Advanced State Space Methods for Neural and Clinical Data

ZHE CHEN

Dedicated to ...

Contents

List of illustrations *page* iv
List of tables vi
List of contributors vii

1 Identifying outcome-discriminative dynamics 1
1.1 Background 1
1.2 Time Series Classification and Switching Vector Autoregressive modeling 3
1.2.1 Marginals-based Learning via Error Backpropagation 5
1.3 Experiments 8
1.3.1 Decoding Local Field Potentials 11
1.4 Discussion and Conclusion 12
1.5 APPENDIX 15
1.5.1 Gradient of inference operations and hidden state marginals 15

Bibliography 20
Author index 23
Subject index 24

Illustrations

1.1 Marginals-based learning in dynamic Bayesian networks. (a) Graphical model representation of the switching vector autoregressive (switching VAR) model, including a layer of discrete latent variables (square nodes) with Markovian transition dynamics matrix Z, and autoregressive observations (round nodes) with mode-specific parameters θ_j. (b) Unrolled representation (with respect to time and inference steps) of the switching VAR model, with an added logistic regression layer (elliptic nodes). Note that due to parameter tying across time the number of model parameters is independent of time series length. Inference in the switching VAR model involves a forward-backward algorithm yielding a sequence of filtered (S_t^f) and smoothed (S_t^s) marginals. 4

1.2 Two examples of simulated bivariate time series with switching dynamics. The time series were divided into 4 categories, each having different proportions of four modes. These dynamical modes recur within each time series and are shared across the different time series. Here, we introduced an offset of 2 in one of the channels of each time series for improved visualization. 7

1.3 Classification performance over ten folds using expectation maximization (EM) versus marginals-based learning via backpropagation (BP). Panel (a) shows accuracy of classification (chance level is at 25%), and panel (b) is the multinomial probability of the outcomes. Each panel represents a fixed number of EM-based pre-training (5, 8, 10, 15, and 20 iterations) followed by supervised learning with early stopping. The figure demonstrates the effects of generative pre-training, and the tendency of EM to overfit to artifacts with increased number of iterations. 8

1.4 Examples of heart rate and mean blood pressure from a tilt-table experiment. Tilting or standing up results in an increased activity of the sympathetic nervous system, which operates at lower frequencies than the parasympathetic nervous system. This manifests itself as lower frequency oscillations in heart rate time series within the non-supine segments. 9

1.5 An example of a filtered time series of heart rate (HR) and mean blood pressure (MAP) from the tilt-table experiment (panel a). The inferred marginal probabilities of each of the four modes using the EM and the outcome-discriminative approaches are shown in panels (b) and (c), respectively. 10

1.6 Comparison of EM and outcome-discriminative learning (BP) on the tilt-table dataset. Panel (a) shows 10-fold cross-validated performance of the EM versus BP (using 30 iteration of the BFGS algorithm with early stopping). Panel (b) shows a comparison of the two techniques in terms of classification accuracy. 11

1.7 Physiological interpretation of learned dynamics. LF/HF ration for the EM and outcome-discriminative learning are shown in panels (a) and (b), respectively.The $\star$ symbol indicates a significant change from baseline ($p<0.05$; Kruskal-Wallis nonparametric ANOVA test). 12

1.8 Effects of EM-based pre-training on the performance of supervised learning on the LFP decoding experiment. Each panel shows classification performance over ten folds (testing set performance) based on 0, 1, 5, 10, and 20 iterations of expectation maximization (EM) followed by 30 iterations of the supervised learning via backpropagation (BP). The cost function that is being optimized (Bernoulli probability of outcomes) is shown in the top row, and the area under receiver operating curve (AUC) is presented in the bottom row. 13

Tables

List of contributors

Shamim Nemati
Harvard School of Engineering and Applied Sciences, Cambridge, MA

Ryan P. Adams
Harvard School of Engineering and Applied Sciences, Cambridge, MA

1 Identifying outcome-discriminative dynamics in multivariate physiological cohort time series

Shamim Nemati and Ryan P. Adams

1.1 Background

Physiological control systems typically involve multiple interacting variables operating in feedback loops that enhance an organism's ability to self-regulate and respond to internal and external disturbances. The resulting multivariate time series often exhibit rich dynamical patterns that are altered under pathological conditions, and are therefore informative of health and disease (Ivanov, Rosenblum, Peng, Mietus, Havlin, Stanley & Goldberger 1996, Costa, Goldberger & Peng 2002, Stein, Domitrovich, Huikuri, Kleiger & 2005, Nemati, Edwards, Sands, Berger, Wellman, Verghese, Malhotra & Butler 2011). Previous studies using nonlinear (Ivanov et al. 1996, Costa et al. 2002) indices of HR variability (i.e., beat-to-beat fluctuations in HR) have shown that subtle changes to the dynamics of HR may act as an early sign of adverse cardiovascular outcomes (e.g., mortality after myocardial infarction (Stein et al. 2005)) in large patient cohort. However, these studies fall short of assessing the multivariate dynamics of the vital signs (such as heart rate, blood pressure, respiration, etc.), and do not yield any mechanistic hypotheses for the observed deteriorations of normal variability. This shortcoming is in part due to the inherent difficulty of parameter estimation in physiological time series, where one is confronted by nonlinearities (including rapid regime changes), measurement artifacts, and/or missing data, which are particularly prominent in ambulatory recordings (due to patient movements) and bedside monitoring (due to equipment malfunction).

