Ensemble Coupled Hidden Markov Models for Joint Characterisation of Dynamic Signals

Iead Rezek, Stephen Roberts, Peter Sykacek
Pattern Analysis & Machine Learning Research Group,
Information Engineering,
University of Oxford, UK.
{irezek,sjrob,psyk}@robots.ox.ac.uk

Abstract

How does one model data with the aid of labels, when the labels themselves are noisy, unreliable and have their own dynamics? How does one measure interactions between variables that are so different in their nature that a direct comparison using, say cross-correlations, is meaningless? In this paper these problems are approached using Coupled Hidden Markov Models which are estimated in the Variational Bayesian framework. Signals can be diverse since each chain has its own observation model. Signals can have their own dynamics and may temporally lag or lead one another by allowing linking edges in the network topology to be estimated and chosen according to the most probable posterior model. Integrated feature extraction and modelling is accomplished by providing the Markov models models with linear observations models. We derive Coupled Hidden Markov Models estimators, apply and compare them with sampling based approaches found in the literature.

1 Motivation

When describing signals, ideally the models should mimic the underlying processes that are thought to generate them. The advantage is that the results are easily interpretable and applicable. In the absence of such models, or if such models are mathematically or numerically intractable, simpler models are used, which might have the disadvantage of producing results that do not easily relate to our understanding. One can, however use markers or scores (manual or otherwise) which guide the model in the desired direction. How one chooses to apply such labels varies. In classification, labels are frequently assumed to be noise-less and exact. They are given almost dictatorial power which might be appropriate depending on the mechanism that generated the class labels. Human scores, on the other hand, are everything but noiseless. Humans are subjective in their interpretation of the system they study. This frequently requires the combination of several human scores to a consensus score. Human scores even have their own dynamical properties. In the case of human sleep stage scoring, it is typical for the human scorers to lag somewhat behind a sleep state change before they label it as such, as if to make sure the state has indeed changed. To complicate matters even further, scores might have a different sampling rate than the actual signal and might characterise a state based only on a short event which occurred within the sampling period.

We believe therefore, particularly when dealing with manually generated “class labels”, one should really avoid treating labels as such and instead consider them as manifestations of the same underlying system which also produced the observed time series - only with a different mapping onto the observation space. Thus, labels are observables which are permitted to be noisy. Characterising the system is, in this case, done by fusing a set of observations (e.g. labels and signal) at a higher (state space) level.

2 Problem Approach

There is very little knowledge with regard to the biological processes governing most of the applications described in this paper. Hence, we cannot resort to specific models such as those found in the literature of nonlinear-dynamic theory (we’re thinking along the lines of Van der Pol oscillators and the like). Here we make use of another kind of state space models, that of Markov models. A kind of Markov model which incorporate class labels are Markov Decision Trees. From a generative model perspective, however, the labels uniquely define the state and the signal, i.e. they hard-sectioning the observation space and do not allow for label uncertainty. Label dynamics are also typically not accounted for. Here we suggest a different class of Markov models: coupled Markov models, in particular, coupled hidden Markov models (CHMM) [3] in which the labels are modelled as observables generated via one chain and signals by other neighbouring chains. The CHMM thus accounts for individual dynamic properties of labels and signals. Labels and signals are then fused in state space. The time delay between them is estimated to allow for lagging or lead behaviour between the chains. Labels in the classical sense are discrete. By allowing multivariate discrete observation models for the labels, an optimal concensus score can be statistically estimated within the overall framework of parameter estimation. Indeed, labels need not be conventional scores, but can be markers of any kind, e.g. muscle activity signals to describe the state of hand movement in brain computer interface applications.

A typical way of using hidden Markov models (HMM) is by extracting features from the observation time series and then using these as the "observed" time series to train the HMM parameters. Thus, one in fact conditions on the features and thus expects the model to perform worse than when allowing for feature uncertainty and integrating them out in a Bayesian sense. Such an approach, using Markov Chain Monte Carlo sampling, has already been shown to give significantly improved classification results [13]. In this paper we perform full Bayesian analysis using a variational learning framework. In variational learning theory all aspects of HMMs can be estimated within the same framework, be it belief propagation, parameter estimation or model selection. By approximating the full posterior distribution of the parameters, we avoid many of the problems of the maximum likelihood framework. In addition, the free energy provides a measure for the choice of the best model.

3 Variational learning of Coupled Hidden Markov Models

Variational learning aims to minimise the variational free energy [7] between the (intractable) model posterior P and a simpler (analytic) approximating distribution Q. The free energy is given as the Kullback-Leibler (KL) divergence between Q and P, where the distribution Q is defined over the hidden variables, such as parameters or hidden states. Being a directed divergence, it is an upper bound to the true log-probability of the data, i.e. the evidence. The divergence is maximised with respect to the individual distributions.

