

---

# Ensemble Hidden Markov Models with Extended Observation Densities for Biosignal Analysis

Ieab Rezek and Stephen Roberts

University of Oxford, Dept. of Engineering Science, Oxford, U.K.  
{irezek,sjrob}@robots.ox.ac.uk

**Summary.** Hidden Markov Models (HMM) have proven to be very useful in a variety of biomedical applications. The most established method for estimating HMM parameters is the maximum likelihood method which has shortcomings, such as repeated estimation and penalisation of the likelihood score, that are well known. This paper describes an Variational learning approach to try and improve on the maximum-likelihood estimators. Emphasis lies on the fact, that for HMMs with observation models that are from the exponential family of distributions, all HMM parameters and hidden state variables can be derived from a single loss function, namely the Kullback-Leibler Divergence. Practical issues, such as model initialisation and choice of model order, are described. The paper concludes by application of 3 types of observation model HMMs to a variety of biomedical data, such as EEG and ECG, from different physiological experiments and conditions.

## 1 Introduction

Hidden Markov models (HMM) are well established tools with widespread use in biomedical applications. They have been applied to sleep staging [3], non stationary biomedical signal segmentation [15], brain computer interfacing [12], and of course speech [23]. Their strength lies in the fact that they are capable of modelling sudden changes in the dynamics of the data, a situation frequently observed in physiological data. Another crucial factor is their ability to segment the data into informative states in a completely unsupervised fashion. When very little about the underlying physiological state is known *a priori*, unsupervised methods are a useful way forward.

The most established method of training the parameters of a HMM uses the maximum-likelihood framework [16, 23]. The method leads to simple parameter update equations. There are, however, some problems with the maximum-likelihood approach. Leaving the Bayesian argument aside, maximum-likelihood estimators are, comparatively slow in their convergence. From our experience, maximum *aposteriori* estimators are considerably faster, presumably because the priors somewhat smooth the likelihood surface and thus optimisation homes in quicker on a local maximum. Also, in maximum-likelihood estimation, special care must be taken not to over-fit the data. To avoid this, models of different complexity must be repeatedly

estimated and the residuals computed and penalised for complexity of the model until an optimal model is found. While some heuristics may well be available to overcome such a computationally expensive approach they nonetheless remain expensive [22] and the finiteness of training data means that larger models cannot be computed without running into computational problems, such as near-singular covariance matrices.

These drawbacks of maximum-likelihood base estimation of HMM parameters have lead us to investigate Bayesian approaches to learning model parameters and structures. Unlike Maximum likelihood estimators which estimate a single value for the parameters, a Bayesian estimator results in a distribution for the HMM parameters and thus confidence measure for the model. In computing this distribution, which is known as the posterior, the Bayesian estimator requires the likelihood of the data under the model and the prior distributions of the model parameters. The priors in Bayesian approaches play a role similar to information gathered from previously collected data [7] and thus allow us to overcome the problems of singularities in the presence of limited actual training data. In addition, if new data does not support the hypothesised model complexity, posterior densities of the model parameters will be no different from their prior densities, which in effect results in an automatic pruning process. The main difficulty of Bayesian methods tends to be of a mathematical nature. Complex models can quickly become mathematically intractable, i.e. integrating out nuisance parameters becomes impossible. This is often resolved by the use of sampling approaches to estimate the posterior densities [19]. Analytic approximations also exist. They might lead to slightly inferior solutions when compared to sampling, as they often seek only *local* maxima. Being analytic, however, we expect them but converge must faster than sampling estimators.

In this paper we describe the use of an analytical approximation to the exact Bayesian posterior densities by means of the variational framework. This framework provides a unified view of estimating all unknown HMM variables - parameters and hidden states. We will describe how update equations can be obtained for a wide range of HMMs with various observation densities. We finally apply the HMMs to biomedical data, such as electrocardiogram derived R-wave interval series, respiration recordings and electroencephalographic signals (EEG).

## 2 Principles of Variational Learning

Chapters on variational calculus can be found in many mathematics text books, yet, their use in statistics is relatively recent (see [8] and [10] for excellent tutorials). In essence, the aim is to minimise a cost function which, in this case, is the Kullback-Leibler (KL) divergence [2]. The KL divergence measures a distance between two distributions, say  $Q$  and  $P$ , by the integral<sup>1</sup>

$$\mathcal{F} = D(Q(H)||P(H, V)) = \int Q(H) \log \frac{Q(H)}{P(H|V)} dS + \log P(V) . \quad (1)$$

---

<sup>1</sup> Physicists have noted that the same function occurs in statistical mechanics and therefore often refer to it as minimising the so-called variational free energy [9].

Here, the distributions  $Q(H)$  and  $P(H|V)$  are defined over the set of all hidden variables  $H$ , such as parameters or hidden states, conditioned on the observed data  $V$ .

The choice of cost function, like that of the squared error, is primarily influenced by its practicability, i.e. it often results in simple solutions. One way of achieving this, in the case of the KL divergence, is to stay within the group of the exponential family and to approximate the full but intractable posterior probability density  $P(H|V)$ , which parameterises the true model, by a simpler but tractable distribution,  $Q(H)$ . Minimising or differentiating the KL divergence with respect to the approximate posterior distribution,  $Q(H)$ , results in simple equations, which are often coupled and thus need to be re-estimated in an iterative fashion.

By approximating the full posterior, one can enjoy some of the benefits of Bayesian analysis, such as full Bayesian model estimation and automatic penalties for over-complex models to avoid over-fitting. Note, the first term on the right-hand side in equation (1) is always non-negative, and thus the divergence is a bound to the true log-probability of the data  $P(V)$ . This means that optimal model selection takes place within the class of approximated and thus suboptimal models.

The simplest of all approximations to the true posterior distribution,  $P(H|V)$ , can be obtained by assuming that, if one is given a set of hidden variables  $H = \{H_1, \dots, H_N\}$ , the Q-distribution factorises

$$Q(H) = \prod_{i=1}^T Q(H_i), \quad (2)$$

with the additional obvious constraint that the distributions integrate to unity, that is  $\int Q(H_i) dH_i = 1$ . This assumption is known as the “mean-field” assumption. Under the mean-field assumption, the distributions  $Q(H_i)$  which maximise the KL divergence (1) can be shown [6] to have the general form

$$Q(H_i) = \frac{1}{Z} \exp \int Q(\bar{H}_i) \log P(H_i|\bar{H}_i) d\bar{H}_i, \quad (3)$$

where  $\bar{H}_i = H \setminus H_i$  is the set of all hidden variables  $H$  excluding  $H_i$  and  $Z$  is just a normalisation constant. For completeness, we repeat the derivation of the model-free form (3) of [6] in Appendix A.

There is, in principle no reason to restrict oneself to the independence assumption of the  $H_i$ . One can easily define  $H_i$  to be actually a subset of variables forming a partition of  $H$ . In this paper for instance, we form such a partition by grouping all HMM variables into the set of hidden state variables and the set of HMM model parameters (governing the probability of observing the datum at a particular time instance and the probability of transit to the next time step). The actual set of model parameters we denote by  $\theta = \{\theta_1, \dots, \theta_M\}$  and the set of hidden state variables by  $S = \{S_0, \dots, S_T\}$ . The hidden state variables,  $S$  form one partition within the overall set of variables and are updated jointly. In contrast the members of the set of HMM model parameters,  $\theta$ , are assumed to be independent from one another. With these assumptions, the approximating distribution  $Q(H)$  may be expressed as

$$Q(H) \triangleq Q(S)Q(\theta) = Q(S_0) \prod_{t=1}^T Q(S_t|S_{t-1}) \prod_{j=1}^M Q(\theta_j). \quad (4)$$

This is useful because the distribution  $Q(S)$  has the structure of a simple chain. This makes it tractable and an exact updating scheme can be found jointly for all  $Q(S_t)$ .

Finally, we make one further assumption. Models of different structures or sizes, e.g. state space dimensions or observation model order, are taken to be independent from one another. The set of all model sizes is denoted by  $A$  and a particular model size is indexed with  $a$ . We can then write the posterior probability of all model sizes  $A$  and unknown variables in the following form

$$P(S, \theta, A) = \prod_{a=1}^A P(S|a)P(\theta|a)P(a), \quad (5)$$

where each factor,  $P(S|a)P(\theta|a)P(a)$ , corresponds to a different model size. Assuming that all model structures in  $A$  are equally likely, the posterior model probability can be computed as

$$Q(a) \propto \exp\{-\mathcal{F}_a\}, \quad (6)$$

where  $\mathcal{F}_a$  is the KL divergence of the entire model for a fixed model size, i.e.

$$\mathcal{F}_a = \iint Q(S|a)Q(\theta|a) \log \frac{Q(S|a)Q(\theta|a)}{P(S, V, \theta, a)} dS d\theta. \quad (7)$$

The assumption is primarily chosen to make it easier to select a particular model, which will be the model with the smallest KL divergence between the true and approximate model distribution. In general, such an assumption should be used with caution as models are very likely to overlap significantly. However, this assumption used widely though in a different form. It is identical to the assumption of uniform priors over model structures.

### 3 Variational Learning of Hidden Markov Models

Consider a set of  $T$  random variables,  $S = \{S_1, \dots, S_t, \dots, S_T\}$ . Each random variable can take on one of, say  $M$ , discrete values,  $S_t = \{s_1, \dots, s_M\}$ . If we impose, for  $t > 0$  a probability on observing the variable  $S_t$  that is conditioned on the value of the variable  $S_{t-1}$  and denote this probability by  $P(S_t|S_{t-1})$ , one obtains a Markov Chain since  $P(S_t|S_{t-1})$  is known as the Markov property. A Hidden Markov process supposes further that what is actually observed are not the individual  $S_t$ , but a corrupted version of them. The variables  $S_t$  are thus hidden from the observer and the observations, say  $X_t$  are dependent on the variable  $S_t$ . This probability distribution over the entire sequence of observations  $X = \{X_1, \dots, X_t, \dots, X_T\}$  and states  $S = \{S_1, \dots, S_t, \dots, S_T\}$  then takes the mathematical form of

$$P(S, X) = P(S_0) \prod_{t=1}^T P(S_t|S_{t-1})P(X_t|S_t). \quad (8)$$

The state transition probability,  $P(S_t|S_{t-1})$  is a multinomial (i.e. discrete) distribution, encoded in an  $M \times M$  matrix since there are  $M$  possible values  $S_t$  can take for every value  $S_{t-1}$  has taken. These probabilities are denoted by  $\pi_{t,t-1}$  and are assumed to be independent of the time  $t$ , i.e. we assume the Markov Chain is homogeneous. The initial state probability  $S_0$  is parameterised by  $\pi_0$  and is also a multinomial with  $M$  probability values. The value  $M$  is called the state space dimension. The probability  $P(X_t|S_t)$  is called the observation probability and is parameterised by  $\theta_{\text{Obs}}$ . Its form depends on the assumed observation model, which in the case of a Gaussian corrupted Markov Chain is simply a set of  $M$  Gaussian densities, one for each value of  $S_t$ .