In the previous chapter we developed a framework for unsupervised discovery of shared dynamics in multivariate physiological time series from large patient cohorts. A central premise of our approach was that even within heterogeneous cohorts (with respect to demographics, genetic factors, etc.) there are common *phenotypic dynamics* that a patient's vital signs may exhibit, reflecting underlying pathologies (e.g., detraction of the baroreflex system) or temporary physiological state changes (e.g., postural changes or sleep/wake related changes in physiology). We used a switching state-space model (in particular, a switching vector autoregressive) to automatically segment the time series into regions with similar dynamics, i.e., time-dependent rules describing the evolution of the system state. The state-space modeling approach allows for incorporation of physiologically-constrained linear models (e.g., via linearization of the nonlinear

dynamics around equilibrium points of interest) to derive mechanistic explanations of the observed dynamical patterns, for instance, in terms of directional influences among the interacting variables (e.g., baroreflex gain or chemoreflex sensitivity).

Although we may assume *a priori* knowledge of the underlying physiology to constrain the state-space models, the model parameters and latent variables have to be learned from the data. While in many problems model fitting and prediction of the future values of a time series are the primary quantities of interest, unsupervised learning is often used to learn *features* for a downstream supervised (classification) task (Marlin, Kale, Khemani & Wetzel 2012, Lasko, Denny & Levy 2013). For instance, Lasko et al. applied a *deep learning*-based approach to unsupervised learning of phenotypical features in longitudinal sequences of serum uric acid measurements. The resulting unsupervised phenotypic features were passed to a classifier to distinguish the uric acid signatures of gout vs. acute leukemia, with a performance level competitive with the gold-standard features engineered by domain experts. In practice, this two-stage procedure – unsupervised feature extraction followed by supervised learning for outcome discrimination — may be suboptimal, since the latent dynamics that are important to the supervised target may only be weakly related to those that are best for explaining the raw statistics of the time series. Additionally, generative approaches to unsupervised feature learning (Lehman, Adams, Mayaud, Moody, Malhotra, Mark & Nemati 2014) may be hamstrung by the shortcomings of approximate inference, or the underlying models may be underspecified with respect to the nuanced features associated with the outcomes of interest. For instance, in a neurophysiological experiment involving EEG recordings, it may be the case that only a single low amplitude oscillation is the distinguishing feature of successful trials, and therefore a reduced-model specifically trained to capture that oscillation may provide a more parsimoneous solution to the problem of predicting outcomes of each trial. It is therefore desirable to learn models of time series dynamics in which the latent variables are directly tuned towards the supervised task of interest.

In this chapter, we present a learning algorithm specifically designed to learn dynamical features of time series that are directly predictive of the associated labels. Rather than depending on label-free unsupervised learning to discover relevant features of the time series, we build a system that expressly learns the dynamics that are most relevant for classifying time series labels. Our goal is to obtain compact representations of nonstationary and multivariate time series (*representation learning*)(Bengio, Courville & Vincent 2013). To accomplish this we use a connection between dynamic bayesian networks (e.g., the switching VAR model) and artificial neural networks (ANNs) to perform inference and learning in state-space models in a manner analogous to backpropagation in neural networks (Rumelhart, Hinton & Williams 1988). This connection stems from the observation that the directed acyclic graph structure of a state-space model can be unrolled both as a function of time and inference steps to yield a deterministic neural network with efficient parameter tying across time (see Fig. 1.2). Thus, the parameters governing the dynamics and observation model of a state-space model can be learned in a manner analogous to that of a neural network. Indeed, the resulting system can be viewed as a compactly-parameterized recurrent neural network (RNN)

(Sutskever 2013). Although the standard use of RNNs has been for time series prediction (network output is the predicted input time series in the future) or sequential labeling (when output is a label sequence associated with the input data sequence), with additional processing layers one may obtain a time series classifier from this class of models (Graves, Fernández, Gomez & Schmidhuber 2006). Nevertheless, RNNs have proven hard to train, since the optimization surface tend to include multiple local minima. Moreover, standard RNN are 'black box' algorithms(as apposed to 'model-based') and therefore do allow for incorporation of physiological models of the underlying systems. The framework proposed here addresses both these shortcomings. First, knowledge of the underlying physiology can be directly incorporated into the state-space models that constitute the basic building blocks of a dynamic Bayesian network. Secondly, equipped with a generative model, we can rely on unsupervised pre-training (via expectation maximization) to systematically initialize the parameters of the equivalent RNN; in a manner analogous to pre-training of very large neural networks (*deep learning*) (Erhan, Bengio, Courville, Manzagol, Vincent & Bengio 2010).

Discriminative approaches to learning in graphical models can be broadly classified into discrete versus continuous latent variable models. Some of the recent works within the first category include: *structured output classification* (Memisevic 2006) where the hidden discrete states of an HMM are designed to correspond to target labels which are observed in the training data and thus can be learned using outcome-discriminative learning, and *approximate marginal inference* in conditional random fields (Eaton & Ghahramani 2009), (Stoyanov, Ropson & Eisner 2011),(Domke 2013). Supervised learning techniques for learning of HMMs and related conditional random fields have been shown to outperform generative maximum likelihood learning in many tasks (McCallum, Freitag & Pereira 2000, Lafferty, McCallum & Pereira 2001, Woodland & Povey 2000). More recently, It has empirically been shown that *marginalization-based* learning via empirical risk minimization gives better results than likelihood based approximations in the presence of model mis-specification (Domke 2013). In the continuous domain, (Kim & Pavlovic 2009) used a gradient based approach to learning parameters of a conditional state-space model. They assumed ground truth for continuous latent state is known during the learning phase, and provided analytical gradients for the conditional likelihood of the latent state variables with respect to the state-space model parameters. In contrast, here we propose a framework for gradient-based learning in hybrid discrete and continuous state-space models, and given differentiable but otherwise arbitrary cost functions.

