Abstract. Coupled Hidden Markov Models (CHMM) are a new tool which model interactions in state space rather than observation space. Thus they may reveal coupling where classical tools such as correlation fail. In this paper we derive the maximum likelihood equations for the CHMM parameters using the Expectation Maximisation algorithm. The use of the models is demonstrated in simulated data, as well as in biomedical signal analysis.

INTRODUCTION

Analysis of physiological systems frequently involves studying the interactions between them. In fact, the interactions themselves can give us clues about the state of health of the patient. For instance, interactions between cardio-respiratory variables are known to be abnormal in autonomic neuropathy [1]. Conversely, if the type of interaction is known, a departure from this state is, by definition, an anomaly. For instance, in multichannel EEG analysis we would expect coupling between the channels. If the states of the channels are divergent one may suspect some recording error (e.g. electrode failure). Traditionally, such interactions would be measured using correlation or, more generally, mutual information measures [2]. There are, however, major limitations associated with correlation, namely, the linearity and stationarity assumptions. Mutual information, on the other hand, can handle non-linearity. However, it is unable to infer anti-correlation and it also assumes stationarity (for the density estimation).

Some recent work has focused on Hidden Markov models (HMM) in order to avoid some of the stationarity assumptions (the chain is still assumed to be ergodic) and model the dynamics explicitly [3]. HMMs handle the dynamics of the time-series through the use of a discrete state-space. These dynamics are taken from the state transition probabilities from one time step to the next. HMMs, however, typically have a single observation model and observation sequence and thus, by
default, cannot handle the composite modelling of multiple time-series. This can be overcome through the use of multivariate observation models, such as multivariate autoregressive models [4]. The latter implies that interactions are modelled in the observation-space and not in state-space. This is unsatisfactory in some biomedical applications where the signals have completely different forms and thus spectra/bandwidths, e.g. heart-rate and respiration. Another approach models the interactions in state-space by coupling the state-spaces of each variate [5]. This allows interaction analysis to proceed on arbitrary multi-channel sequences.

**HMMS: MARKOV CONDITION AND COUPLING**

**Classical HMMS.**
A standard HMM with $K$ states and length $N$ is defined by the parameter set $\theta = \{A, X, \pi\}$, where $\pi$ is the initial state probability parameter set$^1$, $A$ is the set of observation probabilities conditioned on the state, and $X$ is the set of state transition probabilities. Graphically, it is represented as an unfolded mixture model whose states at time $t$ are conditioned on those at time $t-1$ (see Figure 1[a]), also known as the first order Markov property. Estimation is often performed using the maximum likelihood framework [6, 7], where the hidden states are iteratively re-estimated by their expected values and, using these, the model parameters are re-estimated. HMMS may be extended by changing the dimensions of $A$ and $X$ or by changing the distributions of $A$ and $X$. In this paper, however, we are interested in systems with different topologies, and in particular, with extensions to more than one state variable.

![Figure 1: Hidden Markov Model Topologies](image)

**Coupled HMMS.**
One possible extension to multi-dimensional systems is the coupled hidden Markov model (CHMM), as shown in Figure 1[b]. Here the current state is dependent on the states of its own chain ($X$) and that of the neighbouring chain ($Y$) at the previous time-step. These models have been used in computer vision [5] and digital communication [8]. Several things are of note:

---

$^1$This measure is often taken to be the probability of the state at time $t = 0$. This is of little use as it is not the probability of the states themselves which, for a homogenous Markov Chain can be estimated from the eigenvectors of the transition probabilities.
The model, though still a DAG (directed acyclic graph) contains a loop which forms the essential difficulty in estimating parameters in CHMMs. One approach to solve this problem is clustering [9]: all nodes of a loop are merged to form a single higher dimensional node. Though this may be a valid approach for the simple model shown in Figure 1[b] (the state variables \(X_t, Y_t, X_{t+1}, Y_{t+1}\) would be merged in this case) it is infeasible once the “lag-link” (link from \(X_t \rightarrow Y_{t+1}\) or \(Y_t \rightarrow X_{t+1}\)) is moved forward in time. For instance, the DAG shown in figure 1[c] requires merging of the state variables \(X_t, Y_t, X_{t+1}, Y_{t+1}, X_{t+2}, Y_{t+2}\) and thus causes an immediate explosion of node dimensionality. Another approach, that of cut-sets [9], is also impractical for dynamic Bayes nets, since the number of cut-set variables increases with the chain length \(N\).

