Statistica Sinica **17**(2007), 481-504

EMPIRICAL BAYES DENSITY REGRESSION

David B. Dunson

National Institute of Environmental Health Sciences

Abstract: In Bayesian hierarchical modeling, it is often appealing to allow the conditional density of an (observable or unobservable) random variable Y to change flexibly with categorical and continuous predictors $\mathbf{X}$. A mixture of regression models is proposed, with the mixture distribution varying with $\mathbf{X}$. Treating the smoothing parameters and number of mixture components as unknown, the MLE does not exist, motivating an empirical Bayes approach. The proposed method shrinks the spatially-adaptive mixture distributions to a common baseline, while penalizing rapid changes and large numbers of components. The discrete form of the mixture distribution facilitates flexible classification of subjects. A Gibbs sampling algorithm is developed, which embeds a Monte Carlo EM-type stage to estimate smoothing and hyper-parameters. The method is applied to simulated examples and data from an epidemiologic study.

Key words and phrases: Conditional density estimation, Dirichlet process, EM algorithm, Gaussian mixture sieve, Gibbs sampling, nonlinear regression, nonparametric Bayes, smoothing.

1. Introduction

In assessing the relationship between a response variable $Y \in \mathcal{Y}$ and predictors $\mathbf{X} = (X_1, \ldots, X_p)' \in \mathcal{X}$, one typically relies on a mean or quantile-based regression model with a constant residual density, possibly up to a scale factor $\tau(\mathbf{X})$ allowing heteroscedasticity. In many applications, this structure may be overly restrictive, because scientific interest focuses on identifying the features of the response density which vary across $\mathcal{X}$. For example, in epidemiologic applications, differences in susceptibility due to unmeasured environmental and genetic factors can lead to changes in the shape of the distribution of a health outcome with changing dose of a drug or chemical. Such changes can result in increasing skewness or additional modes at higher exposure levels.

In recent years, there has been an active interest in the development of methods for conditional density estimation, often motivated by time series applications. For example, Yu and Jones (1998) applied the double-kernel, local linear approach of Fan, Yao and Tong (1996) to the problem. Hall, Wolff and Yao (1999) proposed improvements based on local logistic and adjusted Nadaraya-Watson estimators. Also using the double-kernel approach, Fan and Yim (2004)

proposed a cross validation method to address the important problem of bandwidth selection. For related articles, refer to Hyndman, Bashtannyk and Grunwald (1996), Bashtannyk and Hyndman (2001) and Hyndman and Yao (2002).

The Bayesian literature on the topic of conditional density estimation, referred to in this article as density regression, is sparse. Müller, Erkanli and West (1996) proposed an innovative Bayesian approach that relies on specifying a Dirichlet process mixture of normals for the joint distribution of Y and $\mathbf{X}$, and then deriving the resulting conditional distribution of Y given $\mathbf{X}$. The method essentially results in a locally weighted mixture of normal regression models, which Müller, Erkanli and West (1996) used to estimate the mean regression function but not the conditional density function.

From a Bayesian perspective, one can allow a distribution function to be unknown by choosing a prior distribution with support on the space of probability measures (refer to Müller and Quintana (2004) for a recent review). The most common choice of prior for an unknown distribution is the Ferguson (1973) Dirichlet process (DP). Letting F denote the random distribution, the typical notation expresses the DP prior as $F \sim DP(\alpha F_0)$, where α is a precision parameter and F_0 is the base measure. In the simple case in which $Y_i \overset{i.i.d.}{\sim} F$, for $i = 1, \ldots, n$, the posterior is $(F \,|\, Y_1, \ldots, Y_n) \sim DP\big(\alpha F_0 + nF_n\big)$, where $F_n = (1/n)\sum_{i=1}^n \delta_{Y_i}$ is the empirical probability measure, with δ_ϕ the Dirac measure concentrated at ϕ. Hence, the DP prior is conjugate, and α controls the degree of shrinkage towards F_0.

Due to the discreteness constraint, the DP tends to be too inflexible as a prior for F directly, but is a good choice for a mixture distribution. DP mixture models have been widely studied in the Bayesian literature (Escobar and West (1995, 1998), MacEachern and Müller (1998), among many others) and have broad applications. However, as in frequentist density estimation, the results can be sensitive to the choice of smoothing parameter (α). For this reason, Escobar and West (1998) recommend choosing a hyperprior distribution for α to allow the data to inform about its value. As discussed in MacEachern and Müller (1998), one can also allow uncertainty in F_0 by choosing a parametric form (e.g, Gaussian) with unknown parameters (mean, variance). An alternative is to use an empirical Bayes approach to estimate α (Liu (1996)) or both α and F_0 (McAuliffe, Blei and Jordan (2006)).

To address the density regression problem, it is necessary to consider priors for a collection of dependent, random distributions $(F_x, x \in \mathcal{X})$. A simple approach is to use a DP or DP mixture for each F_x, allowing for dependence through a regression in the base measure (Cifarelli and Regazzini (1978), Mira and Petrone (1996) and Giudici, Mezzetti and Muliere (2003)). Although this approach is limited by only allowing dependence in features captured by

the base parametric model, flexibility can be improved somewhat by allowing the hyperparameters to have an unknown distribution (Tomlinson and Escobar (1999)). MacEachern (1999) proposed an alternative dependent Dirichlet process (DDP) approach based on defining a stochastic process for the atoms in Sethuraman's (1994) stick-breaking representation of the DP. The DDP was recently applied to ANOVA (De Iorio, Müller, Rosner and MacEachern (2004)) and spatial (Gelfand, Kottas and MacEachern (2005)) applications.

The DDP-based approaches are limited somewhat in flexibility by assuming a fixed number of atoms, with constant probability weights on these atoms. Griffin and Steel (2006) relaxed this assumption through use of an order-based Dirichlet process, while Duan, Guindani and Gelfand (2005) instead define a spatial stick-breaking process that generalizes the DP. An alternative is to incorporate dependency through mixtures of independent DP components. Müller, Quintana and Rosner (2004) used this approach to define a hierarchical dependency structure and borrow information across studies. Dunson (2006) generalized the idea to a time series setting, and Dunson, Pillai and Park (2007) proposed a kernel-weighted mixture of DPs (WMDP) motivated by interest in conditional density estimation. They used a nonparametric mixture of linear regression models for the conditional density of Y given $\mathbf{X}$, allowing the unknown collection of mixture distributions to vary with predictors through a WMDP prior.

Although the Dunson et al. (2007) approach is very flexible, and can be implemented with a straightforward Markov chain Monte Carlo (MCMC) algorithm, a potential criticism is sensitivity to subjectively-chosen hyperparameters. In particular, smoothing, borrowing of information across $\mathcal{X}$, and clustering of subjects is controlled by weight parameters $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)'$ on the different basis locations, a kernel precision parameter ψ and the DP parameter α. As a default analysis that avoids sensitivity to subjectively-chosen hyperparameters, it is appealing to consider an empirical Bayes approach for estimating the hyperparameter values. Generalized maximum likelihood estimation (GMLE) (Wecker and Ansley (1983), Wahba (1985) and Stein (1990)) derived from an empirical Bayes framework is a common approach for estimating smoothing parameters in nonlinear regression.

This article develops an empirical Bayes approach for density regression, relying on a local mixture of parametric regression models. We borrow information across the predictor space using a kernel-weighted urn scheme, which is motivated by the WMDP prior of Dunson et al. (2007). This urn scheme incorporates two smoothing parameters, α and ψ, which control the generation of new clusters and borrowing of information. Focusing on location-scale mixtures of normal linear regression models, a Gibbs sampling algorithm is developed for posterior computation, initially considering hyper-parameters, including α, ψ, as known.

Methods are then described for empirical estimation of hyper-parameters using a Monte Carlo EM-type procedure.

Section 2 describes the mixture model, motivates the empirical Bayes approach, and considers theoretical properties. Section 3 proposes the estimation algorithm. Section 4 applies the method to simulated data examples. Section 5 contains an application to epidemiologic data, and Section 6 discusses the results.

2. Mixture Models for Density Regression

2.1. Mixture structure and background

The probability density function of the response Y conditional on predictors $\mathbf{x} \in \mathcal{X}$ is expressed as a mixture of parametric densities as

$$f(y \mid \mathbf{x}) = \int f(y \mid \mathbf{x}, \phi)\, dG_{\mathbf{x}}(\phi), \tag{1}$$

where $f(y \mid \mathbf{x}, \phi)$ is a known density on $\mathcal{Y}$ that depends on the finite-dimensional parameter $\phi = (\phi_1, \ldots, \phi_q)' \in \Phi$, and $G_{\mathbf{x}}$ is a random mixing distribution on Φ indexed by the predictor $\mathbf{x} \in \mathcal{X}$. Mixtures of Gaussian or exponential family densities have been widely used to obtain flexible density estimators. Most of the theoretical work does not include covariates in the mixture formulation (i.e., replace (1) with $f(y) = \int f(y \mid \phi)\, dG(\phi)$). A recent focus has been on Gaussian mixture sieves, which use a location or location-scale mixture of Gaussian densities, with the number of components in G increasing with sample size (Ghosal, Ghosh and Ramamoorthi (1999), Genovese and Wasserman (2000) and Ghosal and Van der Vaart (2001)).

In the general setting, $f(y \mid \mathbf{x}, \phi)$ could be chosen to correspond to a linear or generalized linear regression model. The special case in which $G_{\mathbf{x}} = \sum_{h=1}^{k} p_h(\mathbf{x})\delta_{\theta_h}$, with the weights $p_h(\mathbf{x})$ modeled using a probabilistic decision tree, corresponds to the hierarchical mixture of experts (HME) model proposed in the neural computing literature (Jacobs, Jordan, Nowlan and Hinton (1991) and Jordan and Jacobs (1994)). Jiang and Tanner (1999) showed that the HME can be used to approximate exponential family densities with arbitrary smooth mean regression functions. In addition, Viele and Tong (2000) showed posterior consistency for a narrower class of mixtures of Gaussian linear models with $p_h(\mathbf{x}) = p_h$ and k fixed.

