Variational Bayesian Analysis for Hidden Markov Models

C.A. McGrory
Queensland University of Technology
Corresponding author: School of Mathematical Sciences, Queensland University of Technology, GPO Box 2434, Brisbane, Queensland, 4001, Australia, Tel. (61)7-3864-1287, Fax.: (61)7-3864-2310 c.mcgrory@qut.edu.au

D.M. Titterington
University of Glasgow
Department of Statistics, University of Glasgow, Glasgow, G12 8QQ, Scotland, UK, Tel.: (44)-330-5022, Fax.:(44)-330-4814, mike@stats.gla.ac.uk

Running title: Variational Bayes for HMMs

Abstract

The variational approach to Bayesian inference enables simultaneous estimation of model parameters and model complexity. An interesting feature of this approach is that it also leads to an automatic choice of model complexity. Empirical results from the analysis of hidden Markov models with Gaussian observation densities illustrate this. If the variational algorithm is initialised with a large number of hidden states, redundant states are eliminated as the method converges to a solution, thereby leading to a selection of the number of hidden states. In addition, through the use of a variational approximation, the Deviance Information Criterion for Bayesian model selection can be extended to the hidden Markov model framework. Calculation of the Deviance Information Criterion provides a further tool for model selection which can be used in conjunction with the variational approach.

Keywords: Hidden Markov model, Variational approximation, Deviance Information Criterion (DIC), Bayesian analysis

1 Introduction

Markov models are a valuable tool for modelling data that vary over time and can be thought of as having been generated by a process that switches between different phases or states at different time-points. However, in many situations the particular sequence of states which gave rise to an observation set is unobserved, i.e. the states are 'hidden'. We can imagine that there is a missing set of indicator variables that describe which state gave rise to a particular observation. These missing indicator variables are not independent, but are governed by a stationary Markov chain. This framework represents a hidden Markov Model (HMM). HMMs have found application in a wide range of areas. Examples include speech recognition (the tutorial by Rabiner (1989) provides a good introduction), biometrics problems such as DNA sequence segmentation (see Boys et al. (2004), for example), econometrics (see Chib (1996), for instance) and finance (see Rydén, Teräsvirta & Åsbrink (1998)). For a recent text on the subject of hidden Markov modelling see MacDonald & Zucchini (1997).

Variational Bayes is a computationally efficient deterministic approach to Bayesian inference. The speed and efficiency of the variational approach makes it a valuable alternative to Markov chain Monte Carlo. As such it is gaining popularity in the machine-learning literature but it remains relatively unexplored by the statistics community. In this paper we describe how the variational approximation method can be used to perform Bayesian inference for hidden Markov models (HMMs) with Gaussian observation densities. The resulting algorithm is a modified version of the well-known forward-backward/Baum-Welch algorithm (Baum et al. (1970)). This extends previous research (McGrory & Titterington (2006)) in which we considered how variational methods can be used to perform model-selection automatically for mixtures of Gaussians. Empirical results indicate that using variational methods for model selection in the case of an HMM with Gaussian observation densities also leads to an automatic choice of model complexity. The reader is also referred to MacKay (2001), Attias (1999) and Corduneanu & Bishop (2001) for a discussion of the component or state removal effect connected with using variational Bayes for mixture and HMM analysis.

The variational approximation can also be used to extend the Deviance Information

Criterion (DIC) model selection criterion (Spiegelhalter et al. (2002)) to latent variable models. We show how the DIC can be approximated for an HMM and we use it as a model selection tool together with variational Bayes in our applications.

This paper focuses on performing variational Bayesian inference for HMMs when the number of hidden states is unknown and has to be estimated along with the model parameters. Other approaches to this problem include the computationally intensive classical approach of Rydén, Teräsvirta & Åsbrink (1998) which uses a parametric bootstrap approximation to the limiting distribution of the likelihood ratio. There are also Bayesian approaches such as those presented in Robert, Rydén, & Titterington (2000) and Boys & Henderson (2004) that are based on the Reversible Jump Markov Chain Monte Carlo (RJMCMC) technique (see Green (1995) and Green & Richardson (2002)) for assessing the number of hidden states.

MacKay (1997) was the first to propose applying variational methods to HMMs, considering only the case where the observations are discrete. Despite the lack of understanding of the state-removal phenomenon, variational methods are beginning to be applied to HMMs in the machine learning community. For instance, Lee, Attias & Deng (2003) propose a variational learning algorithm for HMMs applied to continuous speech processing. Variational methods have been shown to be successful in other areas but their full potential in HMM analysis is yet to be explored.

In Section 2 we outline the variational approach to Bayesian inference. Section 3 describes the DIC and how it can be approximated using variational Bayes. In Section 4 we apply the variational Bayes algorithm to an HMM with Gaussian observation densities, Section 5 considers synthetic and real-data applications and Section 6 gives concluding remarks.

2 The Variational Approach to Approximate Bayesian Inference

In this section we review how variational methods can be applied to approximate quantities required for Bayesian inference. We assume a parametric model with parameters

θ, where z denotes latent or unobserved values in the model. In this paper the z will be discrete variables. Given observed data y, Bayesian inference focuses on the posterior distribution $p(\theta|y)$ of θ given y. The posterior distribution $p(\theta|y)$ is the appropriate marginal of $p(\theta, z|y)$. The variational Bayes approach allows us to approximate the complex quantity $p(\theta, z|y)$ by a simpler density, $q(\theta, z)$. The approximating density q that we introduce is obtained by constructing and maximising a lower bound on the observed-data log-likelihood using variational calculus:

$$\begin{aligned} \log p(y) &= \log \int \sum_{\{z\}} q(\theta, z) \frac{p(y, z, \theta)}{q(\theta, z)} d\theta \\ &\geq \int \sum_{\{z\}} q(\theta, z) \log \frac{p(y, z, \theta)}{q(\theta, z)} d\theta, \end{aligned} \tag{1}$$

by Jensen's Inequality.

