

Variational Bayesian Analysis for Hidden Markov Models

C.A. McGrory ¹

Queensland University of Technology

D.M. Titterton ²

University of Glasgow

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 appears also to lead 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 an automatic selection of the number of hidden states. In addition, through the use of a variational approximation, the Deviance Information Criterion (DIC) for Bayesian model selection can be extended to the hidden Markov model framework. Calculation of the DIC 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

¹Corresponding 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

²Department of Statistics, University of Glasgow, Glasgow, G12 8QQ, Scotland, UK, Tel.: (44)-330-5022, Fax.:(44)-330-4814, mike@stats.gla.ac.uk

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 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 et al. (1998)). For a recent text on the subject of hidden Markov modelling see MacDonald and 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 and Titterton (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 a 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 and 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 a 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 et al. (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 et al. (2000) and Boys and Henderson (2004) that are based on the Reversible Jump Markov Chain Monte Carlo (RJMCMC) technique (see Green (1995) and Richardson and Green (1997)) for assessing the number of hidden states.

MacKay (1997) was the first to propose applying variational methods to HMMs, considering only the case when 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 et al. (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 a 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. 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 and Titterton (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. 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 and Bishop (2005) for a description of the software.

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 a 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 McGroary and Titterton (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

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

where $\tilde{\theta}$ is the posterior mean or mode 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

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

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

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

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

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

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. A 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 $y_i : i = 1, \dots, n$ 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 j^{th} observation density. These probabilities 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

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

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_j 's 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

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

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 which removes the computational difficulties attached to likelihood calculation and parameter estimation for HMMs. The variational Bayes algorithm for HMMs is in fact a modification of this algorithm.

4.1 Model Specification

Assigning the Prior Distributions

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

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

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,

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

where τ_j is the precision and is equal to $\frac{1}{\sigma_j^2}$, 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

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

and

$$p(\boldsymbol{\tau}) = \prod_{j=1}^K \text{Ga}(\tau_j | \frac{1}{2} \gamma_j^{(0)}, \frac{1}{2} \delta_j^{(0)}),$$

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

Form of the Variational Posterior Distributions

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

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

where

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

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

and

$$q(\tau_j) = \text{Ga}(\tau_j | \frac{1}{2} \gamma_j, \frac{1}{2} \delta_j),$$

with hyperparameters given by

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

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

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

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

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

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

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

$$a_{j_1 j_2}^* = \exp\{\langle \log \pi_{j_1 j_2} \rangle\} = \exp\{\Psi(\alpha_{j_1 j_2}) - \Psi(\sum_{j=1}^K \alpha_{j_1 j})\},$$

$$b_{ij}^* = \exp\{\langle \log p_j(y_i | \phi_j) \rangle\},$$

where Ψ is the digamma function. The notation ‘ $\langle \cdot \rangle$ ’ is used to denote expectation with respect to the variational distribution q and

$$\langle \log p_j(y_i | \phi_j) \rangle = \frac{1}{2} \Psi\left(\frac{1}{2} \gamma_j\right) - \frac{1}{2} \log \frac{\delta_j}{2} - \frac{1}{2} \left(\frac{\gamma_j}{\delta_j}\right) (y_i - m_j)^2 - \frac{1}{2\beta_j}.$$

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:

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

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

Variational approximation to p_D and the DIC

Our variational approximation to p_D is

$$\begin{aligned}
p_D &\approx -2 \int q_\theta(\theta) \log \left\{ \frac{q_\theta(\theta)}{p(\theta)} \right\} d\theta + 2 \log \left\{ \frac{q_\theta(\tilde{\theta})}{p(\tilde{\theta})} \right\} \\
&= -2 \left[\sum_{j_1} \sum_{j_2} \left\{ \sum_{i=1}^{n-1} q_z(z_i = j_1, z_{i+1} = j_2) \right\} \{ \Psi(\alpha_{j_1, j_2}) - \Psi(\alpha_{j_1, \cdot}) \} \right. \\
&\quad \left. + \sum_{j=1}^K \left(\sum_{i=1}^n q_{ij} \right) \left\{ \frac{1}{2} \{ \Psi(\frac{1}{2} \gamma_j) - \log \frac{\delta_j}{2} \} - \frac{1}{2\beta_j} \right\} \right] \\
&\quad + 2 \left[\sum_{j_1} \sum_{j_2} \left\{ \sum_{i=1}^{n-1} q_z(z_i = j_1, z_{i+1} = j_2) \right\} \log \left(\frac{\alpha_{j_1 j_2}}{\sum_{j_2} \alpha_{j_1 j_2}} \right) + \frac{1}{2} \sum_{j=1}^K \left(\sum_{i=1}^n q_{ij} \right) \log \left(\frac{\gamma_j}{\delta_j} \right) \right].
\end{aligned}$$