An alternative to Viele and Tong (2000) would be to assume $G_{\mathbf{x}} = G \sim DP(\alpha G_0)$, with α the DP precision parameter and G_0 an initial guess at the mixture distribution. From Sethuraman's (1994) stick-breaking representation, this is equivalent to letting $G = \sum_{h=1}^{\infty} p_h \delta_{\theta_h}$, with $\theta_h \overset{i.i.d.}{\sim} G_0$ and $p_h / \prod_{l=1}^{h-1}(1 - p_l) \overset{i.i.d.}{\sim} \text{beta}(1, \alpha)$. Given data $y_i \overset{ind}{\sim} f(\cdot \mid \mathbf{x}_i)$ for $i = 1, \ldots, n$, this DP mixture structure will allocate the n subjects into k groups, each with a common value of

ϕ. The clustering of subjects into groups is clear from the Pólya urn scheme of Blackwell and MacQueen (1973) that prescribes that the conditional distribution of ϕ_i given $\boldsymbol{\phi}^{(i)} = (\phi_1, \ldots, \phi_{i-1}, \phi_{i+1}, \ldots, \phi_n)'$ as

$$\big(\phi_i \,|\, \boldsymbol{\phi}^{(i)}, \alpha\big) \sim \left(\frac{\alpha}{\alpha + n - 1}\right) G_0 + \left(\frac{1}{\alpha + n - 1}\right) \sum_{j \neq i} \delta_{\phi_j}, \tag{2}$$

so that subject i is either assigned to one of the existing clusters by letting $\phi_i = \phi_j$, for some $j \neq i$, or assigned to a new cluster with $\phi_i \sim G_0$. Because the number of clusters (k) increases with n, the resulting model is a type of Gaussian mixture sieve, generalizing the location and location-scale mixtures to a mixture of regression models.

Unfortunately, this specification is not sufficiently flexible due to the assumption of a constant mixture distribution. For example, consider the simple case in which the true conditional densities of Y follow finite normal mixture models:

$$f(y \,|\, \mathbf{x}) = \sum_{h=1}^{k(\mathbf{x})} p_h(\mathbf{x})\, \mathcal{N}\big(y; \mu_h(\mathbf{x}), \sigma_h^2\big), \quad \text{for all } y \in \mathcal{Y} \text{ and } \mathbf{x} \in \mathcal{X},$$

with $p_h(\cdot)$ and $\mu_h(\cdot)$ smooth functions of $\mathbf{x}$. In general, it is not possible to accurately approximate $f(y \,|\, \mathbf{x})$ over $\mathcal{Y} \equiv \Re$ and $\mathcal{X}$ using a mixture of Gaussian linear models without allowing the mixture distribution to vary with $\mathbf{x}$.

2.2. Generalized maximum likelihood

Before placing structure on the collection of unknown mixture distributions, it is informative to consider a maximum likelihood approach. Under mixture model (1), the conditional likelihood of $\mathbf{y} = (y_1, \ldots, y_n)'$ given $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ is

$$L\big(G_{\mathbf{X}}; \mathbf{y}, \mathbf{X}\big) = \left\{ \prod_{i=1}^{n} \int f(y_i \,|\, \mathbf{x}_i, \phi_i)\, dG_{\mathbf{x}_i}(\phi_i) \right\}, \tag{3}$$

where $G_{\mathbf{X}} = \{G_{\mathbf{x}_i}, i = 1, \ldots, n\}$ denotes the collection of unknown mixing distributions at the observed predictor values.

Lemma 1. *If one allows a distinct $G_{\mathbf{x}_i}$ for each $\mathbf{x}_i \in \mathcal{X}$, with $\mathcal{X}$ a continuous sample space, then the nonparametric MLE of $G_{\mathbf{X}}$ under $L(G_{\mathbf{X}}; \mathbf{y}, \mathbf{X})$ does not exist for $q > 1$.*

Noting the inequality $\mathrm{E}_\phi\{f(y \,|\, \mathbf{x}, \phi)\} \leq f(y \,|\, \mathbf{x}, \widehat{\phi})$, for $\widehat{\phi} = \arg\sup_\phi f(y \,|\, \mathbf{x}, \phi)$, the MLE is

$$\widehat{G}_{\mathbf{X}} = \left\{ \widehat{G}_{\mathbf{x}_1}, \ldots, \widehat{G}_{\mathbf{x}_n} \right\} = \left\{ \delta_{\widehat{\phi}_1}, \ldots, \delta_{\widehat{\phi}_n} \right\},$$

where $\widehat{\phi}_i = \arg\sup_{\phi_i} f(y_i \,|\, \mathbf{x}_i, \phi_i)$. Lemma 1 follows directly, because the solution for $\widehat{\phi}_i$ involves maximizing a multi-parameter likelihood based on a single data point. This result shows that nonparametric maximization of the mixture likelihood (3) results in over-fitting.

To obtain reasonable estimates, it is necessary to place restrictions on $G_{\mathbf{X}}$, say by penalizing the rate of change in $G_{\mathbf{x}}$ as $\mathbf{x}$ moves across $\mathcal{X}$. Dunson et al. (2007) proposed a WMDP prior for the uncountable collection of mixture distributions, $\mathcal{G}_{\mathcal{X}} = \{G_{\mathbf{x}} : \mathbf{x} \in \mathcal{X}\}$, having the form

$$G_{\mathbf{x}} = \sum_{j=1}^{n} \pi_j(\mathbf{x}) G^*_{\mathbf{x}_j}, \quad G^*_{\mathbf{x}_j} \sim DP(\alpha G_0), \tag{4}$$

where $\boldsymbol{\pi}(\mathbf{x}) = [\pi_1(\mathbf{x}), \ldots, \pi_n(\mathbf{x})]'$ is a vector of probability weights, with $\sum_j \pi_j(\mathbf{x}) = 1$, for all $\mathbf{x} \in \mathcal{X}$. This formulation introduces independent DP random basis distributions at each of the predictor values in the sample, and then mixes across these basis distributions to obtain a prior for the unknown mixture distribution, $G_{\mathbf{x}}$, at each possible predictor value, $\mathbf{x} \in \mathcal{X}$.

Suppose that $(\phi_i \,|\, \mathbf{x}_i) \overset{ind}{\sim} G_{\mathbf{x}_i}$, for $i = 1, \ldots, n$, with $\mathcal{G}_{\mathcal{X}}$ given a WMDP prior. Then, relying on Theorem 4 in Dunson et al. (2007), we obtain the following generalization of the DP Pólya urn scheme upon marginalizing out the infinite-dimensional WMDP prior:

$$(\phi_i \,|\, \boldsymbol{\phi}^{(i)}, \mathbf{X}, \alpha) = \left(\frac{\alpha}{\alpha + w_i}\right) G_0 + \sum_{j \neq i} \left(\frac{w_{ij}}{\alpha + w_i}\right) \delta_{\phi_j}, \tag{5}$$

where $\{w_{ij}\}$ is a set of weights between 0 and 1 that depend on the function, $\boldsymbol{\pi}$, the DP parameter, α, and the predictors, $\mathbf{X}$, and $w_i = \sum_{j \neq i} w_{ij} \leq n$. Hence, instead of considering the different subjects as exchangeable, as in (2), weights are incorporated characterizing the distance between subjects.

In order to simplify modeling and obtain a more parsimonious and interpretable form, we propose to avoid an explicit specification of $\boldsymbol{\pi}$, instead relying on the generalization of the Pólya urn scheme in (5), with $w_{ij} = w_\psi(\mathbf{x}_i, \mathbf{x}_j)$. Here, $w_\psi : \mathcal{X} \times \mathcal{X} \to [0, 1]$ is a bounded kernel measuring how close two predictors are in terms of a distance measure d, with ψ a smoothing parameter controlling how rapidly $w_\psi(\mathbf{x}_1, \mathbf{x}_2) \to 0$ as $d(\mathbf{x}_1, \mathbf{x}_2)$ increases. In the limit as $\psi \to 0$, $w_\psi(\mathbf{x}, \mathbf{x}') = 0$ for any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ having $d(\mathbf{x}, \mathbf{x}') > 0$. In addition, for all $\psi > 0$, $\lim_{\mathbf{x} \to \mathbf{x}'} w_\psi(\mathbf{x}, \mathbf{x}') = 1$.

Note that the prior at (5) automatically allocates the n subjects into $k \leq n$ clusters (or mixture components) according to their ϕ_i values. Because subjects located close together are more likely to be clustered together, the prior tends to penalize changes across $\mathcal{X}$ in the parameter values. In addition, the prior tends to

favor introducing new clusters slowly with increasing n in a manner controlled by parameters, α and ψ, with new clusters added more rapidly as α and ψ increase. As noted by Genovese and Wasserman (2000), inconsistency can result when the number of components increases too rapidly. The prior for k in terms of α, ψ and n is not available in closed form.

Letting $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$ denote the unique ϕ values, (3) is replaced by

$$L\big(\boldsymbol{\theta}, k, \alpha, \psi; \mathbf{y}, \mathbf{X}\big) = \left\{ \prod_{i=1}^{n} \sum_{h=1}^{k} p_h(\mathbf{x}_i; \alpha, \psi)\, f(y_i \,|\, \mathbf{x}_i, \phi_i = \theta_h) \right\}, \tag{6}$$

where $p_h(\mathbf{x}_i; \alpha, \psi)$, the probability that a subject with predictors $\mathbf{x}_i$ is allocated to cluster h, is a function of the predictor values and the smoothing parameters α and ψ, with $\sum_{h=1}^{k} p_h(\mathbf{x}_i; \alpha, \psi) = 1$. There is no closed form expression for $p_h(\mathbf{x}_i; \alpha, \psi)$ and the unknown number of clusters k and the cluster allocation probabilities depend in a complex manner on smoothing parameters and the relative values of the predictors.

