1, \dots, M_j,$$

which confirms that $g_j'(\tau_{mj}) = s_{mj}$.

3.2. Eliminating s_j

The fact that $g_j(w)$ is a natural cubic spline implies that $g_j''(\tau_{1j}) = 0 = g_j''(\tau_{M_jj})$ and that the second derivatives are continuous at the knot points. These conditions place restrictions on the s_{mj}. If we define $\omega_{mj} = h_{mj}/(h_{mj}+h_{m+1,j})$, and $\mu_{mj} = 1-\omega_{mj}$ for $m = 2, \dots, M_j$, then Lancaster and Šalkauskas (1986, Sec. 4.2) show that the ordinates and slopes are related by the relations $C_jf_j = A_js_j$, or

$$s_j = A_j^{-1}C_jf_j,$$

where

$$A_j = \begin{pmatrix} 2 & 1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ \omega_{2j} & 2 & \mu_{2j} & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & \omega_{3j} & 2 & \mu_{3j} & 0 & \dots & 0 & 0 & 0 \\ \vdots & \dots & \ddots & \ddots & \ddots & \dots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \dots & \omega_{M_j-1,j} & 2 & \mu_{M_j-1,j} \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & 1 & 2 \end{pmatrix},$$

and C_j is given in Box I. It follows that $g_j(w_{lj})$ in (3.3) can be re-expressed as

$$\begin{aligned} g_j(w_{lj}) &= \sum_{m=1}^{M_j} \left(\Phi_{mj}(w_{lj})f_{mj} + \Psi_{mj}(w_{lj})s_{mj}\right) \\ &= \left(\Phi_j(w_{lj})' + \Psi_j(w_{lj})'A_j^{-1}C_j\right)f_j \\ &= z_{lj}'f_j, \end{aligned}$$

where

$$\Phi_j(w_{lj})' = \left(\Phi_{1j}(w_{lj}), \dots, \Phi_{M_jj}(w_{lj})\right),$$
$$\Psi_j(w_{lj})' = \left(\Psi_{1j}(w_{lj}), \dots, \Psi_{M_jj}(w_{lj})\right),$$

and

$$z_{lj}' = \Phi_j(w_{lj})' + \Psi_j(w_{lj})'A_j^{-1}C_j = (z_{l1j}, \dots, z_{lM_jj}).$$

Applying this spline representation to each of the unknown functions in our model, we can write the data generating process for the observed y_i in (2.1) as

$$y_i = x_{i0}'\beta_0 + \sum_{j=1}^{q} z_{ij}'f_j + \varepsilon_i.$$

In vector form, for a sample of n observations $y = (y_1, \dots, y_n)'$, the generating process can be expressed as

$$y = X_0\beta_0 + Zf + \varepsilon, \tag{3.6}$$

$$C_j = 3\begin{pmatrix} -\frac{1}{h_{2j}} & \frac{1}{h_{2j}} & 0 & 0 & \dots & 0 & 0 & 0 \\ -\frac{\omega_{2j}}{h_{2j}} & \frac{\omega_{2j}}{h_{2j}} - \frac{\mu_{2j}}{h_{3j}} & \frac{\mu_{2j}}{h_{3j}} & 0 & \dots & 0 & 0 & 0 \\ 0 & -\frac{\omega_{3j}}{h_{3j}} & \frac{\omega_{3j}}{h_{3j}} - \frac{\mu_{3j}}{h_{4j}} & \frac{\mu_{3j}}{h_{4j}} & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & -\frac{\omega_{M_j-1,j}}{h_{M_j-1}} & \frac{\omega_{M_j-1,j}}{h_{M_j-1}} - \frac{\mu_{M_j-1,j}}{h_{M_j}} & \frac{\mu_{M_j-1,j}}{h_{M_j}} \\ 0 & 0 & 0 & 0 & \dots & 0 & -\frac{1}{h_{M_jj}} & \frac{1}{h_{M_jj}} \end{pmatrix}$$

Box I.

where $X_0 = (x_{10}, \dots, x_{n0})'$ and

$$Z = \begin{pmatrix} z'_{11} & z'_{12} & \dots & z'_{1q} \\ z'_{21} & z'_{22} & \dots & z'_{2q} \\ \vdots & \vdots & \vdots & \vdots \\ z'_{n1} & z'_{n2} & \dots & z'_{nq} \end{pmatrix} = (Z_1, Z_2, \dots, Z_q),$$

and Z_j is an $n \times M_j$ matrix.

3.3. Identification

Since the M_s columns of Z_s and the M_t columns of Z_t are both bases for a cubic spline, an $n \times 1$ vector of ones is contained in each of their column spaces. The columns of Z_s and Z_t are, therefore, not linearly independent, which implies that the elements of f_s and f_t are not identified. Although this problem does not arise if there is only one unknown function, to unify the discussion we assume that x_0 always contains an intercept, so that the identification problem arises even then. To obtain identification, we impose the identifying constraints $\sum_m f_{mj} = 0, j = 1, \dots, q$, which implies that $f_{1j} = -(f_{2j} + \dots + f_{M_jj})$. Accordingly, we have

$$\begin{aligned} z'_{ij}f_j &= f_{1j}z_{i1j} + f_{2j}z_{i2j} + \dots + f_{M_jj}z_{iM_jj} \\ &= -(f_{2j} + \dots + f_{M_jj})z_{i1j} + f_{2j}z_{i2j} + \dots + f_{M_jj}z_{iM_jj} \\ &= f_{2j}(z_{i2j} - z_{i1j}) + \dots + f_{M_jj}(z_{iM_jj} - z_{i1j}) \\ &= x'_{ij}\beta_j, \end{aligned}$$

where

$$x'_{ij} = \big((z_{i2j} - z_{i1j}), \dots, (z_{iM_jj} - z_{i1j})\big) \tag{3.7}$$

and

$$\beta_j = \begin{pmatrix} f_{2j} \\ f_{3j} \\ \vdots \\ f_{M_jj} \end{pmatrix}. \tag{3.8}$$

Now, rewrite the model for the observed data with identified parameters as

$$y_i = x'_i\beta + \varepsilon_i,$$

where

$$x'_i = (x'_{i0}, x'_{i1}, \dots, x'_{iq}) \quad \text{and} \quad \beta = (\beta'_0, \beta'_1, \dots, \beta'_q)'.$$

In vector–matrix form, the final form of our model is, therefore,

$$y = X\beta + \varepsilon, \tag{3.9}$$

where $X = (X_0, X_1)$,

$$X_1 = \begin{pmatrix} x'_{11} & x'_{12} & \dots & x'_{1q} \\ x'_{21} & x'_{22} & \dots & x'_{2q} \\ \vdots & \vdots & \vdots & \vdots \\ x'_{n1} & x'_{n2} & \dots & x'_{nq} \end{pmatrix} = \big(X_{11}, X_{12}, \dots, X_{1q}\big),$$

and X_{1j} is an $n \times (M_j - 1)$ matrix. In this final form, there are k_0 regression parameters in β_0 and $M_j - 1$ regression parameters in $\beta_j, j = 1, \dots, q$, a total of $k = k_0 + k_1$ regression parameters, where $k_1 = \sum M_j - q$. In the ordinal outcome model, the same formulation prevails at the level of the latent data: setting $y^* = (y^*_1, \dots, y^*_n)'$, we have

$$y^* = X\beta + \varepsilon.$$

A consequence of this identification scheme is that the levels of the individual functions are not identified, but in practice this causes no difficulty. First, since the intercept is included when computing the predictive distribution, the overall level of the sum of the functions is identified. Second, the effect of a change in a covariate is identified because levels of the functions are irrelevant for the derivatives.

4. Prior distribution

We now provide a prior distribution on the β_j that incorporates the assumption of a priori smoothness. We do this by reasoning in terms of the differences in ordinates at the end knots and in terms of the second differences of ordinates at interior knots. We find it appealing to think in terms of first and second differences of the ordinates, because these are quantities about which one may have prior knowledge and because the prior we suggest reinforces the assumption of smoothness without introducing strong a priori information.

For the first knot, we assume that, conditioned on the variance σ^2_{ej},

$$\frac{f_{2j} - f_{1j}}{h_{2j}} \sim \text{N}(0, \sigma^2_{ej})$$

or, because $f_{1j} = -(f_{2j} + \dots + f_{M_jj})$, that

$$\frac{2f_{2j} + f_{3j} + \dots + f_{M_jj}}{h_{2j}} \sim \text{N}(0, \sigma^2_{ej}). \tag{4.1}$$

We treat the last two knots similarly and let

$$\frac{f_{M_jj} - f_{M_j-1,j}}{h_{M_jj}} \sim \text{N}(0, \sigma^2_{ej}). \tag{4.2}$$

As for the interior knots, conditioned on a different variance σ^2_{dj}, we assume that the differences in slopes are normal with expectation of zero:

$$\frac{f_{m+1,j} - f_{mj}}{h_{m+1,j}} - \frac{f_{mj} - f_{m-1,j}}{h_{mj}} \sim \text{N}(0, \sigma^2_{dj}), \quad m = 3, \dots M_j - 1. \tag{4.3}$$

These assumptions on the prior of the f_{mj} reinforce the smoothness assumptions embodied in the cubic spline formulation. The parameters σ^2_{ej} and σ^2_{dj} are smoothness parameters in the sense that small variances have the effect of smoothing the function because the differences are presumed to be small, while large variances have the opposite effect. The variances are analogous to the weight put on the penalty function in some implementations of cubic splines. Our Bayesian approach allows these variances to be determined as part of the inferential procedure, so that the degree of smoothness achieved depends on the data as well as the prior.

The foregoing assumptions imply that for each $j = 1, 2, \ldots, q$,

$$\Delta_j\beta_j|\sigma_{ej}^2, \sigma_{dj}^2 \sim \mathrm{N}_{M_j-1}(0, T_j),$$

or that

$$\beta_j|\sigma_{ej}^2, \sigma_{dj}^2 \sim \mathrm{N}_{M_j-1}(0, \Delta_j^{-1}T_j(\Delta_j^{-1})'), \tag{4.4}$$

where N_{M_j-1} is the $M_j - 1$ variate normal distribution, Δ_j is given in Box II, and

$$T_j = \begin{pmatrix} \sigma_{ej}^2 & 0 & 0 \\ 0 & \sigma_{dj}^2 I_{M_j-3} & 0 \\ 0 & 0 & \sigma_{ej}^2 \end{pmatrix}.$$

Remark 1. In this framework, nondogmatic prior knowledge about functional form can be introduced by assuming nonzero values for the means of the differences. For example, if one believes that the function is increasing for small values of the covariate, a positive value can be assigned to the mean of $(f_{2j} - f_{1j})/h_{2j}$. Nonzero prior means permit one to influence the shape of the functions while allowing the data to not support that belief. Other prior assumptions about the levels or changes in the ordinates may be adopted subject to the requirement that the Δ_j matrix is nonsingular.