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

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

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

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

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. The hyperparameters are chosen to correspond to broad 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 a HMM with Gaussian noise leads to an automatic choice of model complexity. We found that in some cases it was necessary to place a lower bound on values of a^* to ensure convergence for certain numbers of initial states. For the synthetic data we considered this corresponded to instances where the algorithm was initialised with too many states. 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 500 observations from a 4-state HMM with transition matrix

$$\pi = \begin{bmatrix} 0.2 & 0.2 & 0.3 & 0.3 \\ 0.3 & 0.2 & 0.2 & 0.3 \\ 0.2 & 0.3 & 0.3 & 0.2 \\ 0.3 & 0.3 & 0.2 & 0.2 \end{bmatrix}$$

and Gaussian noise distributions with means -1.5, 0, 1.5 and 3, and equal standard deviations of 0.25.

Without a lower bound on the values of a^* the variational algorithm would only converge to a solution when initialised with the true number of states, 4, or a number smaller than this. By bounding the a^* values we can force the algorithm to converge when initialised with any number of states. Table 1 reports the variational estimates of the posterior means and standard deviations obtained in this way. Table 1 also shows the relevant DIC and p_D values.

We were able to recover a 4-state solution with good posterior estimates for the initial number of states ranging from 4-6. For 7 initial states, the solution we obtained had 5

states, with two of the states having noise models that are close together. One of these noise models had an estimated posterior weight that was much smaller than the other, but not small enough to be removed automatically at our cut-off level. Here the DIC is higher when the number of states fitted is incorrect, so that the DIC appears to be useful in indicating the appropriateness of the fitted model. For the recovered 4-state solution, the estimated posterior transition matrix was

$$\begin{bmatrix} 0.26 & 0.21 & 0.24 & 0.29 \\ 0.34 & 0.20 & 0.18 & 0.28 \\ 0.14 & 0.34 & 0.29 & 0.23 \\ 0.25 & 0.34 & 0.25 & 0.16 \end{bmatrix}.$$

Table 1: Results for the 500-observation dataset with a lower bound on a^* values

No. of Initial States	No. of States Found	Estimated Posterior Means	Estimated Posterior st. dev.	p_D	DIC
7	5	-1.47, -0.01, 0.04, 1.53, 3	0.26, 0.24, 0.22, 0.23, 0.27	35.09	540
6	4	-1.47, -0.005, 1.5, 3	0.26, 0.24, 0.23, 0.27	20.03	513
5	4	-1.47, -0.005, 1.5, 3	0.26, 0.24, 0.23, 0.27	20.03	513
4	4	-1.47, -0.005, 1.5, 3	0.26, 0.24, 0.23, 0.27	20.03	513
3	3	-1.5, 0.68, 3	0.23, 0.97, 0.24	12.00	729
2	2	-0.67, 2.24	0.85, 0.85	5.99	936
1	1	0.7	1.68	1.99	1025

5.2 Application to Real Datasets

The three datasets used in this section were analysed by Robert et al. (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.

Wind Data

This dataset comprises a series of 500 hourly wind velocity measurements taken in Athens in January 1990. This dataset has had a small uniform disturbance added to avoid exact zeroes. These data were also analysed by Francq and Roussignol (1997) (without the added disturbance).

Geomagnetic Data

The second dataset is made up of 2700 residuals from a fit of an autoregressive moving average model to a planetary geomagnetic activity index. Francq and Roussignol (1997) also analysed this dataset.

Daily Returns Data

The third 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 by Rydén et al. (1998) and was the dataset referred to as subseries E in their paper.

Our variational analysis of the wind dataset, with 7 initial states, resulted in the selection of a posterior solution with 2 states. By placing a lower bound on the a^* values we could force the variational algorithm to converge to a solution with more than 2 states. However, here the DIC was also lowest for the 2-state solution and therefore a 2-state solution seemed the most appropriate. The solution has variational posterior transition matrix

$$\begin{bmatrix} 0.955 & 0.045 \\ 0.074 & 0.926 \end{bmatrix}$$

and variational posterior estimates given in Table 2. The fitted density is plotted in Figure 1.