To illustrate this briefly, the same CHMM in redrawn in figure 2 as a nested sequence of loops. To cut-open each loop the values of \(X_t\) and \(Y_t\) must be clamped, thus increasing the number of cut-set variables linearly with \(T\), the length of the chain.

There are 2 chains with 2 possibly different observation models. Unless the observation models \(A, B\) are identical, the chains cannot be simply transformed to a single HMM by forming a Cartesian product of the observation-spaces.

![Figure 2: Illustration of the nested loops in a CHMM](image)

While keeping close to the original model, estimation can be performed using variational learning (as shown in [5] although not recognised as such). We have derived the full forward-backward equations for the maximum-likelihood estimation method. When doing so, it becomes apparent that a full approach without “explicit”-node merging\(^2\) is only possible for “lag-one” models. To be precise, a smoother (belief propagation both temporally causal and anti-causal) cannot be derived without node clustering, while filtering (only temporally causal belief propagation), can be derived for higher lags also.

---

\(^2\)The nodes are effectively merged due to the conditional dependence of the states on their past values.
MAXIMUM LIKELIHOOD PARAMETER ESTIMATION

We use the following notation:

- \( O = \{A^T_1, B^T_1\} \): the observation sequence, where \( A^T_1 = \{a_1, \ldots, a_T\} \) are the observations of the first chain; similarly we define \( B^T_2 \) for the second chain;
- \( S = \{X^T_1, Y^T_1\} \): the state sequence, where \( X^T_1 = \{x_1, \ldots, x_T\} \) and \( x \in 1 \cdots M \) are the states of the first chain; similarly we define \( Y^T_1 \);
- \( P(X_{t+1} | X_t, Y_t) \): state transition probabilities of the first chain;
- \( P(Y_{t+1} | X_t, Y_t) \): state transition probabilities of the second chain;
- \( P(X_1) \) and \( P(Y_1) \): prior probabilities of first and second chain, respectively;
- \( P(A_t | X_t) \) and \( P(B_t | Y_t) \): observation densities of the chains, which are here assumed to be multivariate Gaussian with mean vectors \( \mu_x, \mu_y \) and covariance matrices \( \Sigma_x, \Sigma_y \);
- \( K_x \) and \( K_y \): state space dimensionalities;

The likelihood function of the CHMM is then given as

\[
\mathcal{L}(\theta) = P(X_1)P(Y_1) \prod_{t=1}^{N} P(A_t | X_t)P(B_t | Y_t) \\
\quad P(X_{t+1} | X_t, Y_t)P(Y_{t+1} | X_t, Y_t)
\]

where the parameter vector \( \theta \) contains parameters of the transition probabilities, the prior probabilities, and the parameters of the observation densities.

In the maximum likelihood (ML) or maximum a posteriori estimation methods (MAP), clearly the non-standard part is the belief propagation in the E-Step. Once the expected values of the hidden variables have been obtained, the maximisation equations for the observation models are identical to those in the standard HMM case [4, 10]. Thus we will only demonstrate here the derivation of the forward-backward recursions for the CHMM.

From standard HMMs we write:

\[
\alpha_{t+1} \equiv P(O_{t}^{t+1}, S_{t+1}) = P(O_{t+1} | S_{t+1}) \int P(O_{t}^{t}, S_{t})P(S_{t+1} | S_{t})dS_{t}
\]

\[
= P(O_{t+1} | S_{t+1}) \int P(S_{t+1} | O_{t}^{t})dS_{t} \tag{1}
\]

\[
= P(O_{t+1} | S_{t+1})P(S_{t+1} | O_{t}^{t})
\]

and by conditional independence

\[
= P(O_{t}^{t+1}, S_{t+1}) \tag{2}
\]
For CHMMs, we define a variable $\alpha$ for each chain: $\alpha^A_t$ and $\alpha^B_t$. At $t = 1$,
\begin{align}
\alpha^A_t &= P(A_1|X_1)P(X_1), \quad (3) \\
\alpha^B_t &= P(B_1|Y_1)P(Y_1). \quad (4)
\end{align}

At $t = 2$,
\begin{align}
\alpha^A_2 &= P(A_2|X_2) \int \int P(A_1, X_1)P(B_1, Y_1)P(X_2|X_1, Y_1)dX_1dY_1 \\
&= P(A_2|X_2) \int \int P(A_1, B_1, X_1, Y_1)P(X_2|X_1, Y_1)dX_1dY_1 \tag{5} \\
&= P(A_2|X_2)P(A_1, B_1, X_2)
\end{align}

since $A_2$ is independent of $A_1$ given $X_2$, and so
\begin{equation}
\alpha^A_2 = P(A^T_1, B_1, X_2). \tag{6}
\end{equation}

Repeating the steps for the second chain and for arbitrary $t$ gives,
\begin{align}
\alpha^A_{t+1} &= P(A_{t+1}|X_{t+1}) \int \int \alpha^A_t \alpha^B_t P(X_{t+1}|X_t, Y_t)dX_tdY_t \quad (7) \\
\alpha^B_{t+1} &= P(B_{t+1}|Y_{t+1}) \int \int \alpha^B_t \alpha^A_t P(Y_{t+1}|X_t, Y_t)dX_tdY_t \quad (8)
\end{align}

Thus, there are two “un-coupled” prediction equations and no node clustering required.

As for the backwards recursions, the loop prevents a split of the recursions into two un-coupled ones. We are thus forced to perform a clustering operation of all the nodes in a loop, namely, $X_t, Y_t, X_{t+1}, Y_{t+1}$, which simply leads to the following recursion,
\begin{align}
\beta_{t+1} &= P(O^T_{t+1}|S_t) \\
&= \int \int P(A_{t+1}, B_{t+1}|X_{t+1}, Y_{t+1})P(Y_{t+1}|X_{t+1}, Y_{t+1})dX_{t+1}dY_{t+1} \\
&= \int \int P(A_{t+1}, B_{t+1}|X_{t+1}, Y_{t+1}) \beta_{t+1} P(X_{t+1}, Y_{t+1}|X_t, Y_t)dX_{t+1}dY_{t+1} \tag{9} \\
&= \int \int P(A^T_{t+1}, B^T_{t+1}|X_{t+1}, Y_{t+1})P(X_{t+1}, Y_{t+1}|X_t, Y_t)dX_{t+1}dY_{t+1}.
\end{align}

This also demonstrates that it becomes quickly infeasible to explore different coupling topologies, as they will increase the dimensionality of the clustered loop exponentially\(^3\).

\(^3\)As a quick alternative to estimating different lags one could time shift one sequence relative to the other and then run the estimation. One of the links, however, then models anti-causal (reverse time) interactions.
RESULTS

Synthetic Data.

Synthetic data was generated by drawing 1024 samples from a CHMM. Figure 3 shows the input state sequence together with the estimated joint state and marginal sequences. Incorrect state-labels are marked with asterisks - in this case there were no incorrect classifications. The lower two traces in the figure depict the individual state sequences, estimated by marginalisation of the joint transition probabilities (using equal priors) and feeding these into a standard single HMM model. Note that a simple Cartesian product of the two marginal Viterbi paths resembles closely the joint Viterbi path. To highlight the benefits of the interaction model, figure 4 also shows the correlation function for the two observation sequences - clearly, no structure can be seen.