Remark 2. The prior we propose has some similarities to the Markov process prior but the similarities are superficial because our prior is on the coefficients of the cubic spline, whereas the Markov process prior is on the function g_j itself. Essentially, the latter approach is most useful when certain values of w_j are replicated many times in the sample but if each value of the covariate is unique the Markov process prior can behave poorly. In the same vein, a matrix similar to Δ_j appears in the penalized *B*-spline approach (see Eilers and Marx, 1996) but in connection with the *B*-spline coefficients rather than the ordinates at the knot points.

For the parameters β_0 of the linear part of the model, we assume that, independently of $\{\beta_j\}$, $\beta_0 \sim \mathrm{N}_{k_0}(b_{00}, B_{00})$. The prior distribution of $\beta = (\beta_0, \beta_1, \ldots, \beta_q)$ conditioned on $\sigma_e^2 = (\sigma_{e1}^2, \ldots, \sigma_{eq}^2)$ and $\sigma_d^2 = (\sigma_{d1}^2, \ldots, \sigma_{dq}^2)$ is, therefore, given by

$$\beta|\sigma_e^2, \sigma_d^2 \sim \mathrm{N}_k(b_0, B_0), \tag{4.5}$$

where

$$b_0 = \begin{pmatrix} b_{00} \\ 0_{1\times k_1} \end{pmatrix} \quad \text{and}$$

$$B_0 = \begin{pmatrix} B_{00} & 0 & \ldots & 0 \\ 0 & \Delta_1^{-1}T_1(\Delta_1^{-1})' & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & \Delta_q^{-1}T_q(\Delta_q^{-1})' \end{pmatrix}.$$

We complete the prior model by assuming that, independently,

$$(\sigma_e^2, \sigma_d^2) \sim \prod_{j=1}^{q} \text{inv gamma}\left(\frac{\alpha_{ej0}}{2}, \frac{\delta_{ej0}}{2}\right) \text{inv gamma}\left(\frac{\alpha_{dj0}}{2}, \frac{\delta_{dj0}}{2}\right),$$

$$\alpha \sim \text{gamma}(a_0, d_0), \tag{4.6}$$

for given values of the hyperparameters $\left\{\alpha_{ej0}, \delta_{ej0}, \alpha_{dj0} \text{ and } \delta_{dj0}\right\}_{j=1}^q$ and (a_0, d_0). The parameters of α, following Basu and Chib (2003), are chosen to favor small values in order to differentiate between the DPM and the Student-*t* model, since the latter is approached as $\alpha \to \infty$.

In the ordinal model, the prior distribution is supplemented by a prior on the cut-points. Rather than work with the free ordered cut-points $c = (c_1, \ldots, c_{J-2})$, however, we follow Albert and Chib (2001) by parameterizing the cut-points as $a = (a_1, \ldots, a_{J-2})$, where

$$a_1 = \log c_1, \qquad a_j = \log(c_j - c_{j-1}), \quad 2 \le j \le J-2,$$

with the inverse map $c_1 = \exp(a_1)$, $c_j = \sum_{l=1}^{j} \exp(a_l)$. Since the cut-points are unordered in this parameterization, it can be assumed that, a priori,

$$a \sim N(a_{00}, A_{00}) \tag{4.7}$$

for given values of the hyperparameters.

4.1. Choice of hyperparameters

In this section, we discuss an approach for specifying the hyperparameters of the prior distribution. Consider first the hyperparameters α_{ej0}, δ_{ej0}, α_{dj0}, and δ_{dj0} that define the prior distribution on the function ordinates $X_{1j}\beta_j$ corresponding to the value of the jth covariate $w_{\bullet j}$. We can isolate the effect of these hyperparameters on the distribution of the functions by a straightforward simulation approach. For given value of these hyperparameters we simulate many times

$$(\sigma_e^2, \sigma_d^2) \sim \text{inv gamma}\left(\frac{\alpha_{ej0}}{2}, \frac{\delta_{ej0}}{2}\right) \text{inv gamma}\left(\frac{\alpha_{dj0}}{2}, \frac{\delta_{dj0}}{2}\right) \quad \text{and}$$

$$\beta_j|\sigma_{ej}^2, \sigma_{dj}^2 \sim \mathrm{N}_{M_j-1}(0, \Delta_j^{-1}T_j(\Delta_j^{-1})')$$

and evaluate and store the function values $X_{1j}\beta_j$. The resulting sample can be summarized to produce the prior expectation of $X_{1j}\beta_j$, which is the zero vector under the assumed conditional prior mean of β_j, and pointwise prior quantiles of $X_{1j}\beta_j$ or other measures of spread. If the implied distribution of $X_{1j}\beta_j$ does not satisfactorily represent the a priori beliefs, the process is repeated with a revised set of hyperparameters. We illustrate this approach in Section 8.2.

The same idea can be used to examine the role played by the other hyperparameters of the model: (b_{00}, B_{00}), (a_0, d_0), and (g, a, b) in the continuous outcome model and (b_{00}, B_{00}), (a_0, d_0), ν, and (a_{00}, A_{00}) in the ordinal outcome model. For specified values of the hyperparameters, we simulate (σ_e^2, σ_d^2) and α from (4.6), followed by β from (4.5) and a from (4.7) in the ordinal model. A value of G is then sampled by the "stick-breaking" construction of the Dirichlet process as given by Sethuraman (1994):

$$G(\cdot) = \sum_{l=1}^{\infty} p_l \delta_{\phi_l}(\cdot), \quad \text{where } \phi_l \overset{\text{iid}}{\sim} G_0, l = 1, 2, \ldots, \text{ and}$$

$$p_1 = V_1, \qquad p_l = V_l \prod_{j=1}^{l-1}(1 - V_j), \quad l = 2, \ldots, \tag{4.8}$$

$$\text{with } V_l \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha), l = 1, 2, \ldots,$$

and where δ_a is the unit point mass at a and the sum in (4.8) is truncated at a large integer N as in Ishwaran and James (2001). For each $i \le N$, we sample a ϕ_i from the mixture distribution in (4.8) and then sample $\varepsilon_i|\phi_i$. After drawing β and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ in this way, the outcomes are $y = X\beta + \varepsilon$ in the continuous case and $y^* = X\beta + \varepsilon$ and a in the ordinal model. This process is repeated many times, and the hyperparameters adjusted as necessary, until the sampled values approximately reflect one's prior beliefs.

$$\Delta_j = \begin{pmatrix} \frac{2}{h_{2j}} & \frac{1}{h_{2j}} & \frac{1}{h_{2j}} & \frac{1}{h_{2j}} & \frac{1}{h_{2j}} & \dots & \frac{1}{h_{2j}} & \frac{1}{h_{2j}} \\ \frac{1}{h_{2j}} & -\left(\frac{1}{h_{2j}}+\frac{1}{h_{3j}}\right) & \frac{1}{h_{3j}} & 0 & 0 & \dots & 0 & 0 \\ 0 & \frac{1}{h_{3j}} & -\left(\frac{1}{h_{3j}}+\frac{1}{h_{4j}}\right) & \frac{1}{h_{4j}} & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \dots & -\frac{1}{h_{Mj}} & \frac{1}{h_{Mj}} \end{pmatrix}$$

Box II.

5. Posterior distributions and MCMC algorithms

5.1. Continuous outcome model

Given the data (y, X) and the distribution G_0, the goal is to summarize the posterior distribution of the unknown parameters $\theta = (\beta, \sigma_e^2, \sigma_d^2, \alpha)$ and the location–variance parameters $\phi = (\phi_1, \dots, \phi_n)$. Following Escobar and West (1995), we include ϕ in the prior-posterior analysis because that simplifies the computations. Let $\pi(\theta)$ denote the product of the distributions in (4.5) and (4.6), and let $\pi(\phi|\alpha, G_0)$ be the prior density of ϕ. Then, the posterior distribution of interest is proportional to

$$\pi(\phi, \theta|y, X, G_0) \propto \pi(\theta)\,\pi(\phi|\alpha, G_0)p(y|X, \phi, \theta)$$
$$\propto \pi(\theta)\ \pi(\phi|\alpha, G_0) \prod_{i=1}^{n} \mathrm{N}(y_i|\mu_i + x_i'\beta, \sigma_i^2).$$

This density can be sampled by MCMC methods by using the following facts.

- Conditioned on ϕ, the set-up resembles that of a heteroskedastic regression model, which implies that the conditional distributions of β, σ_e^2, and σ_d^2 are easily derived and of known form.
- Conditioned on θ, the set-up corresponds to a DPM model for which the sampling problem has been well studied. The idea is to update ϕ by revising its components one at a time. From the prequential property of the Dirichlet process, the distribution of ϕ_i conditioned on ϕ_{-i} (the set of ϕ_i's excluding ϕ_i), θ, and G_0, but marginalized over G, is

  $$\phi_i|\phi_{-i}, \theta, G_0 \sim E(G|\phi_{-i}, \theta, G_0)$$
  $$\sim \frac{\alpha}{\alpha+n-1} G_0 + \frac{1}{\alpha+n-1} \sum_{r=1}^{k_{-i}} n_{-i,r}\delta_{\phi^*_{-i,r}}, \qquad (5.1)$$

  where $\{\phi^*_{-i,1}, \dots, \phi^*_{-i,k_{-i}}\}$ are the k_{-i} unique values in ϕ_{-i} and $n_{-i,r}$ is the number of times $\phi^*_{-i,r}$ is repeated in ϕ_{-i}. The conditional prior of ϕ_i in Eq. (5.1) can be combined with the $N(y_i|\mu_i + x_i'\beta, \sigma_i^2)$ density to produce the conditional posterior of ϕ_i. The latter update can be calculated in closed form because the density of y_i marginalized over the conditional distribution of ϕ_i is available as

  $$p(y_i|x_i, \phi_{-i}, \theta, G_0) = \frac{\alpha}{\alpha+n-1} \mathrm{t}_a\left(y_i|x_i'\beta, \frac{b(1+g)}{a}\right) \qquad (5.2)$$
  $$+ \frac{1}{\alpha+n-1} \sum_{r=1}^{k_{-i}} n_{-i,r}\mathrm{N}\left(y_i|\mu^*_{-i,r} + x_i'\beta, \sigma^{*2}_{-i,r}\right), \qquad (5.3)$$

  where $\mathrm{t}_\nu(\cdot|\mu, \sigma^2)$ is the Student-t density with ν degrees of freedom, location μ, and dispersion σ^2.
- Conditioned on ϕ, the sampling of α proceeds as described in Escobar and West (1995).