Robert et al.(2000)'s method selected a 3-state model, and Francq and Roussignol's (1997) analysis resulted in a 2-state model for the wind dataset. The posterior estimates we obtain for the transition probabilities and the state density standard deviations for our 2-state model resemble those found in the analysis by Francq and Roussignol (1997), who estimated the parameters as $\pi_{12} = 0.048$, $\pi_{21} = 0.079$, $\sigma_1 = 0.437$ and $\sigma_2 = 2.393$. They

Table 2: Estimated posterior parameters for the wind, geomagnetic and daily returns datasets

	Dataset		
	Wind	Geomagnetic	Daily Returns
Posterior means	-0.015, 0.0069	-0.209, 1.769	0.00084, -0.00145
Posterior standard deviations	0.43, 2.37	1.997, 5.408	0.00453, 0.00898
Posterior weights	0.62, 0.38	0.911, 0.089	0.63, 0.37

suggest that their two-state model distinguishes between periods having fairly constant wind and periods with gusty wind.

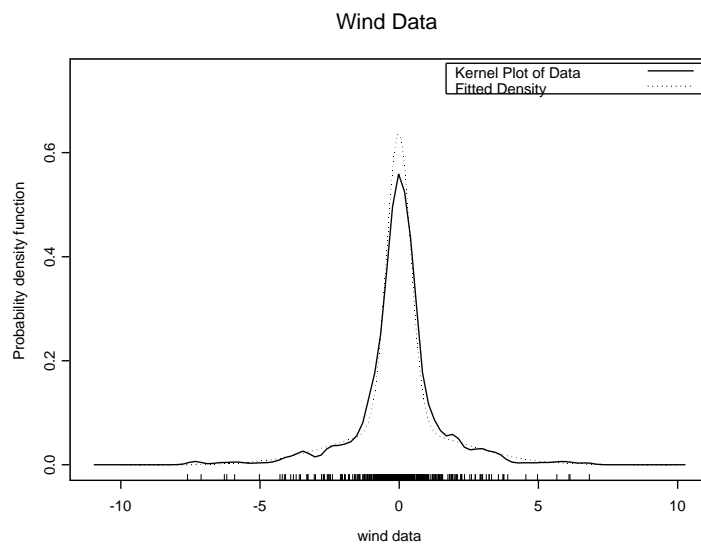


Figure 1: Kernel plot and fitted density for the wind dataset

For the geomagnetic data, the variational algorithm finds 2 states. If we placed a bound on a^* values it was possible to fit a model with a larger number of states and the DIC was lowest for a more complex 4-state solution in this case. 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 2. The fitted density is plotted in Figure 2.

The analysis by Robert et al.(2000) resulted in a 3-state model for these data and Francq and Roussignol's (1997) analysis selected a 2-state model. 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 and Roussignol (1997), whose estimated parameters were $\pi_{12} = 0.014$, $\pi_{21} = 0.16$, $\sigma_1 = 2.034$ and $\sigma_2 = 5.840$. Francq and 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.

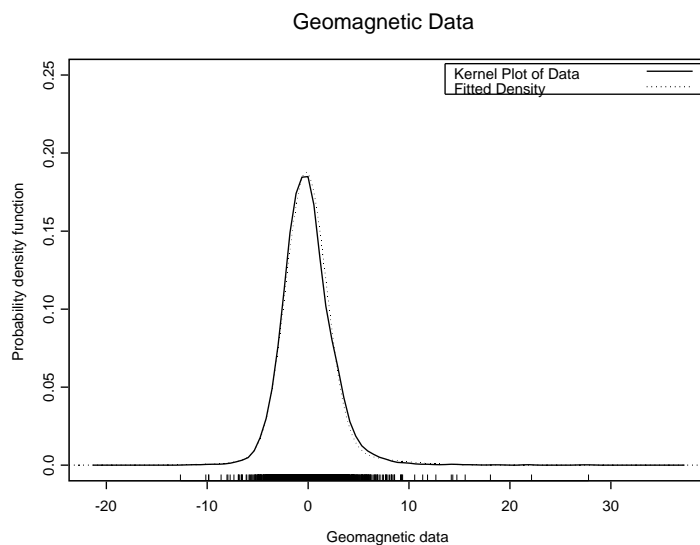


Figure 2: Kernel plot and fitted density for the geomagnetic dataset

For the daily returns data, the variational algorithm only converges for models with an initial number of states smaller than or equal to 2 if we do not impose a lower bound on posterior transition probability estimates. The lowest DIC was for 4 states. The fitted 2-state solution had variational posterior transition matrix

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

and variational posterior estimates given in Table 2. The fitted density is plotted in Figure 3.