The speech community represents a HMM graphically, shown in figure (1[a]) for a HMM with  $M = 2$ , by representing each state a variable  $S_t$  can take by a vertex. The transitions from one state into the next are depicted by arcs and labelled with the probability of making the transition. Unlike the state space representation of the HMM, the graphical model representation, shown in figure (1[b]), depicts the HMM using vertices for the state variables  $S$  and observations  $X$ . This has its origin in interpreting the HMM as a Bayesian network in which all random variables are assigned a vertex. Observed (a.k.a. instantiated) random variables vertices have shaded vertices. Arrows between vertices represent statistical relationships between the random variables. The functional form of the relationship is often left out and results in ambiguities. A factor graph makes the functional form of variable dependencies clearer. Such a graph is shown in figure (1[c]).

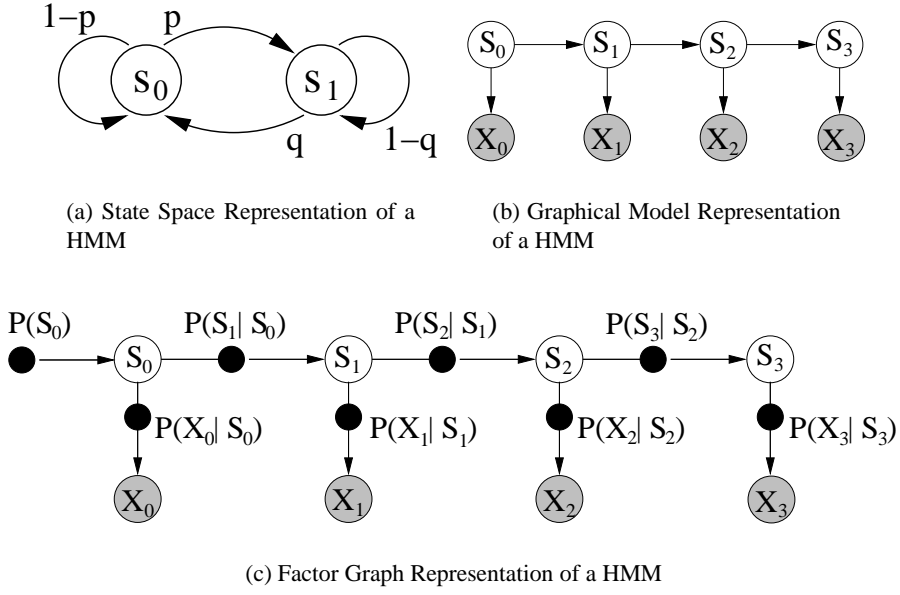
The concepts of variational learning described in the previous section can be readily applied to a first order hidden Markov model. We assume a HMM of length  $T$ , state space dimension  $M$ , hidden state variables,  $S = \{S_1, \dots, S_T\}$ , and observations  $X = \{X_1, \dots, X_T\}$ , transition model and the observation model. The HMM parameters  $\theta$ , consist of  $\pi_{t,t-1}$  which determine the state transition probability  $P(S_t|S_{t-1})$ ,  $\pi_0$  which parameterise the initial state probability  $P(S_0)$  and  $\theta_{\text{Obs}}$  which describe the observation probabilities  $P(X_t|S_t)$ . The full true posterior probability of the model is then given by

$$P(S, X, \theta) = P(S, X|\theta)P(\theta) \quad (9)$$

$$\triangleq P(S, X|\pi_{t,t-1}, \pi_0, \theta_{\text{Obs}})P(\pi_{t,t-1}, \pi_0, \theta_{\text{Obs}}) \quad (10)$$

$$= P(S_0|\pi_0) \prod_{t=1}^T P(S_t|S_{t-1}, \pi_{t,t-1})P(X_t|S_t, \theta_{\text{Obs}}) P(\pi_{t,t-1})P(\pi_0)P(\theta_{\text{Obs}}) . \quad (11)$$

What remains to apply the cost function (1) is the distribution which approximates the full posterior probability  $P(S, X, \theta)$ . First, we assume that all the HMM model parameters,  $\theta$  (i.e. all variables but the state space variables) are independent (mean field assumption). Second, all the hidden state variables are grouped together and are governed by one distribution  $Q(S)$ . With these assumptions, the resulting KL divergence (1) then becomes



**Fig. 1.** A HMM represented graphically. Subfigures [b] and [c] represent the HMM with 4 unrolled time slices. Shaded nodes denote observed random variables.

$$\mathcal{F} = \int \dots \int Q(S) Q(\pi_{t_{-1}}) Q(\pi_0) Q(\theta_{\text{Obs}}) \log \frac{Q(S) Q(\pi_{t_{-1}}) Q(\pi_0) Q(\theta_{\text{Obs}})}{P(S, X, \pi_{t_{-1}}, \pi_0, \theta_{\text{Obs}})} dS d\theta . \quad (12)$$

This KL divergence can now be minimised individually with respect to the distributions,  $Q(S)$ ,  $Q(\pi_{t_{-1}})$ ,  $Q(\pi_0)$  and  $Q(\theta_{\text{Obs}})$ . In the two following sections we first update divergence (12) with respect to the distribution of the hidden states,  $Q(S)$ , to show that it leads to the well-known Baum-Welch or forward-backward recursions. Then the cost function (12) is minimised with respect to  $Q(\theta_{\text{Obs}})$ . The minimisation depends on the functional form of the observation model,  $Q(\theta_{\text{Obs}})$ , and is shown for different types of observation models, such as Gaussian and Linear observation models.

### 3.1 Learning the HMM Hidden State Sequence

We begin by optimising the KL divergence (12) with respect to the distribution over all hidden states,  $Q(S)$ . Using the mean-field update equation (3), the optimal posterior distribution over all hidden states,  $Q(S)$ , can be computed by replacing all HMM parameters  $\theta$  for  $\bar{H}_i$  in equation (3), to give

$$Q(S) \propto \exp \int Q(\theta) \log P(S, X, \theta) d\theta \quad (13)$$

Of interest, however, is not so much the global distribution,  $Q(S)$ , but the marginal distributions,  $Q(S_t)$ , and the joint hidden state probabilities,  $Q(S_t, S_{t+1})$ . They can be calculated using the well-known Baum-Welch or forward-backward recursions.

The justification for using the forward-backward recursions comes from the fact that they are the result of minimising equation (12) with respect to the individual hidden state probabilities,  $Q(S_t)$ , and the joint hidden state probabilities,  $Q(S_t, S_{t+1})$ . To see this, assume all other Q-distributions are held constant and parameterised by  $\tilde{\theta}$  which are the values calculated during previous iterations. Then, the KL-divergence (12) simplifies to

$$\mathcal{F} = \int Q(S) \log \frac{Q(S)}{\tilde{P}(S, X)} dS + const, \quad (14)$$

where

$$\tilde{P}(S, X) = \int Q(\theta) \log P(S, X, \theta) d\theta, \quad (15)$$

and the constant includes all terms, not involving  $S$ . To further minimise the divergence (14), additional consistency and normalisation constraints are needed. One set of constraints simply ensures that all Q-distributions normalise to 1. In addition, consistency (or holonomic) constraints are required which ensure that marginalising  $Q(S_t, S_{t-1})$  with respect to  $Q(S_t)$  gives the same result as marginalising  $Q(S_t, S_{t+1})$ . These additional constraints are added to the KL-divergence (14) with the usual Lagrangian multipliers. The simple structure of the hidden state chain permits an analytic solution to equation (14) in form of iterative computations of the Lagrangian multipliers. The detailed steps starting from optimising the KL divergence (14) leading up to the Baum-Welch recursions are given in Appendix B.

For simplicity, in the following we assume the Baum-Welch recursions have been applied and yield the marginal and joint distributions  $Q(S_t)$  and  $Q(S_t, S_{t-1})$ , respectively. We make use of the notation introduced in [16], specifically we denote the probability of the state variable  $S_t$  taking one of the  $M$  values  $m = 1, 2, \dots, M$ , by

$$\gamma_t(m) = Q(S_t = m|X). \quad (16)$$

Further, we denote the joint probability of variable  $S_t$  taking value  $n$  and  $S_{t-1}$  taking value  $m$ , by

$$\xi_t(m, n) = Q(S_t = n, S_{t-1} = m|X). \quad (17)$$

### 3.2 Learning HMM Parameters

The updated the probabilities of the hidden state variables result in values of  $\gamma_t(m)$  and  $\xi_t(m, n)$  (see previous section) for each time instance  $t$ . In order to obtain an analytic solution for the HMM parameters after substituting into equation (3), we

need to choose appropriate prior distributions,  $P(\theta)$ , for the HMM parameters,  $\theta$ . To obtain analytic solutions, the prior distributions,  $P(\theta)$ , are required to be conjugate distributions. The approximate posterior distributions  $Q(\theta)$  will then be functionally identical to the prior distributions (i.e. a Gaussian prior density is mapped to a Gaussian posterior density). Apart from the parameters of the observation model, which will be discussed later, for the HMM these conjugate prior distributions are, as given in [1]. For the initial state probability  $\pi_0$ , we use an  $M$ -dimensional Dirichlet density

$$Dir(\pi_0) = \frac{\Gamma(\sum_l \kappa_l)}{\prod_l \Gamma(\kappa_l)} \prod_{m=1}^M \pi_{0_m}^{\kappa_m - 1}, \quad (18)$$

and, for the transition probabilities,  $\pi_m$ ,  $M \times M$ -dimensional Dirichlet densities

$$Dir(\pi_{t-1}) = \prod_{m=1}^M \frac{\Gamma(\sum_l \lambda_{m_l})}{\prod_l \Gamma(\lambda_{m_l})} \prod_{n=1}^M \pi_{m_n}^{\lambda_{m_n} - 1}. \quad (19)$$

Based on these assumptions, minimisation of the KL divergence with respect the posterior distributions  $Q(\pi_{t-1})$  and  $Q(\pi_0)$ , leads to posterior initial state and transition probabilities that are again Dirichlet distributed and have parameters, respectively,

$$\tilde{\kappa}_m = \gamma_{t=0} = m + \kappa_m; \quad (20)$$

$$\tilde{\lambda}_{m_n} = \sum_t \xi_t(m, n) + \lambda_{m_n}. \quad (21)$$

The hyper-parameters  $\kappa$  and  $\lambda$  are fixed and typically set to integer values just greater than 1, reflecting the fact that little is known *a priori* about the initial state and state transition probabilities.