1.2 Time Series Classification and Switching Vector Autoregressive modeling

Assume we are given a collection of N multivariate time series and the associated outcome Variables: $\{(y^{(1)}, O^{(1)}), (y^{(2)}, O^{(2)}), \cdots, (y^{(N)}, O^{(N)})\}$, where the n-th time series $y^{(n)}$ is of length T_n, and may include M channels. The corresponding label $O^{(n)}$ can be a scalar such as a discrete patient outcome, or it may itself be be a length-T_n time series

vector that assigns a label to each instant[1]. Our objective is to find shared dynamical features across the different time series that are predictive of the labels.

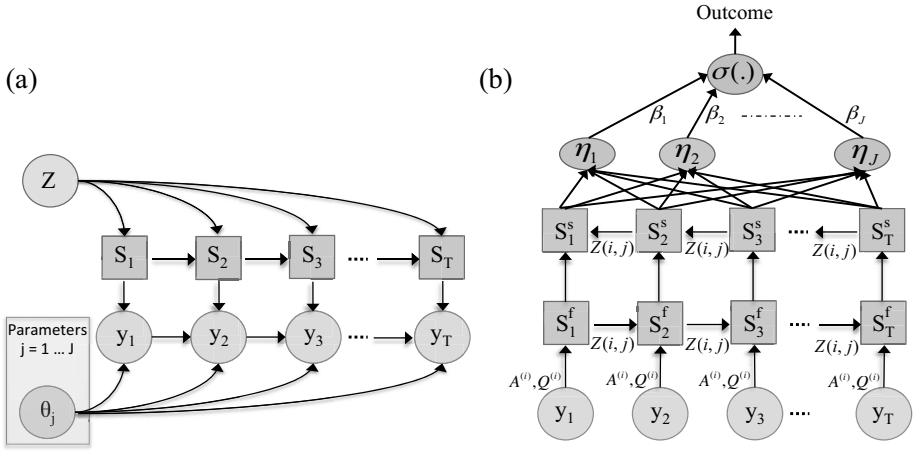

Figure 1.1 Marginals-based learning in dynamic Bayesian networks. (a) Graphical model representation of the switching vector autoregressive (switching VAR) model, including a layer of discrete latent variables (square nodes) with Markovian transition dynamics matrix Z, and autoregressive observations (round nodes) with mode-specific parameters θ_j. (b) Unrolled representation (with respect to time and inference steps) of the switching VAR model, with an added logistic regression layer (elliptic nodes). Note that due to parameter tying across time the number of model parameters is independent of time series length. Inference in the switching VAR model involves a forward-backward algorithm yielding a sequence of filtered (S_t^f) and smoothed (S_t^s) marginals.

In the previous chapter we used a switching vector autoregression (VAR) to model a time series cohort. The switching VAR models time series using a single layer of hidden discrete random variables (see Fig. 1.2), describing the evolution of a set of J latent states according to a Markovian dynamic. Each of these states correspond to a unique VAR model that generates the observed time series. The generative model is as follows: a latent process for each time series $s_t^{(n)} \in \{1, \cdots, J\}$ evolves according to a Markovian dynamic with initial distribution $\pi^{(n)}$ and $J \times J$ transition matrix Z. The n-th time series $y_t^{(n)}$ evolves according to VAR model with parameters determined by the current latent state $s_t^{(n)}$. The jth VAR model has dynamics and noise parameters $A^{(j)}$ and $Q^{(j)}$, respectively:

$$y_t^{(n)} = \sum_{p=1}^{P} \mathbf{a}_p^{(s_t^{(n)})} y_{t-p}^{(n)} + e_t^{(n)}\,, \qquad e_t^{(n)} \sim \mathcal{N}(0, Q^{(s_t^{(n)})})\,, \tag{1.1}$$

with the multivariate autoregressive model coefficient matrices $\mathbf{a}_p^{(j)}$ of size $M \times M$,

[1] A closely related problem considered in natural language processing under three categories of *temporal classification, segment classification, and sequence classification*(Graves 2012)

with maximal time lag $p = 1 \ldots P$, and noise term w_t with covariance $Q^{(j)}$. The set of parameters $\Delta^{(j)} = \{\mathbf{a}_1^{(j)}, \cdots, \mathbf{a}_P^{(j)}, Q^{(j)}\}$ define a dynamical *mode*. Fig. 1.2 (a) depicts the graphical model representation of a switching VAR model, which is equivalent to an HMM with continuous-valued autoregressive observations. Henceforth we use $\Theta = \{\{\Delta^{(j)}\}_{j=1}^{J}, Z, \pi^{(n)}\}$ to denote the set of all parameters defining a SVAR model.

A comprehensive treatment of the expectation maximization (EM) algorithm for learning the parameters of SVAR can be found elsewhere (Murphy 1998). Briefly, in practice we neither know the set of switching variables nor the parameters that define the modes. EM is a two-pass iterative algorithm: (1) in the expectation (E) step we obtain the expected values of the latent variables $\{\{s_t^{(n)}\}_{t=1}^{T_n}\}_{n=1}^{N}$ using a forward-backward algorithm (Murphy 1998, Heskes & Zoeter 2002), and (2) in the maximization (M) step we find the model parameters Θ that maximize the expected complete data log likelihood. In our implementation of the EM algorithm, we achieve *shared* dynamics by pooling together all subjects' inferred latent variables in the M step. It is also possible to impose physiological constraints on the model parameters using a constrained least square approach within the M step. Iteration through several steps of the EM algorithm results in learning a set of J shared modes and a global transition matrix Z for all the patients.