As a result of the relationship

$$\begin{aligned} \log p(y) &= \int \sum_{\{z\}} q(\theta, z) \log \frac{p(y, z, \theta)}{q(\theta, z)} d\theta + \int \sum_{\{z\}} q(\theta, z) \log \frac{q(\theta, z)}{p(\theta, z|y)} d\theta \\ &= \int \sum_{\{z\}} q(\theta, z) \log \frac{p(y, z, \theta)}{q(\theta, z)} d\theta + KL(q|p), \end{aligned}$$

any q which maximises the lower bound, (1), also minimises the Kullback-Leibler(KL) divergence between q and $p(\theta, z|y)$. The KL divergence is zero when $q(\theta, z) = p(\theta, z|y)$, but to make calculation feasible $q(\theta, z)$ is restricted to have the factorised form $q(\theta, z) = q_\theta(\theta) q_z(z)$. The lower bound, (1), can then be maximised with respect to the variational distributions, resulting in a set of coupled equations for $q_\theta(\theta)$ and $q_z(z)$. The hyperparameters can then be found using an EM-like algorithm.

For an introductory tutorial on variational methods see Jordan et al. (1999) or Jaakkola (2000), for example, and for discussion of some theoretical aspects of the variational Bayes algorithm see Wang & Titterington (2006), who explore its convergence

properties in the context of mixture models.

An engine called VIBES (Variational Inference for BayEsian networkS) has recently been developed for performing variational inference for certain types of model such as mixtures of factor analysers and bayesian state space models. It allows users to input their Bayesian network model in the form of a directed acyclic graph, and it derives and solves the corresponding variational equations. See Winn & Bishop (2005) for a description of the software. A recent addition to the software is code for implementing the variational analysis of HMMs described in the report by MacKay (1997). The framework used in the report deals with discrete observations and does not involve inference for hyperparameters, these are fixed.

3 The Deviance Information Criterion (DIC)

The Deviance Information Criterion, or DIC (Spiegelhalter et al. (2002)), is a model-selection criterion that is based on the premise of trading off Bayesian measures of model complexity and fit. For complex hierarchical models, the DIC provides an useful alternative to the widely used Bayes factor approach as the computation is comparatively straightforward and the number of unknown parameters in the model does not have to be known in advance. As modern applications become more complex these issues are increasingly relevant in statistical inference as is the availability of a selection criterion which is straightforward to calculate.

The initial derivation of the DIC focused on exponential-family models. It has since been extended to deal with incomplete data models by Celeux et al. (2006), using MCMC methods, and also by McGrory & Titterington (2006) using the variational approximation in a mixture model setting. Here, using the variational approximation in conjunction with the forward algorithm, we can extend the criterion further by approximating it within the HMM framework.

The DIC is defined as

$$\text{DIC} = \overline{D(\theta)} + p_D,$$

where $\overline{D(\theta)}$ measures the fit of the model and p_D is the model complexity measure. The above definition is based on the deviance

$$D(\theta) = -2\log p(y|\theta),$$

and $\overline{D(\theta)}$ corresponds to the expectation with respect to $p(\theta|y)$. The complexity measure is defined as

$$\begin{aligned} p_D &= \overline{D(\theta)} - D(\widetilde{\theta}) \\ &= \mathrm{E}_{\theta|y}(-2\log p(y|\theta)) + 2\log p(y|\tilde{\theta}), \end{aligned} \tag{2}$$

where $\widetilde{\theta}$ is the posterior mean of the parameters of interest.

A model which better fits the data will have a larger likelihood and hence a smaller deviance. Model complexity is penalised by the complexity term, p_D, which measures the effective number of parameters in the model. Intuitively then, the model of choice would be the one with the lowest DIC value.

The DIC can easily be re-expressed as

$$\mathrm{DIC} = -2\log p(y|\tilde{\theta}) + 2p_D. \tag{3}$$

In the HMM setting, $\mathrm{E}_{\theta|y}\{-2\log p(y|\theta)\}$, required in the calculation of p_D, has to be approximated. The variational approach leads to the approximation

$$p(y|\theta) = \sum_{\{z\}} p(y,z|\theta) = \sum_{\{z\}} \frac{p(\theta,z|y)p(y)}{p(\theta)} \approx \frac{q_\theta(\theta)p(y)}{p(\theta)}.$$

Substituting this approximation into the formula (2) for p_D gives us the form

$$p_D \approx -2\int q_\theta(\theta)\log\left(\frac{q_\theta(\theta)}{p(\theta)}\right)d\theta + 2\log\left(\frac{q_\theta(\tilde{\theta})}{p(\tilde{\theta})}\right), \tag{4}$$

where $\tilde{\theta}$ is the posterior mean. Then $p(y|\tilde{\theta})$ can be obtained using the forward algorithm

for HMMs for substitution into (3).

4 Variational Bayesian Inference for Gaussian Hidden Markov Models with an Unknown Number of States

A Markov model assumes that a system can be in one of K states at a given time-point i, and at each time-point the system either changes to a different state or stays in the same state. An HMM is a stochastic process generated by a stationary Markov chain whose state sequence cannot be directly observed. Instead, what we actually observe is a distorted version of the state sequence. A discrete first-order Markov model has the property that the probability of occupying a state, z_i, at time i, given all previous states, depends only on the state occupied at the immediately previous time-point. Here we fix the first state in the sequence by setting $z_1 = 1$. The remaining states are not fixed and the probability of moving from one state to another is characterised by a transition matrix

$$\pi = \{\pi_{j_1 j_2}\}, \quad 1 \leq j_1, j_2 \leq K,$$

where $\pi_{j_1 j_2} = p(z_{i+1} = j_2 | z_i = j_1)$, $\pi_{j_1 j_2} \geq 0$ and $\sum_{j_2=1}^{K} \pi_{j_1 j_2} = 1$, for each j_1. Suppose we have n observations, corresponding to n time-points, i.e. data $y_i : i = 1, \ldots, n$ generated by such a Markov process. The probability density for y_i at time-point i, given that the system is in state j, is given by

$$p(y_i | z_i = j) = p_j(y_i | \phi_j),$$

where the $\{\phi_j\}$ are the parameters within the jth observation density. These densities are often called the emission densities. We shall assume that the y_i are univariate, and in fact Gaussian, but other cases can be easily dealt with. The model parameters are given by $\theta = (\pi, \phi)$ with $\phi = \{\phi_j\}$. The prior densities are assumed to satisfy