### 3.3 HMM Observation Models

What remains to specify is the probability of the observation given the state at time  $t$ ,  $P(X_t|S_t)$ . Determining it will also force our hands in determining the prior distributions of the parameters governing the observation model,  $P(X_t|S_t)$ . The choice of  $P(X_t|S_t)$  is also very much dependent on the application. The two classic and simplest of all cases assume the Markov chain is corrupted by additive Gaussian noise or that the observations are discrete. In the latter, the observation model is simply a multinomial while in the former the observation model is (multivariate) Gaussian. Many others have, of course, been suggested in the HMM's long history. Among them are Poisson densities for count processes [11], and linear models (spectral or autoregressive) for use in, say EEG modelling [15], along with more complex models such as "Independent Component" models [14].

#### The Gaussian Observation Model

For observations which, conditional on the state, are Gaussian distributed,

$$P(X_t|S_t = m) = P(X_t|\mu_m, C_m) \quad (22)$$

where  $\mu = \{\mu_1, \dots, \mu_M\}$  and  $C = \{C_1, \dots, C_M\}$  are the normal distribution mean vectors and precision matrices, respectively. The conjugate densities [1] for the means  $\mu_m$ , are  $K$ -dimensional Normal densities ( $m = 1, \dots, M$ )

$$P(\mu_m) \propto e^{-\frac{1}{2}(\mu - \mu_{m0})^\top C_{m0}(\mu - \mu_{m0})}, \quad (23)$$

and for the precisions  $C_m$ ,  $K$  dimensional Wishart densities ( $m = 1 \dots, M$ )

$$P(C_m) \propto |C_m|^{\alpha_m - \frac{K+1}{2}} e^{-\text{tr}(B_m C_m)} \quad (24)$$

Inserting these densities into (12) and subsequent minimisation leads to update equations for the parameters of the posterior densities of the means and precisions. The posterior means follow normal densities  $q(\mu_m) \sim \mathcal{N}(\tilde{\mu}_{m0}, \tilde{C}_{m0})$ , with parameters

$$\tilde{\mu}_{m0} = (\tilde{\gamma}_m \tilde{\alpha}_m \tilde{B}_m^{-1} + C_{m0})^{-1} (\tilde{\alpha}_m \tilde{B}_m^{-1} \bar{x}_m + C_{m0} \mu_{m0}); \quad (25)$$

$$\tilde{C}_{m0} = (\tilde{\gamma}_m \tilde{\alpha}_m \tilde{B}_m^{-1} + C_{m0}) \quad (26)$$

where  $\bar{x}_m = \sum_{t=1}^T \gamma_t(m) x_t$  and  $\tilde{\gamma}_m = \sum_{t=1}^T \gamma_t(m)$ .

Similarly, the posterior precisions follow a Wishart density,  $q(C_m | \tilde{\alpha}_m, \tilde{B}_m) \sim \mathcal{W}(\tilde{\alpha}_m, \tilde{B}_m)$ , with parameters

$$\tilde{\alpha}_m = \frac{1}{2} \tilde{\gamma}_m + \alpha_m; \quad (27)$$

$$\tilde{B}_m = \frac{1}{2} \sum_t \gamma_t(m) (x_t - \tilde{\mu}_{m0})(x_t - \tilde{\mu}_{m0})^\top + \frac{1}{2} \tilde{\gamma}_m \tilde{C}_{m0}^{-1} + B_m \quad (28)$$

### The Poisson Observation Model

The choice of Gaussian observation model depends on the data and, hence, might not be appropriate. Particularly, when dealing with count data, such as the RR interval series obtained from the ECG signal, Gaussian observation models are only applicable after some preprocessing of the data, such as interpolation [20]. To avoid this we demonstrate the variational estimation of a Poisson observation model for HMMs. In this case, the observation density is

$$P(X_t|S_t = m, \mu) = e^{-x_t \mu_m} (x_t \mu_m)^{y_t} \frac{1}{y_t!} \quad (29)$$

where, each of the  $M$  states has a Poisson distribution with parameter  $\mu_m$ . The data points  $y_1, \dots, y_T$  are counts while the values  $x_t$  are called the exposure of the  $t$ -th unit, i.e. a fraction of the unknown parameter of interest  $\mu_m$  [4]. The prior for the parameter of the Poisson distribution is chosen to be conjugate, which is a Gamma density, ( $m = 1, \dots, M$ )

$$P(\mu_m) \propto \mu_m^{\alpha_0 - 1} \exp\{\beta_0 \mu_m\}. \quad (30)$$

The optimised Q-distribution for the Poisson rate of state  $m$  are also Gamma, with parameters

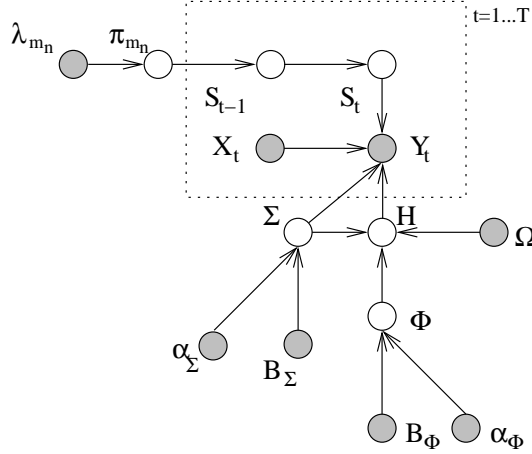
$$Q(\mu_m) \sim \mathcal{G}(\tilde{\alpha}_0, \tilde{\beta}_0)$$

$$\tilde{\alpha}_0 = \sum_t \gamma_{tm} y_t + \alpha_0 \quad (31)$$

$$\tilde{\beta}_0 = \sum_t \gamma_{tm} x_t + \beta_0 \quad (32)$$

### The Multivariate Linear Observation Model

The final model described here is the multivariate linear observation model. It is particularly useful when, for instance, auto-regressive features are to be extracted from EEG signals and later segmented using and HMM with Gaussian observation models. It is clear, that if the process of feature extraction and segmentation can be combined, the results are much improved [21]. Further, if the linear model has sinusoids as basis functions, one can use the HMM to segment the signal based on its spectral content. Without any major mathematical complications, we can also assume that a short data segment can be described with the same linear model. The model then becomes a matrix variate linear model, in which a segment of multivariate data forms an observation matrix that is modelled by a linear model. The model is best described as a graphical model shown in figure (2).



**Fig. 2.** Directed graph of a HMM with a matrix-variate linear observation model for observations  $Y$  and basis functions  $X$ .

For observation  $Y_t$  and basis functions  $X_t$ , the linear observation model is given as

$$P(Y_t - H_m X_t | S_t = m) = \mathcal{N}_{d,u}(0, \Sigma_m, I_u) \quad (33)$$

where  $\mathcal{N}_{d,u}$  is a  $d \times u$ -matrix variate normal density function [5],  $H = \{H_m\} \quad \forall m = \{1, \dots, M\}$  and  $I_u$  is a  $u \times u$  identity matrix. For example, assume a multivariate autoregressive (AR) model of order  $p$  which models a  $d$ -variate observation  $\mathbf{y}_t$ . The past samples can be concatenated in a matrix,  $x_t \stackrel{\text{def}}{=} [\mathbf{y}_{t-1}^\top, \dots, \mathbf{y}_{t-p}^\top]^\top$ . Thus  $\mathbf{y}_t \in \mathbb{R}^d$  is a  $d$ -variate response vector at time  $t$ ,  $\mathbf{x}_t \in \mathbb{R}^{dp}$  is a  $dp$ -variate basis vector at time  $t$ .  $\tilde{H}_p \in \mathbb{R}^{d \times dp}$  is the matrix of model coefficients,  $H \in \mathbb{R}^{d \times dp}$  is a partitioned matrix composed of the coefficient matrices at lag  $p$ . If, furthermore, the samples are grouped into segments, indexed by  $n$ , each of which contains  $u$  samples of  $\mathbf{y}_t$ , one can construct a matrix with  $u$  response- and  $u$  basis-vectors, i.e.

$$Y_n = [\mathbf{y}_{(n-1)u}, \dots, \mathbf{y}_{nu}] \quad (34)$$

$$X_n = [\mathbf{x}_{(n-1)u}, \dots, \mathbf{x}_{nu}] \quad (35)$$

The final form of the linear model takes the form  $Y_n = H X_n$  the residuals of which are Gaussian distributed, thus giving equation (33), with index  $t$  replaced by  $n$ .

The prior densities for the coefficient matrices  $H_m$  are assumed to be a  $d \times dp$ -matrix variate normal densities,  $\mathcal{N}_{d,dp}(\Omega, \Sigma_m, \Phi_m)$ , with mean  $\Omega$  and precisions  $\Sigma_m$  and  $\Phi_m$ . The prior for residual precisions  $\Sigma_m$  and the coefficient precisions  $\Phi_m$  are Wishart densities [1],  $\mathcal{W}_d(\alpha_\Sigma, B_\Sigma)$  and  $\mathcal{W}_{dp}(\alpha_\Phi, B_\Phi)$ , with shape/scale parameters  $\alpha_\Sigma/B_\Sigma$  and  $\alpha_\Phi/B_\Phi$ , respectively.