Note that (5) can be written as

$$\big(\phi_i \,|\, \boldsymbol{\phi}^{(i)}, \mathbf{X}, \alpha, \psi\big) \sim \left(\frac{\alpha}{\alpha + w_i(\psi)}\right) G_0 + \left(\frac{1}{\alpha + w_i(\psi)}\right) \sum_{h=1}^{k^{(i)}} w_{ih}^{*}(\psi) \delta_{\theta_h^{(i)}}, \tag{7}$$

where $\boldsymbol{\theta}^{(i)} = (\theta_1^{(i)}, \ldots, \theta_{k^{(i)}}^{(i)})$ denotes unique values of $\boldsymbol{\phi}^{(i)}$, $w_{ih}^{*}(\psi) = \sum_{j \neq i} 1(\phi_j = \theta_h^{(i)}) w_\psi(\mathbf{x}_i, \mathbf{x}_j)$ and $w_i(\psi) = \sum_{j \neq i} w_\psi(\mathbf{x}_i, \mathbf{x}_j)$. Potentially, one can maximize (6) using (7) to define the relationship between the cluster allocation probabilities $\{p_h(\mathbf{x}_i; \alpha, \psi)\}$ and the unknown parameters α and ψ, incorporating the identifiability constraint $\theta_1 < \cdots < \theta_k$. However, Theorem 1 notes that similar overfitting problems result as with (3). This is due to the fact that, although additional structure has been placed on the unknown mixture distributions to reduce dimensionality, there is still no penalty to limit growth in the number of clusters.

Theorem 1. *For $q > 1$ (sometimes for $q = 1$) and continuous $\mathcal{X}$, the MLE at* (6)*, with the cluster allocation probabilities defined by* (7)*, does not exist.*

To prove this result, first recall that the smoothing parameter ψ controls the rate at which $w_\psi(\mathbf{x}_i, \mathbf{x}_j) \to 0$ with increasing distance $d(\mathbf{x}_i, \mathbf{x}_j)$. In the limit as $\psi \to 0$, $w_\psi(\mathbf{x}_i, \mathbf{x}_j) = 0$ for all i, j. Hence, in this case, (7) prescribes that all subjects are assigned to their own cluster, so that $k = n$ and $\boldsymbol{\theta} = (\phi_{[1]}, \ldots, \phi_{[n]})'$, letting $\phi_{[1]}, \ldots, \phi_{[n]}$ denote the order statistics of $\boldsymbol{\phi}$ to preserve the identifiability constraint on $\boldsymbol{\theta}$. In addition, in the limit as $\psi \to 0$, the conditional MLE of α

does not exist, as α has no impact on the likelihood, and the conditional MLEs of k and $\boldsymbol{\theta}$ are

$$\widehat{k}_{[\psi=0]} = n \quad \text{and} \quad \widehat{\boldsymbol{\theta}}_{[\psi=0]} = \big(\widehat{\phi}_{[1]}, \ldots, \widehat{\phi}_{[n]}\big)', \tag{8}$$

where $\widehat{\phi}_i = \arg\sup_{\phi_i} f(y_i \,|\, \mathbf{x}_i, \phi_i)$, for $i = 1, \ldots, n$, and $\widehat{\phi}_{[1]}, \ldots, \widehat{\phi}_{[n]}$ are the order statistics. For positive ψ and $k < n$, the likelihood can only decrease, so from a similar argument to that used in Lemma 1, Theorem 1 follows directly.

From Theorem 1, it is clearly necessary to incorporate a penalty to limit the number of clusters. A natural approach is to utilize additional structure from the conditional prior at (7). In particular, as in the DP Pólya urn scheme in (2), new cluster-specific parameters are generated by independently sampling from the base measure G_0. Hence, a natural penalty arises by multiplying (6) by

$$P(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \Big\{ \prod_{h=1}^{k} g_0(\theta_h; \boldsymbol{\gamma}) \Big\}, \tag{9}$$

with $g_0(\theta; \boldsymbol{\gamma})$ denoting the probability density function corresponding to G_0. Additional flexibility is accommodated by parameterizing g_0 in terms of parameters $\boldsymbol{\gamma}$.

To demonstrate how the penalty works, again consider the limiting case as $\psi \to 0$. The conditional generalized MLE (GMLE) obtained by maximizing $L(\boldsymbol{\theta}, k, \alpha, \psi; \mathbf{y}, \mathbf{X})\, P(\boldsymbol{\theta}, \boldsymbol{\gamma})$ (initially for fixed $\boldsymbol{\gamma}$) has the same form as in (8), but with

$$\widehat{\phi}_i = \arg\sup_{\phi_i} \Big\{ f(y_i \,|\, \mathbf{x}_i, \phi_i)\, g_0(\phi_i; \boldsymbol{\gamma}) \Big\}, \tag{10}$$

resulting in a shrinkage estimator for ϕ_i. For example, for $f(y_i \,|\, \mathbf{x}_i, \phi_i) = \mathcal{N}(y_i; \mathbf{x}_i'\boldsymbol{\beta}_i, \tau_i^{-1})$ with $g_0(\boldsymbol{\beta}_i, \tau_i) = \mathcal{N}_{q-1}(\boldsymbol{\beta}_i; \boldsymbol{\beta}_0, \tau_i^{-1}\mathbf{V})\mathcal{G}(\tau_i; a_\tau, b_\tau)$ and $\mathcal{G}(z; a, b) = C(a, b)\, z^{a-1} \exp(-zb)$ denoting the gamma density with mode $(a-1)/b$, the conditional GMLE is

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_i &= \big(\mathbf{V}^{-1} + \mathbf{x}_i\mathbf{x}_i'\big)^{-1}\big(\mathbf{V}^{-1}\boldsymbol{\beta}_0 + \mathbf{x}_i y_i\big), \\ \widehat{\tau}_i &= \frac{2a + q - 2}{2b + y_i^2 + \boldsymbol{\beta}_0'\mathbf{V}^{-1}\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}_i'(\mathbf{V}^{-1} + \mathbf{x}_i\mathbf{x}_i')\widehat{\boldsymbol{\beta}}_i}, \end{aligned}$$

which shrinks $(\boldsymbol{\beta}_i, \tau_i)$ towards the mode of the normal-gamma density, g_0.

Although shrinkage allows estimation of cluster-specific parameters even when there is a single subject per cluster ($k = n$), there is strong reliance for $k = n$ on the choice of G_0. In particular, unless the variance of G_0 is high, the resulting procedure behaves similarly to the model assuming $G_{\mathbf{x}_i} \equiv G_0$. Some degree of robustness can be gained by estimating parameters $\boldsymbol{\gamma}$ characterizing G_0,

but the structure is still highly restrictive. However, as k decreases and the number of subjects per cluster grows, the degree of shrinkage of the cluster-specific parameters towards G_0 decreases, allowing lack of fit of the base parametric model $f(y \mid \mathbf{x}) = \int f(y \mid \mathbf{x}, \phi)\, dG_0(\phi; \boldsymbol{\gamma})$. Because allowing lack of fit will tend to result in a higher likelihood, the global maximum is typically not achieved at the conditional MLE given $k = n$ and $\psi \to 0$, but instead at $k << n$.

Some comments are in order. First, when the base parametric model provides an excellent approximation, the global maximum may be achieved for k approaching n. However, in this case there is no problem with overfitting due to the high degree of shrinkage toward G_0, which is parameterized in terms of the relatively low-dimensional $\boldsymbol{\gamma}$. Even in the more typical case in which $k << n$, there is still some degree of shrinkage toward G_0. This provides a mechanism of stabilizing estimation through global smoothing. The structure of (7) also allows local smoothing, because subjects having similar predictor values are much more likely to be assigned to the same cluster than subjects with widely different predictor values. The degree of borrowing of information in local neighborhoods of $\mathcal{X}$ and the size of these neighborhoods is controlled by the smoothing parameter ψ. It is appealing to allow ψ to vary with sample size, so that neighborhoods are relatively small in large samples. By decreasing α and ψ as n increases, one can limit the increase in k with n, while reducing neighborhood sizes. Because it is difficult to choose an optimal sequence $\{\alpha_n, \psi_n\}$ in advance, the recommendation is to estimate α, ψ based on the data.

2.3. Choice of kernel

In applying the approach, it is necessary to choose an explicit form for the bounded kernel function, $w_\psi(\mathbf{x}, \mathbf{x}')$. For continuous $\mathbf{x}$, a natural choice is the Gaussian kernel $w_\psi(\mathbf{x}, \mathbf{x}') = \exp(-\psi^{-1}||\mathbf{x} - \mathbf{x}'||^2)$, where $||\mathbf{x} - \mathbf{x}'||^2$ denotes the L_2 distance between $\mathbf{x}$ and $\mathbf{x}'$. A criticism of this choice for $p > 1$ is the use of a single smoothing parameter, ψ, for each of the predictors. To avoid sensitivity to differences in scale for the different elements of $\mathbf{x}$, one can normalize the predictors and then transform back to the original scale in performing inferences.

After normalization, it is possible to choose plausible values of ψ without prior knowledge of the variability in the predictors. As a default choice in fully Bayes analyses, one can take $\psi = 25/n$, with 25 changed to 10 or 50 in sensitivity analyses. In smaller sample sizes, this approach borrows information more broadly across the predictor space, focusing on increasingly narrow regions as the sample size grows. The empirical Bayes approach avoids possible sensitivity to this choice by estimating ψ.