Given a set of hidden variables $A = \{A_1, \cdots, A_L\}$, the method known as "Mean Field" variational approximation assumes that the Q-distributions factorise, i.e. $Q(A) = \prod_{l=1}^{L} Q(A_l)$ with the additional constraint that $\int Q(A_l)\ dA_l = 1$. In this paper, the CHMM parameters are assumed to be independent form each other, i.e. $Q(\boldsymbol{\theta}) = \prod_{j=1}^{M} Q(\boldsymbol{\theta}_j)$ while the hidden state sequence maintains its chain-like structure. In the case of 2 coupled chains, the hidden states are composed of subsets, $S = \{S_1, \cdots, S_N\}$ and $T = \{t_1, \cdots, T_N\}$ which form a joint distribution of the form

$$Q(S,T) = Q(S_1)Q(T_1) \prod_{n=2}^{N} Q(S_t|S_{t-1}, T_{t-l_S})Q(T_t|T_{t-1}, S_{t-l_T}) \tag{1}$$

where l_S is a time delay or lag from chain T onto S and l_T is the lag from chain S onto T.

The value of the free energy varies depending on the number of state space dimensions, the lag between the state chains and the observation model order. Indexing the free energy for each possible setting of these quantities by a, we can choose the most probable model from the relation [7]

$$\alpha_a \propto \exp\{-\mathcal{F}_a\}. \tag{2}$$

In effect the form reduces to computation of the posterior model probability in the event of flat priors over model structures [2].

One example of a directed graph of the CHMM is shown in figure 1. The graph has 2 observation models, one multinomial for the discrete set of label observations and one multivariate linear observation model for stochastic signals. The discrete observation model's random variable λ_Z has a Dirichlet prior with prior counts κ_z. The update formulae for this model are particularly simple and consist of simply adding all the probabilities of the states T_n for all values falling within a category. The linear observation model is a multivariate autoregressive (AR) model, i.e. $\vec{y}_t = \sum_{l=1}^{p} \bar{\boldsymbol{H}}_l \vec{y}_{t-l} + \vec{\boldsymbol{e}}_t$, where $\vec{y}_t \in \mathbb{R}^{d \times 1}$ is a d-variate response vector at time t, $\bar{\boldsymbol{H}}_p \in \mathbb{R}^{d \times d}$ is the matrix of model coefficients, and $\vec{\boldsymbol{e}}_n \in \mathbb{R}^{d \times 1}$ is a stochastic noise vector. The model is simplified by concatenating the p coefficient matrices into one partitioned matrix, $\boldsymbol{H} \in \mathbb{R}^{d \times dp}$, and thus $\vec{y}_t = \boldsymbol{H}\vec{x}_t + \vec{\boldsymbol{e}}_t$, where $\vec{x}_t \in \mathbb{R}^{dp \times 1}$ is a dp-variate basis vector at time t which, in the AR-model case, is just a vector with stacked lagged samples of original time series, $x_t = [\vec{y}_{t-1}{}^{\mathsf{T}} \cdots, \vec{y}_{t-p}{}^{\mathsf{T}}]^{\mathsf{T}}$. To overcome the problem of different sampling rates between the two chains, we assume the same linear model for a set or segment of multivariate observations. This has the advantage of implicitly applying some smoothing to the observations and is mathematically simpler than tackling the problem in the state space. Thus, in a segment of observations, indexed by n, u samples of $\vec{y}_t$ are concatenated to a single matrix $Y_n = [\vec{y}_{(n-1)u}, \cdots, \vec{y}_{nu}]$. The same number u of basis vectors x_t are concatenated to give $X_n = [\vec{x}_{(n-1)u}, \cdots, \vec{x}_{nu}]$. It follows that $Y_n \in \mathbb{R}^{d \times u}$ and $X_n \in \mathbb{R}^{dp \times u}$.

The discrete nature of the state space means that the linear model distribution has discrete parents which results in a mixture of linear models . The model, given the state $S_n = m$, becomes

$$p(Y_n - \boldsymbol{H}_m X_n) = \mathcal{N}_{d,u}(0, \boldsymbol{\Sigma}_m, I_u) \tag{3}$$

where $\mathcal{N}_{d,u}$ defines a $d \times u$-matrix variate normal density function [4] with precision $\boldsymbol{\Sigma}_m$ and I_u is a $u \times u$ identity matrix. The prior densities for the coefficient matrices $\boldsymbol{H}_m$ are assumed to be a $d \times dp$-matrix variate normal densities, $\mathcal{N}_{d,dp}(\Omega, \boldsymbol{\Sigma}_m, \boldsymbol{\Phi}_m)$, with mean Ω and precisions $\boldsymbol{\Sigma}_m$ and $\boldsymbol{\Phi}_m$. The priors for residual precisions $\boldsymbol{\Sigma}_m$ and the coefficient precisions $\boldsymbol{\Phi}_m$ are Wishart densities [1], $\mathcal{W}_d(\alpha_\Sigma, B_\Sigma)$ and $\mathcal{W}_{dp}(\alpha_\Phi, B_\Phi)$, with shape/scale parameters α_Σ/B_Σ and α_Φ/B_Φ, respectively.