$$p(\pi, \phi) = p(\pi)p(\phi).$$

Then the joint density of all of the variables is

$$p(y, z, \theta) = \prod_{i=1}^{n} \prod_{j=1}^{K} (p_j(y_i|\phi_j))^{z_{ij}} \prod_{i=1}^{n-1} \prod_{j_1} \prod_{j_2} (\pi_{j_1 j_2})^{z_{ij_1} z_{i+1 j_2}} p(\phi)p(\pi),$$

where z_{ij} indicates which state the chain is in for a given observation and is equal to the Kronecker delta, i.e. $z_{ij} = 1$, if $z_i = j$, and $z_{ij} = 0$, if $z_i \neq j$. The ϕ_js are distinct and we assume prior independence, so that

$$p(\phi) = \prod_{j=1}^{K} p_j(\phi_j).$$

As mentioned in Section 2, we assume that our variational posterior is of the form $q(z, \theta) = q_z(z)q_\theta(\theta)$. We also assume prior independence among the rows of the transition matrix, and therefore $q_\theta(\theta)$ takes the form

$$q_\theta(\theta) = \prod_{j=1}^{K} q_{\phi_j}(\phi_j) \prod_{j_1} q_{j_1}(\pi_{j_1}),$$

where

$$\pi_{j_1} = \{\pi_{j_1 j_2} : j_2 = 1, \ldots, K\}.$$

If $p_j(y_i|\phi_j)$ represents an exponential family model and $p_j(\phi_j)$ is taken to be from an appropriate conjugate family then the optimal variational posterior for ϕ_j will also belong to the conjugate family.

Finding $q_z(z)$ in our variational scheme involves the forward and backward variables from the Baum-Welch procedure (Baum et al. (1970)). The Baum-Welch algorithm removes the computational difficulties attached to likelihood calculation and parameter estimation for HMMs and leads to an expectation-maximization algorithm. The Baum-Welch algorithm has two steps: based on some initial estimates, the first involves calcu-

lating the so-called forward probability and the backward probability for each state (see the Appendix for a description of the forward and backward algortihms), and the second determines the expected frequencies of the paired transitions and emissions. These are obtained by weighting the observed transitions and emissions by the probabilities specified in the current model. These expected frequencies then provide the new estimates, and iterations continue until there is no improvement. The method is guaranteed to converge to at least a local maximum, and estimates of the transition probabilities and parameter values can be obtained. Note that evaluation of the likelihood only involves the forward part of the algorithm. See Rabiner (1989) for a detailed description.

The variational Bayes algorithm for HMMs is in fact a modification of this algorithm. The forward-backward algorithm of the Baum-Welch procedure is used to obtain the variational posterior transition probabilities and estimates of the indicator variables for the states. These estimates can then be used to update the variational posterior estimates for the model parameters. The variational Bayes algorithm is also guaranteed to converge to at least a local maximum.

4.1 Model Specification

Assigning the Prior Distributions

For each state j_1, we assign an independent Dirichlet prior for the transition probabilities $\{\pi_{j_1 j_2} : j_2 = 1, \ldots, K\}$, so that

$$p(\pi) = \prod_{j_1} \mathrm{Dir}(\pi_{j_1} | \{{\alpha_{j_1 j_2}}^{(0)}\}),$$

for given hyperparameters $\{{\alpha_{j_1 j_2}}^{(0)}\}$. We assign univariate Gaussians with unknown means and variances to represent the emission densities $p_j(y_i|\phi_j)$. Therefore,

$$p_j(y_i|\phi_j) = \mathrm{N}(y_i|\mu_j, {\tau_j}^{-1}),$$

where μ_j is the mean, τ_j is the precision and $\phi_j = (\mu_j, \tau_j)$.

The means are assigned independent univariate Gaussian conjugate priors, conditional on the precisions. The precisions themselves are assigned independent Gamma prior distributions so that

$$p(\mu|\tau) = \prod_{j=1}^{K} \mathrm{N}(\mu_j|{m_j}^{(0)}, ({\beta_j}^{(0)}\tau_j)^{-1})$$

and

$$p(\tau) = \prod_{j=1}^{K} \gamma(\tau_j|\frac{1}{2}{\eta_j}^{(0)}, \frac{1}{2}{\delta_j}^{(0)}),$$

where $\mu = (\mu_1, \ldots, \mu_K)$ and $\tau = (\tau_1, \ldots, \tau_K)$ for given hyperparameters $\{m_j^{(0)}, \beta_j^{(0)}, \eta_j^{(0)}, \delta_j^{(0)}\}$.

Form of the Variational Posterior Distributions

The variational posteriors for the model parameters turn out to have the following forms:

$$q_{j_1}(\pi_{j_1}) = \mathrm{Dir}(\pi_{j_1}|\{\alpha_{j_1 j_2}\}),$$

where

$$\alpha_{j_1 j_2} = {\alpha_{j_1 j_2}}^{(0)} + \sum_{i=1}^{n-1} q_z(z_i = j_1, z_{i+1} = j_2);$$

$$q(\mu_j|\tau_j) = \mathrm{N}(\mu_j|m_j, (\beta_j\tau_j)^{-1})$$

and

$$q(\tau_j) = \gamma(\tau_j|\frac{1}{2}\eta_j, \frac{1}{2}\delta_j),$$

with hyperparameters given by

$$\beta_j = {\beta_j}^{(0)} + \sum_{i=1}^{n} q_{ij}$$

$$\eta_j = \eta^{(0)} + \sum_{i=1}^{n} q_{ij}$$

$$\delta_j = \delta^{(0)} + \sum_{i=1}^{n} q_{ij} {y_i}^2 + {\beta_j}^{(0)} {m_j}^{(0)^2} - \beta_j m_j^2$$

$$m_j = \frac{{\beta_j}^{(0)} {m_j}^{(0)} + \sum_{i=1}^{n} q_{ij} y_i}{\beta_j},$$

where $q_{ij} = q_z(z_i = j)$. For each of the j states, $\sum_{i=1}^{n} q_{ij}$ provides a 'weighting' in the form of a 'pseudo' number of observations associated with that state.