This solution shows similarities to that of Robert et al.(2000), whose analysis favoured 2 or 3 states, and that of Rydén et al. (1998), who fitted a 2-state model. In the analysis by Robert et al.(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 et al. (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 $\sigma_2 = 0.0092$.

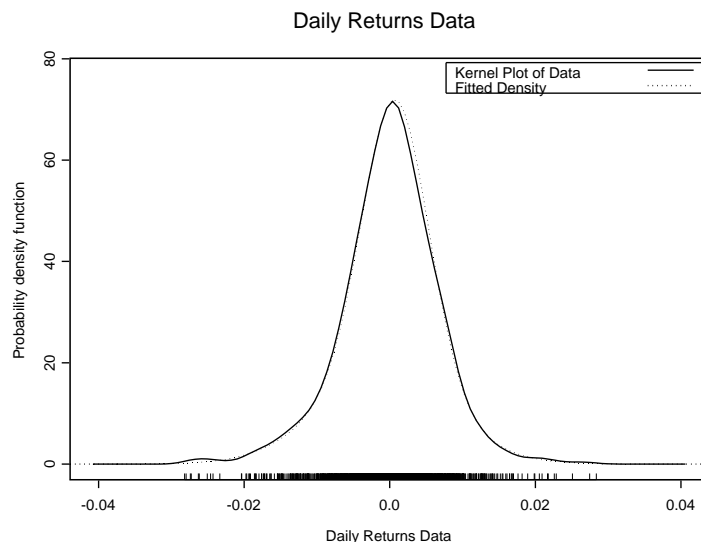


Figure 3: Kernel plot and fitted density for the daily returns dataset

Modelling the Real datasets as Mixture Models, Ignoring the Time Dependency

We also analysed the three real datasets as mixture data, using the variational approach from McGrory and Titterton (2006). For the wind dataset, initialising the algorithm with 15 components resulted in a solution with 7 components. Therefore, ignoring the dependency between observations and treating them as identically and independently distributed resulted in a more complex model than that which was obtained by modelling the dependency. We also initialised the mixture program with two components to allow a comparison of the fit with that obtained from the HMM analysis. Results are given in Table 3. There are similarities between these posterior estimates and those obtained from the HMM analysis (Table 2).

Table 3: Estimated posterior parameters from mixture analysis of the wind, geomagnetic and daily returns data with the algorithm initialised with 2 components

	Dataset		
	Wind	Geomagnetic	Daily Returns
Posterior means	0.0049, -0.028	-0.232, 2.658	0.00058, -0.00138
Posterior standard deviations	0.47, 2.45	2.02, 5.69	0.00457, 0.00978
Posterior weights	0.64, 0.36	0.93, 0.07	0.70, 0.30

Interestingly, the DIC value for the 2-state model (DIC=1571.98) was higher than that for the 7-state model obtained by initialising with 15 states (DIC=1487.43). The DIC gives more support to the more complex model in this instance.

For the geomagnetic data, initialising the algorithm with 15 states resulted in a solution with 5 states. This is more complex than the result we obtained in the HMM analysis. To compare results with the HMM analysis, the results found by fitting a mixture model with two components are reported in Table 3. The posterior estimates here are similar to those found in the HMM analysis (Table 2). Again, the DIC value for the 2-state model (DIC=12289.25) was higher than that for the 5-state model (DIC=12261.81) and so the DIC favours the more complex model.

For the daily returns data, initialising the algorithm with 15 components resulted in a solution with 2 components; results are presented in Table 3. In this case variational Bayes

selects a model with the same complexity as that found by the HMM analysis. Again, these posterior estimates are similar to those obtained in the HMM analysis (Table 2).

In the case of the wind and geomagnetic datasets, the mixture analysis resulted in a more complex model than did the HMM analysis. In these instances, adding model complexity has led to a smaller number of components. However, for the daily returns data, the level of complexity found was the same. In all cases there were similarities between the posterior estimates found by the two types of analysis.

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.

For the three real datasets analysed we compared results obtained from the HMM analysis with those we find by treating the data as if they had come from a mixture distribution and ignoring the extra dependencies. In two cases the mixture analysis resulted in a model with higher complexity than did the HMM analysis. However, for one dataset, the level of complexity found was the same for both the mixture and the HMM set-up. In two instances, therefore, adding model complexity led to a smaller number of components, but there is insufficient evidence to suggest that this would be the case in general. For all three datasets the posterior estimates found by both types of analysis showed similarities.