The posterior density of the model coefficients,  $H_m$ , is a  $d \times dp$  matrix variate Normal density with mean  $\tilde{\Omega}$  and precision matrices  $\tilde{\Sigma}_m, \tilde{\Phi}_m$  computed by

$$\tilde{\Omega}_m^\top = \tilde{\Phi}_m^{-1} \left( \sum_n \gamma_{nm} X_n Y_n^\top + \tilde{\alpha}_{\Phi_m} \tilde{B}_{\Phi_m}^{-1} \Omega^\top \right) \quad (36)$$

$$\tilde{\Sigma}_m = \tilde{\alpha}_{\Sigma_m} \tilde{B}_{\Sigma_m}^{-1} \quad (37)$$

$$\tilde{\Phi}_m = \sum_n \gamma_{nm} X_n X_n^\top + \tilde{\alpha}_{\Phi_m} \tilde{B}_{\Phi_m}^{-1} \quad (38)$$

The posterior of the residual variances,  $\Sigma_m$ , is a Wishart density with shape and scale parameters computed by

$$\begin{aligned} \tilde{\alpha}_{\Sigma_m} &= \frac{1}{2} \left( \sum_n u \gamma_{nm} + dp + 2\alpha_\Sigma \right) \\ \tilde{B}_{\Sigma_m} &= \frac{1}{2} \sum_n \gamma_{nm} \left[ (Y_n - \tilde{\Omega}_m X_n)(Y_n - \tilde{\Omega}_m X_n)^\top + \text{tr} \left( X_n X_n^\top \tilde{\Phi}_m^{-1} \right) \tilde{\Sigma}_m^{-1} \right] \\ &\quad + \frac{1}{2} \left( \tilde{\Omega}_m - \Omega_m \right) \tilde{\alpha}_{\Phi_m} \tilde{B}_{\Phi_m}^{-1} \left( \tilde{\Omega}_m - \Omega_m \right)^\top + \frac{1}{2} \text{tr} \left( \tilde{\alpha}_{\Phi_m} \tilde{B}_{\Phi_m}^{-1} \tilde{\Phi}_m^{-1} \right) \tilde{\Sigma}_m^{-1} + B_\Sigma \end{aligned} \quad (39)$$

The posterior model coefficient variances,  $\Phi_m$ , also follow a Wishart density with parameters

$$\begin{aligned}
\tilde{\alpha}_{\Phi_m} &= \frac{d}{2} + \alpha_{\Phi} \\
\tilde{B}_{\Phi_m} &= \frac{1}{2}(\tilde{\Omega}_m - \Omega)^\top \tilde{\alpha}_{\Sigma_m} \tilde{B}_{\Sigma_m}^{-1} (\tilde{\Omega}_m - \Omega) + \frac{1}{2} \text{tr}(\tilde{\alpha}_{\Sigma_m} \tilde{B}_{\Sigma_m}^{-1} \tilde{\Sigma}_m^{-1}) \tilde{\Phi}_m^{-1} + B_{\Phi}
\end{aligned} \tag{40}$$

### 3.4 Estimation

Having obtained the update equations for the observation models and the state transition probabilities, estimation then follows the familiar fashion of iteratively computing the HMM hidden state sequence (equivalent to the E-Step in the Max. Likelihood EM framework) and the HMM parameter posterior distributions (equivalent to M-Step). This is repeated until convergence is reached. Convergence is measured by the actual value of the KL-divergence (12) and the estimation is terminated when (12) no longer changes significantly. The mathematical form of the KL-divergence (12) will obviously depend on the type of model, which here means the type of observation model. The complete formulae, for each of the 3 observation model HMMs, are listed in Appendix C.

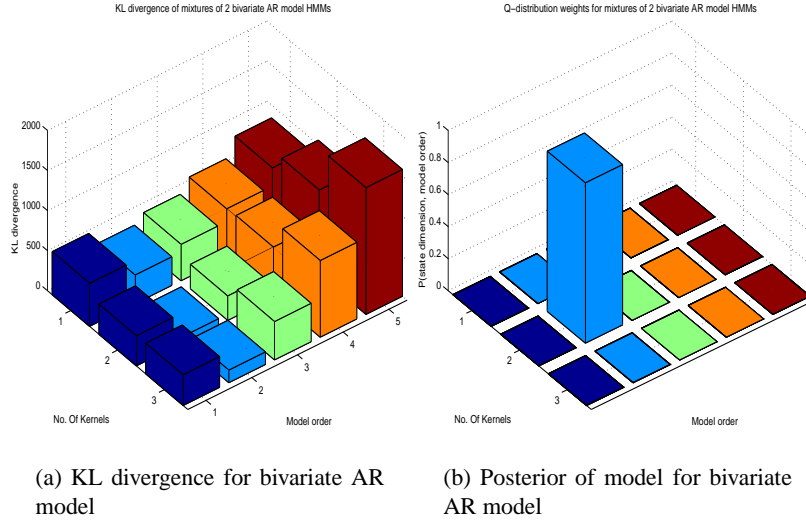
### Choice of Model Size

The iterations are run for a fixed HMM state space dimension and observation model order (number of basis function coefficients in the linear observation model). Estimation is then repeated for different settings and the value of the KL-divergence (12) recorded. The smallest achievable value of (12), according to equation (6), results in the highest probability for the particular choice of model.

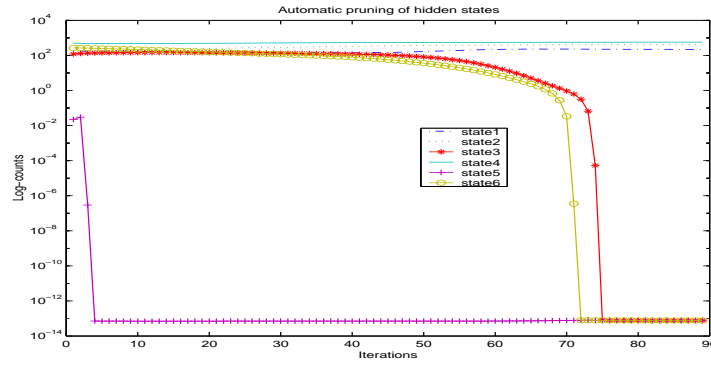
As an example, data was generated from two bi-variate AR processes, with AR-coefficient matrices of the first model set to  $A_{lag=1} = [0.4, 1.2; 0.3, 0.7]$  and  $A_{lag=2} = [0.35, -0.3; -0.4, -0.5]$ . The second model coefficient matrices were set to  $A_{lag=1} = [0.4, 0; 0, 0.7]$  and  $A_{lag=2} = [0.35, -0.3; -0.4, -0.5]$ . The intercept vector was set to a zero vector and the noise variance matrix to  $C = [1.00, 0.50; 0.50, 1.50]$ . Figure (3[a]) shows the KL-divergence values for different settings of linear model order and hidden state space dimension. The most probable model size is shown in figure (3)(b) and peaks at the expected model size.

Under certain conditions, the state space dimension need not be estimated as above. If the observation model is correct, for example the dimensions of the Gaussian observation model match those of the clusters in the data or the linear model order is the true order, one can exploit the fact that the state transition probabilities factorise, i.e.  $P(S_t|S_{t-1}) = \prod_{m=1}^M P(S_t|S_{t-1} = m)$ . The state space dimension,  $M$ , can now be set to an arbitrarily large value and the HMM is estimated just once. States, that are not visited by the model will automatically collapse to be equal to their prior distributions. One can thus read out most likely hidden state dimension from the number of distinct state values in the hidden state sequence.

This effect is shown in figure (4). The HMM was started with  $M = 6$  and the number of distinct states monitored at each step of the iteration.



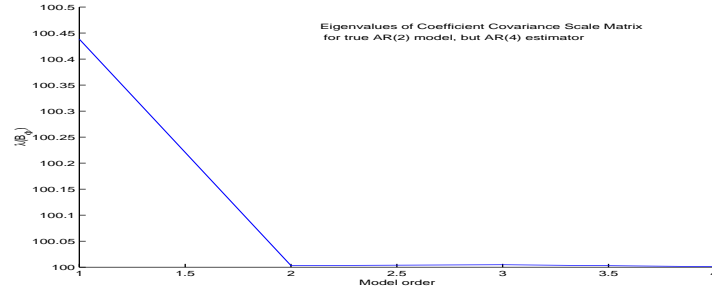
**Fig. 3.** KL divergences for mixtures of AR models



**Fig. 4.** The collapse of the state space dimension during the estimation

Observation model parameter distributions can also be automatically set to their prior distributions, provided the mathematical formulation of model is in a factorised form. For instance, if the linear model description was in terms of reflection rather than autoregressive coefficients, the mean field assumptions leads to an estimator which automatically prunes out all states and model coefficients (and thus model orders) not supported by the data. This is not always possible, however. So some engineering short-cuts can be used to quickly home-in on the true observation model size. For example, in the linear observation model's current form one can investigate the eigen-spectrum of the linear model parameters' scale matrix  $\tilde{B}_{\Phi_m}$ . The number

of non-zero eigenvalues are an indication of the model order. Also, looking at the distance between the model coefficient matrices of each state,  $\|H_i - H_j\|$  for different  $i \neq j$  and  $i, j = 1, \dots, M$ , might give some indication as to the preferred number of hidden states. An example of the use of the eigen-spectrum is shown in figure (5). As the model order is increased, the eigenvalues clearly shallow off at the correct model order. The coefficients themselves are relatively close to zero - below 10% of the largest coefficient. Ignoring numerical stability, the Bayesian treatment avoids again the singularities of maximum likelihood methods.



**Fig. 5.** Eigen-spectrum of the coefficient precision posterior density scale parameter  $B_\phi$  for a univariate AR model.

### Model Initialisation

Evidently, there are various ways of initialising the HMM and all depend on the observation model. The simplest initialisation is the random initialisation of the hidden state probabilities. The estimation then begins by estimating the HMM model parameters, thus avoiding to a large extent manual setting of HMM parameters. In practice, however, there is often the desire to keep the number of iterations, till convergence is reached, to a minimum. To achieve that, more educated “guesses” have to be made as far as the model parameters are concerned. This can be done in some cases by making use of the data to set the parameters and begin with the estimation of the hidden state probabilities. We illustrate this below for the Gaussian and linear Autoregressive observation models.

*The Gaussian observation model* can be initialised by running a few iterations of the (much faster) K-means algorithm or Gaussian mixture EM on the training data. The obtained cluster centres can be assigned to the posterior means,  $\tilde{\mu}_m$ , of the HMM observation model means. The means’ posterior precision matrices,  $\tilde{C}_{m_0}$ , are set all equal to the covariance of the total training data. The parameters of the posterior Wishart density for the observation model precision matrices consist of the shape parameter  $\tilde{\alpha}_m$  and the scale parameter  $\tilde{B}_m$ . The value of  $\tilde{\alpha}_m$  is set to the half the dimensionality of the training data, while  $\tilde{B}_m$  is set to the total training data covariance matrix scaled up by  $\tilde{\alpha}_m$ . Practice has shown this to be the most robust

initialisation procedure; robust, that is, in the sense that is least sensitive to cluster shapes and data range.

The observation model priors are generally set to be as flat as possible. The observation model means have an associated Gaussian prior with mean and covariance parameters set to the median and squared range of the training data, respectively.

*The Poisson observation model* is initialised at random. Conditioned on each state, we randomly select a sample from the data. The sample size itself is also drawn randomly. The shape,  $\tilde{\alpha}_0$  and scale parameters,  $\tilde{\beta}_0$ , of the posterior Gamma distribution are then set to the sum of the sample counts and exposure, respectively.

For the experiment described below, the observation model prior shape and scale parameters are set to a minimum of 1 count per interval, i.e.  $\alpha_0 = 1$ , and an average rate of 60 beats per minute, respectively.

*The Linear Autoregressive observation model* is initialised in three steps. First, an initial data segment is used to calculate the values of the AR-coefficient variances and the residual noise variance. Second, with the use of the variances obtained in the first step, a multivariate Kalman filter is applied to the entire training data<sup>2</sup> Finally, the so obtained AR-coefficients are segmented using a few iterations of K-means, say.

