Supplementary materials for this article are available online.
Please go to www.tandfonline.com/r/JCGS

Regression Density Estimation With Variational Methods and Stochastic Approximation

David J. NOTT, Siew Li TAN, Mattias VILLANI, and Robert KOHN

Regression density estimation is the problem of flexibly estimating a response distribution as a function of covariates. An important approach to regression density estimation uses finite mixture models and our article considers flexible mixtures of heteroscedastic regression (MHR) models where the response distribution is a normal mixture, with the component means, variances, and mixture weights all varying as a function of covariates. Our article develops fast variational approximation (VA) methods for inference. Our motivation is that alternative computationally intensive Markov chain Monte Carlo (MCMC) methods for fitting mixture models are difficult to apply when it is desired to fit models repeatedly in exploratory analysis and model choice. Our article makes three contributions. First, a VA for MHR models is described where the variational lower bound is in closed form. Second, the basic approximation can be improved by using stochastic approximation (SA) methods to perturb the initial solution to attain higher accuracy. Third, the advantages of our approach for model choice and evaluation compared with MCMC-based approaches are illustrated. These advantages are particularly compelling for time series data where repeated refitting for one-step-ahead prediction in model choice and diagnostics and in rolling-window computations is very common. Supplementary materials for the article are available online.

Key Words: Bayesian model selection; Heteroscedasticity; Mixtures of experts; Stochastic approximation; Variational Bayes.

1. INTRODUCTION

Regression density estimation is the problem of flexibly estimating a response distribution assuming that it varies smoothly as a function of covariates. An important approach to regression density estimation uses finite mixture models and our article considers flexible regression models where the response distribution is a normal mixture, with the component

David J. Nott is Associate Professor, Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546 (E-mail: *standj@nus.edu.sg*). Siew Li Tan is Ph.D. student, Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546 (E-mail: *g0900760@nus.edu.sg*). Mattias Villani is Professor, Division of Statistics, Department of Computer and Information Science, Linköping University, SE-58183 Linköping, Sweden (E-mail: *mattias.villani@liu.se*). Robert Kohn is Professor, Australian School of Business, University of New South Wales, Sydney 2052, Australia (E-mail: *r.kohn@unsw.edu.au*).

Journal of Computational and Graphical Statistics, Volume 21, Number 3, Pages 797–820
DOI: 10.1080/10618600.2012.679897

means, variances, and mixing weights all varying with covariates. The component means and variances are described by a heteroscedastic linear model, and the component weights by a multinomial logit model. These mixtures of heteroscedastic regression (MHR) models extend conventional mixtures of regression models (Jacobs et al. 1991; Jordan and Jacobs 1994) by allowing the component models to be heteroscedastic. Mixtures of regression models are usually referred to as mixtures of experts models in machine learning, and this terminology is sometimes used in statistics as well. The term "expert" refers to the individual mixture components, and the term "gating function" is used for the model for the mixing weights. Very often, the experts are generalized linear models, and such models (with or without covariate-dependent mixing weights) are sometimes referred to as mixtures of generalized linear models. In marketing, the term "concomitant variable mixture regression model" is sometimes used (e.g., see Wedel 2002).

The heteroscedastic extension of mixture of regression models is important in modern applications since simulations by Villani, Kohn, and Giordani (2009) showed that, when used to model heteroscedastic data, the performance of models with homoscedastic components deteriorates as the number of covariates increases, and there comes a point when their performance cannot be improved simply by increasing the number of mixture components. They also observed that with MHR models, fewer mixture components are required, which makes the estimation and interpretation of mixture models easier. In an analysis of the benchmark LIDAR dataset, Li, Villani, and Kohn (2010a) showed that a model with homoscedastic thin-plate components requires three components to achieve approximately the same performance as an MHR model with a single thin-plate component, providing further evidence that MHR models help in reducing the number of mixture components required. Moreover, Villani, Kohn, and Giordani (2009) demonstrated that the fit of an MHR model to homoscedastic data is comparable with that of a model with homoscedastic components.

Our article makes three contributions. First, fast variational methods are developed for fitting MHR models and a variational lower bound is obtained in closed form. Second, stochastic approximation (SA) optimization techniques (e.g., see Spall 2003) are used to improve the variational approximation (VA) to attain higher accuracy. The computational methods developed here can also be applied beyond the context of MHR models. Third, in situations where it is necessary to fit a number of complex models to the data or if cross-validation is used to select the best models, we demonstrate that VA methods provide an attractive alternative to Markov chain Monte Carlo (MCMC) methods, which may not be feasible due to the computational complexity involved. For example, in model selection for complex time series models, repeated one-step-ahead prediction is often carried out (Geweke and Amisano 2010), with the parameters of each model reestimated at each step. Another popular time series approach to studying model performance is the use of rolling windows (Pesaran and Timmermann 2002). Suppose the dataset consists of T observations, with each window of fixed size M. In the first instance, each model is fitted to the first M observations. Next, each model is fitted to observations 2 to $M + 1$, etc., with the final window consisting of observations $T - M + 1$ to T. Variational methods are ideally suited to this kind of repeated refitting and can benefit from a "warm start" obtained from the previous fit. In Section 6.2, we quantify in an example the computational speedup that comes from initializing the variational optimization based on the fit to the previous window rather than treating each fit as an independent computation.

The use of VAs may be particularly compelling where the main focus is predictive inference. Bayesian predictive inference is based on the predictive distribution

$$p(y^*|y) = \int p(y^*|\theta, y)p(\theta|y)d\theta,$$

where y^* denotes a future response, y is the observed data, and θ are parameters. There are two components to our uncertainty in $p(y^*|y)$. The first component is the inherent randomness in y^* that would occur even if θ were known: this is captured by the term $p(y^*|\theta, y)$ in the integrand. The second component is related to parameter uncertainty and is captured by the term $p(\theta|y)$. In general, with large datasets, the parameter uncertainty is small, and an approximate treatment of $p(\theta|y)$ for predictive purposes may be attractive, provided that the approximate posterior provides good point estimation. In general, plugging in an approximate variational posterior instead of $p(\theta|y)$ in the expression for the predictive density can result in excellent predictive inference. Furthermore, this still accounts, to some extent, for parameter uncertainty, hence improving on simple plug-in predictive density estimates.

Bayesian approaches to inference in mixtures of regression models were first considered in Peng et al. (1996). Wood, Jiang, and Tanner (2002) and Wood et al. (2008) considered models with flexible terms for the covariates for continuous and binary responses, respectively. Geweke and Keane (2007) also took a Bayesian approach to inference in mixture models with homoscedastic components using a multinomial probit model for the mixing weights, which allows a convenient Gibbs sampling MCMC scheme. MHR models have been considered previously by Villani, Kohn, and Giordani (2009). They used a Bayesian approach to inference using MCMC methods for computation and considered general smooth terms for the covariates and variable selection in the mean and variance models and the gating function. Norets (2010) considered approximation results for approximating quite general conditional densities in the Kullback–Leibler sense using various kinds of normal mixtures. Approximation error bounds are derived and some interesting insights are obtained about when additional flexibility might be most usefully employed in the mean, variance, and gating functions. The approximation results of Jiang and Tanner (1999) are concerned, on the other hand, with conditional densities in a one-parameter exponential family in which the components come from the same exponential family. Their results are also useful for conditional densities that are discrete.

Methods related to finite mixture models are currently under active development in the area of Bayesian nonparametric approaches to regression density estimation. Rather than considering a finite mixture of regressions, it is possible to put a flexible prior on a mixing distribution that varies over the space. For common priors, the resulting models might be considered to be mixtures with an infinite number of components. There are both advantages and disadvantages to this kind of approach. On the positive side, the difficult question of model choice for the number of mixture components is avoided. However, a finite mixture may be more interpretable and the nature of the model may be easier to communicate to scientific practitioners. In addition, in the finite mixture framework, it is easier to incorporate some very natural extensions, such as the components being of qualitatively different types. We do not discuss these methods any further, but refer readers to MacEachern (1999), De Iorio et al. (2004), Griffin and Steel (2006), and Dunson, Pillai, and Park (2007) and the references therein for a summary of relevant methodology and recent developments.

In terms of computational methodology, early approaches to fitting mixtures of regression models (Jordan and Jacobs 1994, for example) used maximum likelihood and the expectation–maximization (EM) algorithm. Modern Bayesian strategies for inference use MCMC computational methods, although there are a number of authors who consider variational approaches similar to those described in this article (Waterhouse, MacKay, and Robinson 1996; Ueda and Ghahramani 2002; Bishop and Svensén 2003). However, these authors did not consider heteroscedastic components. Moreover, our article also explores the advantages of variational methods in repeated estimations of the model, as in model comparison by cross-validation or rolling-window estimates to check for model stability. Outside the regression context, there are a number of innovative approaches to model selection in fitting Gaussian mixture models that follow a variational approach. Corduneanu and Bishop (2001) considered a variational lower bound on the log marginal likelihood with all parameters, except the mixing coefficients, integrated out, and estimated the mixing coefficients by maximizing the lower bound. This leads to some of the coefficients being set to zero and an automated model selection approach. McGrory and Titterington (2007) considered a variational extension of the deviance information criterion (DIC) (Spiegelhalter et al. 2002) and a variational optimization technique where the algorithm is initialized with a large number of components and mixture components, whose weightings become sufficiently small as the optimization proceeds and thus are dropped out. Blei and Jordan (2006) considered VA for Dirichlet process mixture models. Recently, Wu, McGrory, and Pettitt (2012) considered variational methods for fitting mixtures for data that require models with narrow, widely separated mixture components. They discussed sophisticated split-and-merge algorithms, building on earlier related methods, such as those in Ghahramani and Beal (2000), Ueda and Ghahramani (2002), and Constantinopoulos and Likas (2007), for simultaneous model selection and parameter estimation as well as on novel criteria for model evaluation.