These observations lead to the fitting scheme that is summarized as Algorithm 1 in the Appendix.

We repeat the steps of Algorithm 1 until a desired sample of size N is acquired after a burn-in of N_0 iterations. In the applications below, we set $N_0 = 2000$ and $N = 20{,}000$. Given the MCMC sample of the parameters $\left\{\beta_1^{(g)}, \dots, \beta_q^{(g)}\right\}_{g=1}^{N}$, we find the distribution of the jth function $X_{1j}\beta_j$ on the grid $w_{\bullet j}$ from the values

$$\left\{X_{1j}\beta_j^{(g)}\right\}_{g=1}^{N}, \quad j = 1, 2, \dots, q,$$

which are draws from the posterior distribution of $g(w_{\bullet j})$. This sample can, therefore, be used to estimate the posterior mean of $g(w_{\bullet j})$ as

$$\hat{g}_j(w_{\bullet j}) = N^{-1} \sum_{g=1}^{N} X_{1j}\beta_j^{(g)}$$

and to estimate the 95% credibility interval of $g(w_{\bullet j})$, component by component, from the 0.025 and 0.975 sample quantiles of that sample. It is important to note that this interval estimate of the function incorporates uncertainty from the estimation of all the unknowns in the model. The derivatives and the credibility interval of the derivatives are estimated in much the same way.

5.2. Ordinal outcome model

The sampling approach just described can be extended to the ordinal model by including the latent variables $y^* = \{y_i^*\}$ as additional unknowns in the inferential process, following Albert and Chib (1993). Upon redefining $\theta = (a, \beta, \sigma_e^2, \sigma_d^2, \alpha)$ and $\phi = (\phi_1, \dots, \phi_n)$, where $\phi_i = \lambda_i$ is the precision parameter, the posterior distribution of interest is

$$\pi(y^*, \phi, \theta|y, X, G_0) \propto \pi(\theta)\,\pi(\phi|\alpha, G_0)p(y^*|\phi, \theta)p(y|y^*, X, \phi, \theta)$$
$$\propto \pi(\theta)\ \pi(\phi|\alpha, G_0) \prod_{i=1}^{n} \mathrm{N}(y_i^*|x_i'\beta, \lambda_i^{-1})$$
$$\times \prod_{j=0}^{J-1} I\left\{c_{j-1} < y_i^* < c_j\right\} I\left\{y_i = j\right\}.$$

The advantage of adopting this framework is that, conditioned on y^*, the term $p(y|y^*, X, \phi, \theta)$ drops out, and the posterior distribution resembles the one in the continuous outcome case with y_i replaced by y_i^*, μ_i set to zero, and σ_i^2 set to λ_i^{-1}. Then, conditioned on y^*, the updates of ϕ and θ_{-a} (the parameters excluding the cut-points a) are essentially the same. For instance, the update of $\phi_i = \lambda_i$ occurs in a similar fashion. As before, the distribution of ϕ_i conditioned on ϕ_{-i}, θ, and G_0, but marginalized over G, has the form

$$\phi_i|\phi_{-i}, \theta, G_0 \sim E(G|\phi_{-i}, \theta, G_0)$$
$$\sim \frac{\alpha}{\alpha+n-1} G_0 + \frac{1}{\alpha+n-1} \sum_{r=1}^{k_{-i}} n_{-i,r}\delta_{\phi^*_{-i,r}},$$

where $\{\phi^*_{-i,1}, \dots, \phi^*_{-i,k_{-i}}\}$ are the k_{-i} unique values in ϕ_{-i} and $n_{-i,r}$ is the number of times $\phi^*_{-i,r}$ is repeated in ϕ_{-i}. This distribution can be combined with $N(y^*_i|x_i\beta, \lambda_i^{-1})$ density to produce the closed-form conditional posterior of ϕ_i, making use of the fact that

$$p(y^*_i|x_i, \phi_{-i}, \theta) = \frac{\alpha}{\alpha+n-1} \mathrm{t}_\nu \left(y_i|x_i'\beta, 1\right) + \frac{1}{\alpha+n-1}\sum_{r=1}^{k_{-i}} n_{-i,r} \mathrm{N}\left(y^*_i|x_i'\beta, \lambda^{*-1}_{-i,r}\right). \quad (5.4)$$

Conditioned on ϕ, one must also sample the cut-points a and y^*. The sampling of the cut-points is exactly as in Albert and Chib (2001). Specifically, a is sampled marginalized over y^* from the conditional posterior distribution

$$\pi(a|y, \phi, \theta_{-a}) \propto \pi(a) \Pr(y|\phi, \beta, a),$$

where

$$\Pr(y|\phi, \beta, a) = \prod_{j=0}^{J-1} \prod_{i:y_i=j} \left[\Phi((c_j - x_i'\beta)\lambda^{1/2}) - \Phi((c_{j-1} - x_i'\beta))\lambda^{1/2}\right].$$

We can sample this conditional distribution by an MH step with a proposal density that is tailored to $\log \Pr(y|\phi, \beta, a)$ as described in Chib and Greenberg (1994, 1995). Finally, the sampling of y^* is from truncated normal distributions, exactly as in Albert and Chib (1993).

A summary of the resulting MCMC steps is contained in Algorithm 2 in the Appendix.

6. Marginal likelihood

In practice, one would be interested in the comparison of competing models (defined, for example, with fewer or additional unknown functions, with fewer or more knots, or parametric distributions of the error). Such comparisons are possible from the marginal likelihood/Bayes factor perspective because the computation of the marginal likelihood is straightforward using the approach of Chib (1995) as extended by Basu and Chib (2003) for the DPM model.

6.1. Continuous outcome model

Our starting point is the identity

$$m(y|X, G_0) = \frac{f(y|X, \theta^*, G_0)\,\pi(\theta^*)}{\pi(\theta^*|y, X)}, \quad (6.1)$$

where the first term is the likelihood, the second is the prior ordinate, and the last is the posterior ordinate, each evaluated at a particular point $\theta^* = (\beta^*, \sigma^{2*}_e, \sigma^{2*}_d, \alpha^*)$, which we take to be the estimated posterior mean. The prior ordinate is easily calculated. Calculation of the posterior ordinate is also straightforward. One begins with the decomposition

$$\pi(\theta^*|y, X) = \pi(\sigma^{2*}_e|y, X)\,\pi(\sigma^{2*}_d|y, X)\,\pi(\beta^*|y, X, \sigma^{2*}_e, \sigma^{2*}_d) \times \pi(\alpha^*|y, X, \beta^*)$$

from which the first three ordinates can be estimated by the Rao–Blackwell method by taking the output from Algorithm 1 and averaging the densities from which these draws are sampled. The third ordinate requires an additional MCMC run in which the parameters (σ^2_e, σ^2_d) are held fixed at their starred values and sampling is only over (ϕ, β, α). The draws from this reduced MCMC run on ϕ are used to average the normal density from which β is sampled. Finally, the last ordinate is calculated as the average of the density from which α is sampled, where the draws on ϕ come a second reduced run in which $(\beta, \sigma^2_e, \sigma^2_d)$ are all fixed at their starred values.

The likelihood computation requires the approach developed by Basu and Chib (2003) in the context of the general DPM model. The details are given in Algorithm 3 in the Appendix.

6.2. Ordinal outcome model

In the ordinal model, we start with the same identity as above, with $\theta^* = (\sigma^{2*}_e, \sigma^{2*}_d, a^*, \beta^*, \alpha^*)$. On letting

$$\pi(\theta^*|y, X) = \pi(\sigma^{2*}_e|y, X)\,\pi(\sigma^{2*}_d|y, X)\,\pi(a^*|y, X, \sigma^{2*}_e, \sigma^{2*}_d) \times \pi(\beta^*|y, X, \sigma^{2*}_e, \sigma^{2*}_d, a^*)\pi(\alpha^*|y, X, \beta^*, a^*),$$

we can estimate the first two ordinates as in the continuous outcome model. To estimate the ordinate of a, we use a result from Chib and Jeliazkov (2001) to write that

$$\pi(a^*|y, X, \sigma^{2*}_e, \sigma^{2*}_d) = \frac{E_1\left\{\alpha_{MH}(a, a^*|y, \beta, \lambda)\mathrm{t}_\xi(a^*|m, V)\right\}}{E_2\left\{\alpha_{MH}(a^*, a|y, \beta, \lambda)\right\}},$$

where E_1 is the expectation over the posterior distribution of (a, β, ϕ, α) given $(y, \sigma^{2*}_e, \sigma^{2*}_d)$. This expectation can be calculated from the output of a first reduced run of the MCMC algorithm by holding (σ^2_e, σ^2_d) fixed at their starred values and sampling (ϕ, a, β, α). The denominator expectation E_2 and the ordinate of β are both calculated from a second reduced run in which $(\sigma^2_e, \sigma^2_d, a)$ are fixed at their starred values. The E_2 expectation is calculated by averaging $\alpha_{MH}(a^*, a|y, \beta, \lambda)$ over draws of (ϕ, β, α) from this second reduced run and over draws of a drawn from the proposal density $\mathrm{t}_\xi(a|m, V)$ calculated in this reduced run. In this same run, the ordinate of β is obtained by averaging the normal density from which β is sampled. The last ordinate is calculated as the average of the density from which α is sampled, where the draws on ϕ come a third reduced run in which $(a, \beta, \sigma^2_e, \sigma^2_d)$ are all fixed at their starred values and sampling is over (ϕ, α).