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.

References

- Attias, H. (1999). Inferring parameters and structure of latent variable models by variational Bayes. *Proceedings of Conference on Uncertainty in Artificial Intelligence (UAI)*.
- Baum, L.E., Petrie, T., Soules, G. and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, **41** 164-171.
- Berg, A., Meyer, R. and Yu, J. (2004). Deviance information criterion for comparing stochastic volatility models. *Journal of Business and Economic Statistics*, **22** 107-119.
- Boys, R.J. and Henderson (2004). A Bayesian approach DNA sequence segmentation (with Discussion). *Biometrics*, **60** 573-588.
- Celeux, G., Forbes, F., Robert, C. and Titterton, D.M. (2006). Deviance Information Criteria for missing data models (with Discussion). *Bayesian Analysis*, to appear.
- Chib, S. (1996). Calculating posterior distributions and modal estimates in Markov mixture models. *Journal of Econometrics*, **75** 79-97.
- Corduneanu, A. and Bishop, C.M. (2001). Variational Bayesian model selection for mixture distributions. In *Artificial Intelligence and Statistics* (T. Jaakkola and T. Richardson, eds.), pp.27-34. Morgan Kaufmann.
- Francq, C. and Roussignol, M. (1997). On white noise driven by hidden Markov chains. *Journal of Time Series Analysis*, **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. and Richardson, S. (2002). Hidden Markov models and disease mapping. *Journal of the American Statistical Association*, **97** 1055-1070.
- Han, C. and Carlin, B.P. (2001). Markov chain Monte Carlo methods for computing Bayes factors: a comparative review. *Journal of the American Statistical Association*, **96** 1122-1132.
- Jaakkola, T.S. (2000). Tutorial on variational approximation methods. In *Advanced Mean Field Methods* (M. Opper and D. Saad, eds.) pp. 129-159. MIT Press, Cambridge, MA.

- Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, **37** 183-233.
- Lee, L., Attias, H. and Deng, L. (2003). Variational inference and learning for segmental switching state space models of hidden speech dynamics. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Hong Kong, Apr, 2003*.
- MacDonald, I.L. and 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. and Titterton, D.M. (2006). Variational Approximations in Bayesian Model Selection for Finite Mixture Distributions. *Computational Statistics and Data Analysis*, to appear.
- Rabiner, L.R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, **77** 257-284.
- Richardson, S. and Green, P.J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society, Series B*, **59** 731-792.
- Robert, C.P., Celeux, G. and Diebolt, J. (1993). Bayesian estimation of hidden Markov chains: a stochastic implementation. *Statistics and Probability Letters*, **16** 77-83.
- Robert, C.P. and Titterton, D.M. (1998). Resampling schemes for hidden Markov models and their application for maximum likelihood estimation. *Statistical Computing*, **8** 145-158.
- Robert, C.P., Rydén, T. and Titterton, D.M. (2000). Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method. *Journal of the Royal Statistical Society, Series B*, **62** 57-75.
- Rydén, T., Teräsvirta, T. and Åsbrink, S. (1998). Stylized facts of daily return series and the hidden Markov model. *Journal of Applied Econometrics* **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). *Journal of the Royal Statistical Society, Series B*, **64** 583-639.

Wang, B. and Titterton, D.M. (2006). Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model. *Bayesian Analysis*, **1** 625-650.

Winn, J. and Bishop, C.M. (2005). Variational message passing. *Journal of Machine Learning Research*, **6** 661-694.

Zhu, L. and Carlin, B.P. (2000). Comparing hierarchical models for spatio-temporally misaligned data using the deviance information criterion. *Statistics in Medicine*, **19** 2265-2278.

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 $\text{fvar}_i(j_1) = p(y_1, y_2, \dots, y_i, z_i = j_1)$ and the algorithm proceeds as follows.

1. Calculate $\text{fvar}_1(j_1) = \pi_{j_1} p(y_1 | z_1 = j_1)$ for j_1 such that $1 \leq j_1 \leq K$, 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 $i = 1, \dots, n - 1$ and each j_2 , calculate

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

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

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

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 n^{th} iteration.

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

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

where

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

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 $\text{bvar}_i(j_1) = p(y_{i+1}, y_{i+2}, \dots, y_n | z_i = j_1)$, 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 $\text{bvar}_n(j_1) = 1$, for all 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 $i = n - 1, n - 2, \dots, 1$,

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

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 $b_{i+1 j_2}^*$.