1.2.1 Marginals-based Learning via Error Backpropagation

As discussed earlier, essentially any standard supervised learning algorithm can incorporate the latent variable marginals as features for time series classification or sequential labeling. Here we examine two significant cases of interest: where there is a global label for the time series, and where the supervised target is itself an aligned time series. We describe the classification setting, but these approaches would generalize directly to continuous labels and more structured settings.

Global Label from Hidden State Proportions

We assume that each label $O^{(n)}$ can take on one of K possible outcomes, and can be modeled using a softmax classifier with parameters β. The inputs to the logistic regressor are the marginal estimates of expected proportion of the time that is spent in each of the latent discrete states[2]:

$$\mu_k^{(n)}(P_\Theta(s_1^{(n)}), \cdots, P_\Theta(s_{T_n}^{(n)})) = \frac{\exp\{\beta_{k,0} + \beta_k^{\mathsf{T}} \eta^{(n)}\}}{\sum_{k'=1}^{K} \exp\{\beta_{k',0} + \beta_{k'}^{\mathsf{T}} \eta^{(n)}\}}, \quad \eta_j^{(n)} = \frac{1}{T_n} \sum_{t=1}^{T_n} P_\Theta(s_t^{(n)} = j)$$

where the β_k are length J weight vectors, $\beta_{k,0}$ are biases, and the $\eta^{(n)}$ are length J vectors of hidden state proportions, which are weighed in a softmax function with row vector parameters β_k. We take the classification cost function to be the negative log likelihood (*negentropy*) of the outcome labels, given the time series:

[2] In the remainder of the paper, we will write the data-conditional marginals as $P_\Theta(s_t^{(n)})$ for compactness.

$$-\log \Pr(\mathbf{O}|\mu(\Theta,\beta)) = -\sum_{n=1}^{N}\sum_{k=1}^{K} \mathbf{O}_k^{(n)} \log \mu_k^{(n)}(\Theta,\beta). \tag{1.2}$$

Training can then be performed using the gradient of the logistic regression log likelihood, learning the β_k as well as backpropagating through the $\eta^{(n)}$ to fit the dynamics parameters Θ proxied by the marginals .

Sequential Labels from Local Marginals

Some tasks require a time-aligned sequence of labels, in a similar fashion to a conditional random field, i.e., $O^{(n)}$ is a sequence of size T_n, with each label taking one of K discrete values. Here, the marginal-based predictor produces a label at each time step, which is the result of a softmax applied to the marginal estimates:

$$\mu_{t,k}^{(n)}(P_\Theta(s_1^{(n)}),\cdots,P_\Theta(s_{T_n}^{(n)})) = \frac{\exp\{\beta_{k,0}+\beta_k^\top \eta_t^{(n)}\}}{\sum_{k'=1}^{K}\exp\{\beta_{k',0}+\beta_{k'}^\top \eta_t^{(n)}\}}, \quad \eta_{t,j}^{(n)} = P_\Theta(s_t^{(n)}=j)\,.$$

Here the $\eta_t^{(n)}$ are length-J marginal estimates at each time t, being weighed in a softmax classifier with parameters β_k. We take the classification objective to be the negative log likelihood (*negentropy*) of the outcome labels, given the time series:

$$-\log \Pr(\mathbf{O}|\mu(\Theta,\beta)) = -\sum_{n=1}^{N}\sum_{t=1}^{T_n}\sum_{k=1}^{K} \mathbf{O}_{t,k}^{(n)} \log \mu_{t,k}^{(n)}(\Theta,\beta). \tag{1.3}$$

Again, the standard logistic regression likelihood can be used for training, with gradients of β directly available and gradients of Θ available via backpropagation.

As noted earlier, within the EM framework unsupervised learning of the dynamics is treated separately from the discriminative learning of a mapping between switching states and outcome labels. The objective of outcome-discriminative learning is to design purely-supervised learning algorithm that discovers dynamical features in series that are predictive of the outcome variables. The key insight of the proposed learning algorithm is that the gradient of the objectives calculated in Eqs. (1.2) and (1.3) can be backpropagated through the network architecture depicted in Fig. 1.2(b) to efficiently calculate the gradient with respect to all latent variables and model parameters.

The analytic gradients of the above cost function in terms of β and Θ can be calculated using a two-pass algorithm. The forward pass involves running inference to approximate the marginal distributions over the latent variables, and subsequently evaluating the predictor $\mu(\cdot)$. The backward pass utilizes the chain rule (reverse mode differentiation) from calculus to obtain the gradients of the overall loss. Since SVAR inference algorithms involve a sequence of differentiable operations, the derivative of the loss function with respect to the discrete marginals, and finally model parameters Θ can be calculated efficiently. To accomplish this, it helps to visualize an unrolled version of the SVAR forward-backward inference procedure, in which snapshots of a random variable at times t and $t+1$ are distinct deterministic (fixed at the values determined by

the inference step) nodes in a feedforward neural network (see Fig. 1.2(b)). Note that since the overall gradient over a time series cohort is the sum of the individual gradients, gradient calculations can be done in parallel for each time series. Analytic expressions for the gradients with respect to parameters of the switching VAR model are presented in the Appendix.