Finally, the likelihood ordinate (as above) emerges from the method given in Basu and Chib (2003). The details are supplied in Algorithm 4 in the Appendix.

7. Special cases

Each of the general models discussed above can be specialized. In the continuous outcome model, one special case occurs when $\varepsilon_i|\sigma^2 \sim \mathrm{N}(0, \sigma^2)$. Another is the Student-t model defined by the hierarchical formulation

$$\varepsilon_i|\lambda_i, \sigma^2 \sim \mathrm{N}(0, \lambda_i^{-1}\sigma^2),$$
$$\lambda_i|\nu \sim \mathrm{gamma}(\nu/2, \nu/2),$$
$$\nu \sim U(\nu_1, \dots, \nu_S),$$

where $U(\nu_1, \dots, \nu_S)$ is the discrete uniform distribution. The support of this discrete uniform distribution should be on small values to differentiate it from the Gaussian model.

Subsumed within the ordinal formulation are the ordinal probit and the t-link models. The binary outcome model emerges when $J - 1 = 1$. Each of these models can be fit and the marginal likelihoods calculated from the general discussion above. In the binary outcome model, for example, the cut-points are not sampled. In the likelihood computation, one would set $c^* = (c^*_{-1}, c^*_0, c^*_1) = (-\infty, 0, \infty)$ and leave everything else unchanged.

8. Examples

8.1. Simulated data

As a simple illustration of the model that we have been discussing, consider data generated from the model

$$y_i = \beta_0 + g_1(w_{i1}) + g_2(w_{i2}) + g_3(w_{i3}) + \varepsilon_i, \quad i \le 2000,$$

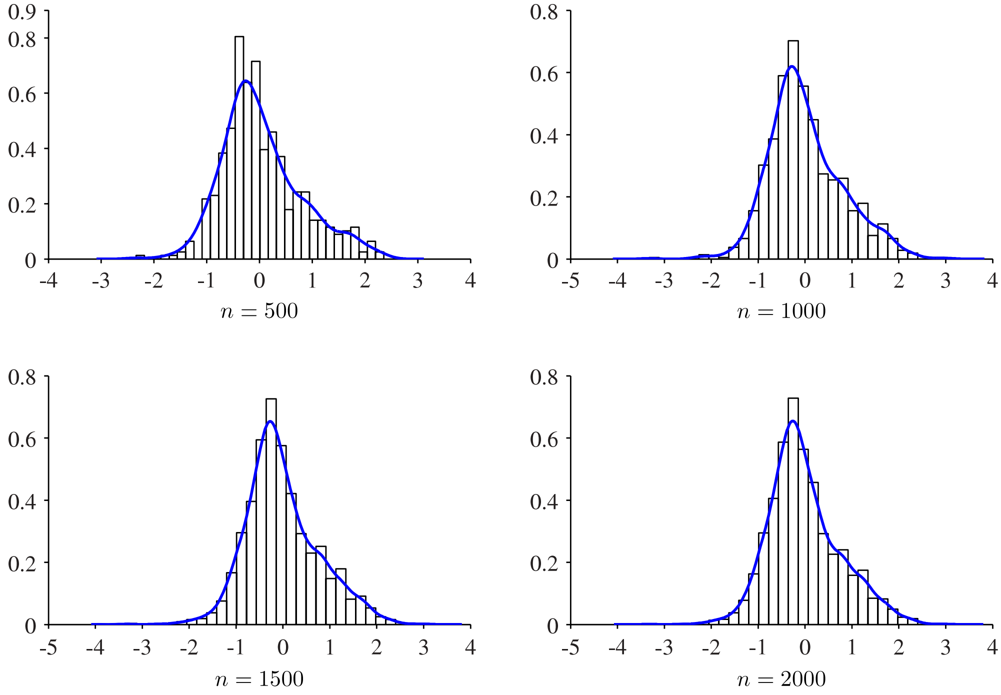

Fig. 2. Distribution of the simulated ε from the DPM prior for samples of sizes $n = 500$, 1000, 1500 and 2000.

where $\beta_0 = 5$, the w_j are created independently as Unif(0, 1), and the functions are given by

$$g_1(w_1) = 8w_1 + \sin(4\pi w_1),$$
$$g_2(w_2) = -1.5 - w_2 + \exp[-30(w_2 - 0.5)^2], \quad \text{and}$$
$$g_3(w_3) = 6w_3^3(1 - w_3).$$

For the error term ε_i, assume that $\varepsilon_i|\phi_i \sim \mathrm{N}\left(\mu_i, \sigma_i^2\right)$, where the ϕ_i are distributed as G and $G \sim \mathrm{DP}(\alpha G_0)$ with $\alpha = 5$ and

$$G_0 = \mathrm{N}(\mu_i|0, \sigma_i^2) \text{ inv gamma} \left(\sigma_i^2 \middle| \frac{4.5}{2}, \frac{1.5}{2}\right).$$

As described in Section 4.1, the ε_i can be generated from the stick-breaking representation of the DP prior of G. That is, a G can be generated from

$$G(\cdot) = \sum_{l=1}^{N} p_l \delta_{\phi_l}(\cdot), \quad \text{where } \phi_l \overset{\text{iid}}{\sim} G_0, \text{ and}$$

$$p_1 = V_1, \qquad p_l = V_l \prod_{j=1}^{l-1} (1 - V_j), \quad \text{with } V_l \overset{\text{iid}}{\sim} \mathrm{Beta}(1, \alpha),$$
$$l = 1, 2, \ldots, N,$$

where we take $N = 150$. Having drawn G, we draw $\phi_i (i = 1, 2, \ldots, 2000)$ from G and simulate $\varepsilon_i|\phi_i$ from $\mathrm{N}\left(\mu_i, \sigma_i^2\right)$. We fit each version of our model for four different sample sizes (500, 1000, 1500 and 2000) and show how the fitting and model comparison algorithms perform. The distributions of the simulated ε are shown in Fig. 2 for the first 500, the first 1000, the first 1500, and finally the entire sample. The positive skewness in all of the distributions is obvious.

To explore the behavior of the algorithm at various sample sizes and to assess its ability to choose between models, we fit both the DPM and Student-t models to these data for each sample size, leading to a total of eight models. We use one set of knots (8, 5, 5) for g_1, g_2 and g_3, respectively. The dimension of β is 16, and there are 6 $\{\sigma_{je,}^2 \sigma_{jd}^2\}$, resulting in 22 parameters in the spline part of the model. The DPM model has an unknown α, and the Student-t model has an unknown ν. Thus, each model has 23 unknown parameters. We assume that the 22 parameters common to the Student-t and DPM models have the same prior distribution. In particular, we assume that $(\alpha_{ej0}, \delta_{ej0}, \alpha_{dj0}, \delta_{dj0}) = (4.125, 2.005, 4.125, 2.005)$ for each j, $(a_0, d_0) = (1.96, 0.28)$, and $(g, a, b) = (1, 4.003, 1.083)$. For ν we assume a discrete uniform distribution on the values {5, 10, 15, 20}. For each model specification, we generate 20,000 draws from the posterior distribution following a burn-in of 2500 draws. The Student-t version with 2000 observations required six minutes to draw the MCMC sample and compute the marginal likelihood; the DPM version required three hours. The log (base 10) marginal likelihoods of the models are given in Table 1. The evidence in favor of the correct DPM model is very strong for each sample size. This shows that the marginal likelihood is a valuable tool for picking the correct model. We have checked that the reverse is also true—the Student-t model wins for each sample size when the data are generated from the Student-t model. We do not present those results to conserve space.

Table 1
Log_{10} marginal likelihoods, simulated data.

n	Model	
	t	DPM
500	−271.689	−268.087
1000	−537.961	−527.584
1500	−825.307	−768.195
2000	−1024.756	−1000.546

We next present in detail the results from the fitting of the DPM model with a sample size of $n = 1000$. We select this sample size because it is medium sized and because the results for the other sample sizes are similar, except that the inferences improve as the sample size increases. A summary of the posterior distribution in this case is given in Table 2. The last column of this table contains the inefficiency factors, defined as $[1 + 2\sum_{k=1}^{L} \rho_k(l)]$, where $\rho_k(l)$ is the sample autocorrelation at lag l for the kth parameter and L is the value at which the correlations taper off. This measure may be interpreted as the ratio of the numerical variance of the posterior mean from the MCMC chain to the variance of

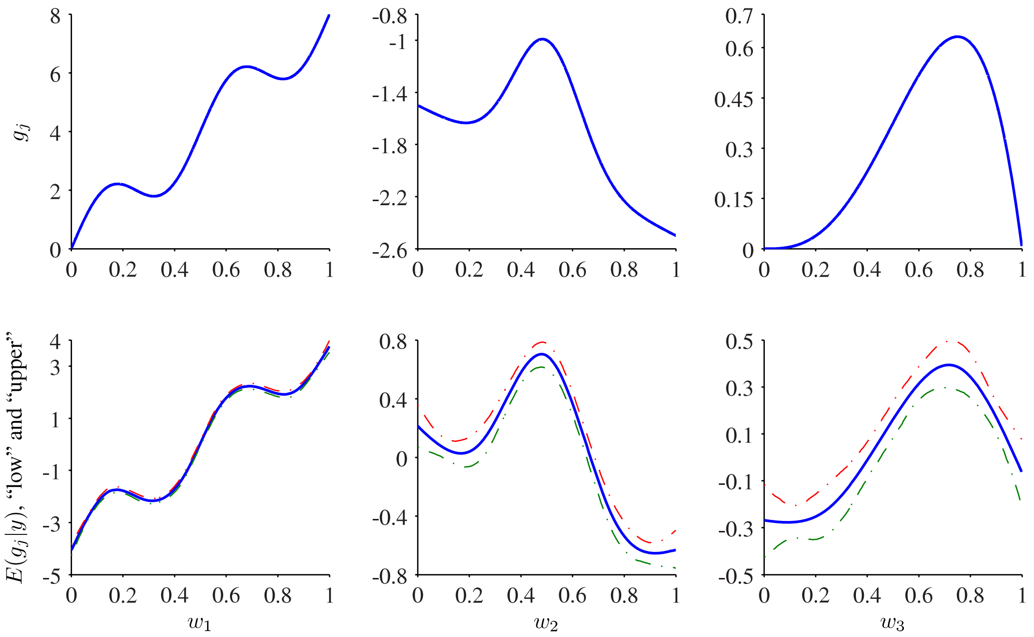

Fig. 3. Simulated data and DPM model with 8, 5 and 5 knots and $n = 1000$. Top panel: true g_1, g_2 and g_3. Bottom panel: posterior expectation of g_1, g_2 and g_3, 0.025 quantile of posterior distribution ("low") and 0.975 quantile ("upper").