The variational posterior for $q_z(z)$ will have the form

$$q_z(z) \propto \prod_i \prod_j {b_{ij}^*}^{z_{ij}} \prod_i \prod_{j_1} \prod_{j_2} {a_{j_1 j_2}^*}^{z_{ij_1} z_{i+1 j_2}},$$

for certain $\{a_{j_1 j_2}^*\}$ and $\{b_{ij}^*\}$. This is the form of a conditional distribution of the states of a hidden Markov chain, given the observed data. From this we need the marginal probabilities

$$q_z(z_i = j),$$

$$q_z(z_i = j_1, z_{i+1} = j_2).$$

These can be obtained by the forward-backward algorithm (see the Appendix for details), based on a^* and b^* quantities given by

$$a_{j_1 j_2}^* = \exp\left(\mathrm{E}_q(\log \pi_{j_1 j_2})\right) = \exp\left(\Psi(\alpha_{j_1 j_2}) - \Psi(\sum_{j=1}^{K} \alpha_{j_1 j})\right),$$

$$b_{ij}^* = \exp\left(\mathrm{E}_q(\log p_j(y_i|\phi_j))\right),$$

where Ψ is the digamma function and

$$\mathrm{E}_q\left(\log p_j(y_i|\phi_j)\right) = \frac{1}{2}\Psi(\frac{1}{2}\eta_j) - \frac{1}{2}\log\frac{\delta_j}{2} - \frac{1}{2}(\frac{\eta_j}{\delta_j})(y_i - m_j)^2 - \frac{1}{2\beta_j}.$$

Here $a^*_{j_1 j_2}$ is an estimate of the probability of transition from state j_1 to j_2 and b^*_{ij} is an estimate of the emission probability density given that the system is in state j at time point i.

One can obtain $q_z(z_i = j)$ and $q_z(z_i = j_1, z_{i+1} = j_2)$ from the following formulae based on the forward and backward variables, which we denote by fvar and bvar, respectively, and which are defined in the Appendix:

$$\begin{aligned} q_z(z_i = j) = p(z_i = j_1|y_1, \ldots, y_n) &\propto \mathrm{fvar}_i(j_1)\mathrm{bvar}_i(j_1) \\ &= \frac{\mathrm{fvar}_i(j_1)\mathrm{bvar}_i(j_1)}{\sum_{j_2}\mathrm{fvar}_i(j_2)\mathrm{bvar}_i(j_2)} \end{aligned}$$

$$\begin{aligned} q_z(z_i = j_1, z_{i+1} = j_2) &\propto \mathrm{fvar}_i(j_1)a^*_{j_1 j_2}b^*_{i+1 j_2}\mathrm{bvar}_{i+1}(j_2) \\ &= \frac{\mathrm{fvar}_i(j_1)a^*_{j_1 j_2}b^*_{i+1 j_2}\mathrm{bvar}_{i+1}(j_2)}{\sum_{j_1}\sum_{j_2}\mathrm{fvar}_i(j_1)a^*_{j_1 j_2}b^*_{i+1 j_2}\mathrm{bvar}_{i+1}(j_2)}. \end{aligned}$$

Variational approximation to p_D and the DIC

Our variational approximation to p_D is

$$p_D \approx -2\int q_\theta(\theta)\log\left(\frac{q_\theta(\theta)}{p(\theta)}\right)d\theta + 2\log\left(\frac{q_\theta(\tilde{\theta})}{p(\tilde{\theta})}\right)$$

$$
\begin{aligned}
= \; - \; & 2\left(\sum_{j_1}\sum_{j_2}\left(\sum_{i=1}^{n-1} q_z(z_i = j_1, z_{i+1} = j_2)\right)\left(\Psi\left(\alpha_{j_1,j_2}\right) - \Psi\left(\alpha_{j_1\cdot}\right)\right)\right. \\
& \qquad \left. + \sum_{j=1}^{K}\sum_{i=1}^{n} q_{ij}\left(\frac{1}{2}\left(\Psi(\frac{1}{2}\eta_j) - \log\frac{\delta_j}{2}\right) - \frac{1}{2\beta_j}\right)\right) \\
+ \; & 2\left(\sum_{j_1}\sum_{j_2}\left(\sum_{i=1}^{n-1} q_z(z_i = j_1, z_{i+1} = j_2)\right)\log\left(\frac{\alpha_{j_1 j_2}}{\sum_{j_2}\alpha_{j_1 j_2}}\right) + \frac{1}{2}\sum_{j=1}^{K}\sum_{i=1}^{n} q_{ij}\log\left(\frac{\eta_j}{\delta_j}\right)\right).
\end{aligned}
$$

The DIC value can then be found via the usual formula,

$$\mathrm{DIC} = 2p_D - 2\log p(y|\tilde{\theta}),$$

in which $p(y|\tilde{\theta})$ is found by summing over the final forward variable in the forward algorithm:

$$p(y|\tilde{\theta}) = \sum_{j=1}^{K} \mathrm{fvar}_n(j).$$

5 Application to Simulated and Real Data Sets

The variational Bayes algorithm is initialised with a larger number of hidden states than one would reasonably expect to find in each of our applications. To initialise the algorithm, the missing indicators z are randomly allocated to correspond one of the initial states. We set all of the $\{{\alpha_{j_1 j_2}}^{(0)}\}$s to be 1 in the Dirichlet prior. This is equivalent to a multivariate uniform prior for the mixing weights. The remaining hyperparameters were also chosen to correspond to non-informative prior distributions. In some cases, as the algorithm progresses, the weighting of one state will dominate those of others whose noise models are similar, causing the latter's weightings to tend towards zero. When a state's weighting becomes sufficiently small, we consider less than 1 observation assigned to the state as sufficient, it is removed from consideration and the algorithm continues

with the remaining states. The algorithm converges to a solution involving a number of states smaller than or equal to the initial number. The DIC and p_D values are also computed at each iteration.

Using variational methods to make inference about an HMM with Gaussian noise leads to an automatic choice of model complexity since a feature of the algorithm is that superfluous states are removed during convergence to the solution. It is also possible to force the algorithm to converge to a solution with fewer states than the number selected automatically by initialising the algorithm with a smaller number of states. In these situations one might use the DIC value to select a suitable model.