The posterior density of the model coefficients,  $H_m$ , is a matrix variate Gaussian. Its mean,  $\hat{\Omega}_m$  is set to the K-means cluster centres. The covariance matrices,  $\tilde{\Sigma}_m$  and  $\tilde{\Phi}_m$  are set, respectively, to the estimates of the residual noise variance and AR-coefficient variances of step one above. The posterior of the residual noise variances,  $\Sigma_m$ , is a Wishart density. Its scale parameter is set to  $\frac{1}{2}d + 1$ , i.e. half the residual noise dimensionality incremented by 1. The shape parameter is set to the residual noise variance multiplied by the Wishart density's scale parameter. Similarly, the posterior of the model coefficients,  $\Phi_m$ , also follow a Wishart density. Its scale parameter is set to  $\frac{1}{2}dp + 1$ , that is half the product of the linear model order and training data dimension and incremented by 1. The shape parameter is set to the AR coefficient variances of step one above, multiplied by the Wishart density's scale parameter.

The priors for the linear observation model assumes a standardised training data set, i.e. the data is detrended and its variance normalised to unity. The prior for the linear model coefficients,  $H_m$ , is thus set to have a mean of zero. The scale coefficient,  $B_\Sigma$ , of the prior for the noise precision  $\Sigma_m$ , is set to unity and the shape coefficient  $\alpha_\Sigma$  to the dimension of the noise precision. The prior of the linear model coefficient precisions is assumed to be more accurate than the noise precision. The scale,  $B_\Phi$ , is set to one order of magnitude larger than  $B_\Sigma$ , while the shape is set to the dimension of the coefficient space (model order  $\times$  data dimensionality).

<sup>2</sup> If the training data is very large experience has shown that a simple segmentation of the data and estimation of the AR coefficient matrices is faster than the Kalman filter approach, yet equally useful.

## 4 Experiments

### 4.1 Sleep EEG with Arousal

The first application to medical time series analysis demonstrates the use of the variational HMM to features extracted from a section of electroencephalogram data. The recordings were taken from a subject during a sleep experiment. The study looked at changes of cortical activity during sleep in response to external stimuli (e.g. from a vibrating pillow under subjects head). The fractional spectral radius (FSR) measure and spectral entropy [18] were computed for consecutive non-overlapping windows of one second length. In figure (6), the model clearly shows a preference for a 3-dimensional state space. Figure (7) shows a 10 minute section of data with the corresponding Viterbi state sequence. The data is segmented into the following regimes: wake (state 2, first 90s of the recording), deeper sleep (state 1) and light sleep (state 3) which is clearly visible at sleep onset ( $t \approx 90s$ ) and at the arousal ( $t \approx 200s$ ).

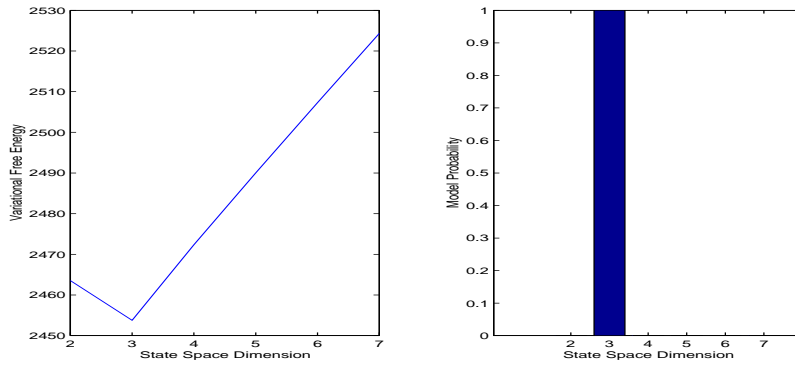


Fig. 6. State Space Dimension Selection

### 4.2 Whole Night Sleep EEG

In the following example we study the use of the HMM in segmenting EEG recordings for sleep staging. The data was recorded from one female subject exhibiting poor sleep quality during an 8 hour session in a sleep laboratory. The data was manually scored by 3 different sleep experts based on the standard Rechtschaffen and Kales (R&K) system [17]. The majority vote as then taken, to overcome disagreements between manual labels, resulting in a consensus score.

The data used here was recorded at the electrode sites  $C4$  [13] using a  $200Hz$  sampling rate. The state space dimension was set to 7, corresponding to the number of states according to R&K. The confusion matrix (table 1) based on the consensus score shows, however, that only 4 states are used by the HMM. One such state corresponds clearly to deep sleep, i.e. sleep stage 4. Corresponding somewhat less

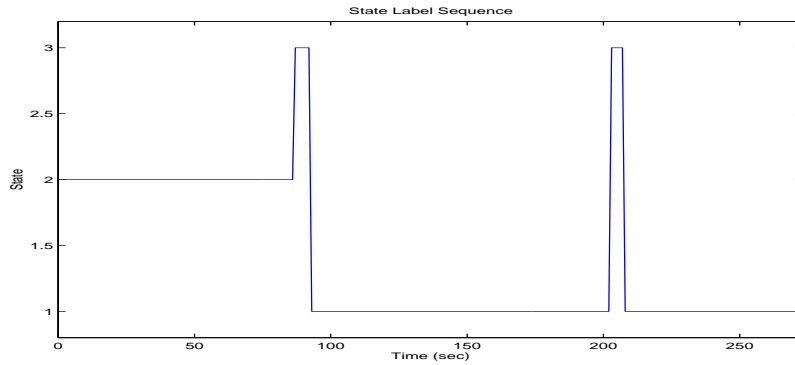


Fig. 7. HMM Sleep EEG Segmentation

strongly, are state 2 of the HMM with the lighter sleep stages 2 and 3. Class 3 of the HMM is predominantly visited when the subject resides in REM sleep, according to the experts. It is interesting to note that the HMM exhibits the same difficulty humans face when trying to distinguish REM sleep from light sleep (stages 1 and 2). Finally, the weakest association between the HMM and human labels is observed in HMM class 4. Seemingly mostly connected with the Wake state, much overlap also exists between it and sleep stage 2, hence the significance of this last class is uncertain.

The table is best summarised by the estimated HMM state sequence. Figure (8) shows the hypnogram with a the filtered Viterbi state sequence (using an 11-th order median filter). The filtering is justified by noting that human labels are over a 30s long data segment and are based on the occurrence of a particular distinctive feature within that time period. The algorithm, on the other hand, calculates the label based on the most frequent class for that segment, on a one second resolution.

Table 1. HMM Gaussian Observation Confusion Matrix

HMM Class	Manual Sleep Score						
	S4	S3	S2	S1	REM	Movement	Wake
Class 1	0.9086	0.0600	0.0092	0	0	0.0222	0
Class 2	0.1014	0.1978	0.6889	0.0060	0	0.0060	0
Class 3	0	0	0.2191	0.1561	0.6027	0.0118	0.0103
Class 4	0	0.1250	0.2159	0.1477	0	0	0.5114

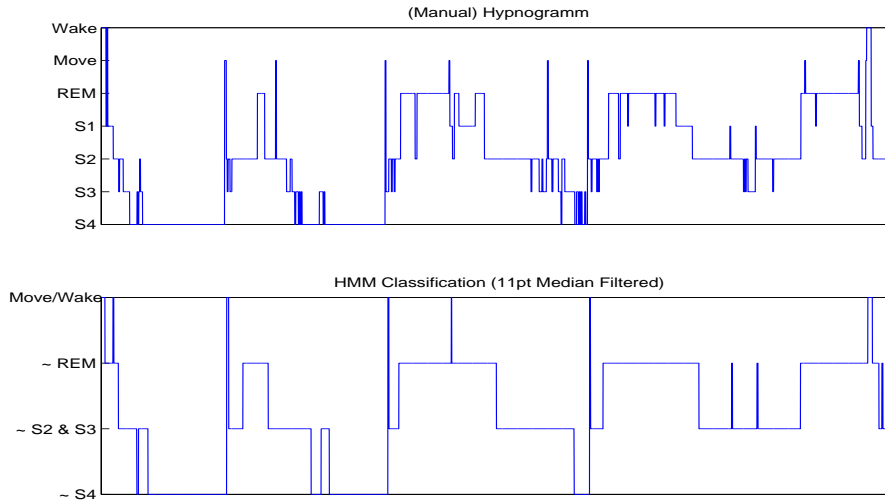


Fig. 8. Manual and Estimated Sleep Segmentation of 1 Night Sleep EEG

### 4.3 Periodic Respiration

We also applied the HMM to features extracted from a section of Cheyne Stokes (CS) Data<sup>3</sup>, consisting of one EEG recording and a simultaneous respiration recording, both sampled at  $128Hz$ . The feature, the fractional spectral radius (FSR) [18], was computed from consecutive non-overlapping windows of two seconds length for the EEG and respiration signals separately. The features thus extracted jointly formed a 2-dimensional feature space to which the HMM was then applied. As seen in figure (9), the model clearly shows a preference for a 4-dimensional state space. Figure (10) shows a data section with the corresponding Viterbi state sequence. The data is segmented predominantly into the following regimes: segments of arousal from sleep, wake state with rapid respiration, and two sleep states different only in the EEG micro-structure.

### 4.4 Heart Beat Intervals

In order to apply the Poisson model of the HMM, we took a sequence of RR-intervals, obtained from the RR-Data base at UPC, Barcelona. A subject (Identifier RPP1, Male, Age: 25 years, Height: 178cm, Weight: 70kg) underwent a controlled respiration experiment. While sitting, the subject took 6 deep breaths between 30 and 90 seconds after recording onset (ECG signal sampling rate is  $1kHz$ ). The optimal segmentation was found to be 3 states and is shown in figure (11). The top plot depicts the original RR-interval time series and the middle plot the state labels resulting from the Viterbi path. The segmentation based on the Viterbi path is shown

<sup>3</sup> A breathing disorder in which bursts of fast respiration are interspersed with breathing absence.



Fig. 9. Model order for CS-data: Variational free energy and model probability.