The above gradient can be directly plugged into a optimizer such as the limited-memory BroydenFletcherGoldfarbShanno (BFGS) algorithm[3] to optimize Θ and β. However, it is necessary to carefully manage the optimization procedure in order to avoid overfitting and local minima. In practice we observed that good initial parameters can easily be found using a few iteration of the expectation maximization algorithm (Murphy 1998) for unsupervised learning from the time series, in the absence of label information. This observation supports the intuition that although likelihood based learning and the resulting features may not be necessarily good for discriminating between classes, they nevertheless capture the structure of the input data and therefore provide a good starting point for discriminative fine-tuning to make rapid progress (Erhan et al. 2010). We also found it useful to implement an early stopping criteria based on classification performance on a held-out validation set.

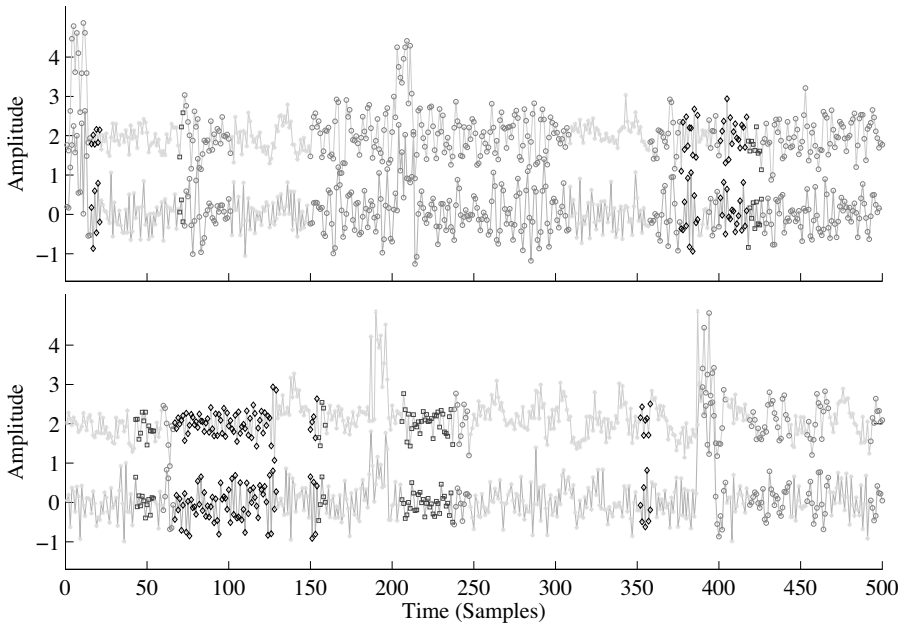

Figure 1.2 Two examples of simulated bivariate time series with switching dynamics. The time series were divided into 4 categories, each having different proportions of four modes. These dynamical modes recur within each time series and are shared across the different time series. Here, we introduced an offset of 2 in one of the channels of each time series for improved visualization.

[3] We used the Matlab implementation of the BFGS algorithm provided in the *minFunc* optimization package: http://www.di.ens.fr/~mschmidt/Software/minFunc.html.

1.3 Experiments

Simulated time series with Switching Dynamics

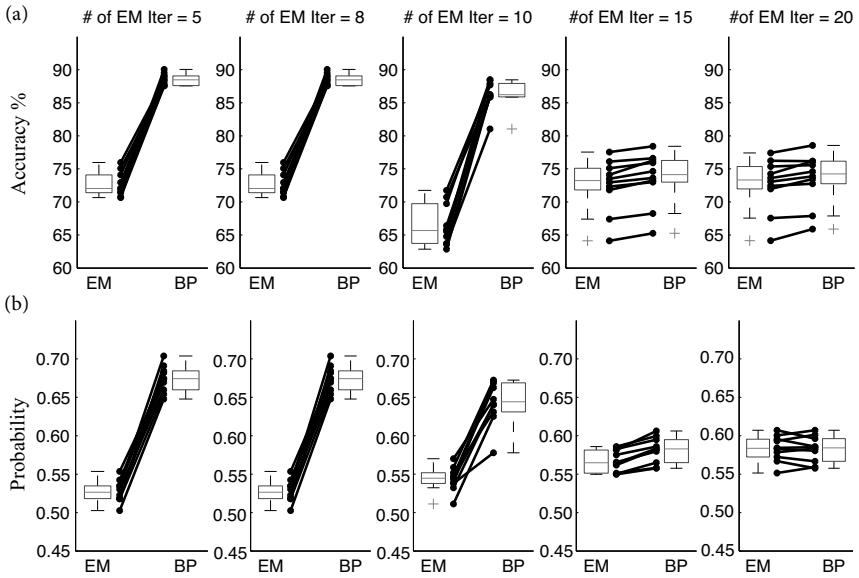

Figure 1.3 Classification performance over ten folds using expectation maximization (EM) versus marginals-based learning via backpropagation (BP). Panel (a) shows accuracy of classification (chance level is at 25%), and panel (b) is the multinomial probability of the outcomes. Each panel represents a fixed number of EM-based pre-training (5, 8, 10, 15, and 20 iterations) followed by supervised learning with early stopping. The figure demonstrates the effects of generative pre-training, and the tendency of EM to overfit to artifacts with increased number of iterations.

We will next demonstrate the idea of outcome-discriminative learning in a sequential labeling task learning, consisting of 200 simulated time series with dynamic switching among four stable dynamical modes (VAR models of order two). To increase the heterogeneity of the dataset, the time series were simulated using four different Markov transition matrices (the stationary distribution of the four categories were $[0.67,\ 0.10,\ 0.10,\ 0.13]$, $[0.14,\ 0.57,\ 0.19,\ 0.10]$, $[0.08,\ 0.16,\ 0.54,\ 0.22]$, and $[0.09,\ 0.09\ ,\ 0.23,\ 0.59]$). Additionally, we introduced approximately 10% variation in the AR coefficients across each realization by adding white Gaussian noise with standard deviation 0.05 to each of the AR coefficients. Finally, all time series included two randomly-placed large-amplitude artifacts (uniform random noise in the interval of $[0, 15]$) of 10 samples duration. Two examples of the simulated time series are shown in Fig. 1.2.1).