One contribution of our article is the development of broadly applicable SA correction methods for VA. From our simulation studies, the SA correction is very helpful in getting an improved approximation and requires less computation time than required by MCMC methods. Ji, Shen, and West (2010) proposed similar SA methods for learning VAs, but we offer a number of improvements on their implementation. In particular, we are able to suggest an improved gradient estimate compared with their approach and consider a strategy of perturbing only the mean and scale in an initial VA obtained using closed-form updates. Perturbing an existing solution allows us to keep the dimension of the optimization low, which is important for a fast and stable implementation. Ji, Shen, and West (2010) suggested a number of other innovative ideas, including the use of MCMC methods, to obtain computationally attractive upper and lower bounds on the marginal likelihood.

The article is organized as follows. Section 2 introduces MHR models. Section 3 describes fast VA methods for MHR models, and Section 4 describes uses of variational methods in model choice. Section 5 discusses improvements on the basic approximation using an SA correction, which also integrates out the mixture component indicators from the posterior. Section 6 considers examples involving both real and simulated data, and Section 7 concludes.

2. HETEROSCEDASTIC MIXTURES OF REGRESSION MODELS

Suppose that responses $y_1, \dots, y_n$ are observed. They are modeled by an MHR model (Jacobs et al. 1991; Jordan and Jacobs 1994) of the form:

$$y_i | \delta_i, \beta, \alpha \sim N\left(x_i^T \beta_{\delta_i}, \exp\left(\alpha_{\delta_i}^T z_i\right)\right),$$

where δ_i is a categorical latent variable with k categories, $\delta_i \in \{1, \dots, k\}$, $x_i = (x_{i1}, \dots, x_{ip})^T$ and $z_i = (z_{i1}, \dots, z_{im})^T$ are vectors of covariates for observation i, and $\beta_j = (\beta_{j1}, \dots, \beta_{jp})^T$ and $\alpha_j = (\alpha_{j1}, \dots, \alpha_{jm})^T$, $j = 1, \dots, k$ are vectors of unknown parameters. The above model says that conditional on $\delta_i = j$, the response follows a heteroscedastic linear model, where the mean is $x_i^T \beta_j$ and the log variance is $z_i^T \alpha_j$. The prior for the latent variables δ_i is

$$P(\delta_i = j | \gamma) = p_{ij} = \frac{\exp\left(\gamma_j^T v_i\right)}{\sum_{l=1}^k \exp\left(\gamma_l^T v_i\right)}, \quad j = 1, \dots, k,$$

where $v_i = (v_{i1}, \dots, v_{ir})^T$ is a vector of covariates and $\gamma_j = (\gamma_{j1}, \dots, \gamma_{jr})^T$ are vectors of unknown parameters, $j = 2, \dots, k$, with γ_1 set to be identically zero for identifiability. With this prior, the responses are modeled as a mixture of heteroscedastic linear regressions, where the mixture weights vary with the covariates. For Bayesian inference, we require prior distributions on the remaining unknown parameters in the model. Independent priors are assumed for the β_j, $\beta_j \sim N(\mu_{\beta j}^0, \Sigma_{\beta j}^0)$, $j = 1, \dots, k$, for α_j, $\alpha_j \sim N(\mu_{\alpha j}^0, \Sigma_{\alpha j}^0)$, $j = 1, \dots, k$, and for $\gamma = (\gamma_2^T, \dots, \gamma_k^T)^T$, $\gamma \sim N(\mu_\gamma^0, \Sigma_\gamma^0)$. We will describe fast methods for variational inference in the above model. Variational inference for mixtures of regression models has been considered before (Waterhouse, MacKay, and Robinson 1996; Ueda and Ghahramani 2002; Bishop and Svensén 2003) but not for the case of heteroscedastic mixture components. In the case of heteroscedastic mixture components, a variational lower bound can still be computed in closed form, thus allowing fast computation.

3. VARIATIONAL APPROXIMATION

We consider a VA to the joint posterior distribution of all the parameters θ of the form $q(\theta|\lambda)$, where λ is a set of variational parameters to be chosen. VA methods originated in statistical physics, and they have been widely used in the machine learning community for some time. Jordan et al. (1999) is an early reference and Bishop (2006, chap. 10) gives a recent summary. Ormerod and Wand (2010) gives an introduction to VA methods that is particularly accessible to statisticians. Variational approximation is a very active area of research in both statistics and machine learning. For additional discussion of the current state of the art in relation to mixture models, see the references in Section 1. Here, a parametric form is chosen for $q(\theta|\lambda)$ (described below), and we attempt to make $q(\theta|\lambda)$ a good approximation to $p(\theta|y)$ by minimizing the Kullback–Leibler divergence between $q(\theta|\lambda)$ and $p(\theta|y)$, that is,

$$\int \log \frac{q(\theta|\lambda)}{p(\theta|y)} q(\theta|\lambda) d\theta = \int \log \frac{q(\theta|\lambda)}{p(\theta)p(y|\theta)} q(\theta|\lambda) d\theta + \log p(y), \qquad (1)$$

where $p(y)$ is the marginal likelihood $p(y) = \int p(y|\theta)p(\theta)d\theta$. Note that since the Kullback–Leibler divergence is positive, we have

$$\log p(y) \geq \int \log \frac{p(\theta)p(y|\theta)}{q(\theta|\lambda)} q(\theta|\lambda)d\theta, \tag{2}$$

which gives a lower bound on the log marginal likelihood, and maximizing the lower bound is equivalent to minimizing the Kullback–Leibler divergence between the posterior distribution and the VA. From (1), the difference between the lower bound and the log marginal likelihood is the Kullback–Leibler divergence between the posterior and the VA, and the lower bound is sometimes used as an approximation to the log marginal likelihood for Bayesian model selection purposes. We discuss the role of the marginal likelihood in Bayesian model comparison later.

The difference between our development of a VA for the MHR model and previous developments of variational methods for mixture models with homoscedastic components lies both in the fact that a closed-form derivation of the variational lower bound is not obvious in the heteroscedastic case and in the need to deal with the variance parameters in the component models during optimization. Optimization of variational parameters for the variational posterior factors cannot be done in closed form, and since numerical optimization over high-dimensional covariance matrices can be time-consuming, we have developed an approximate method for dealing with these parameters that is computationally efficient and effective in practice. For the model in Section 2, we write the parameters as $\delta = (\delta_1, \ldots, \delta_n)^T$, $\beta = (\beta_1^T, \ldots \beta_k^T)^T$, $\alpha = (\alpha_1^T, \ldots \alpha_k^T)^T$, and $\gamma = (\gamma_2^T, \ldots, \gamma_k^T)^T$ so that $\theta = (\delta^T, \beta^T, \alpha^T, \gamma^T)^T$. For convenience, we write our VA $q(\theta|\lambda)$ as $q(\theta)$, suppressing dependence on λ in the notation. We consider a VA to the posterior of the form $q(\theta) = q(\delta)q(\beta)q(\alpha)q(\gamma)$, where

$$q(\delta) = \prod_{i=1}^{n} q(\delta_i), \quad q(\beta) = \prod_{i=1}^{k} q(\beta_i), \quad q(\alpha) = \prod_{i=1}^{k} q(\alpha_i),$$

and $q(\beta_i)$ is normal, $N(\mu_{\beta i}^q, \Sigma_{\beta i}^q)$ say for $i = 1, \ldots, k$, $q(\alpha_i)$ is normal, $N(\mu_{\alpha i}^q, \Sigma_{\alpha i}^q)$ say for $i = 1, \ldots, k$, $q(\gamma)$ is a delta function placing point mass of 1 on μ_γ^q, $q(\gamma) = \delta(\gamma - \mu_\gamma^q)$, and $q(\delta_i = j) = q_{ij}$, where $\sum_{j=1}^k q_{ij} = 1, i = 1, \ldots, n, j = 1, \ldots, k$. We are assuming in the variational posterior that parameters for different mixture components are independent of each other and of all other parameters, that mean and variance parameters are independent, and that the latent variables δ are independent of each other and of all other parameters. We assumed a degenerate point mass variational posterior for γ to make computation of the lower bound tractable. However, following our description below of the variational algorithm that uses the point mass form for $q(\gamma)$, we also suggest a method for relaxing the form of $q(\gamma)$ to be a normal distribution. The assumption of independence between mean and variance parameters can also be relaxed (John Ormerod, personal communication), but this also makes the variational optimization slightly more complex. Although the independence and distributional assumptions made in VAs are typically unrealistic, it is often found that variational approaches give good point estimates, reasonable estimates of marginal posterior distributions, and excellent predictive inferences compared with other approximations, particularly in high dimensions. For example, Blei and Jordan (2006, sec. 5) demonstrated in the context of Dirichlet process mixture models that predictive inference based on a VA is similar to a fully Bayes predictive inference implemented via MCMC. Braun and McAuliffe (2010, sec. 4) reported similar findings in large-scale models of discrete choice.