All approximating Q-distributions over model parameters are assumed to factorise. This is the mean field case and thus, if all prior distributions are chosen to be conjugate, the posterior distributions, $Q(\boldsymbol{\theta})$, are functionally identical to the prior distributions [6]. To avoid confusion, we denote the parameters of $Q(\boldsymbol{\theta})$ with tildes, e.g. $Q(\boldsymbol{\Sigma}_m) = \mathcal{W}_d(\tilde{\alpha}_{\Sigma_m}, \tilde{B}_{\Sigma_m})$. With all the prior and posterior distributions in place, we can compute the update equations using the generic formulae in [5]. The equations for the linear observations models are given in the appendix, as is the free-energy formula.

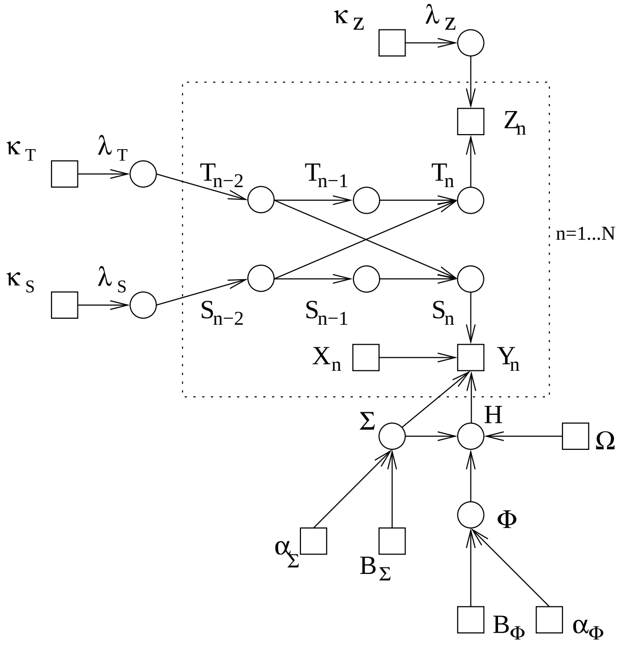


Figure 1: Directed graph of 2 chain CHMM with one chain having a multinomial observation model (for observations Z) and the second chain a multivariate linear model (for observations Y). In this particular example both lags between the two chains are symmetric and have a value of 2.

3.1 Estimation

Model Parameters: By taking the derivatives of the free energy with respect to the distributions of the unknown parameters, we obtain a set of update formulae for the parameters of the distributions, which are given in the appendix.

Hidden States: The hidden variables (i.e. the state sequence) can be estimated using standard forward-backward message passing [9], conditioned on the data and the expectations of the model parameters under the Q-distributions. The use of the forward-backward recursions is justified by the fact that message passing equations are fixed point equations of the free-energy when the Q-distributions are assumed to be of the form given in equation (1) [14] [8]. In the case of CHMMs, however, message passing can only be employed after prior node clustering. With increasing lag, this becomes exponentially expensive. Thus, for lags greater than 1 we sample the hidden state sequence. The process is reasonably fast since one can sample the hidden state space densities directly and need not accumulate a much larger number of state space variable values to compute empirically the parameters of their distributions.

State Space Dimension and Lag: Estimation is performed over several state space dimensions. Given a fixed state space dimension estimation involves iterative application of forward-backward message passing, update of the model parameters, and estimation of the free energies. The free energies, obtained for each state-space dimension are then used to evaluate the highest probability model according to equation (2) and thus the optimal state space dimension. At present the lag estimation is simply a search within the bounds beyond which we believe there is no interaction between the chains. Again, the free energies, obtained for each lag value are used evaluate the highest probability model.