Although other choices are possible, we focus on general use of the Gaussian kernel even in cases involving categorical or mixed predictors, motivated by parsimony and simplicity. For binary predictors and $\psi = 25/n$, such an approach

assigns low prior probability to subjects in different categories being grouped together for moderate to large samples. This effectively results in a DP mixture of normals being fitted separately to the two groups for a single binary predictor, which is a reasonable default.

The issue of what sample size is required for good performance of our approach is worth commenting on. Examining (7) carefully, it is clear that the approach tends to introduce clusters with higher probability in data-sparse regions of the predictor space, while relying on neighboring values more heavily in data-rich regions. Because new clusters are drawn from the base parametric mixture model, this structure effectively relies on global smoothing under the normal linear model to extrapolate across regions with limited data. Hence for small sample sizes, there is a greater reliance on the parametric model, but as more data become available suggesting local lack of fit of the model, deviations are automatically accommodated. Therefore, the method can be recommended for any sample size. Of course, as a practical matter, there will be limited ability to detect interesting deviations from the base model, such as evolving secondary modes, in small samples.

3. Posterior Computation

In this section, a fully Bayes algorithm is developed for posterior computation assuming known smoothing parameters α, ψ and hyperparameters $\boldsymbol{\gamma}$. Using the Gibbs sampler to integrate out the latent cluster allocation indicators, a Monte Carlo EM algorithm is then developed to estimate $\alpha, \psi, \boldsymbol{\gamma}$ via an Empirical Bayes approach.

3.1. Gibbs sampling for fully Bayes inferences

Let $\mathbf{S} = (\mathcal{S}_1, \ldots, \mathcal{S}_n)'$ be a vector of indicators denoting the global configuration of subjects to unique values $\boldsymbol{\theta}$, with $\mathcal{S}_i = h$ if $\phi_i = \theta_h$ indexing the location of the ith subject within the $\boldsymbol{\theta}$ vector. Excluding the ith subject, $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta} \setminus \phi_i$ denotes the $k^{(i)}$ unique values of $\boldsymbol{\phi}^{(i)}$ and $\mathbf{S}^{(i)}$ denotes the configuration of subjects $\{1, \ldots, n\} \setminus i$ to these values. The full conditional posterior distribution of ϕ_i is

$$\left(\phi_i \,|\, \boldsymbol{\phi}^{(i)}, \mathbf{y}, \mathbf{X}, \alpha, \psi, \boldsymbol{\gamma}\right) \propto q_{i,0} G_{i,0} + \sum_{h=1}^{k^{(i)}} q_{i,h} \delta_{\theta_h^{(i)}}, \tag{11}$$

where the posterior obtained by updating prior $G_0(\phi; \boldsymbol{\gamma})$ with likelihood $f(y_i \,|\, \mathbf{x}_i, \phi)$ is

$$G_{i,0}(\phi) = \frac{G_0(\phi; \boldsymbol{\gamma})\, f(y_i \,|\, \mathbf{x}_i, \phi)}{\int f(y_i \,|\, \mathbf{x}_i, \phi)\, dG_0(\phi; \boldsymbol{\gamma})} = \frac{G_0(\phi; \boldsymbol{\gamma})\, f(y_i \,|\, \mathbf{x}_i, \phi)}{h_i(y_i \,|\, \mathbf{x}_i, \boldsymbol{\gamma})},$$

$q_{i,0} = c\,\alpha\, h_i(y_i \,|\, \mathbf{x}_i, \boldsymbol{\gamma})$, $q_{i,h} = c\, w^*_{ih}(\psi) f(y_i \,|\, \mathbf{x}_i, \theta_h)$, and c is a normalizing constant. Note that α and ψ only appear in the expressions for the configuration probabilities $\{q_{i,h}, h = 0, 1, \ldots, k^{(i)}\}$.

Conditional on α and ψ, posterior computation can proceed via a Gibbs sampling algorithm, which alternates between (1) updating $\mathbf{S}, k$ by sampling from the full conditional posterior distribution of each $\mathcal{S}_i$; (2) updating the cluster-specific parameters $\boldsymbol{\theta}$ by sampling from the full conditional posterior given the configuration; and (3) updating $\boldsymbol{\gamma}$ by sampling from its full conditional.

Consider the case in which $f(y_i, \,|\, \mathbf{x}_i, \phi_i) = \mathcal{N}(y_i; \mathbf{x}_i'\boldsymbol{\beta}_i, \tau_i^{-1})$, with $\phi_i = (\boldsymbol{\beta}_i', \tau_i)'$ and $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{ip})'$, so that both the regression coefficients and variance can vary across clusters. A natural choice for g_0 is the multivariate normal-gamma density

$$g_0(\boldsymbol{\beta}_i, \tau_i) = |2\pi\tau_i^{-1}\mathbf{V}|^{-\frac{p}{2}} \exp\Big\{-\frac{\tau_i}{2}(\boldsymbol{\beta}_i - \boldsymbol{\beta})'\mathbf{V}^{-1}(\boldsymbol{\beta}_i - \boldsymbol{\beta})\Big\} C(a_\tau, b_\tau)\tau_i^{a_\tau - 1}\exp(-\tau_i b_\tau),$$

with $\mathbf{V} = \kappa^{-1} n(\mathbf{X}'\mathbf{X})^{-1}$ to correspond to a g-type prior. To allow uncertainty in $\boldsymbol{\gamma} = \{\boldsymbol{\beta}, \kappa\}$, one can choose the hyperprior density $\pi(\boldsymbol{\gamma}) = \mathrm{N}_p(\boldsymbol{\beta}; \boldsymbol{\beta}_0, \kappa^{-1}\mathbf{V}_0)$ $\mathcal{G}(\kappa; a_\kappa, b_\kappa)$.

After standard algebra, the marginal likelihood $h_i(y_i \,|\, \mathbf{x}_i, \boldsymbol{\gamma}) = \int f(y_i \,|\, \mathbf{x}_i, \phi_i)$ $dG_0(\phi_i; \boldsymbol{\gamma})$ is

$$h_i(y_i \,|\, \mathbf{x}_i, \boldsymbol{\beta}, \mathbf{V}, a_\tau, b_\tau) = \frac{C(a_\tau, b_\tau)|\mathbf{V}|^{-\frac{p}{2}}}{\sqrt{2\pi}C(\widetilde{a}_i, \widetilde{b}_i)|\widetilde{\mathbf{V}}_i|^{-\frac{p}{2}}},$$

where $\widetilde{\mathbf{V}}_i = (\mathbf{V}^{-1} + \mathbf{x}_i\mathbf{x}_i')^{-1}$, $\widetilde{a}_i = a_\tau + 0.5(p+1)$, $\widetilde{b}_i = b_\tau + 0.5(y_i^2 + \boldsymbol{\beta}'\mathbf{V}^{-1}\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_i'\widetilde{\mathbf{V}}_i^{-1}\widehat{\boldsymbol{\beta}}_i)$, and $\widehat{\boldsymbol{\beta}}_i = \widetilde{\mathbf{V}}_i(\mathbf{V}^{-1}\boldsymbol{\beta} + \mathbf{x}_i y_i)$. Hence, calculation of conditional configuration probabilities

$$\Pr(\mathcal{S}_i = h \,|\, \mathbf{S}^{(i)}, k^{(i)}, \alpha, \psi, \boldsymbol{\gamma}, \mathbf{y}, \mathbf{X}) = q_{ih}, \quad \text{for } h = 0, \ldots, k^{(i)} \tag{12}$$

in implementing Step 1 of the Gibbs sampler is straightforward.

In addition, letting $\boldsymbol{\theta}_h = (\boldsymbol{\beta}_h', \tau_h)$ denote the value of ϕ_i for subjects in the hth cluster, the full conditional posterior distribution of $\boldsymbol{\theta}_h$ is, after some algebra,

$$\big(\boldsymbol{\beta}_h, \tau_h \,|\, \boldsymbol{\theta}^{(h)}, \mathbf{S}, k, \boldsymbol{\gamma}, \mathbf{y}, \mathbf{X}\big) \sim \mathrm{N}_p(\boldsymbol{\beta}_h; \widehat{\boldsymbol{\beta}}_h, \tau_h^{-1}\widetilde{\mathbf{V}}_h)\, \mathcal{G}(\tau_h; \widetilde{a}_h, \widetilde{b}_h), \tag{13}$$

where $\widetilde{\mathbf{V}}_h = \big(\mathbf{V}^{-1} + \sum_{i:\mathcal{S}_i = h}\mathbf{x}_i\mathbf{x}_i'\big)^{-1}$, $\widehat{\boldsymbol{\beta}}_h = \widetilde{\mathbf{V}}_h\big(\mathbf{V}^{-1}\boldsymbol{\beta} + \sum_{i:\mathcal{S}_i = h}\mathbf{x}_i y_i\big)$, $\widetilde{a}_h = a_\tau + n_h/2$,

$$\widetilde{b}_h = b_\tau + \frac{1}{2}\Big(\sum_{i:\mathcal{S}_i = h} y_i^2 + \boldsymbol{\beta}'\mathbf{V}^{-1}\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_h'\widetilde{\mathbf{V}}_h^{-1}\widehat{\boldsymbol{\beta}}_h\Big),$$

and $n_h = \sum_{i=1}^n 1(\mathcal{S}_i = h)$. The conditional posterior distribution of hyperparameters $(\boldsymbol{\beta}, \kappa)$ is

$$\big(\boldsymbol{\beta}, \kappa \,|\, \mathbf{S}, k, \boldsymbol{\theta}, \mathbf{y}, \mathbf{X}\big) \sim \mathrm{N}_p(\boldsymbol{\beta}; \widehat{\boldsymbol{\beta}}, \kappa^{-1}\widetilde{\mathbf{V}})\, \mathcal{G}(\kappa; \widetilde{a}, \widetilde{b}), \tag{14}$$

where $\widetilde{\mathbf{V}} = (\mathbf{V}_0^{-1} + \sum_{h=1}^k \tau_h \mathbf{X}'\mathbf{X}/n)^{-1}$, $\widehat{\boldsymbol{\beta}} = \widetilde{\mathbf{V}}(\mathbf{V}_0^{-1}\boldsymbol{\beta}_0 + \sum_{h=1}^k \tau_h \mathbf{X}'\mathbf{X}\boldsymbol{\beta}_h/n)$, $\widetilde{a}_h = a_\kappa + kp/2$,

$$\widetilde{b}_h = b_\kappa + \frac{1}{2}\bigg[\boldsymbol{\beta}_0'\mathbf{V}_0^{-1}\boldsymbol{\beta}_0 + \bigg\{\sum_{h=1}^k \tau_h \boldsymbol{\beta}_h'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}_h/n\bigg\} - \widehat{\boldsymbol{\beta}}'\widetilde{\mathbf{V}}^{-1}\widehat{\boldsymbol{\beta}}\bigg].$$

Gibbs sampling proceeds by sequentially sampling from (12), which follows a multinomial closed form in this case, and (13)−(14).