The exactness depends also on the priors used to estimate the marginals. If exact priors are used then the joint Viterbi will be identical to the Cartesian product of the marginals.
Cheyne Stokes Data

We applied the CHMM to features extracted from a section of Cheyne Stokes Data\(^5\), which consisted of one EEG recording and a simultaneous respiration recording sampled at 128Hz. The feature, fractional spectral radius (FSR) measure [11], was computed from consecutive non-overlapping windows of two seconds length. Thus, respiration FSR then formed one channel while EEG FSR the second channel. The data has clearly visible states which can be used to verify the results of the estimated model. Figure 5 shows the FSR traces superimposed with the joint Viterbi path sequence. The CHMM states were set to \(K_x = 2\) and \(K_y = 2\). Essentially, of the four total states the CHMM splits the sequence such that two correspond to respiration and EEG being in phase and two correspond to the transitions states, i.e. respiration on, EEG off and vice versa. The number of states was determined after estimating models with a range of possible configurations of state space dimensionalities. Using Minimum Description Length as a criterion to penalise the likelihood score, the optimum model was given as that with \(K_x = 2\) and \(K_y = 2\).

Sleep EEG Data.

In a final example, we fitted a 5-th order Bayesian auto-regressive (AR) model to a full night EEG recording. The AR-coefficients were extracted from two channels, namely C3 and C4, using sliding non-overlapping windows of 5s length (sampling rate 200Hz). The coefficients, pre-filtered with a moving average filter of order 20, then formed the observation sequence for a CHMM.

The estimation was performed in a semi-supervised manner. For each of the two chains the number of hidden states was set to 3, corresponding to three sleep-
stages [?]: NREM, REM and Wake, a decision based on the prior knowledge of the structure in the data. The training data were state labels which were marked by a human expert as deep sleep (sleep stage 4, S4), REM and Wake. Thus, where the EEG was marked as S4 the CHMM states of both chains were set 0, where the EEG was marked as REM the CHMM states were set to 1, and so on. The training data was then given full weighting whilst the remaining data was given the lower, but equal, weighting of \(1/K\), the number of states in each chain. We note that the degree of supervised learning can be controlled by the weighting of the non-training data: equal weighting of all data is equivalent to fully un-supervised learning while increased weighting of the training data corresponds to an increased degree of supervision.

Figure 6[a] depicts the hypnogram (human labels) together with the estimated state sequence and Figure 6[b] shows the co-occurrences of the Viterbi labels with those of the Hypnogram. Roughly speaking, of the states with the longest duration, states 0/0 and 0/2 correspond to most non-REM sleep, state 1/1 to REM sleep and state 2/2 to wake. Generally, a CHMM estimated using the Maximum Likelihood procedure does make sense of the data, but, unlike in the Bayesian estimation case [12], the chain is more ambiguous in its results.

CONCLUSIONS & FUTURE WORK

The stability of the CHMM estimates is very encouraging. Moreover, the smoothness of the state sequence is much higher than that obtained using single chain HMMs [13]. The main problem lies in the estimation procedure rather than the model itself, as can be easily seen by comparison with Bayesian estimation methods.

The results presented here are clearly only preliminary. The CHMMs themselves may be expanded to more than just two coupled chains and have continuous hidden state spaces or other observation models. One other possible direction is that of trimming unnecessary transition probabilities with the use of entropy priors [14]. More important, however, is the question of lag-estimation. Future work will focus on model-lag selection for which we envisage the use of reversible jump sampling methods [15].

REFERENCES

Figure 6: Comparison between estimated labels and Hypnogram Labels for sleep EEG

Acknowledgement. The authors would like to thank Will Penny, Peter Sykacek and Regina Conradt for helpful discussions. This work is funded by the European Community Commission, DG XII (project SIESTA, Biomed-2 PL962040) whose support we gratefully acknowledge.