4 Experiments

4.1 Model Selection with Synthetic Data

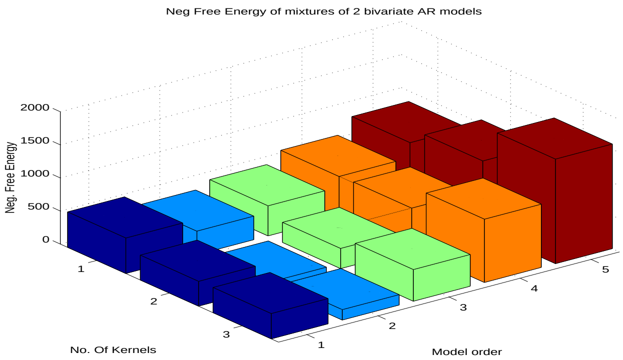
We begin by investigating the free energy behaviour as state-space dimension, linear observation model order and state-space lag are varied. The 1024 samples of synthetic data were drawn from a model with 2 multivariate autoregressive models, with coefficient matrices

$$H_1 = \begin{pmatrix} 0.4 & 1.2 & 0.35 & -0.3 \\ 0.3 & 0.7 & -0.4 & -0.5 \end{pmatrix}$$

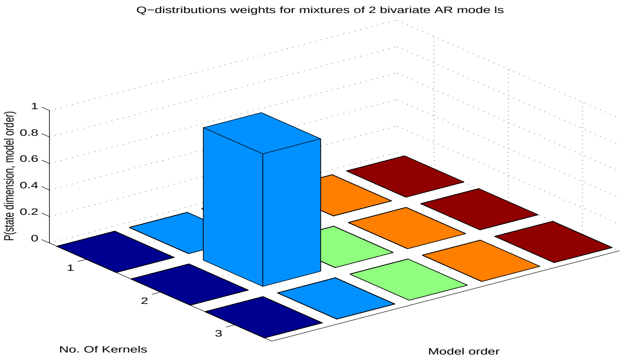
and

$$H_2 = \begin{pmatrix} 0.4 & 0 & 0.35 & 0 \\ 0 & 0.7 & 0 & -0.5 \end{pmatrix}$$

and unit covariance stochastic noise. The estimates of the free energies for a number of settings of model order and size of state spaces are shown in figure 2a. The posterior model probability based on these free energies is shown in figure 2b. The preferred combination is clear. Next, we investigate the free energy for a dual chain CHMMs with different lag topologies, i.e. where the edge between chain S and chain T varies. This time, the data was generated by drawing 1024 samples from a CHMM with additive Gaussian noise and with state lags $l_S = 2$ and $l_T = 2$. Figure 3 shows the free energy values reaching a minimum at the lag for which the data was generated[1].



(a) Free Energy



(b) Model Posterior Probability

Figure 2: Free Energies and resulting Model Posterior Probability for various settings of state space dimensions and linear model orders for synthetic data.

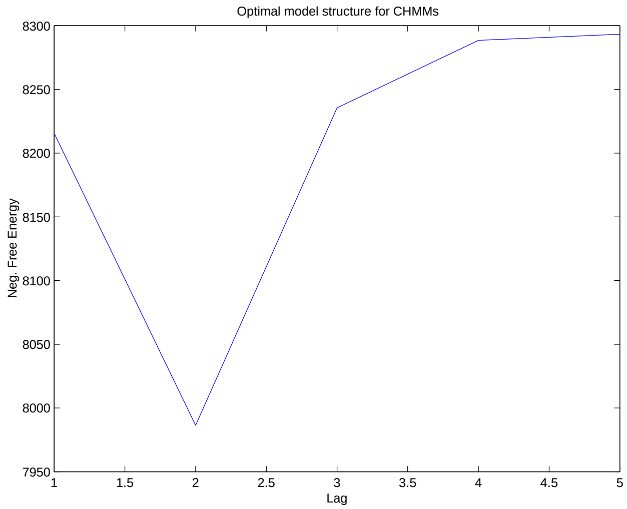


Figure 3: Free Energy for a 2-chain CHMM with different symmetric lags, i.e. $l_S = l_T$.

4.2 Periodic Respiration

We also applied the CHMM to features extracted from a section of Cheyne Stokes Data [2], consisting of one EEG recording and a simultaneous respiration recording, both sampled at $128Hz$. The data was used as it has clearly defined states and thus allows verification of the results of the unsupervised learning algorithm. The feature, the fractional spectral radius (FSR) [11], was computed from consecutive non-overlapping windows of two seconds length for the EEG and respiration signals separately. Each features thus extracted formed the observation sequence of a CHMM chain. A Gaussian observation model was fitted to each chain. The minimum state space dimension is 2 for each chain and thus 4 overall. This was also the dimension the model preferred. As for the lag, the free energy changed little over the range of $l_S = L_T = 1, \cdots, 3$. This suggests that the choice using windows to estimate the features would result and state changes within these to be merged into a single feature measure. With no probabilistic treatment of this feature, conditioning on it does not allow for uncertainty which then has to be absorbed by the Markov model. Overall, the results are very similar to those of a single HMM with 4-dimensional state space using state-space clustering [12], suggesting really no overall lead or lag preference between the channels. Figure (4) 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. Unlike any maximum *apriori* [10] or maximum likelihood method [12], however, the variational implementation results in automatic pruning of states not supported by the data. Thus, we can start the algorithm with any large number states and leave the algorithm to converge to the result shown above.