5.1 Application to a Simulated Example

We simulated 3 datasets comprising 150, 500 and 1000 observations, respectively from a 3-state HMM with transition matrix

$$\pi = \begin{bmatrix} 0.15 & 0.8 & 0.05 \\ 0.5 & 0.1 & 0.4 \\ 0.3 & 0.4 & 0.3 \end{bmatrix}.$$

The Gaussian noise distributions had means of 1, 2 and 3, and standard deviations of 0.25, 0.1 and 0.7, respectively.

We explored the results of a variational analysis with several different numbers of initial states. We also considered how our results changed with the number of observations available. Tables 1-3 report the variational estimates of the posterior means and standard deviations obtained in this way. These tables also show the corresponding estimated DIC and p_D values.

For the 150 and 500-observation datasets we were successfully able to recover a 3-state solution with good posterior estimates of model parameters and transition probabilities for every number of initial states. For the largest dataset (1000 observations), this was the case in all but one of the initial number of states chosen.

It can be seen from the results that, in general, increasing the number of observations in the sample leads to posterior estimates that are closer to the true parameters of

the HMM from which they were simulated. This is to be expected from any inference algorithm. Correspondingly, the better estimates obtained from the larger datasets lead to lower DIC values for the model. This suggests that the DIC is a useful comparison criterion in this setting.

For the smallest data set (150 observations), the number of initial states chosen did not affect the resulting posterior. However, for the 500-observation dataset, the resulting estimates were closer to the true parameters when the initial number of states used was 5 than when it was higher than this. The solution obtained by starting with only 5 initial states also led to a smaller DIC value. This lends support to the assertion that this was a better fit.

These results suggest to us that the initial number of states can affect how observations are classified as the the algorithm converges. Interestingly, in our results this effect was more pronounced when the number of observations was higher. Initialising the analysis of the 1000-observation dataset with 20 states lead to a 4-state solution. However, one of these four states only had 5 observations assigned to it. This suggests that if the initial number of states is too large, some of the observations can become 'stuck' in their initial allocation. Intuitively, we would expect that the larger the dataset, the more opportunity there is for that to happen as there are more observations available to lend support to a superfluous state in the model. It also takes longer for any superfluous states to be removed. With this in mind, it seems that perhaps some caution has be exercised when using an excessively large number of initial states, particularly if the observation set is reasonably large.

5.2 Application to Real Datasets

The two datasets used in this section were analysed by Robert, Rydén, & Titterington (2000) using RJMCMC and have also previously been analysed by other authors. We analyse these datasets using our variational approach and compare our results with other treatments of the data. We used uninformative priors here as we did in the simulation study. We initialised the variational algorithm with 7 states for each application as it did not seem feasible that the number of states would be any larger than this.

Geomagnetic Data

The first dataset is made up of 2700 residuals from a fit of an autoregressive moving average model to a planetary geomagnetic activity index. This dataset is analysed in the paper by Francq & Roussignol (1997) using maximum likelihood techniques. In the context of geomagnetic data, the pattern changes in the residuals of a time series analysis often have useful interpretations.

Daily Returns Data

The second dataset is an extract from the Standard and Poors 500 stock index consisting of 1700 observations of daily returns from the 1950s. It was previously analysed using a computationally intensive classical approach by Rydén, Teräsvirta & Åsbrink (1998) and was the dataset referred to as subseries E in their paper.

For the geomagnetic data, the variational algorithm fitted a 2-state model. The estimated DIC value decreased as the algorithm converged. Here we describe the 2-state model fitted by the variational algorithm. The variational posterior transition matrix is

$$\begin{bmatrix} 0.982 & 0.018 \\ 0.187 & 0.813 \end{bmatrix}$$

and the variational posterior estimates are given in Table 4. The fitted density is plotted in Figure 1.

The analysis by Robert, Rydén, & Titterington (2000) resulted in a 3-state model for these data while Francq & Roussignol's (1997) analysis selected a 2-state model as we did. With this dataset the posterior estimates we obtain for the transition probabilities and the state density standard deviations for our 2-state model are similar to those found by Francq & Roussignol (1997) whose estimated parameters were $\pi_{12} = 0.014$, $\pi_{21} = 0.16$, $\sigma_1 = 2.034$ and $\sigma_2 = 5.840$. Francq & Roussignol (1997) suggest that this two-state model corresponds to tumultuous and quiet states, the tumultuous state being the one with the higher variability. Since their model visits tumultuous states less frequently than it does quiet states, and spends less time in them, they propose that these tumultuous states might correspond to geomagnetic storms. As we obtained a fit similar to this,

the variational solution can be interpreted similarly in this application. Therefore, our variational posterior solution seems plausible in this context.

For the daily returns data, the variational algorithm fitted a 2-state solution which had a variational posterior transition matrix given by

$$\begin{bmatrix} 0.96 & 0.04 \\ 0.07 & 0.93 \end{bmatrix}$$

and variational posterior estimates given in Table 4. The fitted density is plotted in Figure 2.

This solution shows similarities to that of Robert, Rydén, & Titterington (2000), whose analysis favoured 2 or 3 states, as well as that of Rydén, Teräsvirta & Åsbrink (1998), who fitted a 2-state model. Therefore, the variational 2-state posterior is consistent with previous analyses. In the analysis by Robert, Rydén, & Titterington (2000), the estimated transition probabilities for the 2-state model were $\pi_{12} = 0.044$ and $\pi_{21} = 0.083$, and the estimated posterior standard deviations were $\sigma_1 = 0.0046$ and $\sigma_2 = 0.0093$. These were similar to the estimates found by Rydén, Teräsvirta & Åsbrink (1998); their estimates for the transition probabilities were $\pi_{12} = 0.037$ and $\pi_{21} = 0.069$, and their estimated posterior standard deviations were $\sigma_1 = 0.0046$ and $\sigma_2 = 0.0092$. For this application, the variational, RJMCMC and classical analyses have all produced comparable results.

These applications have demonstrated that a variational scheme can produce posterior estimates that are similar to those obtained through RJMCMC and computationally demanding classical techniques. In addition, variational Bayes has the advantage of being fast to implement. Since this is a highly desirable feature for many practical applications, it would be a useful alternative to existing methods in many contexts.