Here, we are considering γ as a fixed-point estimate, and if we write $\theta_{-\gamma}$ for the rest of the unknown parameters (i.e., not including γ), then a lower bound on $\log p(y|\gamma)$, where $p(y|\gamma) = \int p(\theta_{-\gamma}|\gamma)p(y|\theta)d\theta_{-\gamma}$, is

$$\int \log \frac{p(\theta_{-\gamma}|\gamma)p(y|\theta)}{q(\theta_{-\gamma})} q(\theta_{-\gamma})d\theta_{-\gamma}.$$

The lower bound can be computed in closed form, and this gives a lower bound on $\sup_\gamma \log p(\gamma)p(y|\gamma)$ of (see the online supplementary materials)

$$\begin{aligned} L = &-\frac{n}{2}\log 2\pi + \frac{(p+m)k}{2} + \log p(\mu_\gamma^q) - \frac{1}{2}\sum_{j=1}^k \log|\Sigma_{\beta j}^0| - \frac{1}{2}\sum_{j=1}^k \operatorname{tr}({\Sigma_{\beta j}^0}^{-1}\Sigma_{\beta j}^q) \\ &-\frac{1}{2}\sum_{j=1}^k (\mu_{\beta j}^q - \mu_{\beta j}^0)^T {\Sigma_{\beta j}^0}^{-1}(\mu_{\beta j}^q - \mu_{\beta j}^0) - \frac{1}{2}\sum_{j=1}^k \log|\Sigma_{\alpha j}^0| - \frac{1}{2}\sum_{j=1}^k \operatorname{tr}({\Sigma_{\alpha j}^0}^{-1}\Sigma_{\alpha j}^q) \\ &-\frac{1}{2}\sum_{j=1}^k (\mu_{\alpha j}^q - \mu_{\alpha j}^0)^T {\Sigma_{\alpha j}^0}^{-1}(\mu_{\alpha j}^q - \mu_{\alpha j}^0) + \sum_{i=1}^n\sum_{j=1}^k q_{ij}\log\frac{p_{ij}}{q_{ij}} + \frac{1}{2}\sum_{j=1}^k \log|\Sigma_{\beta j}^q| \\ &+\frac{1}{2}\sum_{j=1}^k \log|\Sigma_{\alpha j}^q| - \sum_{i=1}^n\sum_{j=1}^k q_{ij}\left\{\frac{1}{2}z_i^T\mu_{\alpha j}^q + \frac{1}{2}\frac{(y_i - x_i^T\mu_{\beta j}^q)^2 + x_i^T\Sigma_{\beta j}^q x_i}{\exp\left(z_i^T\mu_{\alpha j}^q - \frac{1}{2}z_i^T\Sigma_{\alpha j}^q z_i\right)}\right\}. \end{aligned} \quad (3)$$

Here, $p(\mu_\gamma^q)$ is the prior distribution for γ evaluated at μ_γ^q, and p_{ij} is evaluated by setting $\gamma = \mu_\gamma^q$. The variational parameters to be optimized consist of $\mu_{\beta j}^q$, $\Sigma_{\beta j}^q$, $\mu_{\alpha j}^q$, $\Sigma_{\alpha j}^q$, $j = 1, \ldots, k$, μ_γ^q, and q_{ij} for $i = 1, \ldots, n$, $j = 1, \ldots, k$. We optimize the lower bound with respect to each of these sets of parameters, with the others held fixed, in a gradient ascent algorithm. The updates that are available in closed form are easy to derive using vector differential calculus (e.g., see Wand 2002). To initialize the algorithm, we first generate an initial clustering of the data. Then, for this initial clustering, we perform the following:

Algorithm 1

Initialize: $\mu_{\alpha j}^q = \Sigma_{\alpha j}^q = 0$ for $j = 1, \ldots, k$ and q_{ij} as 1 if the ith observation lies in cluster j, and 0 otherwise.

Do until the change in the lower bound between iterations is less than a tolerance:

- $\Sigma_{\beta j}^q \leftarrow (X^T D_j X + {\Sigma_{\beta j}^0}^{-1})^{-1}$, where X is the design matrix with ith row x_i^T, y is the vector of responses, and D_j is the diagonal matrix with ith diagonal entry $q_{ij}/\exp(z_i^T\mu_{\alpha j}^q - 1/2 z_i^T \Sigma_{\alpha j}^q z_i)$.
- $\mu_{\beta j}^q \leftarrow \Sigma_{\beta j}^q({\Sigma_{\beta j}^0}^{-1}\mu_{\beta j}^0 + X^T D_j y)$.

- Set $\mu_{\alpha j}^q$ to be the conditional mode of the lower bound with other variational parameters fixed at current values. As a function of $\mu_{\alpha j}^q$, the lower bound is the log posterior for a generalized linear model with normal prior $N(\mu_{\alpha j}^0, \Sigma_{\alpha j}^0)$, gamma responses $w_{ij} = (y_i - x_i^T \mu_{\beta j}^q)^2 + x_i^T \Sigma_{\beta j}^q x_i$, coefficients of variation $\sqrt{2/q_{ij}}$, and where the log of the mean is $z_i^T \mu_{\alpha j}^q - 1/2 z_i^T \Sigma_{\alpha j}^q z_i$, where the terms $-1/2 z_i^T \Sigma_{\alpha j}^q z_i$ define an offset. Although the mode has no closed-form expression, it is easily found by an iteratively weighted least-squares approach (West 1985; McCullagh and Nelder 1989) or some other numerical optimization technique.
- $\Sigma_{\alpha j}^q \leftarrow (Z^T W_j Z + \Sigma_{\alpha j}^{0\ -1})^{-1}$, where Z is the design matrix with ith row z_i^T, $i = 1, \dots, n$, and W_j is diagonal with ith diagonal element $q_{ij} w_{ij} \exp(-z_i^T \mu_{\alpha j}^q)/2$. The update is done provided that the replacement leads to an improvement in the lower bound.
- For $i = 1, \dots, n$,

$$q_{ij} \leftarrow \frac{p_{ij} \exp\left(-\frac{1}{2} z_i^T \mu_{\alpha j}^q - \frac{1}{2} \frac{(y_i - x_i^T \mu_{\beta j}^q)^2 + x_i^T \Sigma_{\beta j}^q x_i}{\exp\left(z_i^T \mu_{\alpha j}^q - \frac{1}{2} z_i^T \Sigma_{\alpha j}^q z_i\right)}\right)}{\sum_{l=1}^k p_{il} \exp\left(-\frac{1}{2} z_i^T \mu_{\alpha l}^q - \frac{1}{2} \frac{(y_i - x_i^T \mu_{\beta l}^q)^2 + x_i^T \Sigma_{\beta l}^q x_i}{\exp(z_i^T \mu_{\alpha l}^q - \frac{1}{2} z_i^T \Sigma_{\alpha l}^q z_i)}\right)}.$$

- Set μ_γ^q to be the conditional mode of the lower bound, fixing other variational parameters at their current values. As a function of μ_γ^q, the lower bound is (ignoring irrelevant additive constants) $\log p(\mu_\gamma^q) + \sum_{i=1}^n \sum_{j=1}^k q_{ij} \log p_{ij}$, where p_{ij} is computed here plugging in μ_γ^q for γ. This is the log posterior for a Bayesian multinomial regression with normal prior on μ_γ^q and where the ith response is $(q_{i1}, \dots, q_{ik})^T$. In a typical multinomial regression, only one component of this pseudo-response vector would be 1, with the other terms 0, and although this is not the case here, the usual iteratively weighted least-squares algorithm (or some other numerical optimization algorithm) can be used for finding the mode.

Algorithm 1 requires an initial clustering to initialize the parameters. We consider multiple clusterings to deal with the problem of multiple modes in the optimization. We consider 20 random clusterings where the mixture component for each observation is chosen uniformly at random. For these 20 clusterings, we perform short runs of Algorithm 1 with a very loose stopping criterion (we stop when the increase in the lower bound is less than 1) and only follow the most promising solution with the highest attained value of the lower bound to convergence. This strategy of "short runs" to identify a promising solution to follow to full convergence is similar to the one recommended for initialization of the EM algorithm for maximum likelihood estimation of mixture models by Biernacki, Celeux, and Govaert (2003). Variational approaches to fitting mixture models, such as those of McGrory and Titterington (2007), have made use of the fact that mixture components tend to drop out as fitting proceeds in order to select the number of mixture components. This component elimination feature is something that can happen with our algorithm also, although this is dependent on the initial clustering used.

Unlike the model with homoscedastic components (Bishop and Svensén 2003), not all the parameters have closed-form updates. In Steps 1, 2, and 5, we are able to optimize the lower bound with respect to the parameter in closed form. However, in Steps 3 and 6, we need to use an iterative method, and in Step 4, we have used an approximation and this update step is skipped if it does not improve the lower bound (3). Motivation for the approximation at Step 4 comes from the following: suppose the term $q(\alpha_j)$ in our variational posterior is not a normal distribution, but instead is not subject to any restriction. The optimal choice for this term for maximizing the lower bound, with other terms held fixed, is (e.g., see Ormerod and Wand 2010)

$$q(\alpha_j) \propto \exp\{E(\log p(\theta)p(y|\theta))\}, \tag{4}$$

where the expectation in the exponent is with respect to the variational posterior for all parameters except α_j. The expectation in (4) takes the form, apart from additive constants not depending on α_j, of the log posterior for a gamma generalized linear model (the same model as considered in step 3, except that the offset term in the mean model is omitted). If $\mu_{\alpha j}^q$ is close to the mode, we can get a normal approximation to (4) by taking the mean as $\mu_{\alpha j}^q$ and the covariance matrix as the negative inverse Hessian of the log of (4) at $\mu_{\alpha j}^q$. The inverse of the negative Hessian evaluated at $\mu_{\alpha j}^q$ is of the form $(Z^T W_j Z + \Sigma_{\alpha j}^{0\ -1})^{-1}$, with the ith diagonal element of W_j defined as in step 4. Similar reasoning was used by Waterhouse, Mackay, and Robinson (1996) in approximating the posterior distribution for the parameters in the model for the mixing weights for a homoscedastic mixture model. A referee has pointed out that convergence of the variational Bayes algorithm can be very slow when parameters are highly correlated between the blocks used in the variational factorization. This might occur, for example, when there are two very similar mixture components. We do not see any easy solutions to this problem. One possible remedy is integrating out the mixture indicators and using bigger blocks for the remaining parameters in the blockwise gradient ascent, but this would involve a much greater computational burden and the introduction of new approximations to the variational lower bound.

At convergence, we also replace the delta function variational posterior for γ with a normal approximation, in which the mean is μ_γ^q and the covariance matrix is the inverse of the negative Hessian of the Bayesian multinomial log posterior considered in step 6 at convergence. The justification for this is similar to our justification above for the update of $\Sigma_{\alpha j}^q$ in step 4 of Algorithm 1. Waterhouse, Mackay, and Robinson (1996) outlined a similar idea, but they used such an approximation at every step of their iterative algorithm, whereas we use only a one-step approximation after first using a delta function approximation to the posterior distribution for γ. Write Σ_γ^q for the covariance matrix of the variational posterior for γ. With this variational posterior, the variational lower bound on $\log p(y)$ is the same as (3), except we need to replace $\sum_{i=1}^n \sum_{j=1}^k q_{ij} \log p_{ij} + \log p(\mu_\gamma^q)$ with

$$\sum_{i=1}^n \sum_{j=1}^k q_{ij} E\left(\log\left\{\frac{\exp(v_i^T\gamma_j)}{\sum_{l=1}^k \exp(v_i^T\gamma_l)}\right\}\right) - \frac{1}{2}\log|\Sigma_\gamma^0| - \frac{1}{2}(\mu_\gamma^q - \mu_\gamma^0)^T \Sigma_\gamma^{0^{-1}}(\mu_\gamma^q - \mu_\gamma^0)$$
$$-\frac{1}{2}\mathrm{tr}\left(\Sigma_\gamma^{0^{-1}}\Sigma_\gamma^q\right) + \frac{1}{2}\log|\Sigma_\gamma^q| + \frac{r(k-1)}{2}.$$

The only terms here not in closed form are the expectations in the first term. For the purposes of defining a quantity that might be used as an approximation for $\log p(y)$, we replace

$$E\left(\log\left\{\frac{\exp(v_i^T\gamma_j)}{\sum_{l=1}^k \exp(v_i^T\gamma_l)}\right\}\right) \quad \text{with} \quad \log\left\{\frac{\exp(v_i^T\mu_{\gamma j}^q)}{\sum_{l=1}^k \exp(v_i^T\mu_{\gamma l}^q)}\right\},$$

where $\mu_{\gamma j}^q$ is the subvector of μ_γ^q, corresponding to γ_j, $j = 2, \ldots, r$. As a final note, we observe that variational posterior approximations tend to "lock on" to a single mode of the posterior distribution in mixture models where there are many equivalent modes. Since the gap between the variational lower bound and the log marginal likelihood is the Kullback–Leibler divergence between the variational posterior and the true posterior, the failure to approximate all modes of the true posterior leads to underestimation of the log marginal likelihood by the lower bound; see Bishop (2006) for further discussion. There are related concerns with some MCMC methods when estimating the marginal likelihood (e.g., see Frühwirth-Schnatter 2004 and the references therein). An adjustment to the optimized lower bound to allow for the local nature of the posterior approximation when estimating the marginal likelihood might be considered. If the $k!$ different modes from relabeling are well separated, then adding $\log k!$ would be a reasonable adjustment. However, our experience with this approach when k is large is fairly negative, in the sense that the resulting adjusted lower bound does not provide a good approximation to the true log marginal likelihood useful for model comparison: the $\log k!$ correction tends to be too large when modes overlap, and we usually do not attempt any adjustment. In the examples in Section 6, we do not emphasize the use of the marginal likelihood for model choice or attempt to support the claim that the lower bound estimates the log marginal likelihood accurately.

We also note that for very large datasets, since the posterior is of the same form as the prior for (β, α, γ), it is easy to implement our algorithm sequentially after splitting up the dataset into smaller chunks. One can learn a variational posterior approximation using the data for the first chunk, and then this can be used as the prior for processing the next chunk and so on. There may be difficulties in the naive implementation of this idea, however, as the learning may get stuck in a local mode corresponding to the first reasonable solution found. Honkela and Valpola (2003) presented an online version of variational Bayesian learning based on maintaining a decaying history of previous samples processed by the model, which ensures that the system is able to forget old solutions in favor of new better ones.

4. VARIATIONAL APPROXIMATION AND MODEL CHOICE

4.1 CROSS-VALIDATION

Marginal likelihood is a popular approach to model selection in a Bayesian context. However, we do not use this approach in our article because in our computations, we only have upper and lower approximations to the marginal likelihood, and the accuracy of such approximations may be very problem-dependent. In this section, we briefly outline how we carry out model selection using likelihood cross-validation. In B-fold cross-validation, we split the data randomly into B roughly equal parts and then B different training sets are constructed, $T_1, \ldots, T_B$, by successively leaving out one of the B parts from the complete dataset. The corresponding test sets (the parts of the data that are left out in each case to construct the training set) are denoted $F_1, \ldots, F_B$. Then, one useful measure of predictive performance that can be used for model choice, the log predictive density score (LPDS), is

$$\text{LPDS} = \frac{1}{B}\sum_{i=1}^{B} \log p(y_{F_i}|X_{F_i}, y_{T_i}).$$

Note that

$$\log p(y_F|X_F, y_T) = \log \int p(y_F|X_F, \theta)p(\theta|y_T)d\theta,$$

where θ denotes all the unknown parameters. We have assumed here that y_F and y_T are conditionally independent, given θ. This usually does not hold for time series data and thus modified approaches are appropriate in such a case, as we discuss below. In the context of mixture of regression models, $p(y_F|X_F, \theta)$ is easy to write down as a normal mixture, given the parameters, and we replace $p(\theta|y)$ with our VA $q(\theta)$. To approximate the integral, we generate Monte Carlo samples $\theta_i, i = 1, \ldots, S$, from $q(\theta)$ and then take the average of the values $p(y_F|X_F, \theta_i)$. In later examples, we use $S = 1000$.

4.2 Model Choice in Time Series

Later in the article, we consider autoregressive time series models in the form of MHR models, and in the time series context, the cross-validation approach described above is not very natural. Both Geweke and Keane (2007) and Li, Villani, and Kohn (2010b) considered a training set $y_{\leq T} = (y_1, \ldots, y_T)$ of T initial observations and then measured predictive performance by the logarithmic score for the subsequent T^* observations $y_{>T} = (y_{T+1}, \ldots, y_{T+T^*})$. That is, predictive performance for the purpose of model comparison is measured by

$$\log p(y_{>T}|y_{\leq T}) = \sum_{i=1}^{T^*} \log p(y_{T+i}|y_{\leq T+i-1}), \tag{5}$$

where

$$p(y_{T+i}|y_{\leq T+i-1}) = \int p(y_{T+i}|\theta, y_{\leq T+i-1})p(\theta|y_{\leq T+i-1})d\theta, \tag{6}$$

where $p(\theta|y_{\leq T+i-1})$ denotes the posterior distribution for all unknowns θ based on data at time $T + i - 1$. Note that (5) contains T^* terms, and that from (6), each of these terms involves consideration of a different posterior distribution as successive points from the validation set are added to the observed data. Geweke and Keane (2007) noted that the most reliable and efficient way to compute these T^* terms is to run an MCMC sampler separately for each of the T^* terms to estimate the required posterior distribution. This is extremely computationally demanding, and if T^* is large and if convergence of the MCMC scheme is slow, this may be completely infeasible. While one might consider importance sampling ideas to reuse the MCMC samples for successive terms, such an idea is very difficult to implement reliably (e.g., see Vehtari and Lampinen 2002 for discussion). Li, Villani, and Kohn (2010b) considered a similar approach to Geweke and Keane (2007) for model choice and an approximation where $p(\theta|y_{\leq T})$ is used instead of $p(\theta|y_{\leq T+i-1})$ in (6).

They presented some empirical evidence for the accuracy of this approach by comparison with a scheme where the posterior was updated sequentially every 100th observation in a financial time series example.

We note that our variational approach is very efficient at implementing sequential updating. Apart from the fact that VA is faster than MCMC to begin with, in the time series context, the result of the variational optimization from the last time step can be used to initialize the optimization for the current time step so that the convergence time of the variational scheme is generally small. This makes variational approaches ideally suited to model choice based on one-step-ahead predictions and the logarithmic score for time series data.

5. IMPROVING THE BASIC APPROXIMATION

It is well known that VAs can underestimate the variance of the posterior in the context of mixture models. For example, Wang and Titterington (2005) showed in the case of Gaussian mixtures that the covariance matrices from variational Bayes approximation are too small compared with those obtained by asymptotic arguments from maximum likelihood estimation. Here, we propose a novel approach to improve the estimates obtained from VA using SA, which can result in improved approximation of the posterior and reduced computational cost compared with MCMC. We proceed as follows. First, we integrate out the latent variables δ_i, thereby relaxing the assumption that the latent variables are independent of other parameters in the VA and allowing for an improvement. In approximating the marginal posterior with the indicators integrated out, we fix the posterior correlation structure to that obtained from Algorithm 1 and consider a variational optimization over the choice of the mean and variance in the variational posterior. Fixing the posterior correlation structure keeps the dimension of the optimization problem low. Ji, Shen, and West (2010) independently proposed a Monte Carlo SA that uses a similar approach for maximizing the lower bound numerically. However, we offer a number of improvements on their implementation, which take the form of an improved gradient estimate in the SA procedure and the idea of perturbing only the mean and scale of an initial VA obtained by the approach of Algorithm 1.

Consider once more the general setting with an unknown θ to learn about, prior $p(\theta)$, likelihood $p(y|\theta)$, and VA $q(\theta|\lambda)$, which comes from some parametric family where λ are parameters to be chosen. We will first develop an SA algorithm to maximize the lower bound on the log marginal likelihood. The ideas we describe here are useful quite generally and are not specific to the mixtures context. From (2), the lower bound is

$$L(\lambda) = \int q(\theta|\lambda) \log \frac{p(\theta)p(y|\theta)}{q(\theta|\lambda)} d\theta.$$

Maximizing this lower bound with respect to λ is equivalent to the problem of finding at least one root λ^* such that $g(\lambda) \equiv \frac{\partial}{\partial\lambda} L(\lambda) = 0$. When noisy measurements of $g(\lambda)$ are available, the root-finding SA algorithm introduced by Robbins and Monro (1951) may be used for finding λ^* and one of the conditions for the algorithm to converge is that the noise should have mean zero. Since the lower bound is an expectation with respect to $q(\theta|\lambda)$, unbiased measurements of $g(\lambda)$ at any λ may be computed provided it is valid to interchange the derivative $\frac{\partial}{\partial\lambda}$ and the integral. In this case,

$$g(\lambda) = \int \log\left\{\frac{p(\theta)p(y|\theta)}{q(\theta|\lambda)}\right\} \frac{\partial \log q(\theta|\lambda)}{\partial\lambda} q(\theta|\lambda)d\theta$$

since

$$\int \frac{\partial \log q(\theta|\lambda)}{\partial\lambda} q(\theta|\lambda)d\theta = 0.$$

An unbiased estimate of the gradient $g(\lambda)$ is thus

$$\left(\log\left\{\frac{p(\theta')p(y|\theta')}{q(\theta'|\lambda)}\right\} - c\right) \frac{\partial \log q(\theta'|\lambda)}{\partial\lambda}, \tag{7}$$

where $\theta' \sim q(\theta|\lambda)$ and c can be chosen arbitrarily. Now, note that

$$\log p(y) = \log \frac{p(\theta)p(y|\theta)}{p(\theta|y)},$$

for every θ. This suggests that if $q(\theta|\lambda)$ is a good approximation to $p(\theta|y)$ (as it might be close to the optimal value of λ), then the term

$$\left(\log\left\{\frac{p(\theta')p(y|\theta')}{q(\theta'|\lambda)}\right\} - c\right)$$

in the gradient estimate is nearly constant and equal to $\log p(y) - c$. In this case, this would make the variance of the gradient estimate, when λ is close to the optimal value, roughly equal to

$$(\log p(y) - c)^2 \text{var}\left(\frac{\partial \log q(\theta'|\lambda)}{\partial\lambda}\right).$$

This suggests that when λ is close to the optimal value, taking c close to $\log p(y)$ is a good choice. In our application to mixtures, for the sequence of gradient estimates constructed in the SA procedure, we initialize c to be the variational lower bound obtained from Algorithm 1 and then update it as the algorithm proceeds. This is described more precisely below. Ji, Shen, and West (2010) considered a similar approach, but they effectively used $c = 1$, obtained by differentiating directly under the integral sign. From our experience, choosing $c = 1$ is usually suboptimal, but they counteracted the variability in the gradient estimate by using multiple simulations from $q(\theta|\lambda)$. Clearly, from the above arguments, if $\log p(y)$ is large in magnitude, $c = 1$ could result in a gradient estimate with very high variance [since the factor $(\log p(y) - 1)^2$ is large], and this is supported by the simulations we have conducted (results not shown).

With an unbiased estimate of the gradient, a stochastic gradient algorithm can now be used for optimizing the lower bound. Let $\lambda^{(0)}$ be some initial estimate of λ. We consider the following algorithm:

Algorithm 2

For $k = 0, 1, 2, \ldots$

1. Simulate $\theta^{(k)} \sim q(\theta|\lambda^{(k)})$.

2. Set

$$\lambda^{(k+1)} = \lambda^{(k)} + a_k H(\lambda^{(k)}), \tag{8}$$

where $H(\lambda^{(k)})$ is an unbiased estimate of the gradient $g(\lambda^{(k)})$.

Under regularity conditions (Spall 2003), the $\lambda^{(k)}$ will converge to a local maximum of the lower bound. The a_k, $k \geq 0$, are a sequence satisfying the conditions:

$$a_k \to 0 \quad \sum_k a_k = \infty \quad \sum_k a_k^2 < \infty.$$

In particular, it is important to balance the gain sequence a_k so that it goes to zero fast enough to damp out noise effects when optimal λ is close, but not too fast to avoid false convergence. Step 2 (8) is a stochastic version of a gradient ascent algorithm where the step sizes decrease according to the sequence a_k.

An estimate of the lower bound on the log marginal likelihood from the SA iterates is

$$\frac{1}{N - N_0} \sum_{i=N_0+1}^{N} \log \frac{p(\theta^{(i)})p(y|\theta^{(i)})}{q(\theta^{(i)}|\lambda^{(i)})},$$

which involves negligible additional computation after running the recursion (8). Here, N is the total number of iterations (this is often a fixed number based on our computational budget, but should be large enough to ensure convergence) and N_0 is an initial number of iterates to discard when we are not yet close to the optimal solution. The SA algorithm is easy to implement provided that $q(\theta|\lambda)$ is easy to simulate from. In our gradient estimate, there is a constant c that we have argued should be chosen to be an estimate of the log marginal likelihood. We initialize c as the estimate of the log marginal likelihood from Algorithm 1, and at iteration $k > 1$ of Algorithm 2, we use the above SA log marginal likelihood estimate for c with $N_0 = 0$ and $N = k - 1$. The SA algorithm for our mixture model is described in more detail in the online supplementary materials.

6. EXAMPLES

6.1 EMULATION OF A RAINFALL–RUNOFF MODEL

Our first example is concerned with emulation of a deterministic rainfall–runoff model, a simplification of the Australian Water Balance Model (AWBM) of Boughton (2004). The goal of model emulation for a deterministic model is to develop a computationally cheap statistical surrogate (the emulator) for the original model for some characteristic of the model output of interest. For applications where the deterministic model is expensive to run and where we wish to run the model many times (in model calibration, for example), replacing the deterministic model by the emulator may allow similar results to be achieved with one order of magnitude reduction in computation time. Here, we will be concerned with using a mixture model to emulate the AWBM streamflow response at a time of peak rainfall input (the response y) as a function of the three AWBM model parameters (the covariates). For an overview of the statistical analysis of computer models and model emulation, see

Table 1. Log marginal likelihood (ML) estimated by VA (first row) and LPDS with 10-fold cross-validation estimated by VA (second row) and MCMC (third row) for inverse problem example

	Model A	Model B	Model C	Model D	Model E
ML with VA	−803.4	−688.4	−678.5	−682.8	−729
LPDS with VA	−65.9	−54.5	−51.5	−52.1	−57.2
LPDS with MCMC	−65.5	−54.2	−51.2	−51.4	−57.4

O'Hagan (2006). In the statistical literature, Gaussian process models that interpolate model output are often used for construction of emulators, but it is often recommended to include an independent noise term in such models (Pepelyshev 2010).

The AWBM uses input time series of rainfall and evapotranspiration data to produce estimates of catchment streamflow and is one of the most widely used rainfall–runoff models in Australia for applications such as catchment water yield estimation or design flood estimation. The model has three parameters—the maximum storage capacity S, the base flow index BFI, and the baseflow recession factor K. We have available model simulations for approximately 11 years of average monthly potential evapotranspiration and daily rainfall data for the Barrington River catchment, located in northeastern New South Wales in Australia. The model was run for 500 different values of the parameters (S, K, BFI), which were generated according to a maximin Latin hypercube design. Our goal here is to emulate the streamflow response of the AWBM at a fixed time (the time of peak input rainfall) as a function of the parameters S and K (the model output at this time is fairly insensitive to the value of BFI). We use a mixture model as an emulator where the response is the AWBM output at the time of peak input rainfall (y) and the predictors are S (x_1) and K (x_2). We have added a small amount of independent normal random noise with a standard deviation 0.01 to the response y to avoid degeneracies in the variance model in regions of the space where the response tends to be identically zero.

We considered fitting five models to the data. The first four models are MHR models with both predictors in the mean and variance models and in the model for the mixing weights. The models differ according to the number of mixture components, with models A, B, C, and D in Table 1 having, respectively, two, three, four, and five mixture components. Model E in Table 1 is a model with four mixture components but with only one intercept in the variance model (a homoscedastic mixture). For our normal prior distributions, we have used, in the notation in Section 2, $\mu^0_{\beta j} = 0$, $\Sigma^0_{\beta j} = 10{,}000 \cdot I$, $\mu^0_{\alpha j} = 0$, $\Sigma_{\alpha j} = 100 \cdot I$, $\mu^0_\gamma = 0$, and $\Sigma^0_\gamma = 100 \cdot I$, where mean vectors and covariance matrices have the appropriate dimensions depending on the model fitted. The estimated log marginal likelihoods (from the variational lower bound) for the models fitted are given in Table 1, showing a clear preference for model C, the four-component MHR model. However, it is not our goal in this work to support model choice based on the lower bound, and we discuss further below model choice via the LPDS.

Figure 1 summarizes the fitted model with four heteroscedastic mixture components. Here, we have separated the observations into clusters according to which mixture component each observation is most likely to belong to and plotted observations for each cluster,

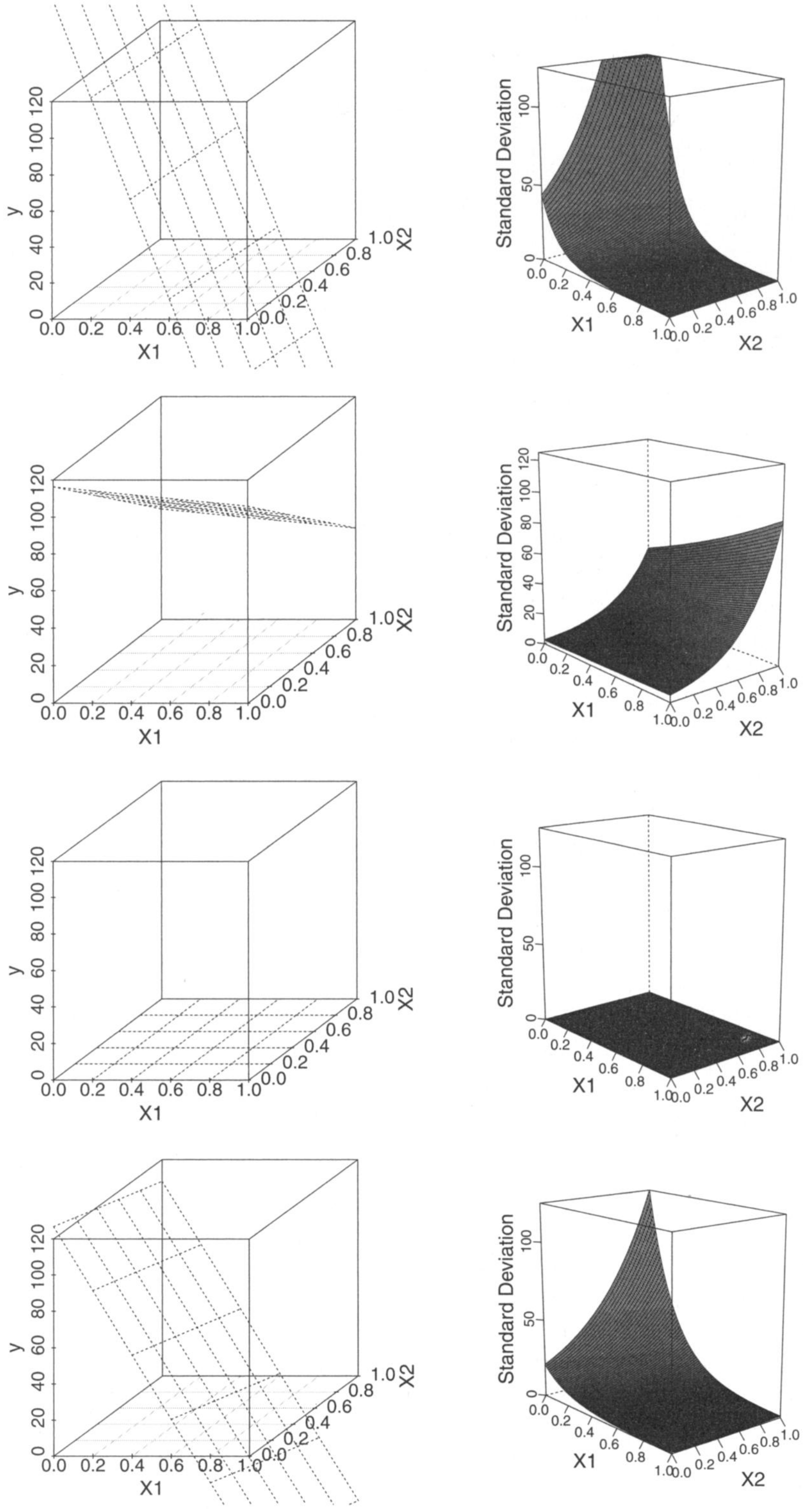

Figure 1. Fitted component means (first column) and standard deviations (second column) for four-component MHR model for rainfall–runoff example.

Table 2. Computation times for VA and MCMC for fitting inverse problem example

	Model A	Model B	Model C	Model D	Model E
Full data					
VA	88	146	215	274	254
MCMC	330	473	650	825	659
Cross-validation					
VA	121	184	281	393	276
MCMC	2941	4409	5979	7626	5929

NOTE: Computation times are in seconds for full data and cross-validation calculations.

together with the fitted mean and standard deviation for each mixture component. Different rows correspond to different mixture components.

We have emphasized that an important advantage of fast variational approaches to inference is the ability to fit many models for model assessment and exploratory analysis. For instance, it is very difficult to use cross-validatory approaches to model comparison if MCMC methods are used for computation since we require repeated MCMC runs for model fits to different parts of the data and for many models. Table 1 shows LPDS values obtained using both the VA and the MCMC. The LPDS values computed by the VA compare well with those obtained by the MCMC, and the results suggest that a model with four mixture components is adequate. We note that the results for the MCMC for model D need to be treated with some caution as there is very slow mixing in the MCMC scheme here due to the use of too many mixture components, and hence, a poorly identified model. For the VA for this case, one of the mixture components effectively drops out, with the mixing weights being very small for all observations for one of the components.

Table 2 shows the computation times taken in fitting the model to the full dataset and in implementing cross-validation using both an MCMC approach and our variational method without SA correction. All code was written in the R language and run on an Intel Core i5-2500 3.30-GHz processor workstation. Some difficulties in comparing MCMC with VA in this way need to be noted. First, the time taken to use the VA will depend on the way the algorithm is initialized and the stopping rule, and the rate of convergence will also depend on the problem. Similarly, for MCMC, the time taken depends on the number of sampling iterations, the number of "burn-in" iterations required to achieve convergence, and the sampling algorithm—these factors also tend to be problem-specific. Here, the method of intialization of the variational method is the one described in Section 3 using short runs for multiple clusterings, and we stop the variational algorithm when the relative change in the lower bound between successive iterations is less than 10^{-6}. The MCMC algorithms were run for 10,000 iterations, with 1000 "burn-in" iterations both in fitting the full dataset and in the cross-validation calculations. Such short run times are only possible because our MCMC scheme actually uses a very good proposal based on the VA itself. We considered a random-walk Metropolis–Hastings algorithm with the mixture component indicators integrated out, where the proposal covariances are taken from the variational method and parameters are updated in blocks that correspond to the variational factorization. This MCMC algorithm generally mixes rapidly and initial values can also be based on the VA so that 1000 "burn-in" iterations are sufficient for all models fitted here. Looking at the

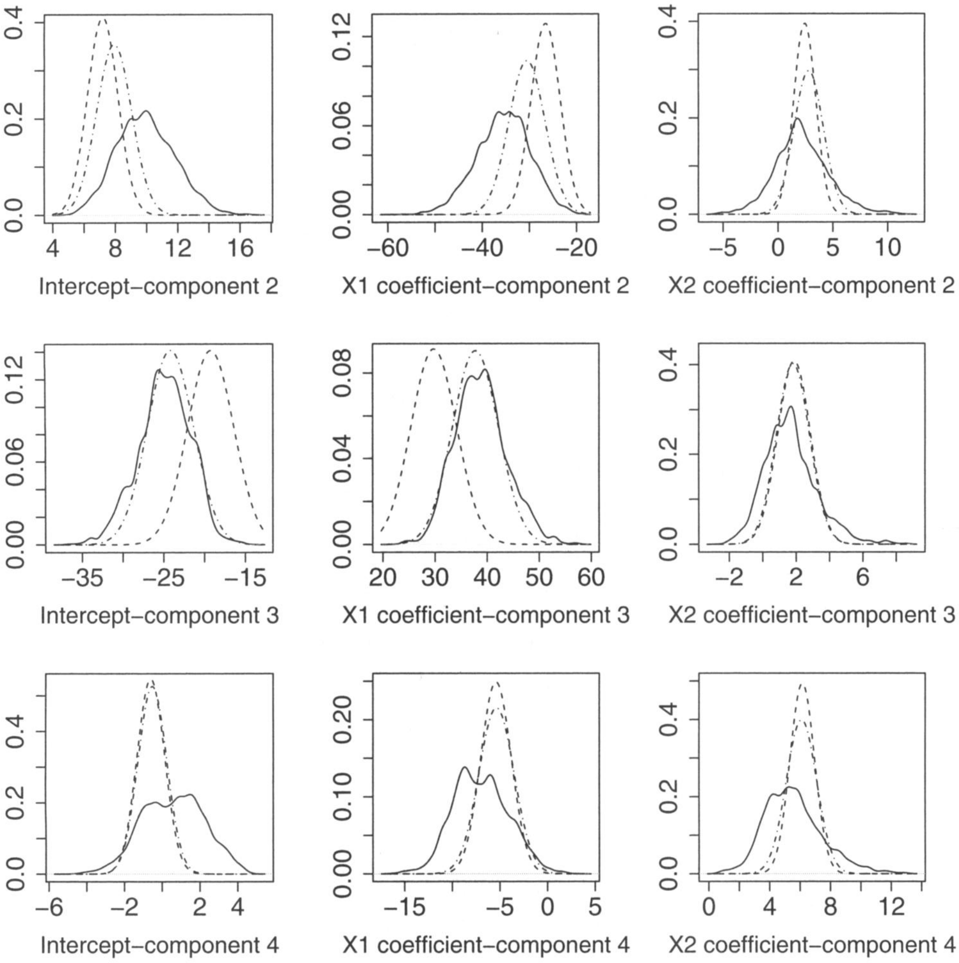

Figure 2. Marginal posterior distributions for parameters in the rainfall–runoff example in the mixing weights estimated by the Monte Carlo method (solid line), simple VA (dashed line), and VA with SA correction (dot-dashed line). Rows are different components, and the left, middle, and right columns, respectively, shows the intercept terms and coefficients for x_1 and x_2.

cross-validation calculations, there is a roughly 20-fold speedup by using VA compared to MCMC in the computations when using just 10,000 iterations in the MCMC sampling for all models. This is a fairly conservative estimate of the benefits and is consistent with other comparisons in the VA literature. The difficulties in convergence assessment using the MCMC approach are also avoided by the variational method. For model C, we also compared posterior distributions obtained via MCMC with both our simple VA and the VA incorporating SA correction. A total of 10,000 iterations of SA were used. Our gain sequence was $a_k = 0.4/(k + 10{,}000)^{0.9}$ for the variance adjustment parameters, and $a_k = 0.4/(x + 10{,}000)^{0.8}$ for the mean adjustment parameters. Computation of the SA correction took approximately 166 sec. Figure 2 shows the marginal posterior distributions for the parameters in the model for the mixing weights. We are looking at just one of the modes here, and there are no issues of label switching in the MCMC as the modes corresponding to relabeling are well separated. The SA correction is helpful in obtaining an improved approximation for at least some of the parameters, with the estimated posterior marginals from SA (dot-dashed lines) generally being closer to the Monte Carlo estimated marginals (solid lines) than the simple variational estimated marginals (dashed lines). There is little improvement in estimation of the marginal posteriors for the mean and variance parameters or in predictive inference by the SA correction (results not shown). Similar benefits in estimation of the mixing weight parameters have been observed in other examples that we have considered. We also conducted a small simulation study to investigate model selection performance for the variational approach using 10-fold cross-validation. In particular, we simulated 50 datasets from the fitted four-component heteroscedastic model, which was chosen as the best in the analysis above. The parameters used for simulating the data were the variational posterior mean values obtained from fitting to the real data. In our simulations, we compared heteroscedastic models with different numbers of mixture components and with x_1 and x_2 in both the mean and variance models (now, model C is the "true" model). When using cross-validation to select the best model, the true model was chosen in 32 of the 50 simulated datasets, with model D (with one extra mixture component) being chosen in 17 cases and a six-component MHR model being chosen once.

6.2 Time Series Example

Geweke and Keane (2007) considered a dataset of returns to the S&P500 stock market index. We consider an analysis of the same data but follow Li, Villani, and Kohn (2010b) and incorporate some more recent observations. The data consist of 4646 daily returns from January 1, 1990, to May 29, 2008, and our response y_t is $\log p_t/p_{t-1}$, where p_t is the closing S&P500 index on day t. Geweke and Keane (2007) and Li, Villani, and Kohn (2010b) considered time series models for the data using mixtures of regressions where the covariates include functions of lagged response values. We refer the reader to Li, Villani, and Kohn (2010b) for a more comprehensive description of the data, and we use their predictors Lastweek (average of returns for the last five trading days), LastMonth (average of returns for the last 20 trading days), and MaxMin95 $[(1-\Phi)\sum_s \phi^s(\log p^{(h)}_{t-1-s} - \log p^{(l)}_{t-1-s})$, with $\Phi = 0.95$, and $p^{(h)}_{t-1-s}$ and $p^{(l)}_{t-1-s}$ being the highest and the lowest values of the index on day t, respectively]. These covariates were found to be significant in the dispersion model in Li, Villani, and Kohn (2010b) in fitting a certain one-component skew t model with dispersion, skewness, and degrees of freedom—all functions of covariates.

For the S&P500 data, we follow Li, Villani, and Kohn (2010b) and take $T = 4646$ training observations and $T^* = 199$ validation observations. Li, Villani, and Kohn (2010b) noted that the choice of the last 199 observations in the series for validation is a difficult test for candidate models because this period covers the recent financial crisis, where unusually high volatility can be observed. We consider MHR models with only one intercept term in the mean model but an intercept term and the covariates LastWeek, LastMonth, and MaxMin95 in the variance model and mixing weights model and $m = 1, 2, 3$, and 4 components. Table 3 shows the LPDS values computed using the VA with sequential updating of the posterior at each time point as well as the those computed using the approximation of Li et al., where the posterior is not updated after the end of the training period (the

Table 3. LPDS values for one, two, three, and four mixture components for the MCMC method with approximation of Li et al. without sequential updating (first row), for the VA with approximation of Li et al. without sequential updating (second row), and for the VA with sequential updating (last row)

	Number of mixture components			
	1	2	3	4
Without sequential updating, MCMC	−477.8	−471.2	−469.0	−470.6
Without sequential updating, VA	−478.0	−470.1	−470.1	−471.7
With sequential updating, VA	−477.7	−470.0	−470.1	−473.3

latter computed using both the VA and the MCMC). Predictive density variates in the expression for the LPDS are approximated averaging over 1000 Monte Carlo draws of the parameters for each method. Based on the largest LPDS, it seems reasonable to choose a two-component mixture as providing an adequate model. Computation times for MCMC and VA are shown in Table 4. The computation time for MCMC is just for the initial fit (based on 10,000 iterations for each of the models with 1000 burn-in iterations), but for VA, we show the computation times for both the initial fit and the initial fit plus sequential updating for validation. The stopping criterion for the variational method is based on a relative tolerance of 10^{-6} for the lower bound. In this case, there would be a roughly 200-fold speed-up by employing the variational method as the complete computations using the variational method (initial fit plus validation) are similar to the computational requirements of just the initial fit for the MCMC method; also recall that the computational cost for the initial fit for MCMC needs to be multiplied by approximately $T^* = 199$ to get the computational cost for the complete computations.

Another application where MCMC methods may not be feasible at all for time series data is where a model is fitted repeatedly within a rolling window of observations. We illustrate this here for our two-component mixture model, where we examine parameter estimates for the model within different windows to investigate the question of structural breaks and model instability. For the S&P500 data, we consider windows of size $M = 500$. First, we fit the model to the first M observations. Next, we advance the rolling window by 50 observations (i.e., we consider observations 51 to $M + 50$) and refit the model. We continue in this way, advancing the rolling window by 50 observations at each step.

Table 4. Computation times (in seconds) for LPDS calculations for models with one, two, three, and four mixture components in the S&P500 example

	Number of mixture components			
	1	2	3	4
Initial fit MCMC	504	2463	3427	4417
Initial fit VA	1	739	1022	1442
Initial fit + validation VA	250	1902	2552	4754

NOTE: Rows 1–3, respectively, are computation times for initial fit for MCMC, initial fit for VA, and initial fit plus sequential updating for validation for VA

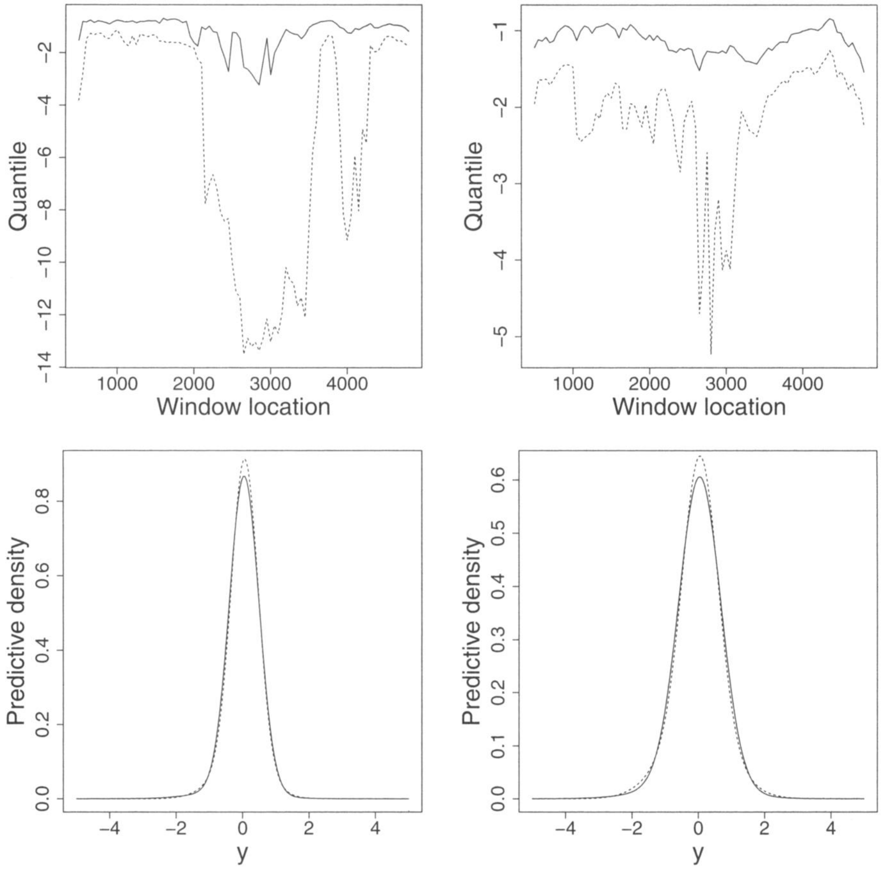

Figure 3. Estimated 5% (solid line) and 1% (dashed line) quantiles of predictive densities based on a rolling window for S&P 500 example and covariate values at $t = 1000$ (top left) and $t = 4000$ (top right). Estimated quantiles are plotted versus the upper edge of the rolling window. Also shown are the estimated predictive densities at covariate values for $t = 1000$ and $t = 4000$ (bottom left and right, respectively,) estimated based on the entire training dataset using MCMC (solid line) and VA (dashed line).

Figure 3 shows the estimated lower 1% and 5% quantiles of the predictive densities for the covariate values for times $t = 1000$ and $t = 4000$ versus the upper edge of the rolling window. There is some evidence of model instability and structural change. Also shown in Figure 3 are estimated predictive densities for the same covariates computed via MCMC (solid lines) and the VA (dashed lines). One can observe that the MCMC and variational predictive densities are nearly indistinguishable, so it can be said that the VA provides excellent predictive inference here.

7. DISCUSSION AND CONCLUSIONS

Our article describes fast VA methods for MHR models and illustrates the benefits of this methodology in problems where repeated refitting of models is required, such as in exploratory analysis and cross-validation approaches to model choice. There are a number of promising avenues for future research. One interesting idea is to combine variational methods (particularly the SA approach in Section 4) with MCMC methods applied to a subset of the data. It might be possible to get a rough idea of the correlation structure in the posterior from an MCMC run for a subset and then to adjust means and variances using SA approaches similar to those that we have described. There are many issues to be addressed in practice with such an approach, however. MCMC methods and variational methods can be complementary, in the sense that variational methods are able to provide good proposal distributions for MCMC schemes, a strategy that is sometimes called variational MCMC (de Freitas et al. 2001). The combination of variational methods with SA has the potential to broaden the applicability of such an approach. Another interesting extension that we have not pursued for MHR models is to allow some of the coefficients in the components to be shared across components. This may be particularly useful in the variance models.

SUPPLEMENTARY MATERIALS

Appendix: Details on the derivation of the variational lower bound, use of the marginal likelihood in Bayesian model choice, and details of the stochastic approximation algorithm. (vmhr.appendix.pdf)

R code and data: An R function to implement our basic variational approximation, together with an example, and the data for the two examples discussed in the article. (vmhr.zip)

ACKNOWLEDGMENTS

The authors thank John Ormerod for comments and suggestions related to this work. The authors also thank Lucy Marshall for supplying the data for the example in Section 6.1. David J. Nott was supported by the Singapore Ministry of Education (MOE) grant R-155-000-068-133. Siew Li Tan was partly supported as part of the Singapore Delft Water Alliance's (SDWA) tropical reservoir research program. Robert Kohn's research was partially supported by the Australian Research Council (ARC) grant DP0988579.

[Received October 2010. Revised September 2011.]

REFERENCES

Biernacki, C., Celeux, G., and Govaert, G. (2003), "Choosing Starting Values for the EM Algorithm for Getting the Highest Likelihood in Multivariate Gaussian Mixture Models," *Computational Statistics and Data Analysis*, 41, 561–575. [804]

Bishop, C. M. (2006), *Pattern Recognition and Machine Learning*, New York: Springer. [801,806]

Bishop, C. M., and Svensén, M. (2003), "Bayesian Hierarchical Mixtures of Experts," in *Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence*, eds. U. Kjaerulffand C. Meek, Waltham, MA: Morgan Kaufmann, p. 57–64. [800,801,805]

Blei, D. M., and Jordan, M. I. (2006), "Variational Inference for Dirichlet Process Mixtures," *Bayesian Analysis*, 1, 121–144. [800,802]

Braun, M., and McAuliffe, J. (2010), "Variational Inference for Large-Scale Models of Discrete Choice," *Journal of the American Statistical Association*, 105, 324–335. [803]

Boughton, W. (2004), "The Australian Water Balance Model," *Environmental Modelling and Software*, 19, 943–956. [810]

Constantinopoulos, C., and Likas, A. (2007), "Unsupervised Learning of Gaussian Mixtures Based on Variational Component Splitting," *IEEE Transactions on Neural Networks*, 18, 745–755. [800]

Corduneanu, A., and Bishop, C. M. (2001), "Variational Bayesian Model Selection for Mixture Distributions," in *Artificial Intelligence and Statistics*, eds. T. Jaakkolaand T. Richardson, Waltham, MA: Morgan Kaufmann, pp. 27–34. [800]

de Freitas, N., Højen-Sørensen, P., Jordan, M. I., and Russell, S. (2001), "Variational MCMC," in *Uncertainty in Artificial Intelligence (UAI): Proceedings of the 17th Conference*, eds. J. Breeseand D. Koller, San Francisco, CA: Morgan Kaufmann, pp. 120–127. [818]

De Iorio, M., Müller, P., Rosner, G. L., and MacEAchern, S. N. (2004), "An ANOVA Model for Dependent Random Measures," *Journal of the American Statistical Association*, 99, 205–215. [799]

Dunson, D. B., Pillai, N., and Park, J.-H. (2007), "Bayesian Density Regression," *Journal of the Royal Statistical Society*, Series B, 69, 163–183. [800]

Frühwirth-Schnatter, S. (2004), "Estimating Marginal Likelihoods for Mixture and Markov Switching Models Using Bridge Sampling Techniques," *The Econometrics Journal*, 7, 143–167. [806]

Geweke, J., and Amisano, G. (2010), "Comparing and Evaluating Bayesian Predictive Distributions of Asset Returns," *International Journal of Forecasting*, 26, 216–230. [798]

Geweke, J., and Keane, M. (2007), "Smoothly Mixing Regressions," *Journal of Econometrics*, 138, 252–291. [799,807,815]

Ghahramani, Z., and Beal, M. J. (2000), "Variational Inference for Bayesian Mixtures of Factor Analysers," in *Advances in Neural Information Processing Systems* (Vol. 12), eds. S. A. Solla, T. K. Leen, and K-R Müller, Cambridge: MIT Press, pp. 831–864. [800]

Griffin, J. E., and Steel, M. F. J. (2006), "Order-Based Dependent Dirichlet Processes," *Journal of the American Statistical Association*, 101, 179–194. [799]

Honkela, A., and Valpola, H. (2003), "On-Line Variational Bayesian Learning," in *Proceedings of the 4th International Symposium on Independent Component Analysis and Blind Signal Separation* (ICA 2003), Berlin: Springer, pp. 803–808. [806]

Jacobs, R., Jordan, M., Nowlan, S., and Hinton, G. (1991), "Adaptive Mixtures of Local Experts," *Neural Computation*, 3, 79–87. [798,801]

Ji, C., Shen, H., and West, M. (2010), "Bounded Approximations for Marginal Likelihoods," Technical Report, ISDS, Duke University. Available at *http://ftp.stat.duke.edu/WorkingPapers/10-05.html* [800,808,809]

Jiang, W., and Tanner, M. (1999), "Hierarchical Mixtures-of-Experts for Exponential Family Regression Models: Approximation and Maximum Likelihood Estimation," *The Annals of Statistics*, 27, 987–1011. [799]

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999), "An Introduction to Variational Methods for Graphical Models," in *Learning in Graphical Models*, ed. M. I. Jordan, Cambridge, MA: MIT Press, pp. 105–158. [801]

Jordan, M. I., and Jacobs, R. A. (1994), "Hierarchical Mixtures of Experts and the EM Algorithm," *Neural Computation*, 6, 181–214. [798,800,801]

Li, F., Villani, M., and Kohn, R. (2010a), "Modeling Conditional Densities Using Finite Smooth Mixtures," in *Mixtures: Estimation and Applications*, eds. K. L. Mengersen, C. P. Robert, and D. M. Titterington, Chichester, UK: John Wiley & Sons, Ltd. [798]

——— (2010b), "Flexible Modeling of Conditional Distributions Using Smooth Mixtures of Asymmetric Student *t* Densities," *Journal of Statistical Planning and Inference*, 140, 3638–3654. [807,815]

MacEachern, S. N. (1999), "Dependent Nonparametric Processes," in *ASA Proceedings of the Section on Bayesian Statistical Science*, Alexandria, VA: American Statistical Association, pp. 50–55. [799]

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman and Hall. [804]

McGrory, C. A., and Titterington, D. M. (2007), "Variational Approximations in Bayesian Model Selection for Finite Mixture Distributions," *Computational Statistics and Data Analysis*, 51, 5352–5367. [800,804]

Norets, A. (2010), "Approximation of Conditional Densities by Smooth Mixtures of Regressions," *The Annals of Statistics*, 38, 1733–1766. [799]

O'Hagan, A. (2006), "Bayesian Analysis of Computer Code Outputs: A Tutorial," *Reliability Engineering and System Safety*, 91, 1290–1300. [811]

Ormerod, J. T., and Wand, M. P. (2010), "Explaining Variational Approximations," *The American Statistician*, 64 (2), 140–153. [801,805]

Peng, F., Jacobs, R. A., and Tanner, M. A. (1996), "Bayesian Inference in Mixtures-of-Experts and Hierarchical Mixtures-of-Experts Models with an Application to Speech Recognition," *Journal of the American Statistician Association*, 91, 953–960. [799]

Pepelyshev, A. (2010), "The Role of the Nugget Term in the Gaussian Process Method," in *MODA 9—Advances in Model-Oriented Design and Analysis: Contributions to Statistics*, New York: Springer, pp. 149–156. [811]

Pesaran, M. H., and Timmermann, A. (2002), "Market Timing and Return Prediction Under Model Instability," *Journal of Empirical Finance*, 9, 495–510. [798]

Robbins, H., and Monro, S. (1951), "A Stochastic Approximation Method," *Annals of Mathematical Statistics*, 22, 400–407. [808]

Spall, J. C. (2003), *Introduction to Stochastic Search and Optimization: Estimation, Simulation and Control*, Hoboken, NJ: Wiley. [798,810]

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van der Linde, A. (2002), "Bayesian Measures of Model Complexity and Fit" (with discussion), *Journal of the Royal Statistical Society,* Series B, 64, 583–616. [800]

Ueda, N., and Ghahramani, Z. (2002), "Bayesian Model Search for Mixture Models Based on Optimizing Variational Bounds," *Neural Networks*, 15, 1223–11241. [800,801]

Vehtari, A., and Lampinen, J. (2002), "Bayesian Model Assessment and Comparison Using Cross-Validation Predictive Densities," *Neural Computation*, 14, 2439–2468. [807]

Villani, M., Kohn, R., and Giordani, P. (2009), "Regression Density Estimation Using Smooth Adaptive Gaussian Mixtures," *Journal of Econometrics*, 153, 155–173. [798,799]

Waterhouse, S., MacKay, D., and Robinson, T. (1996), "Bayesian Methods for Mixtures of Experts," in *Advances in Neural Information Processing Systems*, (Vol. 8), eds. D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, Cambridge, MA: MIT Press, pp. 351–357. [800,801,805]

Wand, M. P. (2002), "Vector Differential Calculus in Statistics," *The American Statistician*, 56, 55–62. [803]

Wang, B., and Titterington, D. M. (2005), "Inadequacy of Interval Estimates Corresponding to Variational Bayesian Approximations," in *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics*, eds. R. G. Cowell and Z. Ghahramani, Society for Artificial Intelligence and Statistics, pp. 373–380. [808]

Wedel, M. (2002), "Concomitant Variables in Finite Mixture Models," *Statistica Neerlandica*, 56, 362–375. [798]

West, M. (1985), "Generalized Linear Models: Outlier Accommodation, Scale Parameters and Prior Distributions," in *Bayesian Statistics 2*, eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, North Holland: Amsterdam, pp. 531–538. [804]

Wood, S. A., Jiang, W., and Tanner, M. A. (2002), "Bayesian Mixture of Splines for Spatially Adaptive Nonparametric Regression," *Biometrika*, 89, 513–528. [799]

Wood, S. A., Kohn, R., Cottet, R., Jiang, W., and Tanner, M. (2008), "Locally Adaptive Nonparametric Binary Regression," *Journal of Computational and Graphical Statistics*, 17, 352–372. [799]

Wu, B., McGrory, C. A., and Pettitt, A. N. (2012), "A New Variational Bayesian Algorithm With Application to Human Mobility Pattern Modeling," *Statistics and Computing*, 22 (1), 185–203. [800]