[1] For simplicity, we only showed values for which $l_S = l_T$. Also, although we have not described the case of a Gaussian observation model, it is clear that by setting the basis vector to be just a scalar of value 1, collapsing the coefficient matrices to a vector, and fixing the number of segments to just $u = 1$, one recovers a simple multivariate Gaussian observation model. The update equations can be found in the appendix

[2] A breathing disorder in which bursts of fast respiration are interspersed with breathing absence.

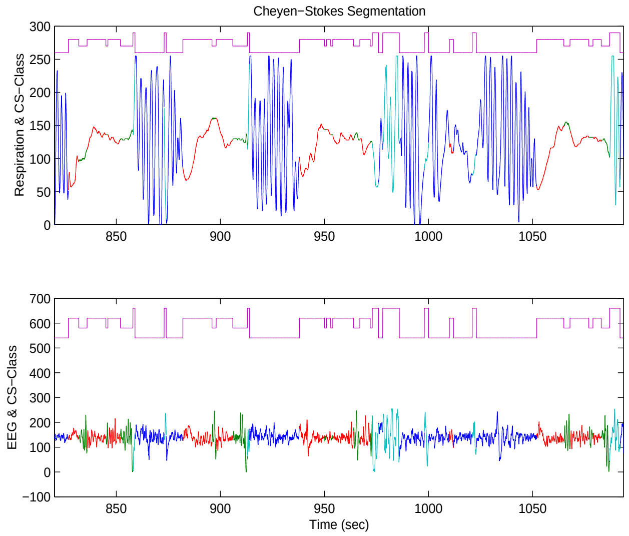


Figure 4: CS-data: Respiration and EEG Signals with their respective segmentation.

4.3 Classification with Synthetic Data

In this experiment we classify an artificial data set. The target labels were drawn form a first order Markov sequence with 2 state values. Conditioned on the current state 12 samples were drawn from a linear model using reflection coefficients $(0.8, -0.8, 0.5)$ and $(0.9, -0.7, 0.6)$ with driving noise of variance 1. To compare our model to that in [13], the labels were also corrupted and 20% of them were replace by white noise which was uncorrelated with the class labels. Table 1 shows the generalisation errors together with those obtained from the comparable model in [13] using sampling for parameter estimation. The variational and sampling results are comparable and it was pointed out in [13] that these sampling results showed a significant improvement to those obtained by conditioning on a priori extracted features.

4.4 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 0.1 Hz and 100 Hz and sampled with 384 Hz and 12 bit resolution. The BCI experiments were done by several young, healthy and untrained subjects, who did two task pairings: auditory-imagination and left-right motor imagination tasks. Each task was done for 7 seconds with an experiment consisting of 10 repetitions of alternating tasks. The recordings are taken from 3 electrode sites: T4, P4 (right tempero-parietal for spatial and auditory tasks), C4', C4" (right motor area for left motor imagination), and C3', C3" (left motor area for left motor imagination). The ground electrode is placed just lateral to the left mastoid process.

We train the CHMM on the EEG of one subject. We used a 7th order linear model order, a state space dimension of 2 and lag of 1 on a total of 10 repetitions the experiments[3]. The results of a 10 fold validation are shown in table 1. Since estimated labels were obtained by integrating out the neighbouring chain's contribution to the state transition probability before computing the Viterbi path on the test data. Again, the results are comparable although weaker than if we had an outlier model included as in [13].

Table 1: Generalisation Accuracies for sampling and variational experiments. Sampling results were taken from [13]

experiment	variational	Sampling
Synthetic	93.6%	87.2%
left vs. right motor	79.1 %	79.5 %
auditory vs. navigation	80.0 %	78.4 %

5 Conclusions and Reflections

The most attractive features of estimating CHMMs in the variational framework is that one avoids many of the pitfalls of maximum likelihood methods, such as singular covariance matrices because no data has been allocated to a particular state. This is neatly avoided when approximating the full posterior probability of the all model parameters. The estimation of linear observation models is also much more robust, since parameters which are not supported by the data are simply set to their prior values [4]. We can also use some post-hoc investigation of the coefficient covariance matrices to determine the effective model order (by finding the knee in the matrix eigen spectrum). Sampling the state sequence within the variational framework empirically proved to be adequate and gave results comparable to other methods. Furthermore, the variational algorithm clearly approaches the best possible solution as confirmed when compared with the results obtained with MCMC sampling. However, it out-performs the sampling algorithm definetly, in terms of speed - convergence is much faster in the variational framework.

Although the free energies correspond nicely to the optimal lag and other parameter settings, the situation in reality is less clear cut. For the lag in particular, one often obtains a large range of small free energy values and one is somewhat free to choose the "optimal". Consider, for instance,

[3] Some sections within which the signal strength resulted in amplifier saturation (i.e. sections of flat lines) had to be removed beforehand. They caused the estimation of determinants to become singular and thus inverses ill-defined.

[4] All within reason, of course, which mathematically means reasonable matrix condition numbers.

when features which are extracted from the original time series using overlapping sliding windows are applied to the Markov model. The features are highly correlated and one cannot expect a sharply peaked free energy spectrum at a particular lag. The estimation processes then becomes more sensitive to noise and outliers.

When comparing our results to those obtained by sampling one thing does become obvious. That is, the effect of outliers or corruption of data. Some data sets used here had short amplitude plateaus as a result amplifier saturation. In contrast to sampling methods, such sections had to be removed as they lead to singular matrix determinants and thus undefined matrix inverses.

Acknowledgement

I.R. & P.S. are supported respectively by the UK EPSRC and the BUPA Foundation to whom we are most grateful. We would like to thank Prof. Maria Stokes of the Royal Hospital for Neurodisability who provided us with the EEG recordings that were used in the experiments section.

A Appendix: Update Equations

In the following we use of the denote the probability of the hidden state variable taking on the value m (out of M possible values) given all the data by $\gamma_{tn} = P(S_n = m|X)$. Also, $\bar{\gamma}_m = \sum_{n=1}^{N} \gamma_{nm}$ and $^\mathsf{T}$ denotes the transpose operation.

The posterior density of the model coefficients, $\boldsymbol{H}_m$, is a $d \times dp$ matrix variate Normal density [4] with mean $\tilde{\Omega}$ and precision matrices $\tilde{\Sigma}_m, \tilde{\Phi}_m$ computed by

$$
\begin{aligned}
\tilde{\Omega}^\mathsf{T} &= \tilde{\Phi}_m^{-1}\left(\Gamma_{XY_m} + \tilde{\alpha}_{\Phi m}\tilde{B}_{\Phi m}^{-1}\Omega^\mathsf{T}\right) \\
\tilde{\Sigma}_m &= \tilde{\alpha}_{\Sigma m}\tilde{B}_{\Sigma m}^{-1} \\
\tilde{\Phi}_m &= \Gamma_{XX_m} + \tilde{\alpha}_{\Phi m}\tilde{B}_{\Phi m}^{-1} \\
\Gamma_{XX_m} &= \sum_n \gamma_{nm} X_n {X_n}^\mathsf{T} \\
\Gamma_{XY_m} &= \sum_n \gamma_{nm} X_n {Y_n}^\mathsf{T}
\end{aligned}
$$

The posterior of the residual variances, $\boldsymbol{\Sigma}_m$, is a Wishart density with shape and scale parameters computed by

$$
\begin{aligned}
\tilde{\alpha}_{\Sigma m} =& \frac{1}{2}\left(\sum_n u\gamma_{nm} + dp + 2\alpha_\Sigma\right) \\
\tilde{B}_{\Sigma m} =& \frac{1}{2}\sum_n \gamma_{nm}\left[(Y_n - \tilde{\Omega}_m X_n)(Y_n - \tilde{\Omega}_m X_n)^\mathsf{T}\right. \\
&+ \mathrm{tr}\left(X_n {X^\mathsf{T}}_n \tilde{\Phi}_m^{-1}\right)\tilde{\Sigma}_m^{-1}\ ] \\
&+ \frac{1}{2}\left(\tilde{\Omega}_m - \Omega_m\right)\tilde{\alpha}_{\Phi m}\tilde{B}_{\Phi m}^{-1}\left(\tilde{\Omega}_m - \Omega_m\right)^\mathsf{T} \\
&+ \frac{1}{2}\mathrm{tr}\left(\tilde{\alpha}_{\Phi m}\tilde{B}_{\Phi m}^{-T}\tilde{\Phi}_m^{-1}\right)\tilde{\Sigma}_m^{-1} + \boldsymbol{B}_\Sigma
\end{aligned}
$$

The posterior model coefficient variances, $\boldsymbol{\Phi}_m$, also follow a Wishart density with parameters

$$
\begin{aligned}
\tilde{\alpha}_{\Phi m} &= \frac{d}{2} + \alpha_\Phi \qquad \forall m = 1, \cdots M \\
\tilde{B}_{\Phi m} &= \frac{1}{2}(\tilde{\Omega}_m - \Omega)^\mathsf{T}\tilde{\alpha}_{\Sigma m}\tilde{B}_{\Sigma m}^{-1}(\tilde{\Omega}_m - \Omega) \\
&+ \frac{1}{2}\mathrm{tr}(\tilde{\alpha}_{\Sigma m}\tilde{B}_{\Sigma m}^{-1}\tilde{\Sigma}_m^{-1})\tilde{\Phi}_m^{-1} + \boldsymbol{B}_\Phi
\end{aligned}
$$

B Appendix: Free Energy

The general variational Free Energy is given as the integral

$$
\begin{aligned}
\mathcal{F} =& \underbrace{-\int q(\boldsymbol{S})q(\boldsymbol{\theta})\log p(Y, \boldsymbol{S}|\boldsymbol{\theta})\ d\boldsymbol{S}\ d\boldsymbol{\theta}}_{\text{Internal Energy}} + \\
& \underbrace{\int q(\boldsymbol{S})\log q(\boldsymbol{S})\ d\boldsymbol{S}}_{\text{Negative Entropy}} + \underbrace{\int q(\boldsymbol{\theta})\log\frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta})}\ d\boldsymbol{\theta}}_{\text{KL-Divergence}}
\end{aligned}
$$

The negative entropy term in the case of coupled hidden Markov models is

$$
\begin{aligned}
E_{CHMM} =& E(S_{n=1}) + E(T_{n=1}) \\
&+ \sum_{n=2}^{N} E(S_n|S_{n-1}, T_{n-l_S}) + E(T_n|T_{n-1}, S_{n-l_T})
\end{aligned}
$$

The internal Energy is given as

$$
\begin{aligned}
\mathcal{I} =& -N\frac{du}{2}\log(2\pi) + \sum_m \bar{\gamma}_m\left(\Psi(\tilde{\rho}_m) - \Psi(\sum_{m=1}^{M}\tilde{\rho}_m)\right. \\
&\left. + \frac{u}{2}\sum_{l=1}^{d}\Psi\left(\frac{1}{2}(2\tilde{\alpha}_{\Sigma m} + 1 - l)\right) - \frac{u}{2}\log|\tilde{B}_{\Sigma m}|\right) \\
&- \sum_{n,m}\frac{\gamma_{nm}\tilde{\alpha}_{\Sigma m}}{2}\mathrm{tr}\left(\tilde{B}_{\Sigma m}^{-1}(Y_n - \tilde{\Omega}_m X_n)(Y_n - \tilde{\Omega}_m X_n)^\mathsf{T}\right) \\
&- \frac{1}{2}\sum_m \tilde{\alpha}_{\Sigma m}\mathrm{tr}\left(\tilde{B}_{\Sigma m}^{-1}\tilde{\Sigma}_m^{-1}\right)\sum_n \gamma_{nm}\ \ \mathrm{tr}\left(X_n {X^\mathsf{T}}_n \tilde{\Phi}_m^{-1}\right)
\end{aligned}
$$

The KL-divergence $D(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}))$ for the variational models is given as

$$\begin{aligned}
&\textstyle\sum_m \left\langle D_{\mathcal{N}}(q(\boldsymbol{H}_m)||p(\boldsymbol{H}_m|\boldsymbol{\Sigma}_m,\boldsymbol{\Phi}_m))\right\rangle_{q(\boldsymbol{\Sigma}_m)q(\boldsymbol{\Phi}_m)} && (4)\\
&\quad + \textstyle\sum_m D_{\mathcal{W}}(q(\boldsymbol{\Sigma}_m)||p(\boldsymbol{\Sigma}_m)) && (5)\\
&\quad + \textstyle\sum_m D_{\mathcal{W}}(q(\boldsymbol{\Phi}_m)||p(\boldsymbol{\Phi}_m)) && (6)\\
&\quad + D_{\mathcal{D}}(q(\boldsymbol{\lambda}_{S,T})||p(\boldsymbol{\lambda}_{S,T)}) && (7)
\end{aligned}$$

where the only unusual term is the divergence (4), which is given by

$$\begin{aligned}
&\frac{1}{2}\Big\{(dp)\log|\tilde{\Sigma}_m| + d\log|\tilde{\Phi}_m| - d^2p\\
&-(dp)\left(\sum_{l=1}^{d}\Psi\left(\frac{1}{2}(2\tilde{\alpha}_{\Sigma m}+1-l)\right) - \log|\tilde{B}_{\Sigma m}|\right)\\
&-d\left(\sum_{l=1}^{dp}\Psi\left(\frac{1}{2}(2\tilde{\alpha}_{\Phi m}+1-l)\right) - \log|\tilde{B}_{\Phi m}|\right)\\
&+\operatorname{tr}\left(\tilde{\alpha}_{\Phi m}\tilde{B}_{\Phi m}^{-1}\tilde{\Phi}_m^{-1}\right)\operatorname{tr}\left(\tilde{\alpha}_{\Sigma m}\tilde{B}_{\Sigma m}^{-1}\tilde{\Sigma}_m^{-1}\right)\\
&+\operatorname{tr}\left(\tilde{\alpha}_{\Sigma m}\tilde{B}_{\Sigma m}^{-1}(\tilde{\Omega}_m-\Omega)\tilde{\alpha}_{\Phi m}\tilde{B}_{\Phi m}^{-1}(\tilde{\Omega}_m-\Omega)^{\mathsf{T}}\right)\Big\}
\end{aligned}$$

Divergences (5) and (6) are standard KL divergences between Wishart densities and divergence (7) are Divergences between Dirichlet densities, all of which can be found in the literature.

C Appendix: Updates for Gaussian Observation Models

For Gaussian observation models we use for the means a conjugate Normal prior with mean and precision $\boldsymbol{\mu}_{m0}$ and $\boldsymbol{C}_{m0}$, and for the precisions a conjugate Wishart prior with shape and scale parameters α_m and $\boldsymbol{B}$, respectively. Thus, the posterior for the means is a Normal distributions $q(\boldsymbol{\mu}_m) \sim \mathcal{N}(\tilde{\boldsymbol{\mu}}_{m0},\tilde{\boldsymbol{C}}_{m0})$, with its parameters computed by

$$\begin{aligned}
\tilde{\boldsymbol{\mu}}_{m0} &= \tilde{\boldsymbol{C}}_{m0}^{-1}(\bar{\alpha}_m\tilde{\boldsymbol{B}}_m^{-1}\bar{\boldsymbol{y}}_m + \boldsymbol{C}_{m0}\boldsymbol{\mu}_{m0});\\
\tilde{\boldsymbol{C}}_{m0} &= (\bar{\gamma}_m\tilde{\alpha}_m\tilde{\boldsymbol{B}}_m^{-1} + \boldsymbol{C}_{m0})
\end{aligned}$$

where $\bar{\boldsymbol{y}} = \sum_{t=1}^{T}\gamma_{tm}\boldsymbol{y}_t$. The posterior of the precisions is a Wishart density, $q(\boldsymbol{C}_m|\tilde{\alpha}_m,\tilde{\boldsymbol{B}}_m) \sim \mathcal{W}(\tilde{\alpha}_m,\tilde{\boldsymbol{B}}_m)$, with its parameters computed by

$$\begin{aligned}
\tilde{\alpha}_m &= \frac{1}{2}\bar{\gamma}_m + \alpha \quad \forall m = 1,\cdots M\\
\tilde{\boldsymbol{B}}_m &= \sum_t^T \frac{\gamma_{tm}}{2}(\boldsymbol{y}_t - \tilde{\boldsymbol{\mu}}_{m0})(\boldsymbol{y}_t - \tilde{\boldsymbol{\mu}}_{m0})' + \frac{\bar{\gamma}_m}{2}\tilde{\boldsymbol{C}}_{m0}^{-1} + \boldsymbol{B}
\end{aligned}$$

References

[1] J.M. Bernardo and A.F.M. Smith. *Bayesian Theory*. John Wiley and Sons, 1994.

[2] C.M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, 1995.

[3] M. Brand. Coupled hidden Markov models for modeling interacting processes. Technical Report 405, MIT Media Lab Perceptual Computing, June 1997.

[4] 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.

[5] 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.

[6] T.S. Jaakkola and M.I. Jordan. Bayesian parameter estimation through variational methods. *Statistics and Computing*, 1997.

[7] 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.

[8] D.J.C. MacKay. Ensemble learning for Hidden Markov models. http://wol.ra.phy.cam.ac.uk/mackay, 1997.

[9] L. R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceeding of the IEEE*, 77(2):257–284, 1989.

[10] I. Rezek, M. Gibbs, and S.J. Roberts. Maximum *A-Postriori* Estimation of Coupled Hidden Markov Models. *Journal of VLSI Signal Processing Systems*, 32:55–66, 2002. invited paper.

[11] I. Rezek and S.J. Roberts. Stochastic Complexity Measures for Physiological Signal Analysis. *IEEE Transactions on Biomedical Engineering*, 44(9):1186–1191, 1998.

[12] I. Rezek, Peter Sykacek, and S.J. Roberts. Learning Interaction Dynamics with Coupled Hidden Markov Models. *IEE Proceedings - Science, Measurement and Technology.*, 147(6):345–350, November 2000. http://www.robots.ox.ac.uk/∼irezek.

[13] P. Sykacek and S.J. Roberts. Baysian Time Series Classification. In T.G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14, 2001.

[14] J.S. Yedidia, W.T. Freeman, and Y. Weiss. Bethe free energy, Kikuchi approximations and Belief Propagation algorithms . In *NIPS*, 2000.