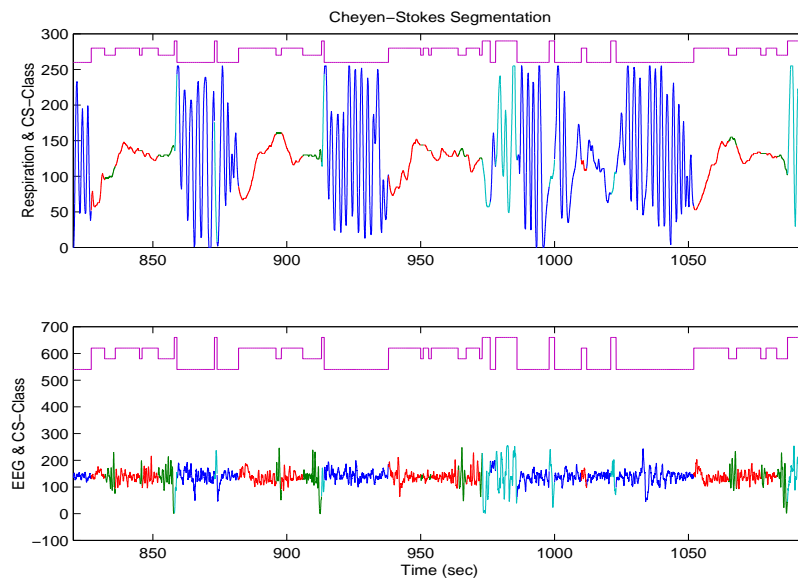
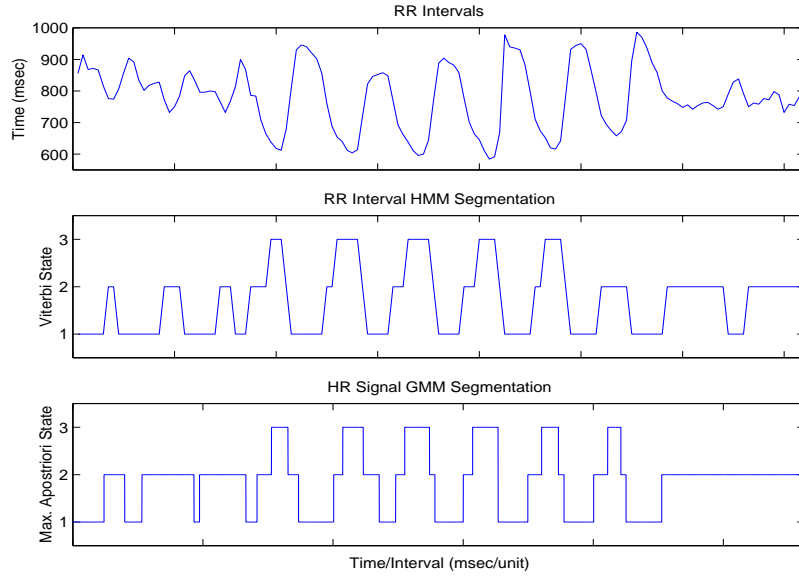


Fig. 10. CS-data: Respiration and EEG Signals with their respective segmentation.

in the bottom plot. A change in state dynamics is clearly visible in the state sequence between 30 and 90 seconds, i.e. the period during with the subject took deep breaths. The state dynamics parallel those observed in the heart-rate signal, obtained from the RR-intervals by interpolation. The difference, however, is that the state sequence is essentially a smoothed version of the heart-rate signal as several heart-rate levels will fall into one state. In addition, no interpolation as such is done, i.e. the HMM derives the heart rate statistically, which is in deep contrast to traditional heart-rate analysis.

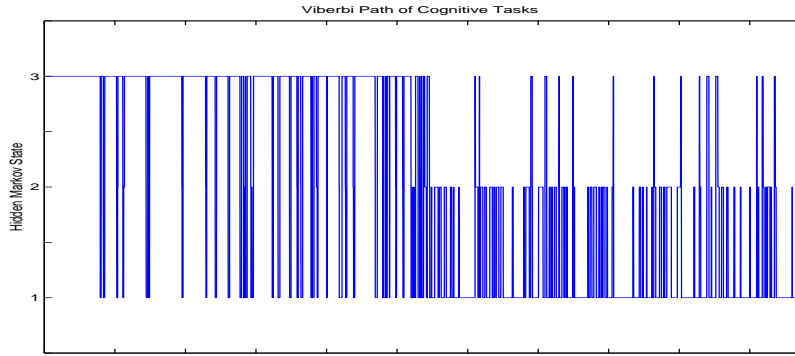


**Fig. 11.** HMM 3-Stage Viterbi Segmentation for subject RPP1 during controlled respiration experiments.

#### 4.5 Segmentation of Cognitive tasks

The idea of the brain computer interface (BCI) experiment is that we infer the unknown cognitive state of a subject from his brain signals which we record via surface EEG. The data in this study were obtained with an ISO-DAM system using a gain of  $10^4$  and a fourth order band pass filter with pass band between  $0.1Hz$  and  $100Hz$  and sampled at  $384Hz$  with 12 bit resolution. The BCI experiments were done by several young, healthy and untrained subjects who performed auditory imagination and imagined spatial navigation tasks. Each task was done for 7 seconds with an experiment consisting of 10 repetitions of alternating these tasks. The recordings are taken from 2 electrode sites: T4, P4. The ground electrode is placed just lateral to the left mastoid process.

We train the HMM with a linear observation model on the EEG of one subject. We used the first 5 repetitions the auditory imagination and imagined spatial navigation task and computed the Viterbi path for 2 further repetitions of each task. Thus 70s of data constituted the training data, while the test data was 28s long. The Viterbi path for the (optimal) three state model is shown in figure (12). The experiment shows that there is a clear change in state dynamics between the different cognitive tasks. We obtain one model almost entirely allocated to the auditory task and two models that are almost exclusively used in the navigation task.



**Fig. 12.** HMM 3-Stage Viterbi Segmentation for two Cognitive tasks, corresponding to the first and second half of the recording, respectively (total duration 28s).

## 5 Conclusion

The goal of the variational approach applied to learning HMMs was, first, to improve in some aspect of the traditional maximum likelihood method and, secondly, to find a unifying view of for deriving all variables, hidden states and parameters. In most cases, the existence of priors densities over the parameters resulted in an improved stability of the model. In some cases the automatic pruning effects of the estimators, which being completely consistent with the theory, is a nice added feature. By deriving all update equations from one single cost function it has also become clear, that the maximum likelihood method is a point estimate in the model parameters, while actually Bayesian in the hidden state space. The variational approach, on the other hand, is consistent from a theoretical point of view and it casts light on the mathematical origin of the Baum-Welch recursions.

While the variational estimators have so far proved to be much more robust than the maximum likelihood based estimators, they also come with a price tag, i.e. the number of parameters increased considerably. It no longer enough to estimate single parameter values, but their distributions. In practice this means that estimator initialisation is considerably more involved.

The examples presented in this paper do not permit any conclusions as to the quality of the estimator, for example with regards to stability of the solutions found or sensitivity to initial conditions and priors. Much is left to understand the estimators better, however, the initial results seem quite promising.

### Acknowledgement

The authors would like to thank R. Conradt for her invaluable contributions to this research and L. Pickup for her help in typesetting. I.R. is supported by the UK EPSRC to whom we are most grateful. We also like to thank Dr. M.A.García González from the Electronic Engineering Department at the Polytechnique University of Catalonia in Barcelona for providing the RR-Interval recording (available at [http://petrus.upc.es/~wwwdib/people/GARCIA\\_MA/database/main.htm](http://petrus.upc.es/~wwwdib/people/GARCIA_MA/database/main.htm)).

### References

- [1] J.M. Bernardo and A.F.M. Smith. *Bayesian Theory*. John Wiley and Sons, 1994.
- [2] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. John Wiley & Sons, New York, 1991.
- [3] A. Flexer, G. Dorffner, P. Sykacek, and I. Rezek. An automatic, continuous and probabilistic sleep stager based on a hidden markov model. *Applied Artificial Intelligence*, 16(3):199–207, 2002.
- [4] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, 2000.
- [5] A.K. Gupta and D.K. Nagar. *Matrix Variate Distributions*. Number 104 in Monographs and Surveys in Pure and Applied Mathematics. Chapman & Hall/CRC, 2000.
- [6] M. Haft, R. Hofmann, and V. Tresp. Model-Independent Mean Field Theory as a Local Method for Approximate Propagation of Information. *Computation in Neural Systems*, 10:93–105, 1999.
- [7] D. Heckerman. A Tutorial on Learning With Bayesian Networks. Technical Report MSR-TR-95-06, Microsoft Research, 1995.
- [8] T.S. Jaakkola. Tutorial on Variational Approximation Methods. In M. Opper and D. Saad, editors, *Advanced Mean Field Methods: Theory and Practice*. MIT Press, 2000.
- [9] T.S. Jaakkola and M.I. Jordan. Improving the Mean Field Approximation Via the Use of Mixture Distributions. In M.I. Jordan, editor, *Learning in Graphical Models*. Kluwer Academic Press, 1997.
- [10] M.I. Jordan, Z. Ghahramani, T.S. Jaakkola, and L.K. Saul. An Introduction to Variational Methods for Graphical Models. In M.I. Jordan, editor, *Learning in Graphical Models*. Kluwer Academic Press, 1997.

- [11] B. Kemp. *Model-based monitoring of human sleep stages*. PhD thesis, Twente University of Technology, The Netherlands, 1987.
- [12] B. Obermeier, C. Guger, C. Neuper, and G. Pfurtscheller. Hidden Markov models for online classification of single trial EEG. *Pattern Recognition Letters*, 22: 1299–1309, 2001.
- [13] C. Pastelak-Price. Das internationale 10-20-System zur Elektrodenplatzierung: Begründung, praktische Anleitung zu den Meßschritten und Hinweise zum Setzen der Elektroden. *EEG-Labor*, 5:49–72, 1983.
- [14] W.D. Penny, R. Everson, and S.J. Roberts. Hidden markov independent component analysis. In M Girolami, editor, *Advances in Independent Component Analysis*. Springer Verlag, 2000.
- [15] W.D. Penny and S.J. Roberts. Dynamic Models for Nonstationary Signal Segmentation. *to appear in Computers and Biomedical Research*, 32(6), 1998.
- [16] L. R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceeding of the IEEE*, 77(2):257–284, 1989.
- [17] A. Rechtschaffen and A. Kales (Eds). *A manual of standardized terminology, techniques and scoring system for sleep stages in human subjects*. U.S. Public Health Service, U.S. Government Printing Office, Washington D.C., 1968.
- [18] I. Rezek and S.J. Roberts. Stochastic Complexity Measures for Physiological Signal Analysis. *IEEE Transactions on Biomedical Engineering*, 44(9):1186–1191, 1998.
- [19] C.P. Robert, T. Rydén, and D.M. Titterton. Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method. *Journal of the Royal Statistical Society, Series B*, 62(1):57–75, 2000.
- [20] O. Rompelman, J.B. Snijders, and C. van Spronsen. The measurement of heart rate variability spectra with the help of a personal computer. *IEEE Transactions on Biomedical Engineering*, 29:503–510, 1982.
- [21] P. Sykacek and S.J. Roberts. Bayesian Time Series Classification. In T.G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14, 2001.
- [22] N. Ueda, R. Nakano, Z. Ghahramani, and G.E. Hinton. SMEM algorithm for mixture models. *Neural Computation*, 12(9):2109–2128, 2000.
- [23] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. *The HTK Book*. Entropic Ltd, 1995.