6 Conclusions

We have seen that applying variational methods in the case of a hidden Markov model with Gaussian noise leads to the automatic removal of components and therefore leads to

an automatic choice of model complexity. Solutions with fewer states than the number automatically selected can be obtained by initialising the algorithm with a number of states smaller than the number obtained automatically. The variational approximation also makes the calculation of the DIC possible and this can be used to choose between competing models.

We have shown that the Variational Bayes approach for HMMs produces good posterior estimates of parameters and can be used when the number of hidden states in the model is unknown. The algorithm is also very fast, making it an attractive option for complex applications. Variational methods have considerable potential for the analysis of HMMs, but there is much scope for further investigation into the state-removal phenomenon which occurs in the implementation.

Acknowledgements

C.A. McGrory wishes to acknowledge the support of a UK Engineering and Physical Sciences Research Council research studentship. This research was also supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778.

We wish to thank two anonymous referees for their helpful comments on an earlier version of this manuscript.

References

ATTIAS, H. (1999). Inferring parameters and structure of latent variable models by variational Bayes. *Proc.* 15*th Conf. on Uncertainty in Artificial Intelligence.*

BAUM, L.E., PETRIE, T., SOULES, G. & WEISS, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Stat.*, **41**, 164-171.

BOYS, R.J. & HENDERSON (2004). A Bayesian approach DNA sequence segmentation (with Discussion). *Biometrics*, **60**, 573-588.

CELEUX, G., FORBES, F., ROBERT, C. & TITTERINGTON, D.M. (2006). Deviance Information Criteria for missing data models (with Discussion). *Bayesian Anal.*, **1**, 651-674.

CHIB, S. (1996). Calculating posterior distributions and modal estimates in Markov mixture models. *J. Econometr.*, **75**, 79-97.

CORDUNEANU, A. & BISHOP, C.M. (2001). Variational Bayesian model selection for mixture distributions. In *Artificial Intelligence and Statistics* (T. Jaakkola & T. Richardson, eds.), pp.27-34. Morgan Kaufmann.

FRANCQ, C. & ROUSSIGNOL, M. (1997). On white noise driven by hidden Markov chains. *J. Time Series Anal.*, **18**, 553-578.

GREEN, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711-732.

GREEN, P.J. & RICHARDSON, S. (2002). Hidden Markov models and disease mapping. *J. Am. Statist. Assoc.*, **97**, 1055-1070.

JAAKKOLA, T.S. (2000). Tutorial on variational approximation methods. In *Advanced Mean Field Methods* (M. Opper & D. Saad, eds.) pp. 129-159. MIT Press, Cambridge, MA.

JORDAN, M. I., GHAHRAMANI, Z, JAAKKOLA, T. S. & Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, **37**, 183-233.

LEE, L., ATTIAS, H. & DENG, L. (2003). Variational inference and learning for segmental switching state space models of hidden speech dynamics. *Proc. of the Intl. Conf. on Acoustics, Speech and Signal Processing, Hong Kong, Apr, 2003.*

MACDONALD, I.L. & ZUCCHINI, W. (1997). *Hidden Markov Models and Other Models for Discrete-valued Time Series.* Chapman & Hall, London.

MACKAY, D.J.C. (1997). Ensemble learning for hidden Markov models. Technical Report, Cavendish Laboratory, University of Cambridge.

MACKAY, D.J.C. (2001). Local minima, symmetry-breaking, and model pruning in variational free energy minimization. Available from:
http://www.inference.phy.cam.ac.uk/mackay/minima.pdf.

MCGRORY, C.A. & TITTERINGTON, D.M. (2007). Variational Approximations in Bayesian Model Selection for Finite Mixture Distributions. *Comp. Statist. Data Anal.*, **51**, 5352-5367.

RABINER, L.R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. of the IEEE*, **77**, 257-284.

ROBERT, C.P., RYDÉN, T. & TITTERINGTON, D.M. (2000). Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method. *J. Roy. Statist. Soc. Ser. B*, **62**, 57-75.

RYDÉN, T., TERÄSVIRTA, T. & ÅSBRINK, S. (1998). Stylized facts of daily return series and the hidden Markov model. *J. Appl. Econometr.* **13**, 217-244.

SPIEGELHALTER, D.J., BEST, N.G., CARLIN, B.P. and VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit (with discussion). *J. Roy. Statist. Soc. Ser. B*, **64**, 583-639.

WANG, B. & TITTERINGTON, D.M. (2006). Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model. *Bayesian Anal.*, **1**, 625-650.

WINN, J. & BISHOP, C.M. (2005). Variational message passing. *J. Mach. Learn. Res.*, **6**, 661-694.

Appendix

The Forward-Backward Algorithm

The forward algorithm calculates the probability of being in state j at time i and the partial observation sequence up until time i, given the model. The forward variable is given by $\text{fvar}_i(j_1) = p(y_1, y_2, \ldots, y_i, z_i = j_1)$ and the algorithm proceeds as follows.

1. Calculate $\text{fvar}_1(j_1) = \pi_{j_1} p(y_1|z_1 = j_1) \quad \text{for} \quad j_1 \quad \text{such that} \quad 1 \leq j_1 \leq K$, and then normalise such that $\sum_{j_1=1}^{K} \text{fvar}_1(j_1) = 1$, i.e. define

$$\widetilde{\text{fvar}}_1(j_1) = \frac{\text{fvar}_1(j_1)}{\sum_{j_1=1}^{K} \text{fvar}_1(j_1)}.$$

2. For $i = 1, \ldots, n-1$ and each j_2, calculate

$$\text{fvar}^*_{i+1}(j_2) = \{\sum_{j_1=1}^{K} \widetilde{\text{fvar}}_i(j_1) p(z_{i+1} = j_2|z_i = j_1)\} p(y_{i+1}|z_{i+1} = j_2).$$

   We then normalise once again, giving

$$\widetilde{\text{fvar}}_i(j_1) = \frac{\text{fvar}^*_i(j_1)}{\sum_{j_1=1}^{K} \text{fvar}^*_i(j_1)}.$$

3. We finally have

$$p(y_1, \ldots, y_n) = \sum_{j_1} \text{fvar}_n(j_1) = \frac{1}{c_n} \sum_{j_1=1}^{K} \widetilde{\text{fvar}}_n(j_1) = \frac{1}{c_n},$$

   since $\sum_{j_1=1}^{K} \widetilde{\text{fvar}}_n(j_1) = 1$, and where c_n is the normalising constant fvar is multiplied by at the nth iteration.