3.2. Estimating smoothing parameters and hyperparameters

Note that the Gibbs sampling algorithm above requires specification of hyperparameters $\boldsymbol{\beta}_0$, $\mathbf{V}_0$, a_τ, b_τ, a_κ, b_κ in addition to the smoothing parameters α, ψ. Following an empirical Bayes approach, one can instead estimate (1) $\boldsymbol{\beta}$, (2) a_τ, b_τ, (3) κ, and (4) α, ψ, eliminating possible sensitivity to subjectively-chosen hyperparameters. Even when prior information is available and the fully Bayes approach is preferred, the empirical Bayes approach provides a useful reference analysis. Here, a hybrid Gibbs/EM-type algorithm is proposed to implement the empirical Bayes approach.

Initially consider Steps (1)−(3), with α, ψ treated as known. Given $\mathbf{S}$, k, $\{\boldsymbol{\theta}_h\}$, the generalized likelihood is proportional to

$$\bigg(\prod_{h=1}^k \bigg[\prod_{i:S_i=h} \tau_h^{\frac{1}{2}} \exp\Big\{-\frac{\tau_h}{2}(y_i - \mathbf{x}_i'\boldsymbol{\beta}_h)^2\Big\}\bigg] |\tau_h^{-1}\mathbf{V}|^{-\frac{p}{2}} \\ \times \exp\Big\{-\frac{\tau_h}{2}(\boldsymbol{\beta}_h - \boldsymbol{\beta})'\mathbf{V}^{-1}(\boldsymbol{\beta}_h - \boldsymbol{\beta})\Big\}\tau_h^{a_\tau - 1}\exp(-\tau_h b_\tau)\bigg),$$

where $\mathbf{V} = \kappa^{-1} n(\mathbf{X}'\mathbf{X})^{-1}$. A standard EM algorithm would iterate between (i) calculate the expected log generalized likelihood (ELGL) with respect to the posterior for $\mathbf{S}, k, \{\boldsymbol{\theta}_h\}$ given current estimates for the parameters; and (ii) update the parameter estimates by maximizing the ELGL. Because the E-step (i) cannot be implemented analytically, one can use the Gibbs sampler described in Subsection 3.1. In addition, because a global maximum cannot be calculated in closed form, a series of conditional maximization steps is used in place of (ii). This combines the Monte Carlo EM (MCEM) algorithm (Wei and Tanner, 1990) with the ECM algorithm (Meng and Rubin (1993)).

Letting $t = 1, \ldots, T$ index iterations of the Gibbs sampler following convergence, the conditional maximization steps are implemented by first taking expectations of sufficient statistics with respect to the posterior distribution of the model unknowns to be integrated out, including $\mathbf{S}$, k, $\{\boldsymbol{\theta}_h\}$. The resulting estimators of $\boldsymbol{\beta}$ and κ at a given iteration of the algorithm have the form

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= \frac{\sum_{t=1}^{T}\sum_{h=1}^{k^{(t)}}\tau_h^{(t)}\boldsymbol{\beta}_h^{(t)}}{\sum_{t=1}^{T}\sum_{h=1}^{k^{(t)}}\tau_h^{(t)}},\\ \widehat{\kappa} &= \frac{\left\{T^{-1}\sum_{t=1}^{T}0.5k^{(t)}p\right\}-1}{T^{-1}\sum_{t=1}^{T}\sum_{h=1}^{k^{(t)}}\tau_h^{(t)}(\boldsymbol{\beta}_h^{(t)}-\widehat{\boldsymbol{\beta}})'(\mathbf{X}'\mathbf{X})(\boldsymbol{\beta}_h^{(t)}-\widehat{\boldsymbol{\beta}})n^{-1}}.\end{aligned} \tag{15}$$

Maximization of the ELGL with respect to a_τ, b_τ proceeds by calculating

$$(\widehat{a}_\tau, \widehat{b}_\tau) = \sup_{a_\tau, b_\tau}\left\{\sum_{t=1}^{T}\sum_{h=1}^{k^{(t)}}\mathcal{G}\big(\tau_h^{(t)}; a_\tau, b_\tau\big)\right\},$$

which can be implemented in standard software for maximizing gamma likelihoods.

This formulation has conditioned on α, ψ. Estimation of α and ψ is more challenging, because it is not possible to maximize the ELGL with respect to α and ψ directly. Potentially, one could pre-specify a finite set of candidate values for α and ψ, run the algorithm separately for each candidate value, and then select the value with the maximum ELGL after convergence. Unfortunately, such a procedure is sensitive to the set of candidate values chosen, and as the dimension increases, the extreme computation involved presents a barrier to implementation. Instead, one can use a procedure based on a Stochastic EM (SEM)-type step. The SEM algorithm was introduced by Celeux and Diebolt (1985) as an approach for computing the MLE for finite mixture models, speeding up convergence and avoiding the problem of staying near an unstable stationary point of the likelihood function. The SEM draws a single value from the posterior distribution of the latent data in place of the E-step, and is a special case of the Monte Carlo EM (MCEM) algorithm (refer to McLachlan and Krishnan (1997), for an overview of EM-type algorithms).

Our proposed procedure adds the following step to the Gibbs/Monte Carlo ECM algorithm: (4a) at each iteration, sample a candidate value (α^*, ψ^*) for (α, ψ) (e.g., by sampling from a distribution centered on the current values); (4b) given this candidate value, generate new values for $\mathbf{S}, k, \{\boldsymbol{\theta}_h\}$ by alternately sampling from the full conditional posterior distributions; (4c) also sample new values for $\mathbf{S}, k, \{\boldsymbol{\theta}_h\}$ given the current (α, ψ); and (4d) set the new value of (α, ψ) equal to the choice that maximizes the LGL. As the algorithm progresses, the

iterates will tend to move stochastically towards values corresponding to a high LGL, converging to a stationary distribution. Simulations and data examples have exhibited show rapid convergence.

3.3. Density estimation

Our interest focuses on estimating the response density for new subjects having a range of different values of $\mathbf{x} \in \mathcal{X}$. In particular, letting $i = n+1$ denote a new subject with predictor value $\mathbf{x}_{n+1}$, the goal is estimation of $f(y_{n+1} \,|\, \mathbf{x}_{n+1})$. Consider two strategies: (1) calculate the posterior mean and credible intervals for $f(y_{n+1} \,|\, \mathbf{x}_{n+1})$, marginalizing across the posterior distribution of $\mathbf{S}, k, \{\boldsymbol{\theta}_h\}$ with estimates of the hyperparameters plugged in (*model-averaged estimator*); (2) plug-in the values of $\mathbf{S}, k, \{\boldsymbol{\theta}_h\}$ that maximize the LGL to obtain a MAP estimator for $f(y_{n+1} \,|\, \mathbf{x}_{n+1})$ (*preferred model estimator*).

The first approach can be implemented utilizing the simple form for the conditional predictive density:

$$f(y_{n+1} \,|\, \mathbf{x}_{n+1}, \mathbf{y}, \mathbf{X}, \mathbf{S}, k, \boldsymbol{\theta}, \boldsymbol{\gamma}, \alpha, \psi) = \left(\frac{\alpha}{\alpha + w_{n+1}(\psi)}\right) h_{n+1}(y_{n+1} \,|\, \mathbf{x}_{n+1}, \boldsymbol{\gamma}) + \left(\frac{1}{\alpha + w_{n+1}(\psi)}\right) \sum_{h=1}^{k} w^*_{n+1,h}(\psi) \mathcal{N}(y_{n+1}; \mathbf{x}'_{n+1}\boldsymbol{\beta}_h, \tau_h^{-1}), \tag{16}$$

where $w_{n+1}(\psi) = \sum_{i=1}^{n} w_{n+1,i}(\psi)$ and $w^*_{n+1,h}(\psi) = \sum_{i=1}^{n} 1(\mathcal{S}_i = h) w_{n+1,i}(\psi)$. One can simply calculate and store (16) for a range of y_{n+1} and $\mathbf{x}_{n+1}$ values of interest at each Gibbs iteration after convergence, basing posterior summaries on a large number of iterations.

Instead of integrating out the number of clusters k, the configuration of subjects to clusters $\mathbf{S}$, and the cluster-specific parameters $\boldsymbol{\theta}$, the second approach plugs in estimates. These estimates are obtained by monitoring the LGL at each iteration of the algorithm after convergence, and selecting the values at the iteration corresponding to the maximum of the LGLs. Less computationally intensive mode finding procedures were also considered. These did not require the Gibbs sampling step. Unfortunately such procedures converge to a local mode, which is typically far from the global mode, given the high degree of multimodality in the generalized likelihood surface. The proposed Monte Carlo approach had much better performance in the cases considered.

Under approach 2, the plug-in density estimator

$$\widehat{f}(y_{n+1} \,|\, \mathbf{x}_{n+1}) = \left(\frac{\widehat{\alpha}}{\widehat{\alpha} + \widehat{w}_{n+1}}\right) \widehat{h}(y_{n+1} \,|\, \mathbf{x}_{n+1}) + \left(\frac{1}{\widehat{\alpha} + \widehat{w}_{n+1}}\right) \sum_{h=1}^{\widehat{k}} \widehat{w}^*_{n+1,h} \mathcal{N}(y_{n+1}; \mathbf{x}'_{n+1}\widehat{\boldsymbol{\beta}}_h, \widehat{\tau}_h^{-1})$$

was used, with estimates of the weight functions obtained by plugging in the estimated ψ and $\mathbf{S}$ values. Note that this expression can be applied very quickly for a wide range of y and $\mathbf{x}$ values of interest without requiring monitoring of the predictive density for each y and $\mathbf{x}$ during the MCMC implementation. Hence for purposes of interpretation, rapid exploration of changes in the density across $\mathcal{X}$, and calculation of summaries such as quantile regression curves, this plug-in estimator is very useful.

4. Simulation Examples

To assess the performance of the fully Bayes and empirical Bayes approaches, a variety of simulation examples were considered. For the fully Bayes analyses, we let $\psi = 25/n$ as recommended above, $\alpha = 0.1$ to favor the introduction of few clusters within a local region, $\boldsymbol{\beta}_0 = \mathbf{0}$, $\mathbf{V}_0 = n(\mathbf{X}'\mathbf{X})^{-1}$, $a_\tau = 1$, $b_\tau = 0.5$, $a_\kappa = 1$ and $b_\kappa = 0.5$. These priors were held fixed across the simulation cases. Results are presented here for $n = 1,000$, though very similar performance was obtained for $n = 500$. The focus here is on the case with a single continuous predictor, $x_i \overset{i.i.d.}{\sim} \mathcal{U}[0,1]$, though similar results were obtained in runs with an additional binary predictor.

To characterize increasing complexity in the model for $f(y_i \,|\, x_i)$, consider (1) a normal linear regression model with heteroscedastic errors, $f(y_i \,|\, x_i) = \mathcal{N}\big(-2+5x_i, (1+x_i)^2\big)$; (2) the same as (1), but with a non-linear mean function $\mathrm{E}(y_i \,|\, x_i) = x_i - 2x_i^3$; and (3), where the density is a finite mixture of normals,

$$f(y_i \,|\, x_i) = \sum_{h=1}^{k} p_h(x_i)\mathcal{N}(y_i; \theta_{h1} + \theta_{h2}x_i, \theta_{h3}^{-1}),$$

with $p_h(x) = x^{\rho_h}/\sum_{l=1}^{k} x^{\rho_l}$, $\rho_h \overset{i.i.d.}{\sim} \mathcal{U}(-1,1)$, $(\theta_{h1}, \theta_{h2}) \overset{i.i.d.}{\sim} \mathcal{N}\big([-2,5], 0.2n(\mathbf{X}'\mathbf{X})^{-1}\big)$, and $\theta_{h3} \sim \mathcal{G}(10,1)$, for $h = 1, \ldots, k$, with $k = 5$.

For each simulated data set, the algorithms proposed in Section 3 were applied to obtain density estimators under the fully Bayes approach, and two empirical Bayes approaches, using a grid of 100 y values spanning the range of the observed data. The conditional densities were estimated for the set of x values corresponding to the $[5, 10, 17.5, 25, 37.5, 50, 62.5, 75, 82.5, 90, 95]$ percentiles of the empirical distribution. The Gibbs sampler was run for 10,000 iterations in each case, updating the hyperparameters every 100 iterations for the empirical Bayes procedure, and discarding the first 5,000 iterations as a burn-in. The hyperparameters and smoothing parameters converged steadily to a stationary distribution in each case, and the burn-in interval was more than sufficient.

For Case 1, the fully Bayes density estimates were very close to the truth across the range of x, with only slight deviations for x near the edge of the range.

In particular, the peak of the density was slightly underestimated for x at the 5th percentile, while the mean was slightly underestimated for x at the 95th percentile. However, the true density was enclosed in 99% pointwise credible intervals in each case. Similar performance was observed for the empirical Bayes estimates, as shown in Figure 1. Performance for each of the approaches was even better in Case 2, with only very slight under-estimation of the peak for x in the 5th percentile, but excellent performance at all other values.

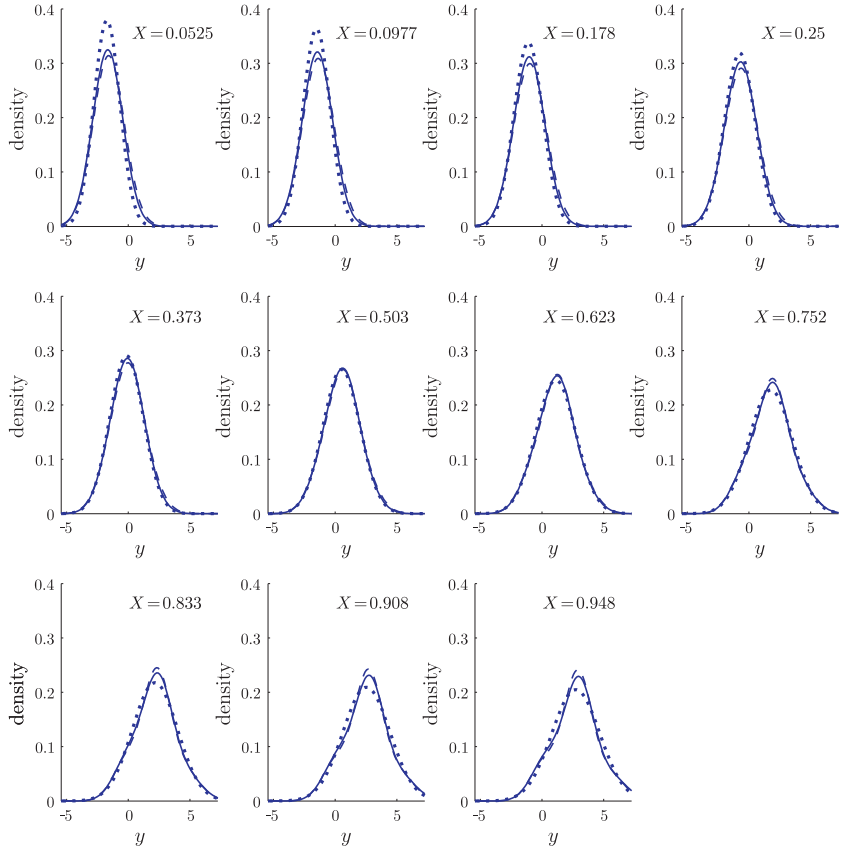

Figure 1. Results for the first simulation example. True (dotted lines), preferred model (dashed lines), and model-averaged (solid lines) density estimators for predictor values at the 5, 10, 17.5, 25, 37.5, 50, 62.5, 75, 82.5, 90, and 95th percentiles of the empirical distribution.

For Case 3, Figure 2 shows the density estimates under the fully Bayes procedure. Clearly, the estimates were close to the truth across the range of x values. The empirical Bayes procedure had similar performance, though it slightly under-estimated the secondary mode appearing at the higher values of

x. The estimated values of the hyper- and smoothing parameters were $\widehat{\boldsymbol{\beta}} = (-2.12, 4.48)'$, $\widehat{\kappa} = 0.25$, $\widehat{\mu}_\tau = 9.34$, $\widehat{\sigma}^2_\tau = 0.001$, $\widehat{\alpha} = 0.08$, $\widehat{\psi} = 0.12$, with $\mu_\tau = a_\tau/b_\tau$ and $\sigma^2_\tau = a_\tau/b^2_\tau$. In addition, the preferred number of mixture components was $\widehat{k} = 8$, with a 95% credible interval [4,12].

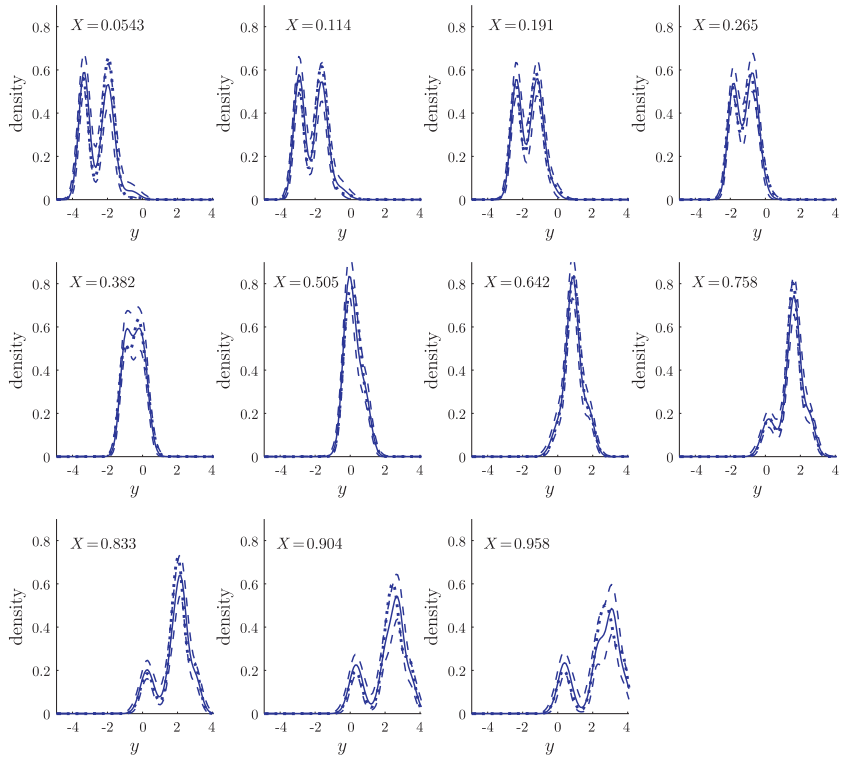

Figure 2. Results for the third simulation example ($n = 1,000$). True (dotted lines), fully Bayes estimate (solid line), and pointwise 99% credible intervals (dashed lines). Conditional density estimates are shown for predictor values at the 5, 10, 17.5, 25, 37.5, 50, 62.5, 75, 82.5, 90, and 95th percentiles of the empirical distribution.

In small samples, the expectation is that there is not enough information in the data to detect subtle local deviations from the base normal linear model. To assess this, simulation Case 3 was repeated for a sample size of $n = 100$. The empirical Bayes estimates are shown in Figure 3. For small values of x, the bi-modal shape was detected, with some underestimation of the peak height. Estimates were accurate across most of the range of x, and at higher values the small secondary mode was detected but flattened out. The credible intervals provided a good measure of uncertainty. The performance of the fully Bayes estimator was not as good: there was more underestimation of the peaks at low values of x, and the secondary mode at high values was missed.

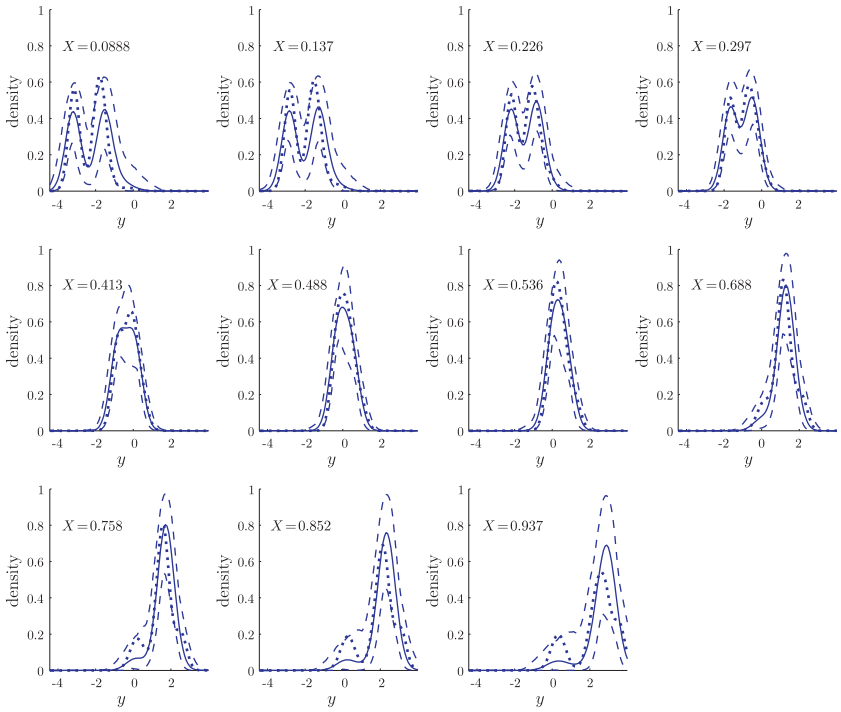

Figure 3. Results for the third simulation example ($n = 100$). True (dotted lines), model-averaged empirical Bayes estimate (solid line), and pointwise 99% credible intervals (dashed lines).

5. Abstinence and Sperm Concentration Application

To illustrate the methods, consider data from a reproductive epidemiology study. The focus is on assessing the relationship between abstinence time (x) and sperm concentration (y) using semen analysis results for $n = 220$ men. Although it is generally believed that sperm concentration increases with abstinence time, at least initially, the relationship is thought to be non-linear. In addition, heterogeneity among men in rates of sperm production may lead to a changing shape of the sperm concentration distribution as abstinence time changes. Assessing this relationship is important in deciding how to control for abstinence time in an epidemiologic study, and in making recommendations to couples attempting pregnancy.

One possibility would be to use the quantile smoothing splines of Koenker, Ng and Portnoy (1994). However, this would involve specifying particular quantiles of interest in advance, or fitting models separately to a sequence of quantiles. By applying our density regression approach, we instead allow smooth, nonlinear effects on all quantiles simultaneously.

Repeating the analyses implemented for the simulated data examples, but for 20,000 iterations, iteration plots for the smoothing parameters α and ψ are shown in Figure 4. The values converge rapidly to a stationary distribution. Figure 5 plots empirical Bayes density estimates and 99% pointwise credible intervals for a range of abstinence times chosen in advance to span the range of values in the sample. We focus on a finite set of values for ease in visualization. The fully Bayes estimates were very similar for an abstinence interval of less than five days, but as abstinence time increases and the data become sparser, the fully Bayes estimates have a lower peak and wider credible intervals.

The sperm concentration distribution is clearly non-normal, and has a shape that is not well-characterized by a log normal distribution. As abstinence times increase, there appears to be a subtle shift in the distribution, though it is difficult to judge based on the density plots. Therefore, empirical Bayes quantile regression curves were estimated for the 5, 10, 25, 50, 75, 90 and 95th percentiles of the sperm concentration distribution as a function of abstinence time. The results are plotted, along with the raw data, in Figure 6. As expected, there is a large amount of heterogeneity among men in sperm concentration, regardless of abstinence time. However, sperm concentration does appear to improve with increasing abstinence time. Interestingly, the effect of abstinence was largest for typical men at the median of the population distribution, with little effect at the high and low concentrations.

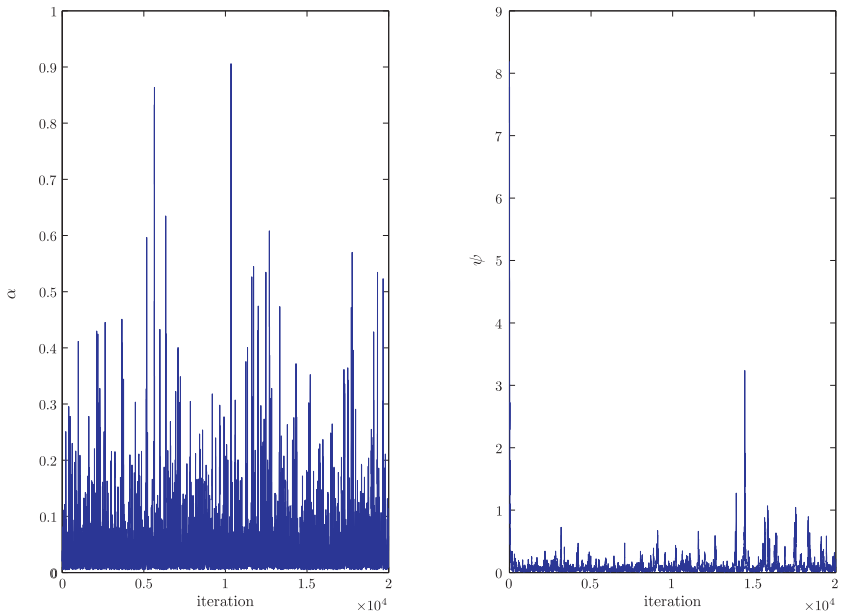

Figure 4. Trace plots of the sampled values of the smoothing parameters α and ψ for the abstinence and sperm concentration example.

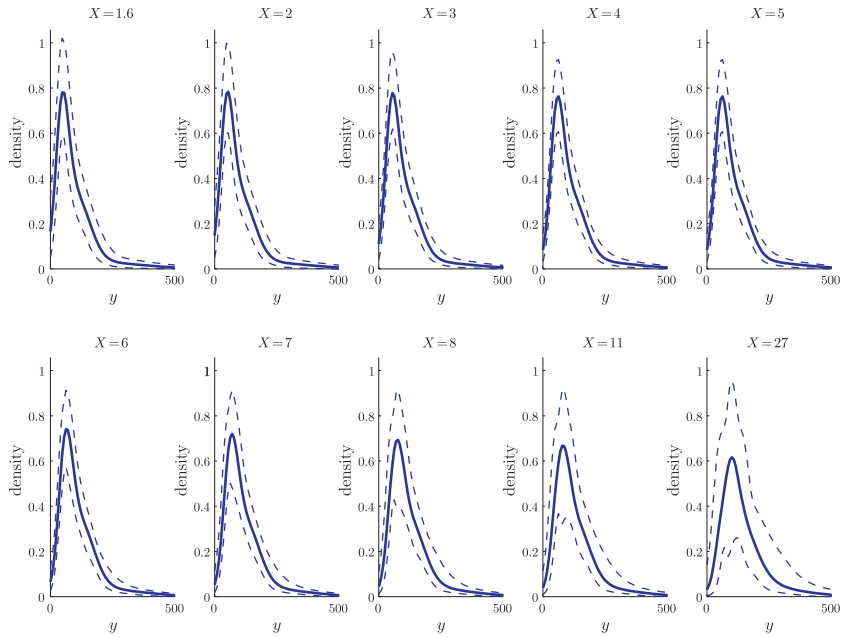

Figure 5. Estimated sperm concentration density among men having different abstinence intervals. The solid line is the model-averaged empirical Bayes estimator and the dashed lines are pointwise 99% credible intervals.

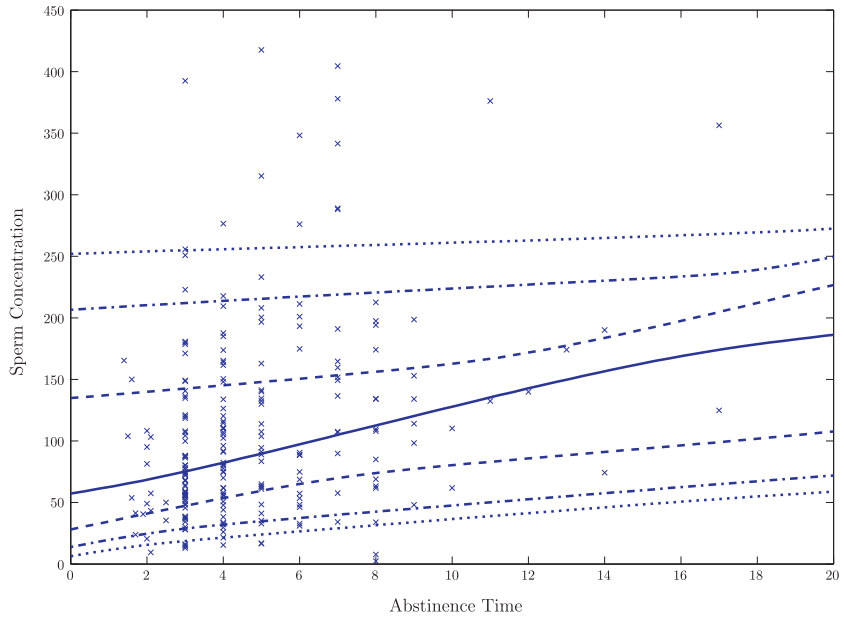

Figure 6. Data values and fitted quantile regression curves for sperm concentration versus abstinence time. Curves for the 5, 10, 25, 50, 75, 90, and 95th percentiles are shown.

It was particularly interesting that the abstinence effect on sperm concentration did not go away after the first two days, as is commonly believed. However, it is important to note that the magnitude of the abstinence effect on sperm concentration is not enough to have a large impact on the probability of pregnancy, as Slama, Kold-Jensen, Scheike, Ducot, Spira and Keiding (2004) estimate only a 15% decrease in fecundability attributable to a 47% decline in sperm concentration. Hence, there appears to be no need for couples attempting pregnancy to reduce intercourse frequency in an attempt to reduce the abstinence effect; the additional sperm introduced with each intercourse act should more than make up for the modest decline in concentration attributable to a higher frequency.

6. Discussion

This article has proposed fully Bayes and empirical Bayes approaches for estimating the density of a response variable in relation to one or more predictors. Based on simulation examples, both approaches appear to have good performance in a variety of cases, with the empirical Bayes approach doing better in small to moderate samples. This reflects the difficulty of choosing good values for the hyperparameters *a priori*. The empirical Bayes approach may also have efficiency advantages, as one can adaptively borrow information over wider regions in cases in which the data do not support a changing mixture distribution.

One advantage over frequentist methods is flexibility, in that it is straightforward to apply this same approach when Y and $\mathbf{X}$ are components of a hierarchical model. For example, Y could be a latent variable in a structural equation or factor analytic model. In addition, the formulation allows discrete predictors, which are not naturally accommodated by previous specifications that model the distribution of $\{Y, \mathbf{X}\}$ using a mixture of normals.

Although the focus here has been on density estimation, the proposed methodology can also be used for classification. For example, there is considerable interest in clustering genes based on differential expression between groups. One could potentially cluster the genes by assigning a Dirichlet process prior to the unknown distribution of the differences in gene expression levels. However this would ignore other information, such as gene function annotations, or covariates such as time or dose of a treatment. The proposed approach can be used to flexibly incorporate such predictors to inform about the classification.

Another interesting area for future research is the development of methods for comparing models with and without a predictor to assess whether that predictor is associated with the response. In the nonparametric case, this problem is one of comparing competing infinite-dimensional models, which results in unique challenges. A simple model comparison criteria would be the maximized generalized log-likelihood, obtained by fitting each of the models using the approach

proposed in this article. Such an approach incorporates an automatic penalty for the difference in model complexity between the competing models.

References

Bashtannyk, D. M. and Hyndman, R. J. (2001). Bandwidth selection for kernel conditional density estimation. *Comput. Statist. Data Anal.* **36**, 279-298.

Blackwell D. and MacQueen J. B. (1973). Ferguson distributions via Pólya urn schemes. *Ann. Statist.* **1**, 353-355.

Celeux, G. and Diebolt, J. (1985). The SEM algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Comput. Statist. Quarterly* **2**, 73-82.

Cifarelli, D. and Regazzini, E. (1978). Problemi statistici non parametrici in condizioni di scambialbilita parziale e impiego di medie associative. Technical Report. Quaderni Istituto Matematica Finanziaria, Turin.

De Iorio, M. D., Müller, P., Rosner, G. L. and MacEachern, S. N. (2004). An ANOVA model for dependent random measures. *J. Amer. Statist. Assoc.* **99**, 205-215.

Duan, J. A., Guindani, M. and Gelfand, A. E. (2005). Generalized spatial Dirichlet process models. ISDS Discussion Paper 2005-23, Duke University, Durham, NC.

Dunson, D. B., Pillai, N. and Park, J-H. (2007). Bayesian density regression. *J. Roy. Statist. Soc. Ser. B* **69**, 163-183.

Dunson, D. B. (2006) .Bayesian dynamic modeling of latent trait distributions. *Biostatistics* **7**, 551-568.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90**, 577-588.

Escobar, M. D. and West, M. (1998). Computing nonparametric hierarchical models. In *Practical Nonparametric and Semiparametric Bayesian Statistics* (Edited by D. Dey, P. Müller and D. Sinha), 1-22. Springer-Verlag, New York.

Fan, J. Q., Yao, Q. and Tong, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika* **83**, 189-206.

Fan, J. Q. and Yim, T. H. (2004). A cross validation method for estimating conditional densities. *Biometrika* **91**, 819-834.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209-230.

Gelfand, A. E., Kottas, A., and MacEachern, S. N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process Mixing. *J. Amer. Statist. Assoc.* **100**, 1021-1035.

Genovese, C. R. and Wasserman, L. (2000). Rates of convergence for the Gaussian mixture sieve. *Ann. Statist.* **28**, 1105-1127.

Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Statist.* **27**, 143-158.

Ghosal, S. and Van der Vaart, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.* **29**, 1233-1263.

Giudici, P., Mezzetti, M. and Muliere, P. (2003). Mixtures of Dirichlet process priors for variable selection in survival analysis. *J. Statist. Plann. Inference* **111**, 101-115.

Griffin, J. E. and Steel, M. F. J. (2006). Order-based dependent Dirichlet processes. *J. Amer. Statist. Assoc.* **101**, 179-194.

Hall, P., Wolff, R. C. L. and Yao, Q. W. (1999). Methods for estimating a conditional distribution function. *J. Amer. Statist. Assoc.* 94, 154-163.

Hyndman, R. J., Bashtannyk, D. M. and Grunwald, G. K. (1996). Estimating and visualizing conditional densities. *J. Comput. Graph. Statist.* **5**, 315-336.

Hyndman, R. J. and Yao, Q. (2002). Nonparametric estimation and symmetry tests for conditional density functions. *J. Nonparametr. Stat.* **14**, 259-278.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J. and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation* **3**, 79-87.

Jiang, W. and Tanner, M. A. (1999). Hierarchical mixtures-of-experts for exponential family regression models: Approximation and maximum likelihood estimation. *Ann. Statist.* **27**, 987-1011.

Jordan, M. I. and Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation* **6**, 181-214.

Koenker, R., Ng, P. and Portnoy, S. (1994). Quantile smoothing splines. *Biometrika* **81**, 673-680.

Liu, J. S. (1996). Nonparametric hierarchical Bayes via sequential imputations. *Ann. Statist.* **24**, 911-930.

MacEachern, S. N. (1999). Dependent nonparametric processes. *Proceedings of the Bayesian Section of the American Statistical Association*, 50-55.

MacEachern, S. N. and Müller, P. (1998). Estimating mixtures of Dirichlet process models. *J. Comput. Graph. Statist.* **7**, 223-238.

McAuliffe, J. D., Blei, D. M. and Jordan, M. I. (2006). Nonparametric empirical Bayes for the Dirichlet process mixture model. *Statist. Comput.* **16**, 5-14.

McLachlan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*. Wiley, New York.

Meng, X. L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80**, 267-278.

Mira A. and Petrone, S. (1996). Bayesian hierarchical nonparametric inference for change point problems. *Bayesian Statist.* **5**, 693-703.

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

Müller, P. and Quintana, F. A. (2004). Nonparametric Bayesian data analysis. *Statist. Sci.* **19**, 95-110.

Müller, P., Quintana, F. and Rosner, G. (2004). A method for combining inference across related nonparametric Bayesian models. *J. Roy. Statist. Soc. Ser. B* **66**, 735-749.

Sethuraman, J. (1994). A constructive definition of the Dirichlet process prior. *Statist. Sinica* **2**, 639-650.

Slama, R., Kold-Jensen, T., Scheike, T., Ducot, B., Spira, A. and Keiding, N. (2004). How would a decline in sperm concentration over time influence the probability of pregnancy. *Epidemiology* **15**, 458-465.

Stein, M. (1990). A comparison of generalized cross validation and modified maximum likelihood estimation for estimating the parameters of a stochastic process. *Ann. Statist.* **18**, 1139-1157.

Tomlinson, G. and Escobar, M. (1999). Analysis of densities. Technical Report. University of Toronto, Toronto.

Viele, K. and Tong, B. (2000). Modeling with mixtures of linear regressions. *Statist. Comput.* **12**, 315-330.

Wahba, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Ann. Statist.* **13**, 1378-1402.

Wecker, M. and Ansley, C. (1983). The signal extraction approach to nonlinear regression and spline smoothing. *J. Amer. Statist. Assoc.* **78**, 81-89.

Wei, G. C. G. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Amer. Statist. Assoc.* **85**, 699-704.

Yu, K. and Jones, M. C. (1998). Local linear quantile regression. *J. Amer. Statist. Assoc.* **93**, 228-237.

Biostatistics Branch, MD A3-03, National Institute of Environmental Health Sciences, P.O. Box 12233, RTP, NC 27709, U.S.A.

E-mail: dunson1@niehs.nih.gov

(Received March 2005; accepted June 2006)