## A Model Free Update Equations

Following [6] we re-derive the model free variational learning functionals for continuous distributions.

In general we aim to estimate the following KL-divergence

$$\begin{aligned} D(q||p) &\triangleq \mathcal{F} = \int Q(S) \log \frac{Q(S)}{P(S|X)} dS \\ &= \int Q(S) \log \frac{Q(S)}{P(S, X)} dS - \log P(X), \end{aligned}$$

where  $\log P(X)$  is the data log-likelihood. The second form makes clear that  $D(q||p)$  is bounded from below by  $\log P(X)$  and attains  $\log P(X)$  only if

$$D(Q(S)||P(S|X)) = 0,$$

i.e.  $Q(S) = P(S|X)$ . Given a set of variables  $S = \{S_1, \dots, S_T\}$ , in the mean field scenario we assume that

$$Q(S) = \prod_i^T Q(S_i).$$

The set  $S$  incorporates all possible variables, hidden variables and “hidden” parameters alike. Without loss, we split the set  $S$  in the form  $S = \{S_i, \bar{S}_i\}$ , where  $\bar{S}_i = \{S_1, \dots, S_{i-1}, S_{i+1}, \dots, S_T\}$ , so that  $Q(S) = Q(S_i)Q(\bar{S}_i)$  and  $P(S, X) = P(S_i, X|\bar{S}_i)P(\bar{S}_i)$ . In all cases we have the additional constraint that  $\int Q(S_i) dS_i = 1$ . Thus, we are seeking to minimise

$$\mathcal{F}(S) = \int Q(S) \log \frac{Q(S)}{P(S, X)} dS + \sum_{i=1}^T \lambda_i \left( \int Q(S_i) dS_i - 1 \right),$$

Under the mean field assumption can be expanded and simplified to

$$\begin{aligned} \mathcal{F}(S) &\triangleq \int Q(\bar{S}_i) \log Q(\bar{S}_i) d\bar{S}_i - \int Q(\bar{S}_i) \log P(\bar{S}_i) d\bar{S}_i + \\ &\int Q(S_i) \log Q(S_i) dS_i - \int Q(\bar{S}_i) \log P(S_i, X|\bar{S}_i) d\bar{S}_i + \\ &\sum_{i=1}^T \lambda_i \left( \int Q(S_i) dS_i - 1 \right). \end{aligned}$$

Thus,

$$\frac{d\mathcal{F}(S)}{dQ(S_i)} = \log(Q(S_i)) + 1 - \int Q(\bar{S}_i) \log P(\bar{S}_i, X|\bar{S}_i) d\bar{S}_i + \lambda_i = 0.$$

Integrating the above expression within symmetric integration bounds we obtain a solution for  $\lambda_i$ ,

$$\exp(-1 - \lambda_i)^{-1} = \int \exp \int Q(\bar{S}_i) \log P(S_i, X | \bar{S}_i) d\bar{S}_i dS_i,$$

which can be inserted into the partial the solution for the functional  $Q(S_i)$  to obtain

$$Q(S_i) = \frac{1}{\int \exp \int Q(\bar{S}_i) \log P(S_i, X | \bar{S}_i) d\bar{S}_i dS_i} \exp \int Q(\bar{S}_i) \log P(S_i, X | \bar{S}_i) d\bar{S}_i,$$

or, in short,

$$Q(S_i) \propto \exp \int Q(\bar{S}_i) \log P(S_i, X | \bar{S}_i) d\bar{S}_i.$$

In deriving those equations we derived the derivative of  $F(S)$  using the total differential and thus partial derivatives. Therefore, in optimising one distribution  $Q(S_i)$  all others, are held constant.

## B Derivation of the Baum-Welch Recursions

We start with the KL Divergence (41), again denoting the entire set of hidden state variables by  $S = \{S_1, \dots, S_T\}$  and all the observations by  $X = \{X_1, \dots, X_T\}$ :

$$\mathcal{F} = \int Q(S) \log \frac{Q(S)}{P(S)} dS \quad (41)$$

$$= \int Q(S) \log Q(S) dS - \int Q(S) \log P(S) dS. \quad (42)$$

where the first integral in equation (42) is just the entropy, denoted by  $H(S)$ . For a HMM,

$$P(S, X) = P(S_0) \prod_{t=1}^T P(S_t | S_{t-1}) P(X_t | S_t) = \prod_{t=1}^T \frac{P(X_t, S_t, S_{t-1})}{P(S_{t-1})}$$

For ease of notation, it is sufficient for the moment to assume that each node  $S_t$  has an associated datum  $X_t$ , and we omit the extra variable  $X_t$  in the notation of the joint distribution. Thus

$$P(S) = P(S_0) \prod_{t=1}^T \frac{P(S_t, S_{t-1})}{P(S_{t-1})},$$

and identically for the  $Q$  joint distribution:

$$Q(S) = Q(S_0) \prod_{t=1}^T \frac{Q(S_t, S_{t-1})}{Q(S_{t-1})}.$$

Substituting into equation (42), and abbreviating  $l(S) = \log P(S)$ , we have

$$\mathcal{F} = H(S) - \sum_{t=1}^T \int Q(S_t, S_{t-1}) l(S_t, S_{t-1}) dS_{t-1}^t + \sum_{t=1}^{T-1} \int Q(S_t) l(S_t) dS_t$$

where the Entropy term is

$$H(S) = \sum_{t=1}^T \int Q(S_t, S_{t-1}) \log Q(S_t, S_{t-1}) dS_{t-1}^t - \sum_{t=1}^{T-1} \int Q(S_t) \log Q(S_t) dS_t.$$

Before minimising  $\mathcal{F}$  there are some additional constraints required to obtain a consistent solution. These relate to the fact that it must be possible to integrate out over one of the variables in the all of the joint distributions and be left with an identical marginal distribution:

$$\int Q(S_t, S_{t-1}) dS_{t-1} = Q(S_t) = \int Q(S_t, S_{t+1}) dS_{t+1}$$

Introducing Lagrange multipliers for each of these constraints, the full expression for  $\mathcal{F}$  becomes

$$\begin{aligned} \mathcal{F} = & \sum_{t=1}^T \int Q(S_t, S_{t-1}) \log Q(S_t, S_{t-1}) dS_{t-1}^t - \sum_{t=1}^{T-1} \int Q(S_t) \log Q(S_t) dS_t - \\ & \sum_{t=1}^T \int Q(S_t, S_{t-1}) l(S_t, S_{t-1}) dS_{t-1}^t + \sum_{t=1}^{T-1} \int Q(S_t) l(S_t) dS_t + \\ & \lambda_{t,t-1}(S_t) \left( Q(S_t) - \sum_{S_{t-1}} Q(S_t, S_{t-1}) \right) + \\ & \mu_{t,t-1}(S_{t-1}) \left( Q(S_{t-1}) - \sum_{S_t} Q(S_t, S_{t-1}) \right). \end{aligned}$$

Differentiating with respect to  $Q(S_t, S_{t-1})$

$$Q(S_t, S_{t-1}) = \frac{1}{z_{t,t-1}} e^{l(S_t, S_{t-1})} e^{\lambda_{t,t-1}(S_t)} e^{\mu_{t,t-1}(S_{t-1})} \quad (43)$$

where the constants of integration have been re-written in the form of a scaling factor  $\frac{1}{z}$ . Likewise, differentiation with respect to the marginals,  $Q(S_t)$ ,

$$Q(S_t) = \frac{1}{z_t} e^{l(S_t)} e^{\lambda_{t,t-1}(S_t)} e^{\mu_{t+1,t}(S_t)} \quad (44)$$

There are two joint distributions defined over  $Q(S_t)$ , namely  $Q(S_t, S_{t-1})$  and  $Q(S_t, S_{t+1})$ . By marginalising out  $S_{t-1}$  and  $S_{t+1}$ , in (43), results in

$$Q(S_t) = \sum_{S_{t-1}} \frac{1}{z_{t,t-1}} e^{l(S_t, S_{t-1})} e^{\lambda_{t,t-1}(S_t)} e^{\mu_{t,t-1}(S_{t-1})} \quad (45)$$

and

$$Q(S_t) = \sum_{S_{t+1}} \frac{1}{z_{t+1,t}} e^{l(S_{t+1}, S_t)} e^{\lambda_{t+1,t}(S_{t+1})} e^{\mu_{t+1,t}(S_t)} \quad (46)$$

Equating the expressions for  $Q(S_t)$  from (44) and (46) one can solve for the first Lagrange multiplier,

$$e^{\lambda_{t,t-1}(S_t)} = z_t e^{-l(S_t)} \sum_{S_{t+1}} \frac{1}{z_{t+1,t}} e^{l(S_{t+1}, S_t)} e^{\lambda_{t+1,t}(S_{t+1})} \quad (47)$$

and, similarly, equating the expressions for  $Q(S_t)$  from (44) and (45):

$$e^{\mu_{t+1,t}(S_t)} = z_t e^{-l(S_t)} \sum_{S_{t-1}} \frac{1}{z_{t,t-1}} e^{l(S_t, S_{t-1})} e^{\mu_{t,t-1}(S_{t-1})} \quad (48)$$

Substituting in (47)  $\beta(t)$  for  $e^{\lambda_{t,t-1}(S_t)}$ , gives

$$\beta(t) = \frac{1}{z_t} \sum_{S_{t+1}} P(S_{t+1}|S_t) \beta(t+1),$$

and substitution of  $\alpha(t) = e^{\mu_{t+1,t}(S_t)} P(S_t)$  in (48), gives

$$\alpha(t) = z_t \sum_{S_{t-1}} P(S_t|S_{t-1}) \alpha(t-1)$$

Finally, restating (43) using  $\alpha(t)$  and  $\beta(t)$  leads to the well known equation of the joint distributions

$$\begin{aligned} Q(S_t, S_{t-1}) &= \frac{1}{z_{t,t-1}} e^{l(S_t, S_{t-1})} e^{\lambda_{t,t-1}(S_t)} e^{\mu_{t,t-1}(S_{t-1})} \\ &= \frac{1}{z_{t,t-1}} P(S_t, S_{t-1}) \beta(t) \left( \frac{\alpha(t-1)}{P(S_{t-1})} \right) \\ &= \frac{1}{z_{t,t-1}} P(S_t|S_{t-1}) \beta(t) \alpha(t-1) \end{aligned}$$

## C Complete KL divergences

To monitor the convergence of the variational algorithm and to test for the best model order/size, requires the calculation of the complete KL divergence (12), given the data. The general overall KL divergence (12) can be split into the following 3 terms,

$$\begin{aligned} \mathcal{F} = & - \underbrace{\int q(S) q(\theta) \log P(X, S|\theta) \, dS \, d\theta}_{\text{Average Log-likelihood}} \\ & + \underbrace{\int q(S) \log q(S) \, dS}_{\text{Negative Entropy}} + \underbrace{\int q(\theta) \log \frac{q(\theta)}{P(\theta)} \, d\theta}_{\text{KL-Divergence}} \end{aligned} \quad (49)$$

All terms vary depending on the observation model, with the exception from the negative entropy term, which only changes if HMM topology changes. The KL divergence measures the divergence between the prior and approximate posterior distributions. Since all the models here are within the exponential family, to mathematical form of many KL divergence terms occur repeatedly. Hence we list these forms separately and refer back to them when needed.