We can calculate the nth normalising constant, c_n, since one can obtain c_{i+1} from c_i in the following way:

$$\begin{aligned}
\widetilde{\text{fvar}}_{i+1}(j_2) &= \frac{\{\sum_{j_1=1}^{K} \widetilde{\text{fvar}}_i(j_1) p(z_{i+1}=j_2|z_i=j_1)\} p(y_{i+1}|z_{i+1}=j_2)}{\sum_{j_2=1}^{K}\{\sum_{j_1=1}^{K} \widetilde{\text{fvar}}_i(j_1) p(z_{i+1}=j_2|z_i=j_1)\} p(y_{i+1}|z_{i+1}=j_2)} \\
&= \frac{c_i\{\sum_{j_1=1}^{K} \text{fvar}_i(j_1) p(z_{i+1}=j_2|z_i=j_1)\} p(y_{i+1}|z_{i+1}=j_2)}{\sum_{j_2=1}^{K}\{\sum_{j_1=1}^{K} \widetilde{\text{fvar}}_i(j_1) p(z_{i+1}=j_2|z_i=j_1)\} p(y_{i+1}|z_{i+1}=j_2)} \\
&= \frac{c_i}{d_i}\text{fvar}_{i+1}(j_2),
\end{aligned}$$

where

$$d_i = \sum_{j_2=1}^{K}\{\sum_{j_1=1}^{K} \widetilde{\text{fvar}}_i(j_1) p(z_{i+1}=j_2|z_i=j_1)\} p(y_{i+1}|z_{i+1}=j_2).$$

Thus,

$$c_{i+1} = \frac{c_i}{d_i}.$$

The backward algorithm works back from the final time-point, n. The backward variable is given by $\text{bvar}_i(j_1) = p(y_{i+1}, y_{i+2}, \ldots, y_n|z_i = j_1)$, i.e. the probability of generating the last $n-i$ observations given state j at time i. The recursive algorithm is as follows.

1. Set $\text{bvar}_n(j_1) = 1, \quad \text{for all} \quad j_1$, and normalise such that $\sum_{j_1=1}^{K} \text{bvar}_n(j_1) = 1$, i.e.

$$\widetilde{\text{bvar}}_n(j_1) = \frac{\text{bvar}_n(j_1)}{\sum_{j_1=1}^{K} \text{bvar}_n(j_1)}.$$

2. For $i = n-1, n-2, \ldots, 1$,

$$\text{bvar}_i^*(j_1) = \sum_{j_2} p(z_{i+1}=j_2|z_i=j_1)\widetilde{\text{bvar}}_{i+1}(j_2) p(y_{i+1}|z_{i+1}=j_2).$$

We normalise again, giving

$$\widetilde{\text{bvar}}_i(j_1) = \frac{\text{bvar}_i^*(j_1)}{\sum_{j_1=1}^{K} \text{bvar}_i^*(j_1)}.$$

In the above algorithms, for $p(z_{i+1} = j_2 | z_i = j_1)$ we use the quantity $a^*_{j_1 j_2}$ and for $p(y_{i+1} | z_{i+1} = j_2)$ we use the quantity $b^*_{i+1 j_2}$.

List of Figures

Table 1: Results for the simulated 150-observation dataset

No. of Initial States	No. of States Found	Estimated Posterior Means	Estimated Posterior st. dev.	p_D	DIC
20	3	1.05, 2.02, 2.86	0.25, 0.12 , 0.52	11.051	-82.67
15	3	1.05, 2.02, 2.86	0.25, 0.12 , 0.52	11.051	-82.67
10	3	1.05, 2.02, 2.86	0.25, 0.12 , 0.52	11.051	-82.67
5	3	1.05, 2.02, 2.86	0.25, 0.12 , 0.52	11.051	-82.67

Table 2: Results for the simulated 500-observation dataset

No. of Initial States	No. of States Found	Estimated Posterior Means	Estimated Posterior st. dev.	p_D	DIC
20	3	1.01, 1.99, 2.83	0.25, 0.09, 0.83	11.01	-162.92
15	3	1.01, 1.99, 2.83	0.25, 0.09, 0.83	11.01	-162.92
10	3	1.01, 1.99, 2.83	0.25, 0.09, 0.83	11.01	-162.92
5	3	1.02, 1.99, 3.10	0.25 , 0.09, 0.69	12.01	-192.33

Table 3: Results for the simulated 1000-observation dataset

No. of Initial States	No. of States Found	Estimated Posterior Means	Estimated Posterior st. dev.	p_D	DIC
20	4	1.00, 2.00, 2.84, 3.21	0.25, 0.11, 0.72, 0.43	18.75	-388.37
15	3	1.00, 2.00, 2.89	0.25, 0.10, 0.69	12.00	-401.00
10	3	1.00, 2.00, 2.89	0.25, 0.10, 0.69	12.00	-401.00
5	3	1.00, 2.00, 2.89	0.25, 0.10, 0.69	12.00	-401.00

Table 4: Estimated posterior parameters for the geomagnetic and daily returns datasets

	Dataset	
	Geomagnetic	Daily Returns
Posterior means	-0.209, 1.769	0.00084, -0.00145
Posterior standard deviations	1.997, 5.408	0.00453, 0.00898
Posterior weights	0.911, 0.089	0.63, 0.37