Table 2
Summary results for selected parameters: simulated data and DPM model with 8, 5, and 5 knots and $n = 1000$.

Parameter	Prior		Posterior						Inefficiency
	Mean	Std. dev.	Mean	Num. S. E.	Std. dev.	Median	Lower 2.5%	Upper 2.5%	
σ^2_{1e}	5.000	20.000	99.860	0.648	91.664	74.690	27.694	329.303	1.000
σ^2_{2e}	5.000	20.000	2.957	0.020	2.879	2.229	0.817	9.568	1.000
σ^2_{3e}	5.000	20.000	3.491	0.025	3.565	2.602	0.940	11.130	1.000
σ^2_{1d}	1.000	20.000	98.825	0.432	61.028	82.907	35.344	262.990	1.000
σ^2_{2d}	1.000	20.000	14.519	0.098	13.850	10.793	3.669	48.251	1.000
σ^2_{3d}	1.000	20.000	2.599	0.019	2.693	1.860	0.568	9.117	1.000

the posterior mean from hypothetical independent draws. The low values of this measure indicate that the sampler mixes well. Also, note that the posterior means of the smoothing parameters for the first function are much larger than those for the other two functions, even though the marginal prior distribution on those parameters is the same. This finding displays the ability of the smoothing parameters to adapt to the data.

In Fig. 3, we depict the true value of the functions and the values of the posterior means of the functions and their credibility intervals (dotted lines). These are computed from the posterior distribution of the functions as discussed below Algorithm 1. The graphs suggest that the posterior expectation of the functions closely follows the general shape of the functions. Results that we do not report in detail show that the widths of the credibility bands for the functions decrease as the sample size is increased.

8.2. Credit rating data

We consider now an example from Verbeek (2008, pp. 215–217) in which the (ordinal) outcome variable of interest is the firm's Standard and Poor credit rating represented on an integer scale ranging from 1 to 7. Based on the original discussion in Altman and Rijken (2004), Verbeek (2008) utilizes an ordinal probit model with 5 continuous predictors, w_1 = booklev (book value of debt divided by total assets), w_2 = ebit (earnings before interest and taxes divided by total assets), w_3 = logsales (log of sales), w_4 = reta (retained earnings divided by total assets), and w_5 = wka (working capital divided by total assets), to analyze the ordinal outcome. The sample contains 921 observations. This problem is particularly interesting because we can model the effect of each w_j nonparametrically in the context of an unknown latent error distribution.

Due to a paucity of observations in the first and last categories, we combine the observations in the first and second categories, and the sixth and the seventh. Accordingly, the outcome consists of five categories that we label from 0 (lowest) to 4 (highest). In addition, we standardize w_3 by subtracting its mean and dividing by its standard deviation. Our main interest is in models of the form

$$y_i^* = \beta_0 + g_1(w_{i1}) + g_2(w_{i2}) + g_3(w_{i3}) + g_4(w_{i4}) + g_5(w_{i5}) + \varepsilon_i,$$

with the outcome y_i determined in conjunction with the cut-points c_1, c_2 and c_3. In our base DPM model, we model each of the g_j s with 5 knots and consider alternative models with one fewer and one additional knot. Thus, in the largest model containing 6 knots for each function, β contains 21 coefficients. In each DPM model, we assume the model of Eq. (2.2) with $\nu = 10$. For comparison, we also consider Student-t error models with 10 degrees of freedom.

Some care is required in formulating the $\mathrm{N}(a_{00}, A_{00})$ prior distribution on the transformed cut-points a. Since we believe that the probabilities of the first and last outcomes are small, we assume that $a_{00} = (0.916, 0.693, 1.386)$, which corresponds to the values $(2.5, 4.5, 8.5)$ on the original scale, and let $A_{00} = \mathrm{diag}(\frac{1}{2}, \frac{1}{4}, \frac{1}{6})$. Letting the prior variance decline in this way is important for the prior on a to be reasonable for the c_j. As discussed before, the implications of the prior on key objects of interest can be determined by simulation. Our prior on β follows our default prescriptions, except that we treat the intercept differently and assume that the prior mean of β_0 is 5. For each j, we assume the

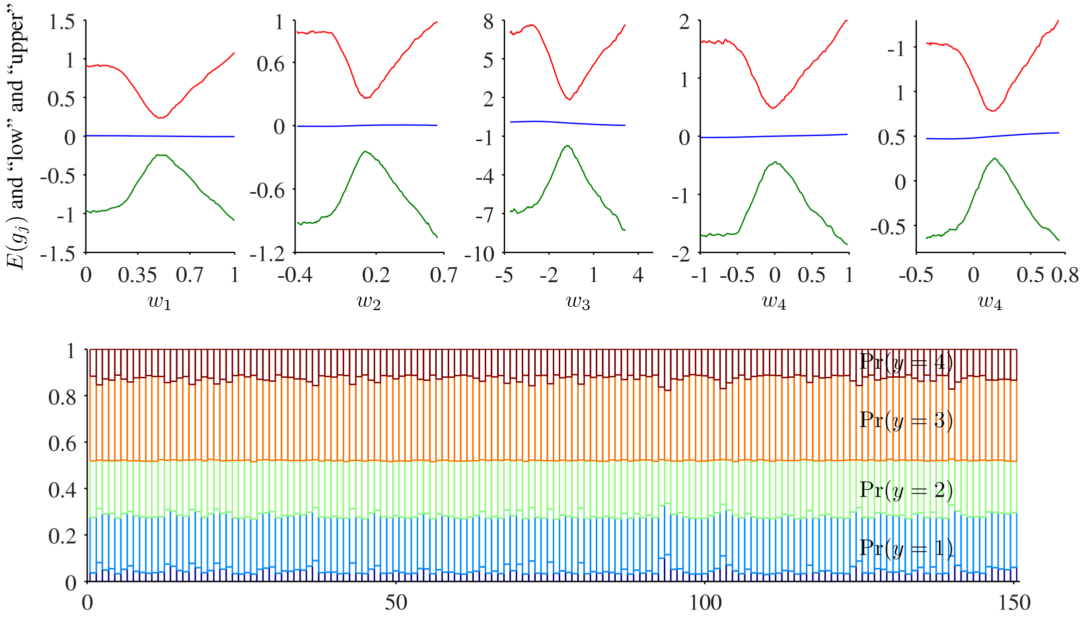

Fig. 4. Implied prior distributions for credit rating model based on DPM errors with 5, 5, 5, 5, and 5 knots. Top panel: distribution of g_1, g_2, g_3, g_4, and g_5, 0.025 quantile of posterior distribution (“low”), 0.975 quantile (“upper”); Bottom panel: implied prior probabilities $\Pr(y_i = j|w_{\bullet j})$ for the first 150 observations in the sample. $\Pr(y_i = 0|w_{\bullet j})$ is the unmarked area.

hyperparameter values

$$\{\alpha_{ej0}, \delta_{ej0}, \alpha_{dj0}, \delta_{dj0}\} = \{4.080, 2.080, 4.080, 2.080\}.$$

We complete our prior modeling by setting the parameters of the gamma prior on α as $(a_0, d_0) = (1.00, 0.20)$. As above, it is of interest to understand the implications of the prior on the implied distribution of the functions and the probabilities of the outcomes. These implied distributions, which are determined by 20,000 draws from the prior and the model, are given in Fig. 4. The top panel of this figure gives the prior mean of the implied prior distribution of g_1, g_2, g_3, g_4, and g_5, each with 5 knots, along with the implied prior 0.025 and 0.975 quantiles. The bottom panel of this figure gives the implied prior probabilities $\Pr(y_i = j|w_{\bullet j})$ for the first 150 observations in the sample. The prior allows for considerable a priori variation in the g_j. For the implied probabilities of the outcomes, which are given in the bottom panel of Fig. 4, the prior implies relatively small probabilities for categories 0 and 4, which is a reasonable a priori belief to hold about these data.

The estimation results are based on 20,000 draws following a burn-in of 2500 cycles. The support for the 6 models we consider (as measured by the log marginal likelihood) is given in Table 3. The marginal likelihood values indicate a preference for the Student-t model with 5 knots for each variable, although the neighboring Student-t models are similarly favored. For these data, the DPM model with the same number of knots is less favored.

Table 4 provides summary statistics for the preferred model. Note how the posterior means of the variance parameters σ^2_{je} and σ^2_{jd} adapt to the data: all start with the same prior value, but the posterior means are different. The inefficiency factors for these parameters are close to one.

In Fig. 5, we display the posterior mean of the function and the derivatives along with the credibility bands. These graphs show that the functions are quite nonlinear, and the first three also non-monotonic. We also find it interesting that the error bands tend to widen for low and high values of the covariates. Although there may be some interest in the g_j functions, the focus of the analysis is to see how the probabilities of the various categories vary as a function of w_j. To present this, we calculate the probability of each outcome at each of 75 different values of w_j, equally spaced from its smallest to largest. For each value of w_j on this grid and at each iteration of the MCMC algorithm, we choose a set of the remaining covariates randomly from among the observed covariates. We then evaluate the probability of being in each category. By averaging these probabilities over the MCMC iterations we obtain the probability of each outcome at each covariate value, marginalized over all other covariate values and all parameters. These probabilities are plotted in Fig. 6. The probabilities for two of the covariates, booklev and wka, display monotonic, although somewhat nonlinear, behavior. We see, for example, that the probability of being in category 0 as a function of booklev is flat at 0.18 from 0 until about 0.3 and then rises to about 0.4 as booklev approaches 1. Other effects display some more interesting patterns. Consider, for example, the strongly nonlinear effects of logsales on the rating probabilities. The probability of category 0 and of category 1 first rises and then falls. A similar nonlinearity appears in the effect of reta on being in category 1. The probability starts just over 0.3, rises to 0.4, and then falls to about 0.18. In the case of ebit, the probabilities of being in categories 0 and 2 are non-monotonic. Nonlinear and non-monotonic patterns such as these may be of interest to researchers of credit ratings.

9. Conclusions

The flexible Bayesian analysis of regression models introduced in this article is applicable to model estimation in many settings. The approach is easy to understand and can be used routinely in regression settings when covariate effects are nonlinear and the error distribution is unknown. Due to its flexibility in modeling regression effects and error distributions, we think of the approach as a step toward a non-parametric Bayesian analysis of the regression model. The framework can be easily extended to clustered and longitudinal data.

The use of cubic splines and the LS basis has several advantages. First, relatively few knots are often needed to approximate even highly nonlinear functions. Second, its coefficients are easily interpretable. This means that a systematic approach can be developed for specifying proper prior beliefs on the coefficients of the basis functions, something that has not been achieved for coefficients of other bases. Our discussion includes an approach

Table 3
Log_{10} of marginal likelihoods for various models, credit rating data.

Number of knots					Error distribution	Log_{10} of marginal likelihood	Numerical std. error
booklev	ebit	logsales	reta	wka			
4	4	4	4	4	*t*	−408.672	0.036
5	5	5	5	5	*t*	−*406.545*	0.059
6	6	6	6	6	*t*	−406.584	0.054
4	4	4	4	4	DPM	−410.014	0.074
5	5	5	5	5	DPM	−407.8318	0.067
6	6	6	6	6	DPM	−408.532	0.074

Table 4
Summary results for selected parameters, credit rating data: Student-*t* model with 5, 5, 5, 5, and 5 knots.

Parameter	Prior		Posterior						Inefficiency
	Mean	Std. Dev.	Mean	Num. S. E.	Std. Dev.	Median	Lower 2.5%	Upper 2.5%	
σ^2_{1e}	1.000	5.000	0.807	0.007	0.983	0.565	0.180	2.887	1.138
σ^2_{2e}	1.000	5.000	1.932	0.034	3.256	1.011	0.227	9.403	2.136
σ^2_{3e}	1.000	5.000	0.915	0.006	0.887	0.679	0.236	3.051	1.000
σ^2_{4e}	1.000	5.000	2.222	0.023	2.562	1.509	0.373	8.579	1.632
σ^2_{5e}	1.000	5.000	1.032	0.012	1.423	0.674	0.201	3.985	1.326
σ^2_{1d}	1.000	5.000	1.456	0.013	1.645	0.992	0.262	5.450	1.195
σ^2_{2d}	1.000	5.000	1.431	0.017	1.889	0.918	0.226	5.671	1.712
σ^2_{3d}	1.000	5.000	0.560	0.004	0.555	0.419	0.154	1.789	1.000
σ^2_{4d}	1.000	5.000	1.560	0.015	1.738	1.082	0.291	5.694	1.425
σ^2_{5d}	1.000	5.000	0.863	0.007	0.938	0.606	0.189	3.055	1.000
c_1	3.185	2.570	1.744	0.002	0.092	1.743	1.567	1.929	7.619
c_2	5.460	2.832	3.208	0.003	0.121	3.207	2.971	3.451	8.838
c_3	9.787	3.367	5.159	0.004	0.192	5.156	4.793	5.546	6.963

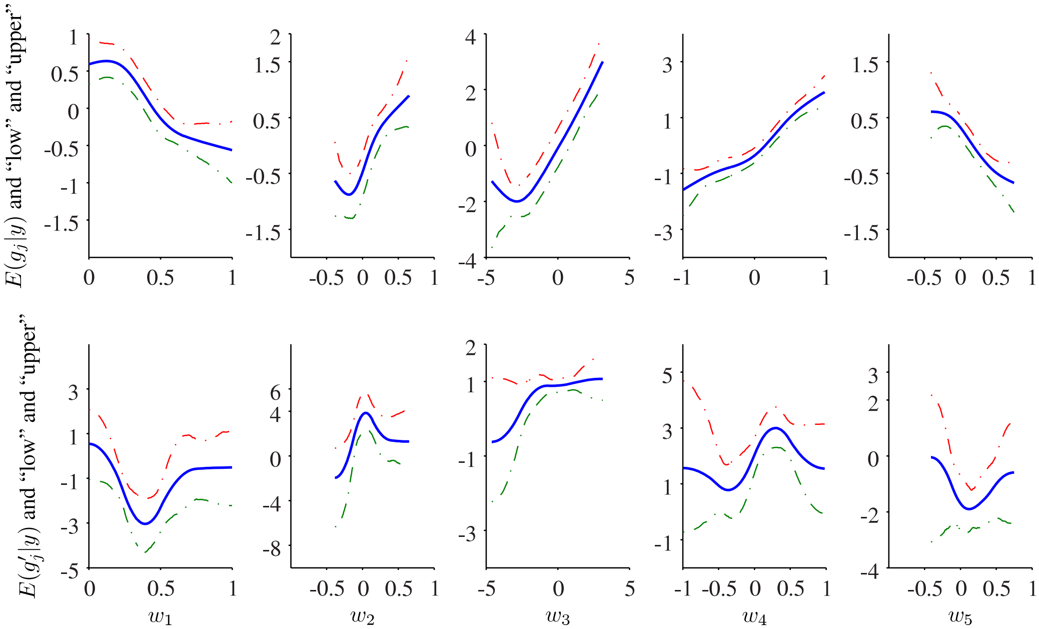

Fig. 5. Credit rating data and Student-*t* model with 5, 5, 5, 5, and 5 knots. Top panel: function estimates. Bottom panel: derivatives. Both panels: 0.025 quantile of posterior distribution ("low"), and 0.975 quantile ("upper").

to the specification of prior distributions for smoothing and other parameters of the models and examples of how to apply the approach to substantive problems. Finally, in our approach, smoothing parameters are estimated along with other parameters, and it is possible to compare models containing different number of knots, covariates, and error distributions.

Appendix

Algorithm 1. MCMC sampling of the posterior distribution in the continuous outcome model

Step 1: Sample β conditioned on $(y, \phi, \theta_{-\beta})$ from

$$\mathrm{N}_k\left(\beta|B\left(B_0^{-1}b_0 + \sum_i \sigma_i^{-2}x_i(y_i - \mu_i)\right), B\right),$$

where $B = (B_0^{-1} + \sum_i \sigma_i^{-2}x_i x_i')^{-1}$

Step 2: Sample σ^2_{ej} and σ^2_{dj}, respectively, conditioned on β_j, for $j = 1, \ldots, q$, from

$$\text{inv gamma}\left(\sigma^2_{ej}\left|\frac{\alpha_{je0}+2}{2}, \frac{\delta_{je0} + \beta_j'\Delta_j' D_{0j}\Delta_j'\beta_j}{2}\right.\right)$$

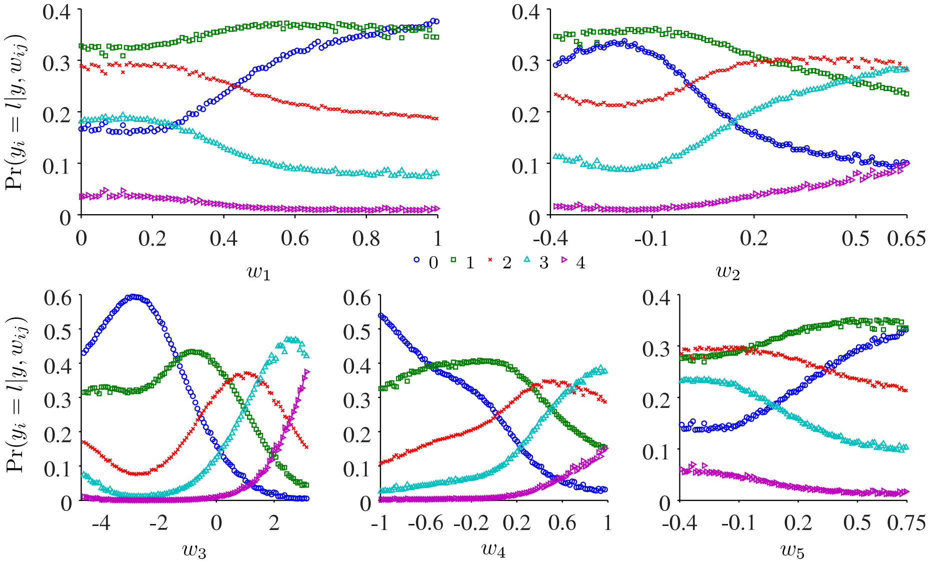

Fig. 6. Credit rating data and Student-*t* model with 5, 5, 5, 5, and 5 knots. Posterior estimates of $\Pr(y_i = k|y, w_{\bullet j})$, $k = 0, 1, 2, 3, 4$. Key for lines: ○ is 0, □ is 1, x is 2, △ is 3, ▷ is 4.

$$\text{inv gamma}\left(\sigma^2_{dj}\,\middle|\,\frac{\alpha_{jd0}+M_j-3}{2}, \frac{\delta_{jd0}+\beta_j'\Delta_j'D_{1j}\Delta_j\beta_j}{2}\right),$$

where

$$D_{0j} = \begin{pmatrix}1 & 0 & 0\\ 0 & 0_{M_j-3} & 0\\ 0 & 0 & 1\end{pmatrix}, \qquad D_{1j} = \begin{pmatrix}0 & 0 & 0\\ 0 & I_{M_j-3} & 0\\ 0 & 0 & 0\end{pmatrix}.$$

Step 3: Sample $\phi_i (i \le n)$ conditioned on $(y, \phi_{-i}, \theta, G_0)$ from the continuous–discrete mixture distribution

$$q_{i,0}\pi_0(\phi_i|y_i, \theta, G_0) + \sum_{r=1}^{k_{-i}} q_{-i,r}\delta_{\lambda^*_{-i,r}},$$

where $q_{i,0} \propto \alpha t_a\left(y_i|x_i'\beta, \frac{b(1+g)}{a}\right)$ and $q_{-i,r} \propto n_{-i,r}\mathrm{N}(y_i|\mu^*_{-i,r} + x_i'\beta, \sigma^{*2}_{-i,r})$, and $\pi_0(\lambda_i|y_i, \theta, G_0)$ is

$$\mathrm{N}(\mu_i|\hat{\mu}_i, \sigma_i^2 V)\ \text{inv gamma}\left(\sigma_i^2\,\middle|\,\frac{a+1}{2}, \frac{b+(y_i-x_i'\beta)^2/(1+g)}{2}\right),$$

where $V = (g^{-1}+1)^{-1}$ *and* $\hat{\mu}_i = V(y_i - x_i'\beta)$

Step 4: Sample α conditioned on ϕ from

$$\frac{a_0+k_n-1}{n(d_0-\log u)}\text{gamma}(a_0+k_n, d_0-\log u) + \left(1-\frac{a_0+k_n-1}{n(d_0-\log u)}\right)\text{gamma}(a_0+k_n-1, d_0-\log u),$$

where k_n denotes the number of distinct values in ϕ and $u \sim \text{beta}(\alpha+1, n)$.

Step 5: Go to 1.

Algorithm 2. MCMC sampling of the posterior distribution in the ordinal outcome model

Step 1: Sample a conditioned on (y, ϕ, θ) by the Metropolis–Hastings algorithm. Calculate

$$m = \arg\max_a \log\Pr(y|\phi, \beta, a)$$

and $V = \left\{-\partial\log\Pr(y|\phi,\theta)/\partial a\partial a'\right\}^{-1}$ the negative inverse of the hessian at m. Propose

$$a^\dagger \sim t_\xi(a|m, V)$$

(where ξ is 15 (say)), calculate the probability of move

$$\alpha_{MH}(a, a^\dagger|y, \phi, \theta_{-a}) = \min\left\{\frac{\pi(a^\dagger)\Pr(y|\phi, \beta, a^\dagger)}{\pi(a)\Pr(y|\phi, \beta, a)}\frac{t_\xi(a|m, V)}{t_\xi(a^\dagger|m, V)}, 1\right\},$$

and move to $a^\dagger$ with probability α_{MH}. Transform the new a to c. Next sample $\{y_i^*\}$ conditioned on (y, ϕ, θ) from the normal distributions

$$\mathrm{N}(x_i'\beta, \lambda_i^{-1})$$

truncated to the interval (c_{j-1}, c_j) if $y_i = j$.

Step 2: Sample β conditioned on (y^*, ϕ) from

$$\mathrm{N}_k\left(\beta|B\left(B_0^{-1}b_0 + \sum_i \lambda_i x_i y_i^*\right), B\right),$$

where $B = (B_0^{-1} + \sum_i \lambda_i x_i x_i')^{-1}$.

Step 3: Sample σ^2_{ej} and σ^2_{dj}, respectively, conditioned on β_j, for $j = 1, \ldots, q$, from

$$\text{inv gamma}\left(\sigma^2_{ej}\,\middle|\,\frac{\alpha_{je0}+2}{2}, \frac{\delta_{je0}+\beta_j'\Delta_j'D_{0j}\Delta_j'\beta_j}{2}\right)$$

$$\text{inv gamma}\left(\sigma^2_{dj}\,\middle|\,\frac{\alpha_{jd0}+M_j-3}{2}, \frac{\delta_{jd0}+\beta_j'\Delta_j'D_{1j}\Delta_j\beta_j}{2}\right),$$

where

$$D_{0j} = \begin{pmatrix}1 & 0 & 0\\ 0 & 0_{M_j-3} & 0\\ 0 & 0 & 1\end{pmatrix} \text{ and } D_{1j} = \begin{pmatrix}0 & 0 & 0\\ 0 & I_{M_j-3} & 0\\ 0 & 0 & 0\end{pmatrix}.$$

Step 4: Sample $\lambda_i (i \le n)$ conditioned on $(y^*, \phi_{-i}, \theta, G_0)$ from the continuous–discrete mixture distribution

$$q_{i,0}\pi_0(\lambda_i|y_i^*, \theta, G_0) + \sum_{r=1}^{k_{-i}} q_{-i,r}\delta_{\lambda^*_{-i,r}},$$

where $q_{i,0} \propto \alpha t_\nu\left(y_i|x_i'\beta, 1\right)$ and $q_{-i,r} \propto n_{-i,r}\mathrm{N}(y_i^*|x_i'\beta, \lambda^{*-1}_{-i,r})$, $\lambda^*_{-i,r}$ is the *rth* unique λ_i in the collection ϕ_{-i}, and $\pi_0(\lambda_i|y_i^*, \theta, G_0)$ is

$$\text{gamma}\left(\lambda_i\,\middle|\,\frac{\nu+1}{2}, \frac{\nu+(y_i^*-x_i'\beta)^2}{2}\right)$$

Step 5: Sample α conditioned on ϕ from

$$\frac{a_0+k_n-1}{n(d_0-\log u)}\text{gamma}(a_0+k_n, d_0-\log u) + \left(1-\frac{a_0+k_n-1}{n(d_0-\log u)}\right)\text{gamma}(a_0+k_n-1, d_0-\log u),$$

where k_n denotes the number of distinct values in ϕ and $u \sim \text{beta}(\alpha+1, n)$.

Step 6: Go to 1.

Algorithm 3. Likelihood computation in the continuous outcome model

Step 1: Initialize $i=1$ and calculate

$$\begin{aligned} u_1 = p(y_1|\beta^*, G_0) &= \int f(y_1|\phi_1, \beta^*)\, \mathrm{d}G_0(\phi_1) \\ &= \mathrm{t}_a\left(y_1|x_1'\beta^*, \frac{b(1+g)}{a}\right), \end{aligned}$$

and set $s_1 = 1$

Step 2: For $i>1$, suppose that there are k_{i-1} clusters and the r^{th} cluster has $n_{i-1,r}$ elements. Calculate the posterior of ϕ_i, $H_{i-1,r}(\phi_i)$, based on $\{y_l : l<i, s_l=r\}$ and given by

$$\begin{aligned} H_{i-1,r}(\phi_i) &= \mathrm{N}(\mu_i|\hat{\mu}_{i-1,r}, \sigma_i^2 V_{i-1,r}) \\ &\quad\times \text{inv gamma}\left(\sigma_i^2 \middle| \frac{a_{i-1,r}}{2}, \frac{b_{i-1,r}}{2}\right), \end{aligned}$$

where

$$\begin{aligned} V_{i-1,r} &= (g^{-1}+n_{i-1,r})^{-1} \\ \hat{\mu}_{i-1,r} &= V_{i-1,r}\sum_{l<i,s_l=r}(y_l - x_l'\beta) \\ a_{i-1,r} &= a + n_{i-1,r} \\ b_{i-1,r} &= b + \sum_{l<i,s_l=r}(y_l - x_l'\beta)^2 \\ &\quad - g\left(\sum_{l<i,s_l=r}(y_l - x_l'\beta)\right)^2 \Big/ (1+gn_{i-1,r}). \end{aligned}$$

Then, calculate the predictive density

$$\begin{aligned} u_i &= f(y_i|y_{(i-1)}, s_{(i-1)}, \beta^*, \alpha^*, G_0) \\ &= \int f(y_i, \phi_i|y_{(i-1)}, s_{(i-1)}, \beta^*, \alpha^*, G_0)\,\mathrm{d}\phi_i \\ &= \frac{\alpha^*}{\alpha^*+i-1}\int f(y_i|\phi_i, \beta^*)\,\mathrm{d}G_0(\phi_i) \\ &\quad + \sum_{r=1}^{k_{i-1}}\int \frac{n_{i-1,r}}{\alpha^*+i-1} f(y_i|\phi_i, \beta^*)\,\mathrm{d}H_{i-1,r}(\phi_i) \\ &= \frac{\alpha^*}{\alpha^*+i-1}\mathrm{t}_a\left(y_i|x_i'\beta^*, \frac{b(1+g)}{a}\right) \\ &\quad + \frac{1}{\alpha^*+i-1}\sum_{r=1}^{k_{i-1}} n_{i-1,r}\mathrm{t}_{a_{i-1,r}}(y_i|\hat{\mu}_{i-1,r} \\ &\quad + x_i'\beta^*, a_{i-1,r}^{-1}\, b_{i-1,r}(1+V_{i-1,r})), \end{aligned}$$

where the notation $z_{(i)} = (z_1, \ldots, z_i)$ denotes the history up to i. Complete this step by sampling s_i from the posterior distribution $\pi(s_i|y_{(i)}, s_{(i-1)}, \beta^*, \alpha^*, G_0)$, letting

$$s_i = \begin{cases} r, \quad \text{with prob } K\dfrac{n_{i-1,r}}{\alpha^*+i-1}\mathrm{t}_{a_{i-1,r}} \\ \quad\times\left(y_i|\hat{\mu}_{i-1,r}+x_i'\beta^*, \dfrac{b_{i-1,r}}{a_{i-1,r}}(1+V_{i-1,r})\right), \\ \quad r \le k_{i-1}, \\ k_{i-1}+1, \quad \text{with prob } K\dfrac{\alpha^*}{\alpha^*+i-1}\mathrm{t}_a \\ \quad\times\left(y_i|x_i'\beta^*, \dfrac{b(1+g)}{a}\right), \end{cases}$$

where K is the normalizing constant

Step 3: Go to 2

Step 4: Calculate $f = u_1\prod_{i=2}^n u_i$

The average of the f's from many loops over Steps 1–4 is an estimate of the likelihood ordinate. In our examples, these steps are repeated 5000 times.

Algorithm 4. Likelihood computation in the ordinal outcome model

Step 1: Initialize $i=1$ and calculate the predictive ordinate of the first observation $y_1 = j (j = 0, 1, \ldots, J-1)$

$$\begin{aligned} u_1 &= \Pr(y_1|\beta^*, G_0) = \mathrm{T}_\nu\left(c_j^* - x_1'\beta^*, 1\right) \\ &\quad - \mathrm{T}_\nu\left(c_{j-1}^* - x_1'\beta^*, 1\right) \end{aligned}$$

where

$$\begin{aligned} c^* &= (c_{-1}^*, c_0^*, c_1^*, \ldots, c_{J-2}^*, c_{J-1}^*) \\ &= (-\infty, 0, c_1^*, \ldots, c_{J-2}^*, \infty) \end{aligned}$$

and then draw (y_1^*, s_1) from its posterior distribution conditioned on $y_1 = j$ by drawing y_1^* from the truncated distribution

$$y_1^*|y_1 = j, \beta^*, G_0 \sim \mathrm{t}_\nu\left(x_1'\beta^*, 1\right) I\left(c_{j-1}, c_j\right),$$

where $\mathrm{t}_\nu(\mu, 1)$ is the Student-t distribution with location μ, dispersion 1 and ν degrees of freedom, and set $s_1 = 1$

Step 2: For $i>1$, suppose that there are k_{i-1} clusters and the r^{th} cluster has $n_{i-1,r}$ elements. Calculate the posterior of λ_i, $H_{i-1,r}(\lambda_i)$, based on $\{y_l : l<i, s_l=r\}$ and given by

$$H_{i-1,r}(\lambda_i) = \text{gamma}\left(\frac{a_{i-1,r}}{2}, \frac{b_{i-1,r}}{2}\right),$$

where

$$\begin{aligned} a_{i-1,r} &= \nu + n_{i-1,r} \\ b_{i-1,r} &= \nu + \sum_{l<i,s_l=r}(y_l^* - x_l'\beta^*)^2. \end{aligned}$$

Then, calculate the predictive density of $y_i = j$

$$\begin{aligned} u_i &= \Pr(y_i = j|y_{(i-1)}, s_{(i-1)}, \beta^*, \alpha^*, G_0) \\ &= \frac{\alpha^*}{\alpha^*+i-1}\left\{\mathrm{T}_\nu\left(c_j^* - x_i'\beta^*, 1\right) - \mathrm{T}_\nu\left(c_{j-1}^* - x_i'\beta^*, 1\right)\right\} \\ &\quad + \frac{1}{\alpha^*+i-1}\sum_{r=1}^{k_{i-1}} n_{i-1,r}\left\{\mathrm{T}_{a_{i-1,r}}\left(c_j^* - x_i'\beta^*, a_{i-1,r}^{-1}b_{i-1,r}\right)\right. \\ &\quad \left. - \mathrm{T}_{a_{i-1,r}}(c_{j-1}^* - x_i'\beta^*, a_{i-1,r}^{-1}b_{i-1,r})\right\}. \end{aligned}$$

Complete this step by sampling (y_i^*, s_i) from the posterior distribution

$$\pi(y_i^*, s_i|y_{(i)}, s_{(i-1)}, \beta^*, \alpha^*, G_0)$$

by drawing y_i^* from the mixture distribution

$$y_i^*|y_i = j, y_{(i)}, s_{(i-1)}, \beta^*, G_0 \sim \frac{\alpha^*}{\alpha^* + i - 1}\mathrm{t}_\nu\left(x_i'\beta^*, 1\right) + \frac{1}{\alpha^* + i - 1}\sum_{r=1}^{k_{i-1}} n_{i-1,r}\mathrm{t}_{a_{i-1,r}}(x_i'\beta^*, a_{i-1,r}^{-1}b_{i-1,r})$$

truncated to the interval (c_{j-1}, c_j) and then draw

$$s_i = \begin{cases} r, & \text{with prob } K\dfrac{n_{i-1,r}}{\alpha^* + i - 1}\mathrm{t}_{a_{i-1,r}}(y_i^*|x_i'\beta^*, a_{i-1,r}^{-1}b_{i-1,r}),\ r \le k_{i-1}, \\ k_{i-1} + 1, & \text{with prob } K\dfrac{\alpha^*}{\alpha^* + i - 1}\mathrm{t}_\nu\left(y_i^*|x_i'\beta^*, 1\right), \end{cases}$$

where K is the normalizing constant

Step 3: Go to 2

Step 4: Calculate $f = u_1 \prod_{i=2}^n u_i$

The average of the f's from many loops over Steps 1–4 is an estimate of the likelihood ordinate. In our examples, these steps are repeated 5000 times.

References

Albert, J.H., Chib, S., 1993. Bayesian analysis of binary and polychotomous response data. Journal of the American Statistical Association 88, 669–679.

Albert, J.H., Chib, S., 2001. Sequential ordinal modeling with applications to survival data. Biometrics 57, 829–836.

Altman, E.I., Rijken, H.A., 2004. How rating agencies achieve rating stability. Journal of Banking & Finance 28, 2679–2714.

Angers, J.F., Delampady, M., 1992. Hierarchical Bayesian curve fitting and smoothing. Canadian Journal of Statistics 20, 35–49.

Antoniak, C.E., 1974. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. The Annals of Statistics 2, 1152–1174.

Basu, S., Chib, S., 2003. Marginal likelihood and Bayes factors for Dirichlet process mixture models. Journal of the American Statistical Association 98 (461), 224–235.

Basu, S., Mukhopadhyay, S., 2000. Binary response regression with normal scale mixture links. In: Dey, D.K., Ghosh, S.K., Mallick, B.K. (Eds.), Generalized Linear Models: A Bayesian Perspective. Marcel Dekker, New York, pp. 231–242.

Chib, S., 1995. Marginal likelihood from the Gibbs output. Journal of the American Statistical Association 90 (432), 1313–1321.

Chib, S., Greenberg, E., 1994. Bayes inference in regression models with ARMA (p, q) errors. Journal of Econometrics 64, 183–206.

Chib, S., Greenberg, E., 1995. Understanding the Metropolis–Hastings algorithm. The American Statistician 49 (4), 327–335.

Chib, S., Hamilton, B.H., 2002. Semiparametric Bayes analysis of longitudinal data treatment models. Journal of Econometrics 110, 67–89.

Chib, S., Jeliazkov, I., 2001. Marginal likelihood from the Metropolis–Hastings output. Journal of the American Statistical Association 96, 270–281.

Chipman, H.A., Kolaczyk, E.D., McCulloch, R.E., 1997. Adaptive Bayesian wavelet shrinkage. Journal of the American Statistical Association 92, 1413–1421.

Clyde, M., Parmigiani, G., Vidakovic, B., 1998. Multiple shrinkage and subset selection in wavelets. Biometrika 85, 391–401.

Congdon, P., 2007. Bayesian Statistical Modelling, second ed. In: Wiley Series in Probability and Statistics, John Wiley & Sons, Chichester.

Denison, D.G.T., Holmes, C.C., Mallick, B.K., Smith, A.F.M., 2002. Bayesian Methods for Nonlinear Classification and Regression. In: Wiley Series in Probability and Statistics, John Wiley & Sons, Ltd., Chichester.

Dunson, D.B., Pillai, N., Park, J., 2007. Bayesian density regression. Journal of the Royal Statistical Society, Ser. B 69 (2), 163–183.

Eilers, P.H.C., Marx, B.D., 1996. Flexible smoothing with B-splines and penalties (with discussion). Statistical Science 11 (2), 89–121.

Escobar, M.D., West, M., 1995. Bayesian density estimation and inference using mixtures. Journal of the American Statistical Association 90, 577–588.

Ferguson, T.S., 1973. A Bayesian analysis of some nonparametric problems. The Annals of Statistics 1, 209–230.

Ferguson, T.S, 1983. Bayesian density estimation by mixtures of normal distributions. In: Rizvi, M.H., Rustagi, J.S., Siegmund, D. (Eds.), Recent Advances in Statistics: Papers in Honor of Herman Chernoff on His Sixtieth Birthday. Academic Press, pp. 287–302.

Geweke, J., Keane, M., 2007. Smoothly mixing regressions. Journal of Econometrics 138 (1), 252–290.

Green, P.J., Silverman, B.W., 1994. Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach. In: Monographs on Statistics and Applied Probability, vol. 58. Chapman & Hall, London.

Griffin, J., Steel, M.F.J., 2007. Bayesian nonparametric modelling with the Dirichlet process regression smoother. CRiSM Working Paper 07-05, University of Warwick.

Härdle, W., 1990. Applied Nonparametric Regression. Cambridge University Press, Cambridge.

Hirano, K., 2002. Semiparametric Bayesian inference in autoregressive panel data models. Econometrica 70, 781–799.

Ishwaran, H., James, L.F., 2001. Gibbs sampling methods for stick-breaking priors. Journal of the American Statistical Association 96, 161–173.

Lancaster, P., Šalkauskas, K., 1986. Curve and Surface Fitting: An Introduction. Academic Press, San Diego.

Leslie, D.S., Kohn, R., Nott, D.J., 2007. A general approach to heteroscedastic linear regression. Statistics and Computing 17, 131–146.

Li, Q., Racine, J.S., 2006. Nonparametric Econometrics: Theory and Practice. Princeton University Press, Princeton.

MacEachern, S.N., Müller, P., 1998. Estimating mixtures of Dirichlet process models. Journal of Computational & Graphical Statistics 7, 223–238.

Müller, P., Erkanli, A., West, M., 1996. Bayesian curve fitting using multivariate normal mixtures. Biometrika 83, 67–79.

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

Pagan, A., Ullah, A., 1999. Nonparametric Econometrics. In: Themes in Modern Econometrics, Cambridge University Press, Cambridge.

Poirier, D.J., 1973. Piecewise regression using cubic spline. Journal of the American Statistical Association 68 (343), 515–524.

Ruppert, D., 2002. Selecting the number of knots for penalized splines. Journal of Computational & Graphical Statistics 11 (4), 735–757.

Ruppert, D., Wand, M.P., Carroll, R.J., 2003. Semiparametric Regression. In: Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge.

Sethuraman, J., 1994. A constructive definition of Dirichlet priors. Statistica Sinica 4, 639–650.

Tiwari, R., Jammalamadaka, S., Chib, S., 1988. Bayes prediction density and regression estimation: A semi parametric approach. Empirical Economics 13, 209–222.

Vannucci, M., Corradi, F., 1999. Covariance structure of wavelet coefficients: Theory and models in a Bayesian perspective. Journal of the Royal Statistical Society, Series B: Statistical Methodology 61, 971–986.

Verbeek, M., 2008. A Guide to Modern Econometrics, third ed. John Wiley & Sons, Ltd., Chichester.

Villani, M., Kohn, R., Giordani, P., September 2007. Nonparametric regression density estimation using smoothly varying normal mixtures. Working Paper 211, Sveriges Riksbank, Stockholm.

Wasserman, L., 2006. All of Nonparametric Statistics. In: Springer Texts in Statistics, Springer, New York.

Wood, S.N., 2006. Generalized Additive Models: An Introduction with R. In: Texts in Statistical Science, Chapman & Hall/CRC, Boca Raton, FL.