### C.1 Negative Entropy

The neg-Entropy term for HMMs is given

$$H_{HMM} = H(S_{t=0}) + \sum_{t=1}^T H(S_t|S_{t-1}) \quad (50)$$

$$= \sum_{t=1}^T H(S_t, S_{t-1}) - H(S_t) \quad (51)$$

### C.2 KL-Divergences

The KL-divergence in equation (49) measures the divergence between the prior and approximate posterior distributions. Many of them occur repeatedly. Given two densities,  $Q$ , and  $P$ , which have parameters indexed by  $q$  and  $p$ , and using the notation of [1] and [5], the KL-divergences between two Dirichlet, Wishart and multi-variate Normal densities are given as follows:

- Between two Dirichlet densities

$$D_{Dir}(q||p) = \log \left( \frac{\Gamma(\sum_{l=1}^k \alpha_{ql})}{\Gamma(\sum_{l=1}^k \alpha_{pl})} \right) + \sum_{l=1}^k \log \frac{\Gamma(\alpha_{pl})}{\Gamma(\alpha_{ql})} \\ + \sum_{l=1}^k (\alpha_{ql} - \alpha_{pl}) \left( \Psi(\alpha_{ql}) - \Psi \left( \sum_{l=1}^k \alpha_{ql} \right) \right) \quad (52)$$

- Between two Gamma densities

$$D_G(q||p) = \log \frac{\Gamma(\alpha_p)}{\Gamma(\alpha_q)} + (\alpha_q - \alpha_p) \Psi(\alpha_q) + \alpha_q \log \frac{\beta_q}{\beta_p} + \alpha_q (\beta_p \beta_q^{-1} - 1) \quad (53)$$

- Between two Wishart densities

$$D_W(q||p) = \sum_{l=1}^k \log \frac{\Gamma(\frac{1}{2}(2\alpha_p + 1 - l))}{\Gamma(\frac{1}{2}(2\alpha_q + 1 - l))} + (\alpha_q - \alpha_p) \sum_{l=1}^k \Psi \left( \frac{2\alpha_q + 1 - l}{2} \right) \\ + \alpha_q \log \frac{|B_q|}{|B_p|} + \alpha_q (\text{tr}(B_p B_q^{-1}) - k) \quad (54)$$

- Between two multivariate variate Normal densities

$$D_{\mathcal{MvN}}(q\|p) = \frac{1}{2} \left( \log \frac{\lambda_q}{\lambda_p} - 1 + \lambda_p \lambda_q^{-1} + (\mu_q - \mu_p)^\top \lambda_p (\mu_q - \mu_p) \right) \quad (55)$$

- Between two  $m \times n$  matrix variate Normal densities

$$D_{\mathcal{MavN}}(q\|p) = \frac{1}{2} \left\{ n \log \frac{|\Sigma_q|}{|\Sigma_p|} + m \log \frac{|\Phi_q|}{|\Phi_p|} + \text{tr}(\Phi_p \Phi_q^{-1}) \text{tr}(\Sigma_p \Sigma_q^{-1}) \right. \\ \left. + \text{tr}(\Sigma_p (M_q - M_p) \Phi_p (M_q - M_p)^\top) - nm \right\} \quad (56)$$

### C.3 Gaussian Observation HMM

The average log-likelihood term in (49) for Gaussian observation models is given as

$$\mathcal{L}_{avg} = \sum_m \gamma_{t_0} \Psi(\tilde{\lambda}_{0_m}) - \bar{\gamma}_0 \Psi\left(\sum_l \tilde{\lambda}_{0_l}\right) + \\ \sum_{m,n} \bar{\xi}(m,n) \Psi(\tilde{\lambda}_{m_n}) - \sum_n \bar{\xi}(n) \Psi\left(\sum_l \tilde{\lambda}_{l_n}\right) + \\ T \log(2\pi)^{\frac{k}{2}} + \frac{1}{2} \bar{\gamma}_m \bar{\Psi}_{\tilde{\alpha}_m} - \frac{1}{2} \bar{\gamma}_m \log |\tilde{B}_m| - \\ \frac{1}{2} \sum_t \gamma_{t_m} (x_t - \tilde{\mu}_{m_0})^\top \tilde{\alpha}_m \tilde{B}_m (x_t - \tilde{\mu}_{m_0}) - \frac{1}{2} \bar{\gamma}_m \text{tr}(\tilde{\alpha}_m \tilde{B}_m \tilde{C}_{m_0}) \quad (57)$$

where

$$\bar{\gamma}_0 = \sum_m \gamma_{0_m} , \\ \bar{\gamma}_m = \sum_t \gamma_{t_m} , \\ \bar{\xi}(m,n) = \sum_t \xi_t(m,n) , \\ \bar{\xi}(n) = \sum_t \sum_m \xi_t(m,n) , \quad (58)$$

and

$$\bar{\Psi}_{\tilde{\alpha}_m} = \sum_k \Psi\left(\frac{1}{2}(2\tilde{\alpha}_m + 1 - k)\right) . \quad (59)$$

### C.4 Poisson Observation HMM

The average log-likelihood term for a Poisson observation HMM can be shown to be

$$\mathcal{L}_{avg} = \sum_{m,n} \gamma_{nm} \left[ -x_n \tilde{\alpha}_m \tilde{\beta}_m + y_n \log(x_n) + y_i \left( \Psi(\tilde{\alpha}_m) - \log(\tilde{\beta}_m) \right) - \log(y_n!) \right] \quad (60)$$

The KL divergence between the approximate posterior density  $Q(\mu_m)$ , with parameters  $\tilde{\alpha}_0, \tilde{\beta}_0$ , and the prior  $P(\mu_m)$ , with parameters  $\alpha_0, \beta_0$ , is a standard Gamma density divergence  $D_G(Q(\mu_m) \| P(\mu_m))$ , given by equation (53).

### C.5 Linear Observation Model HMM

The average log-likelihood term for the HMMs with linear observation models is given as

$$\begin{aligned} \mathcal{L}_{avg} = & -N \frac{du}{2} \log(2\pi) - \frac{u}{2} \sum_m \tilde{\gamma}_m \log |\tilde{B}_{\Sigma m}| \\ & + \sum_m \tilde{\gamma}_m \left( \Psi(\tilde{\rho}_m) - \Psi\left(\sum_{m=1}^M \tilde{\rho}_m\right) + \frac{u}{2} \sum_{l=1}^d \Psi\left(\frac{1}{2}(2\tilde{\alpha}_{\Sigma m} + 1 - l)\right) \right) \\ & - \sum_{n,m} \left\{ \gamma_{nm} \frac{1}{2} \text{tr} \left( \tilde{\alpha}_{\Sigma m} \tilde{B}_{\Sigma m}^{-1} (Y_n - \tilde{\Omega}_m X_n)(Y_n - \tilde{\Omega}_m X_n)^\top \right) \right\} \\ & - \frac{1}{2} \sum_m \tilde{\alpha}_{\Sigma m} \text{tr} \left( \tilde{B}_{\Sigma m}^{-1} \tilde{\Sigma}_m^{-1} \right) \sum_n \gamma_{nm} \text{tr} \left( X_n X_n^\top \tilde{\Phi}_m^{-1} \right) \end{aligned} \quad (61)$$

The KL-divergences are given as

$$D(Q(\theta) \| P(\theta)) = \sum_m \left\langle D_{\mathcal{M}av\mathcal{N}}(Q(H_m) \| P(H_m | \Sigma_m, \Phi_m)) \right\rangle_{Q(\Sigma_m, \Phi_m)} \quad (62)$$

$$+ \sum_m D_{\mathcal{W}}(Q(\Sigma_m) \| P(\Sigma_m)) \quad (63)$$

$$+ \sum_m D_{\mathcal{W}}(Q(\Phi_m) \| P(\Phi_m)) \quad (64)$$

$$+ D_{\mathcal{D}ir}(Q(\kappa) \| P(\kappa)) \quad (65)$$

where divergences (63), (64) and (65) are given by (54), (54) and (52), respectively, and divergence (62) is given by

$$\begin{aligned}
& \left\langle D_{\mathcal{M}a\mathcal{V}\mathcal{N}}(Q(H_m) \| P(H_m | \Sigma_m, \Phi_m)) \right\rangle_{Q(\Sigma_m, \Phi_m)} = \\
& = \frac{1}{2} \left\{ (dp) \log |\tilde{\Sigma}_m| + d \log |\tilde{\Phi}_m| - d^2 p \right. \\
& \quad - (dp) \left( \sum_{l=1}^d \Psi \left( \frac{2\tilde{\alpha}_{\Sigma_m} + 1 - l}{2} \right) - \log |\tilde{B}_{\Sigma_m}| \right) \\
& \quad - d \left( \sum_{l=1}^{dp} \Psi \left( \frac{2\tilde{\alpha}_{\Phi_m} + 1 - l}{2} \right) - \log |\tilde{B}_{\Phi_m}| \right) \\
& \quad + \text{tr} \left( \tilde{\alpha}_{\Phi_m} \tilde{B}_{\Phi_m}^{-1} \tilde{\Phi}_m^{-1} \right) \text{tr} \left( \tilde{\alpha}_{\Sigma_m} \tilde{B}_{\Sigma_m}^{-1} \tilde{\Sigma}_m^{-1} \right) \\
& \quad \left. + \text{tr} \left( \tilde{\alpha}_{\Sigma_m} \tilde{B}_{\Sigma_m}^{-1} (\tilde{\Omega}_m - \Omega) \tilde{\alpha}_{\Phi_m} \tilde{B}_{\Phi_m}^{-1} (\tilde{\Omega}_m - \Omega)^{\top} \right) \right\} \tag{66}
\end{aligned}$$