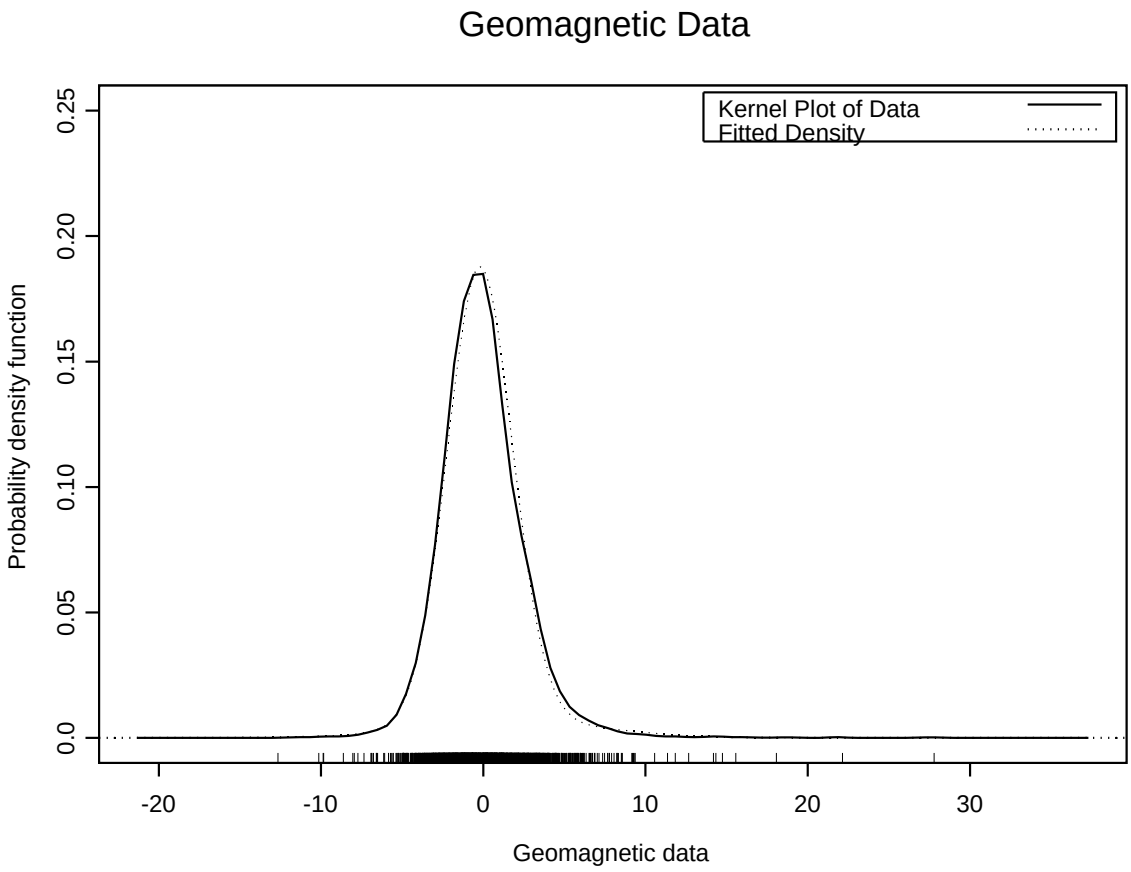
Geomagnetic Data
Kernel Plot of Data
Fitted Density
Probability density function
0.0
0.05
0.10
0.15
0.20
0.25
-20
-10
0
10
20
30
Geomagnetic data

Figure 1:

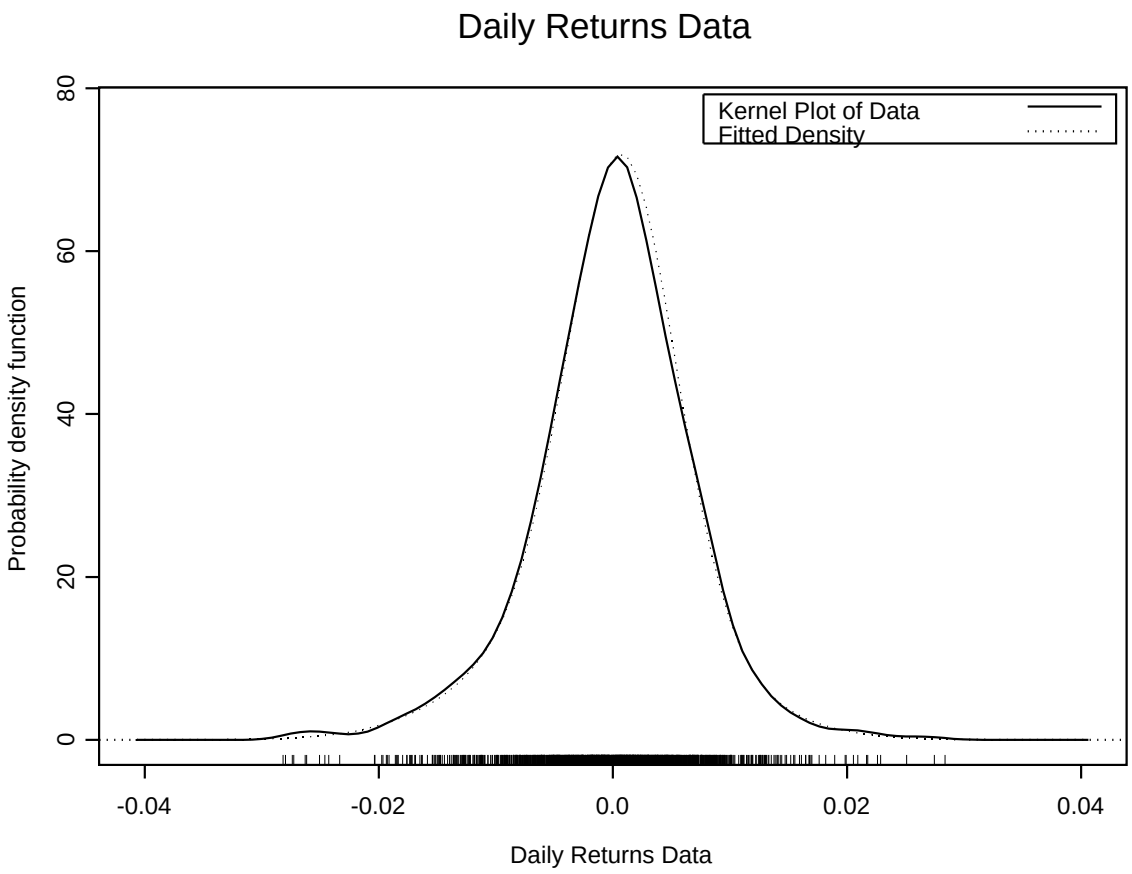
Daily Returns Data
Kernel Plot of Data
Fitted Density
Probability density function
0
20
40
60
80
-0.04
-0.02
0.0
0.02
0.04
Daily Returns Data

Figure 2: