Regression density estimation with variational methods and stochastic approximation

David J. Nott, Siew Li Tan, Mattias Villani and Robert Kohn*

Abstract

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 mixtures of experts models and our article considers flexible mixtures of heteroscedastic experts (MHE) regression 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 methods for inference. Our motivation is that alternative computationally intensive 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 variational approximation for MHE models is described where the variational lower bound is in closed form. Second, the basic approximation can be improved by using stochastic approximation methods to perturb the initial solution to attain higher accuracy. Third, the advantages of our approach for model choice and evaluation compared to 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.

*David J. Nott is Associate Professor, Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546. (email standj@nus.edu.sg). Siew Li Tan is PhD student, Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546 (email g0900760@nus.edu.sg). Mattias Villani is Professor, Division of Statistics, Dept. of Computer and Information Science, Linköping University, SE-58183 Linköping, Sweden. (email: mattias.villani@liu.se). Robert Kohn is Professor, Australian School of Business, University of New South Wales, Sydney 2052 Australia (email r.kohn@unsw.edu.au).
Keywords: Bayesian model selection, heteroscedasticity, mixtures of experts, stochastic approximation, variational approximation.

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 mixtures of experts (ME) models and our article considers flexible regression models where the response distribution is a normal mixture, with the component 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 experts (MHE) models extend conventional mixtures of experts models (Jacobs et al., 1991, Jordan and Jacobs, 1994) by allowing the component models to be heteroscedastic. The terminology mixtures of experts comes from machine learning, but it is frequently 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 mixture of experts is sometimes referred to as the concomitant variable mixture regression model (see, for example, Wedel, 2002).

The heteroscedastic extension of mixture of experts models is important in modern applications since simulations by Villani et al. (2009) showed that, when used to model heteroscedastic data, the performance of ME 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 MHE models, fewer mixture components are required which makes the estimation and interpretation of mixture models easier. In an analysis of the benchmark LIDAR data set Li et al. (2010a) showed that an ME model with homoscedastic thin plate components requires three components to achieve approximately the same performance as an MHE model with a single thin plate component, providing further evidence that MHE models help to reduce the number of mixture components required. Moreover, Villani et al. (2009) demonstrated that the fit of an MHE model to homoscedastic data is comparable to
that of an ME model with homoscedastic components.

Our article makes three contributions. First, fast variational methods are developed for fitting MHE models and a variational lower bound is obtained in closed form. Second, stochastic approximation optimization techniques (see, for example, Spall, 2003) are used to improve the variational approximation to attain higher accuracy. The computational methods developed here can also be applied beyond the context of MHE 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 variational approximation methods provide an attractive alternative to 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 re-estimated at each step. Another popular time series procedure for studying model performance is to use a rolling window (Pesaran and Timmermann, 2002). Suppose the data set 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 speed up that comes from initializing the variational optimization based on the fit to the previous window rather than treating each fit as an independent computation.

Use of variational approximations 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 $\theta$ 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 $\theta$ 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 data sets 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 experts models were first considered in Peng et al. (1996). Wood et al. (2002) and Wood et al. (2008) consider mixtures of experts models with flexible terms for the covariates for continuous and binary responses respectively. Geweke and Keane (2007) also take a Bayesian approach to inference in mixtures of experts models with homoscedastic experts with a multinomial probit for the gating function which allows a convenient Gibbs sampling MCMC scheme. MHE models have been considered previously by Villani et al. (2009). They use a Bayesian approach to inference with MCMC methods for computation and consider general smooth terms for the covariates and variable selection in the mean and variance models and the gating function. Norets (2010) considers 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 expert components come from the same exponential family. Their results are also useful for conditional densities which are discrete.

Methods related to mixtures of experts 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 which varies over the space. For common priors the resulting models might be considered to be mixtures of experts 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 expert components being of qualitatively different types. We do not discuss these methods any further but refer to MacEachern (1999), De Iorio et al. (2004), Griffin and Steel (2006) and Dunson et al. (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 experts models (Jordan and Jacobs, 1994, for example) used maximum likelihood and the 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 paper (Waterhouse, MacKay and Robinson, 1996, Ueda and Ghahramani, 2002, Bishop and Svensén, 2003). However, these authors did not consider heteroscedastic experts. Moreover, our paper 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 which follow a variational approach. Corduneanu and Bishop (2001) consider a variational lower bound on the log marginal likelihood with all parameters, except the mixing coefficients, integrated out, and estimate 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) consider a variational extension of the DIC criterion (Spiegelhalter et al., 2002) and a variational optimization technique where the algorithm is initialised with a large number of components and mixture components whose weightings become sufficiently small are dropped out as the optimization proceeds. Blei and Jordan (2006) have considered variational approximation for Dirichlet process mixture models. Recently Wu et al. (2011) have considered variational methods for fitting mixtures for data which require models with narrow widely separated mixture components. They discuss sophisticated split and merge algorithms, building on earlier related methods such as those of Ghahramani and Beal (2000), Ueda and Ghahramani (2002) and Constantinopoulos and Likas (2007), for simultaneous model selection and parameter estimation as well as novel criteria for model evaluation.

One contribution of our paper is the development of broadly applicable stochastic approximation correction methods for variational approximation. From our simulation studies, the stochastic approximation correction is very helpful for getting an improved approximation and requires less computation time than MCMC methods. Ji, Shen and West (2010) recently proposed similar stochastic approximation methods for learning variational approximations but we offer a number of improvements on their implementation. In particular, we are able to suggest an improved gradient estimate compared to their approach and consider a strategy of perturbing only the mean and scale in an initial variational approximation 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 et al. (2010) suggest 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 paper is organized as follows. Section 2 introduces the mixture of heteroscedastic experts model. Section 3 describes fast variational approximation methods for MHE models and Section 4 uses of variational methods in model choice. Section 5 discusses improvements on the basic approximation using a stochastic approximation 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 experts models

Suppose that responses $y_1, ..., y_n$ are observed. They are modelled by a mixture of experts model (Jacobs et al., 1991, Jordan and Jacobs, 1994) of the form

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

where $\delta_i$ is a categorical latent variable with $k$ categories, $\delta_i \in \{1, ..., k\}$, $x_i = (x_{i1}, ..., x_{ip})^T$ and $z_i = (z_{i1}, ..., z_{im})^T$ are vectors of covariates for observation $i$, and $\beta_j = (\beta_{j1}, ..., \beta_{jp})^T$ and $\alpha_j = (\alpha_{j1}, ..., \alpha_{jm})^T$, $j = 1, ..., k$ 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 $\delta_i$ is

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

where $v_i = (v_{i1}, ..., v_{ir})^T$ is a vector of covariates and $\gamma_j = (\gamma_{j1}, ..., \gamma_{jr})^T$ are vectors of unknown parameters, $j = 2, ..., k$ with $\gamma_1$ set to be identically zero for identifiability. With this prior the responses are modelled 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 $\beta_j$, $\beta_j \sim N(\mu_{\beta j}^0, \Sigma_{\beta j}^0)$, $j = 1, ..., k$, for $\alpha_j$, $\alpha_j \sim N(\mu_{\alpha j}^0, \Sigma_{\alpha j}^0)$, $j = 1, ..., k$, and for $\gamma = (\gamma_2^T, ..., \gamma_k^T)^T$, $\gamma \sim N(\mu_{\gamma}^0, \Sigma_{\gamma})$. We will describe fast methods for variational inference in the above model. Variational inference for mixtures of experts 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 expert components a variational lower bound can still be computed in closed form, allowing fast computation.
3 Variational approximation

We consider a variational approximation to the joint posterior distribution of all the parameters $\theta$ of the form $q(\theta|\lambda)$ where $\lambda$ is a set of variational parameters to be chosen. Variational approximation 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, Chapter 10) gives a recent summary. Ormerod and Wand (2009) give an introduction to variational approximation 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 $p(\theta|y)$, i.e.,

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

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

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

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 variational approximation. From (1), the difference between the lower bound and the log marginal likelihood is the Kullback-Leibler divergence between the posterior and variational approximation, 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 variational approximation for the MHE model and previous developments of variational methods for ME models with homoscedastic experts lies both in the fact that a closed form derivation of the variational lower bound is not obvious in the heteroscedastic case, as well as in the need to deal with the variance parameters in the heteroscedastic linear expert models in the 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 of 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 \( \gamma = (\gamma_1^T, \ldots, \gamma_k^T)^T \) so that \( \theta = (\delta^T, \beta^T, \alpha^T, \gamma^T)^T \). For convenience we write our variational approximation \( q(\theta|\lambda) \) as \( q(\theta) \), suppressing dependence on \( \lambda \) in the notation. We consider a variational approximation 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, \( N(\mu_{\beta_i}, \Sigma_{\beta_i}) \) say for \( i = 1, \ldots, k \), \( q(\alpha_i) \) is normal, \( N(\mu_{\alpha_i}, \Sigma_{\alpha_i}) \) say for \( i = 1, \ldots, k \), \( q(\gamma) \) is a delta function placing point mass of 1 on \( \mu_{\gamma} \), \( q(\gamma) = \delta(\gamma - \mu_{\gamma}) \) 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 expert components are independent and independent of all other parameters, that mean and variance parameters are independent, and that the latent variables \( \delta \) are independent of each other and independent of all other parameters. We assumed a degenerate point mass variational posterior for \( \gamma \) in order to make computation of the lower bound tractable. However, following our description below of the variational algorithm which 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 variational approximations 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 to other approximations, particularly in high dimensions. For example, Blei and Jordan (2005), Section 5, demonstrate in the context of Dirichlet process mixture models that predictive inference based on a variational approximation is similar to fully Bayes predictive inference implemented via MCMC. Braun and McAuliffe (2010), Section 4, report similar findings in large-scale models of discrete choice.

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

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 supplementary materials)

\[
L = -\frac{n}{2} \log 2\pi + \frac{(p+m)k}{2} + \log p(\mu^q) - \frac{1}{2} \sum_{j=1}^{k} \log |\Sigma^0_\beta_j| - \frac{1}{2} \sum_{j=1}^{k} \text{tr}(\Sigma^0_\beta_j^{-1}\Sigma^q_\beta_j) \\
-\frac{1}{2} \sum_{j=1}^{k} (\mu^q_\beta_j - \mu^0_\beta_j)^T \Sigma^0_\beta_j^{-1} (\mu^q_\beta_j - \mu^0_\beta_j) - \frac{1}{2} \sum_{j=1}^{k} \log |\Sigma^0_\alpha_j| - \frac{1}{2} \sum_{j=1}^{k} \text{tr}(\Sigma^0_\alpha_j^{-1}\Sigma^q_\alpha_j) \\
-\frac{1}{2} \sum_{j=1}^{k} (\mu^q_\alpha_j - \mu^0_\alpha_j)^T \Sigma^0_\alpha_j^{-1} (\mu^q_\alpha_j - \mu^0_\alpha_j) + \sum_{i=1}^{n} \sum_{j=1}^{k} q_{ij} \log \frac{p_{ij}}{q_{ij}} + \frac{1}{2} \sum_{j=1}^{k} \log |\Sigma^q_\beta_j| + \frac{1}{2} \sum_{j=1}^{k} \log |\Sigma^q_\alpha_j| \\
- \sum_{i=1}^{n} \sum_{j=1}^{k} q_{ij} \left\{ \frac{1}{2} z_i^T \mu^q_\alpha_j + \frac{1}{2} \frac{(y_i - x_i^T \mu^q_\beta_j)^T (y_i - x_i^T \mu^q_\beta_j)}{\exp(z_i^T \mu^q_\alpha_j - 1/2 z_i^T \Sigma^q_\alpha_j z_i)} \right\}. \tag{3} \]

Here \( p(\mu^q) \) is the prior distribution for \( \gamma \) evaluated at \( \mu^q \) and \( p_{ij} \) is evaluated setting \( \gamma = \mu^q \).

The variational parameters to be optimized consist of \( \mu^q_\beta_j, \Sigma^q_\beta_j, \mu^q_\alpha_j, \Sigma^q_\alpha_j, j = 1, \ldots, k, \mu_q^j \) 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 which are available in closed form are easy to derive using vector differential calculus (see, for example, Wand, 2002). To initialize the algorithm, we first generate an initial clustering of the data. Then for this initial clustering we do the following.

**Algorithm 1:**

Initialize: \( \mu^q_\alpha_j = \Sigma^q_\alpha_j = 0 \) for \( j = 1, \ldots, k \) and \( q_{ij} \) as 1 if the \( i \)th observation lies in cluster \( j \) and 0 otherwise.

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

- \( \Sigma^q_\beta_j \leftarrow \left( X^T D_j X + \Sigma^0_\beta_j^{-1} \right)^{-1} \) where \( X \) is the design matrix with \( i \)th row \( x_i^T \), \( y \) is the vector of responses and \( D_j \) is the diagonal matrix with \( i \)th diagonal entry \( q_{ij}/\exp(z_i^T \mu^q_\alpha_j - 1/2 z_i^T \Sigma^q_\alpha_j z_i) \).

- \( \mu^q_\beta_j \leftarrow \Sigma^q_\beta_j \left( \Sigma^0_\beta_j^{-1} \mu^0_\beta_j + X^T D_j y \right) \).
• 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 $z_i^T \mu_{\alpha j}^q - 1/2z_i^T \Sigma_{\alpha j}^q z_i$ where the terms $-1/2z_i^T \Sigma_{\alpha j}^q z_i$ define an offset. Although the mode has no closed form expression it is easily found by an iteratively weighted least squares approach (McCullagh and Nelder, 1989, West, 1985) or some other numerical optimization technique.

• $\Sigma_{\alpha j}^q \leftarrow \left(Z^T W_j Z + \Sigma_{\alpha j}^0 \right)^{-1}$ where $Z$ is the design matrix with $i$th row $z_i^T$, $i = 1, ..., n$ and $W_j$ is diagonal with $i$th 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 $i = 1, ..., n,$

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

• Set $\mu_{\gamma}^q$ to be the conditional mode of the lower bound fixing other variational parameters at their current values. As a function of $\mu_{\gamma}^q$, 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 $\mu_{\gamma}^q$ for $\gamma$. This is the log posterior for a Bayesian multinomial regression with normal prior on $\mu_{\gamma}^q$ and where the $i$th response is $(q_{i1}, ..., q_{ik})^T$. 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 in order to initialize the parameters. We consider multiple clusterings in order 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 do 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 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 do selection of 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 ME model with homoscedastic components (Bishop and Svensén, 2003), not all the parameters have closed form updates. In steps 1, 2 and 5 above 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 (see, for instance, Ormerod and Wand, 2009)

\[
q(\alpha_j) \propto \exp \left\{ E(\log p(\theta)p(y|\theta)) \right\}
\]

where the expectation in the exponent is with respect to the variational posterior for all parameters except \( \alpha_j \). The expectation in (4) takes the form, apart from additive constants not depending on \( \alpha_j \), of the log posterior for a gamma generalized linear model (the same model as considered in step 3 above, 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 \((Z^T W_j Z + \Sigma_{\alpha_j}^{-1})^{-1}\) with the \( i \)th diagonal element of \( W_j \) defined as in step 4 above. Similar reasoning was used by Waterhouse, Mackay and Robinson (1996) in approximating the posterior distribution for the gating function parameters in the homoscedastic mixture of experts 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 don’t see any easy solutions to this problem. One possible remedy consists of 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 $\gamma$ with a normal approximation, in which the mean is $\mu_q^\gamma$ 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 $\Sigma_q^\alpha_j$ in step 4 of Algorithm 1. Waterhouse, Mackay and Robinson (1996) outline a similar idea but they use 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 $\gamma$. Write $\Sigma_q^\gamma$ for the covariance matrix of the variational posterior for $\gamma$. With this variational posterior the variational lower bound on $\log p(y)$ is the same as (3), except we need to replace $\sum_{i=1}^n \sum_{j=1}^k q_{ij} \log p_{ij} + \log p(\mu_q^\gamma)$ with

$$
\sum_{i=1}^n \sum_{j=1}^k q_{ij} E \left( \log \left\{ \frac{\exp(v_i^T \gamma_j)}{\sum_{l=1}^k \exp(v_i^T \gamma_l)} \right\} \right) - \frac{1}{2} \log |\Sigma_q^\gamma| - \frac{1}{2} (\mu_q^\gamma - \mu_0^\gamma)^T \Sigma_0^{-1} (\mu_q^\gamma - \mu_0^\gamma) - \frac{1}{2} \text{tr} \left( \Sigma_0^{-1} \Sigma_q^\gamma \right) + \frac{1}{2} \log |\Sigma_q^\gamma| + \frac{r(k-1)}{2}.
$$

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

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

where $\mu_q^\gamma_j$ is the subvector of $\mu_q^\gamma$ corresponding to $\gamma_j$, $j = 2, ..., r$. 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 for estimating the marginal likelihood (see, for example, 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 relabelling 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 doesn’t 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 don’t attempt any adjustment. In the examples of 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 data sets, since the posterior is of the same form as the prior for $(\beta, \alpha, \gamma)$, it is easy to implement our algorithm sequentially after splitting the data set up 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 with 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 Vapola (2003) present an on-line 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 favour 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 data set. The corresponding test sets (the parts of the data which 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

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

Note that

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

13
where $\theta$ denotes all the unknown parameters. We have assumed here that $y_F$ and $y_T$ are conditionally independent given $\theta$. This usually does not hold for time series data and modified approaches are appropriate for that case, as we discuss below. In the mixture of experts context $p(y_F|X_F, \theta)$ is easy to write down as a normal mixture given the parameters, and we replace $p(\theta|y)$ with our variational approximation $q(\theta)$. 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 we consider autoregressive time series models in the form of MHE models, and in the time series context the cross-validation approach described above is not very natural. Both Geweke and Keane (2007) and Li et al. (2010b) consider a training set $y_{\leq T} = (y_1, \ldots, y_T)$ of $T$ initial observations and then measure 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

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

where

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

where $p(\theta|y_{\leq T+i-1})$ denotes the posterior distribution for all unknowns $\theta$ 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) note 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 (see, for example, Vehtari and Lampinen, 2002, for discussion). Li et al. (2010b) consider 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 for implementing sequential updating. Apart from the fact that variational approximation 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 variational approximations 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 to those obtained by asymptotic arguments from maximum likelihood estimation. Here, we propose a novel approach to improve the estimates obtained from variational approximation using stochastic approximation (SA) which can result in improved approximation of the posterior with reduced computational cost compared to MCMC. We proceed as follows. First, we integrate out the latent variables \( \delta_i \) thereby relaxing the assumption that the latent variables are independent of other parameters in the variational approximation 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 et al. (2010) independently proposed a Monte Carlo stochastic approximation for maximizing the lower bound numerically which uses a similar approach. 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 variational approximation obtained by the approach of Algorithm 1.

Consider once more the general setting with an unknown \( \theta \) to learn about, prior \( p(\theta) \), likelihood \( p(y|\theta) \) and variational approximation \( q(\theta|\lambda) \) that comes from some parametric family where \( \lambda \) are parameters to be chosen. We will first develop a stochastic approximation
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

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

Maximizing this lower bound with respect to $\lambda$ is equivalent to the problem of finding at least one root $\lambda^*$ 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 $\lambda^*$ and one of the conditions for the algorithm to converge is for the noise to have mean zero. Since the lower bound is an expectation with respect to $q(\theta|\lambda)$, unbiased measurements of $g(\lambda)$ at any $\lambda$ 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

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

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

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

for every $\theta$. This suggests that if $q(\theta|\lambda)$ is a good approximation to $p(\theta|y)$, (as it might be near the optimal value of $\lambda$) 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 $\lambda$ is near 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 $\lambda$ 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 stochastic approximation 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 et al. (2010) considered a similar approach but they effectively use $c = 1$, obtained by differentiating directly under the integral sign. From our experience, choosing $c = 1$ is usually suboptimal, but they counteract 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 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 $\lambda$. We consider the following algorithm:

**Algorithm 2:**
For $k = 0, 1, 2, ...$

- Simulate $\theta^{(k)} \sim q(\theta|\lambda^{(k)})$.
- Set
  
  $\lambda^{(k+1)} = \lambda^{(k)} + a_k H(\lambda^{(k)})$ \hspace{1cm} (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 $\lambda$ is close, but not too fast to avoid false convergence. The step (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 stochastic ap-
proximation 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 where we are not yet close to the optimal solution. The stochastic approximation 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 stochastic approximation log marginal likelihood estimate for \(c\) with \(N_0 = 0\) and \(N = k - 1\). The stochastic approximation algorithm for our mixture model is described in more detail in the 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, the Australian Water Balance Model (AWBM) (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 an order of magnitude reduction in computation time. Here we will be concerned with using a mixture of experts 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 O’Hagan (2006). In the statistical literature Gaussian process models which 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 estimating catchment water yield or design flood estimation. The model has three parameters - the maximum storage capacity $S$, the base flow index $BFI$ and baseflow recession factor $K$. We have available model simulations for approximately eleven years of average monthly potential evapotranspiration and daily rainfall data for the Barrington River catchment, located in north eastern 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 the $BFI$). We use a mixture of experts 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 MHE models with both predictors in the mean and variance models. The models differ according to the number of mixture components with models A, B, C and D in Table 1 having respectively 2, 3, 4 and 5 mixture components. Model E in Table 1 is a model with four mixture components but with only an intercept in the variance model for the experts (a homoscedastic mixture of experts). For our normal prior distributions we have used, in the notation of Section 2, $\mu_{0j} = 0$, $\Sigma_{0j} = 10000 \cdot I$, $\mu_{0j} = 0$, $\Sigma_{aj} = 100 \cdot I$, $\mu_{0j} = 0$ and $\Sigma_{0j} = 100 \cdot I$ 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 4 component MHE 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 4 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 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-
Figure 1: Fitted component means (first column) and standard deviations (second column) for four component MHE model for rainfall-runoff example.
Table 1: Marginal log likelihood estimated by variational lower bound (first row) and LPDS with ten-fold cross validation estimated by variational approximation (second row) and MCMC (third row) for inverse problem example.

<table>
<thead>
<tr>
<th>Model</th>
<th>Model A</th>
<th>Model B</th>
<th>Model C</th>
<th>Model D</th>
<th>Model E</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML with VA</td>
<td>-803.4</td>
<td>-688.4</td>
<td>-678.5</td>
<td>-682.8</td>
<td>-729</td>
</tr>
<tr>
<td>LPDS with VA</td>
<td>-65.9</td>
<td>-54.5</td>
<td>-51.5</td>
<td>-52.1</td>
<td>-57.2</td>
</tr>
<tr>
<td>LPDS with MCMC</td>
<td>-65.5</td>
<td>-54.2</td>
<td>-51.2</td>
<td>-51.4</td>
<td>-57.4</td>
</tr>
</tbody>
</table>

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 from both using the variational approximation and MCMC. The LPDS values computed by the variational approach compare well with those obtained by MCMC, and the results suggest that a model with 4 mixture components is adequate. We note that the results for 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 variational approximation 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 CPU times taken for fitting the model to the full dataset and for implementing cross-validation using both an MCMC approach and our variational method without stochastic approximation correction. All code was written in the R language and run on an Intel Core i5-2500 3.30Ghz processor workstation. Some difficulties in comparing MCMC with variational approximation in this way need to be noted. First, the time taken to use the variational approximation will depend on the way the algorithm is initialized, 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 initialization 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 1,000 iterations burn in both for fitting the full data set 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 variational approximation 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 variational approximation so that 1000 “burn in” iterations was sufficient for all models fitted here. Looking at the cross-validation calculations, there is a roughly 20 fold speed up by using variational approximation 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 consistent with other comparisons in the variational approximation literature. The difficulties of convergence assessment in 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 variational approximation and the variational approximation incorporating stochastic approximation correction. 10,000 iterations of stochastic approximation were used. Our gain sequence was $a_k = 0.4/(k + 10000)^{0.9}$ for the variance adjustment parameters and $a_k = 0.4/(x + 10000)^{0.8}$ for the mean adjustment parameters. Computation of the stochastic approximation correction took approximately 166 seconds of CPU time. Figure 2 shows the marginal posterior distributions for the parameters in the gating function. We are looking at just one of the modes here and there are no issues with label switching in the MCMC as the modes corresponding to relabelling are well separated. The stochastic approximation correction is helpful for obtaining an improved approximation for at least some of the parameters, with the estimated posterior marginals from stochastic approximation(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 stochastic approximation correction (results not shown). Similar benefits

Table 2: CPU times for variational approximation (VA) and MCMC for fitting inverse problem example. Times (in seconds) are given for full data and cross-validation calculations, for models 1-4.

<table>
<thead>
<tr>
<th>Model</th>
<th>Model A</th>
<th>Model B</th>
<th>Model C</th>
<th>Model D</th>
<th>Model E</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full data VA</td>
<td>88</td>
<td>146</td>
<td>215</td>
<td>274</td>
<td>254</td>
</tr>
<tr>
<td>MCMC</td>
<td>330</td>
<td>473</td>
<td>650</td>
<td>825</td>
<td>659</td>
</tr>
<tr>
<td>Cross validation VA</td>
<td>121</td>
<td>184</td>
<td>281</td>
<td>393</td>
<td>276</td>
</tr>
<tr>
<td>MCMC</td>
<td>2941</td>
<td>4409</td>
<td>5979</td>
<td>7626</td>
<td>5929</td>
</tr>
</tbody>
</table>
in estimation of the gating function 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 ten fold cross-validation. In particular, we simulated 50 data sets from the fitted four component heteroscedastic model that was chosen as 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 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 for 32 of the 50 simulated data sets, with model D (with one extra mixture component) being chosen in 17 cases and a six component MHE model chosen once.

6.2 Time series example

Geweke and Keane (2007) consider a data set of returns to the S&P500 stock market index. We consider an analysis of the same data but follow Li et al. (2010b) and incorporate some more recent observations. The data consists of 4646 daily returns from January 1, 1990 to May 29, 2008 and our response \( y_t = \log p_t / p_{t-1} \) where \( p_t \) is the closing S&P500 index on day \( t \). Geweke and Keane (2007) and Li et al. (2010b) consider time series models for the data using mixtures of experts where the covariates include functions of lagged response values. We refer the reader to Li et al. (2010b) for a more comprehensive description of the data, and we use their predictors \( \text{LastWeek} \) (average of return for last 5 trading days) \( \text{LastMonth} \) (average of return for last 20 trading days) and \( \text{MaxMin95} \) \( (1 - \Phi) \sum_s \phi_s (\log p_{t-1-s}^{(h)} - \log p_{t-1-s}^{(l)}) \) with \( \Phi = 0.95 \), \( p_{t-1-s}^{(h)} \) and \( p_{t-1-s}^{(l)} \) the highest and lowest values of the index on day \( t \). These covariates were found to be significant in the dispersion model in Li et al. (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 et al. (2010b) and take \( T = 4646 \) training observations and \( T^* = 199 \) validation observations. Li et al. (2010b) note 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 mixtures of experts models with only an intercept term in the mean model but an intercept and the covariates \( \text{LastWeek}, \text{LastMonth} \) and \( \text{MaxMin95} \) in the variance model and gating function and \( m = 1, 2, 3 \) and 4 experts. Table 3 shows the LPDS values computed using the variational approximation with sequential updating of the posterior at
Figure 2: Marginal posterior distributions for parameters in the rainfall runoff example in the gating function estimated by Monte Carlo method (solid line), simple variational approximation (dashed line) and variational approximation with stochastic approximation 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 $x_2$. 
each time point as well as the LPDS computed using the approximation of Li et al. where the posterior is not updated after the end of the training period (the latter for both the variational method and using 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 variational

Table 3: LPDS values for 1,2,3 and 4 mixture components and MCMC method with approximation of Li et al. without sequential updating (first line) variational method with approximation of Li et al. (second line) and variational method with sequential updating (last line)

<table>
<thead>
<tr>
<th>Number of mixture components</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without sequential updating, MCMC</td>
<td>-477.8</td>
<td>-471.2</td>
<td>-469.0</td>
<td>-470.6</td>
</tr>
<tr>
<td>Without sequential updating, VA</td>
<td>-478.0</td>
<td>-470.1</td>
<td>-470.1</td>
<td>-471.7</td>
</tr>
<tr>
<td>With sequential updating, VA</td>
<td>-477.7</td>
<td>-470.0</td>
<td>-470.1</td>
<td>-473.3</td>
</tr>
</tbody>
</table>

approximation are shown in Table 4. The time for MCMC is just for the initial fit (based on 10,000 iterations for each of the models with 1000 burn in), but for variational approximation we show the times for both the initial fit and 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 from 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 and 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. Figure 3 shows the estimated lower 1%
Table 4: Computation times (seconds) for LPDS calculations for models with 1, 2, 3 and 4 mixture components in the S&P500 example. Rows 1-3 respectively are times for initial fit for MCMC, initial fit for variational approximation, and initial fit plus sequential updating for validation for variational approximation.

<table>
<thead>
<tr>
<th>Number of mixture components</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial fit MCMC</td>
<td>504</td>
<td>2463</td>
<td>3427</td>
<td>4417</td>
</tr>
<tr>
<td>Initial fit VA</td>
<td>1</td>
<td>739</td>
<td>1022</td>
<td>1442</td>
</tr>
<tr>
<td>Initial fit + validation VA</td>
<td>250</td>
<td>1902</td>
<td>2552</td>
<td>4754</td>
</tr>
</tbody>
</table>

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 variational approximation (dashed lines). One can observe that the MCMC and variational predictive densities are nearly indistinguishable, so that the variational approximation provides excellent predictive inference here.

7 Discussion and conclusions

Our article describes fast variational approximation methods for mixtures of heteroscedastic experts 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 stochastic approximation approach of 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 stochastic approximation 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 stochastic approximation has the potential to broaden the applicability of such an approach. Another interesting ex-
Figure 3: Estimated 5% (solid line) and 1% (dashed line) quantiles of predictive densities based on 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 data set using MCMC (solid line) and variational approximation (dashed line).
tension which we have not pursued for MHE models is to allow some of the coefficients in the expert components to be shared across components. This may be particularly useful in the variance models for the expert components.

Acknowledgements

David Nott was supported by Singapore MOE grant R-155-000-068-133. Siew Li Tan was partly supported as part of the SDWA’s tropical reservoir research programme. The authors also wish to thank John Ormerod for comments and suggestions related to this work. Robert Kohn’s research was partially supported by ARC grant DP0988579. We thank Lucy Marshall for supplying the data for the example in Section 6.1.

References