Here we assume that the number of modes and the model order is known *a priori*[4], and test the performance of both the EM and the outcome-discriminative learning on the classification problem of labeling each time series sample as belonging

[4] if the number of modes is not known beforehand, model selection criteria such as the Bayesian

to one of four modes. Fig. 1.3 provides a summary of the performance of EM and outcome-discriminative learning. Notably, the figure demonstrates the dependence of the proposed joint supervised learning on the EM initialization. In particular, outcome-discriminative learning benefits from pre-training with as low as 5 iterations of EM. Further unsupervised pre-training eventually lowers the performance of the outcome-discriminative learning; presumably due to local minima and overfitting of artifacts.

Tilt-Table Experiment

Our next example is based on a tilt-table experiment, and aims at revealing the intricate dependencies among cardiovascular variables. We use this example to illustrate the utility of proposed framework for a model-based approach to pattern recognition in nonstationary physiological time series. In particular, we show how the discovered dynamical patterns in heart rate (HR) and mean arterial blood pressure (MAP) can be interpreted in the light of the underlying cardiovascular control system.

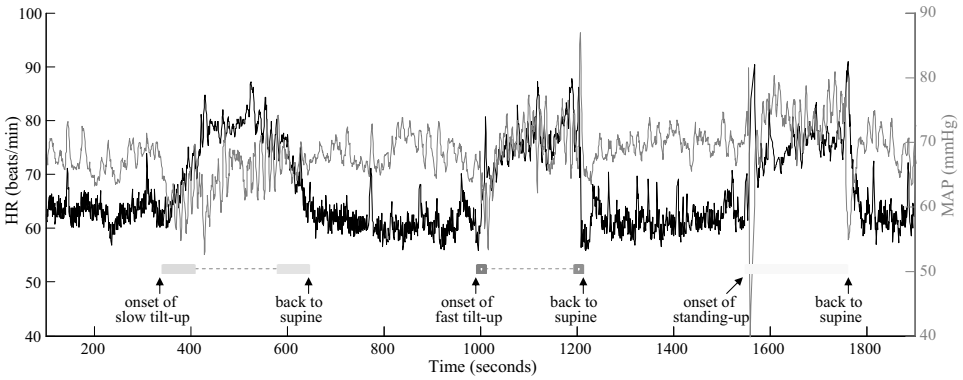

Figure 1.4 Examples of heart rate and mean blood pressure from a tilt-table experiment. Tilting or standing up results in an increased activity of the sympathetic nervous system, which operates at lower frequencies than the parasympathetic nervous system. This manifests itself as lower frequency oscillations in heart rate time series within the non-supine segments.

Time series of HR and mean arterial blood pressure (MAP) were acquired from 10 healthy subjects undergoing a tilt-table experiment. The details of the protocol are described in Heldt et al. (Heldt, Oefinger, Hoshiyama & Mark 2003). Briefly, subjects were placed in the supine position and secured to a table. Tilting was performed at various speeds from the horizontal position to the vertical position and back to supine, generating four postural categories of (1) supine, (2) slow-tilt, (3) fast tilt, and (4) standing (see Fig. 1.3).

Given that we are interested in the interaction between HR and MAP in the frequency range pertinent to sympathetic and parasympathetic regulation (Nemati, Lehman, Adams & Malhotra 2012), we first removed the very low frequency oscillations (slower than

information criteria (BIC) or the nonparametric approaches discussed in the previous chapter can be employed in the pre-training phase.

100 beats) in the associated time series. This filtering was done using a 7th order Butterworth digital filter with cutoff frequency of 0.01 cycles/beat. One example of the resulting time series is shown in Fig. 1.3. Next, a sequential labeling/classification task was constructed, involving the four maneuvers depicted in Fig. 1.3. We used four modes, each corresponding to a VAR model of order three, to model the bivariate time series of heart rate and blood pressure. The supervised learning algorithm was initialized using 10 iterations of the EM algorithm, followed by supervised learning with early stopping. The results shown in Fig. 1.3 indicate that the joint supervised learning significantly improves the multinomial probability over all sequence labels (panel (a)), as well as the accuracy of classification (panel (b)).

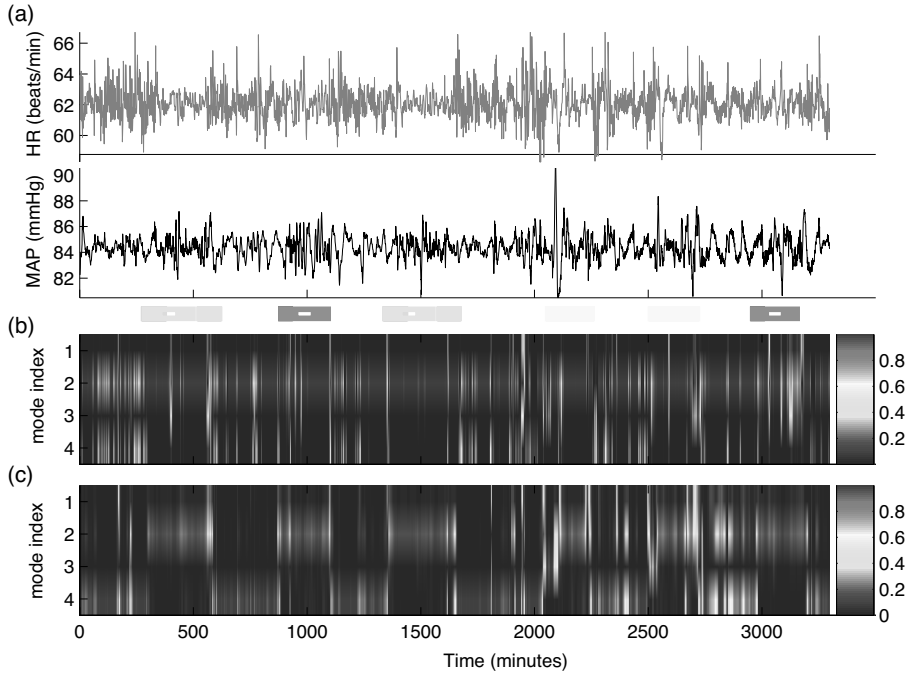

Figure 1.5 An example of a filtered time series of heart rate (HR) and mean blood pressure (MAP) from the tilt-table experiment (panel a). The inferred marginal probabilities of each of the four modes using the EM and the outcome-discriminative approaches are shown in panels (b) and (c), respectively.

As noted earlier, an outcome-discriminative dynamic Bayesian network can be can be viewed as a compactly-parameterized RNN. Therefore, to compare the performance of the algorithm discussed here against the RNN, we experimented with several implementations of RNNs within MATLAB® ANN package, including the layer recurrent neural networks (layrecnet), time delay neural network (timedelaynet), and distributed delay network (distdelaynet), with various number of hidden units and activation functions. The best performance was achieved using a layrecnet architecture with 2 input units, 10 hidden units and 4 output layers (using one-hot coding), which is similar to

feedforward networks, except that each layer has a recurrent connection with a tap delay associated with it. This allows the network to have an infinite dynamic response to time series input data. The best performing layrecnet network achieved a classification AUC of 60.0 [54.0 67.3] on the tilt-table dataset, which is significantly lower than the performance of the outcome-discriminative learning.

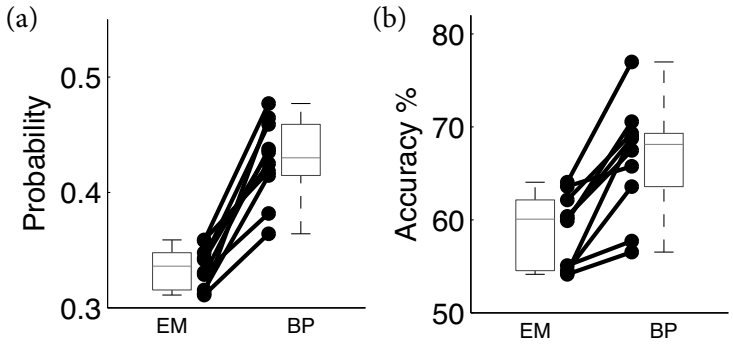

Figure 1.6 Comparison of EM and outcome-discriminative learning (BP) on the tilt-table dataset. Panel (a) shows 10-fold cross-validated performance of the EM versus BP (using 30 iteration of the BFGS algorithm with early stopping). Panel (b) shows a comparison of the two techniques in terms of classification accuracy.

Since we modeled the dynamics using a VAR model, we were able to derive the parametric power spectra corresponding to the individual channels of each time series (Nemati et al. 2011). Notably, we observed a progressive increase in the ratio of the low frequency (LF: periods of 6-20 beats) to the high frequency (HF: periods of 2-5 beats) power of the HR time series (also know as the LF/HF ratio; an index of sympathovagal activation) from supine to slow tilting, fast tilting, and standing. This indicates increased sympathetic modulations. These results were obtained by (1) calculating the parametric power spectrum of the HR for each mode, using its VAR coefficients, and (2) calculating a weighted average of the HR spectrum within the segments corresponding to each postural regime, where the weights were given by the probabilities of belonging to a given mode. The estimated increase in LF/HF ratio from supine to standing was significant with both learning techniques (EM: 4.6 [4.3, 5.4] to 8.4 [8.3, 8.6] † , supervised: 4.4 [3.7 4.8] to 5.53 [5.2 6.5] †, median [interquartiles]; † indicates $p < 0.05$ using Kruskal-Wallis nonparametric ANOVA test).

1.3.1 Decoding Local Field Potentials

Our final example is a binary time series classification task (decoding brain activity), involving bivariate time series of local field potentials (LFP) recorded from the visual area V4 and inferior temporal (IT) cortex of a rhesus macaque while performing an attention task. Each of the 420 trials lasted for 2.6 seconds, starting with the animal gazing at an illuminated location at the center of a computer screen in a dark room.

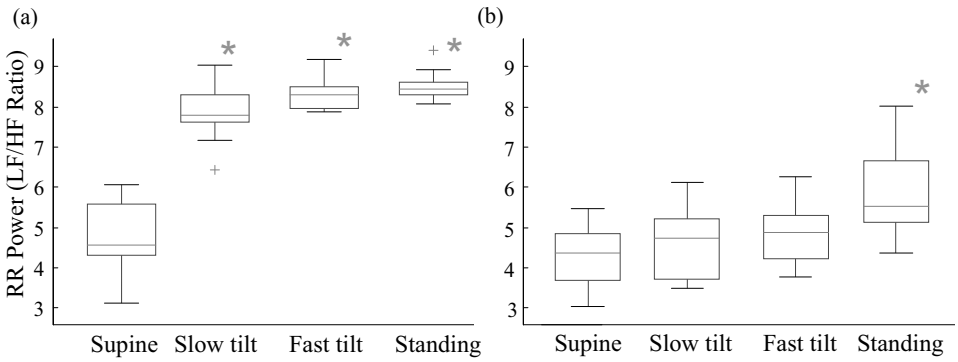

Figure 1.7 Physiological interpretation of learned dynamics. LF/HF ration for the EM and outcome-discriminative learning are shown in panels (a) and (b), respectively.The $\star$ symbol indicates a significant change from baseline ($p<0.05$; Kruskal-Wallis nonparametric ANOVA test).

An arrow then appeared (cue onset) to indicate the location of a target to appear on the screen (one of two possible locations: bottom versus top). Within roughly 500 milliseconds a target object appeared (stimuli onset). After a variable amount of time the target changed color (target change), indicating that the subject should make a saccade within a few tens of milliseconds. LFP time series were recorded at 1000Hz and were downsampled to 200Hz, yielding roughly 520 samples per time series. Selective attention requires communication among multiple brain regions in a timely manner, and consequently the resulting LFP time series are nonstationary. We modeled the data using a switching VAR model with five hidden states, each corresponding to an AR model of order three. Fig. 1.3.1 shows the decoding performance based on EM versus the proposed supervised approach. Consistent with the simulation study presented earlier, again we see that unsupervised pre-training via EM provides a good starting point for fine-tuning by the supervised algorithm.

1.4 Discussion and Conclusion

This chapter introduced a state-space modeling framework for multivariate time series classification and sequential labeling. Our approach was based on the idea of using the inferred marginals of hidden variables as inputs to a gradient-based supervised learner such as a logistic regression classifier. We showed that if the loss function defined on the marginals is differentiable, it will be possible to compute the gradient in terms of these marginals, and then backpropagate the loss gradient through the message passing inference procedure (e.g., the forward-backward algorithm). The resulting algorithm allowed for combining unsupervised pre-training with supervised fine-tuning to design and initialize a new class of RNNs for time series classification and sequential labeling.

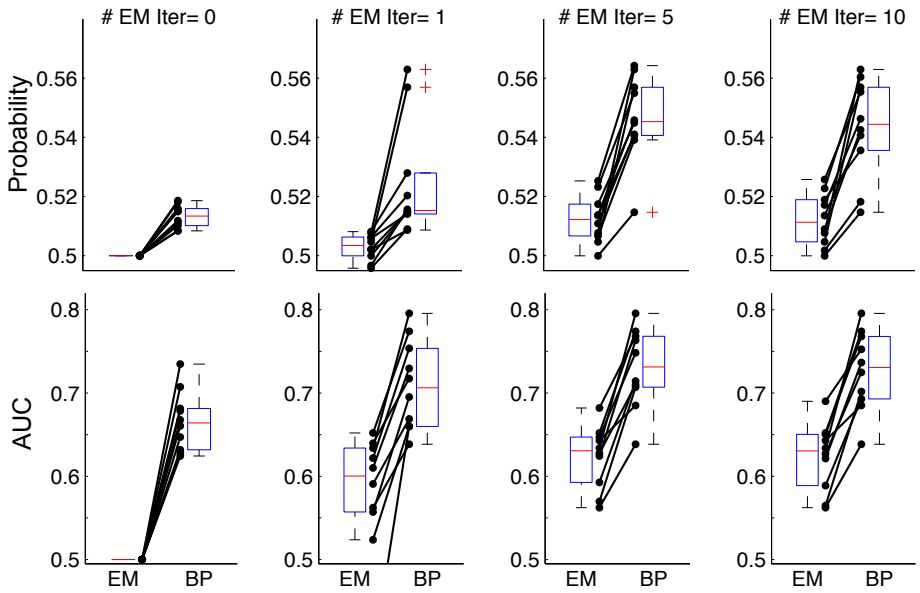

Figure 1.8 Effects of EM-based pre-training on the performance of supervised learning on the LFP decoding experiment. Each panel shows classification performance over ten folds (testing set performance) based on 0, 1, 5, 10, and 20 iterations of expectation maximization (EM) followed by 30 iterations of the supervised learning via backpropagation (BP). The cost function that is being optimized (Bernoulli probability of outcomes) is shown in the top row, and the area under receiver operating curve (AUC) is presented in the bottom row.

In contrast to generative and maximum likelihood-based approaches to feature learning in time series, the outcome-discriminative learning framework provides the learning algorithm with the outcomes (labels) corresponding to each time series sample (e.g., supine, slow-tilt, etc), and learns time series features that are maximally discriminative. In doing so we addressed two shortcomings of the competing neural networks, namely the black box nature of the RNNs and lack of a systematic approach to initialization of network weights in the classical RNNs. The technique developed in this chapter is significant from a theoretical point of view, since one may apply the the backpropagation-based learning described in this chapter to any probabilistic model, define on a directed acyclic graph structure.

Using simulated time series, we showed that outcome-discriminative learning provides a significant improvement over EM-based feature extraction and classification, and moreover benefit from a EM-based initialization. Furthermore, we demonstrated a significant improvement in classification accuracy when decoding postural changes involved in the tilt-table experiment, using the multivariate switching dynamics of HR and BP time series. Since the EM learning objective is the log likelihood of the unlabeled time series, it may learn artifacts and other features that are not relevant to classification. As expected, increasing the number of EM steps in the simulation study

(where high amplitude artifacts were randomly inserted into all time series) did not improve the discriminative performance, even though we observed a significant increase in training log likelihood. Notably, the EM-based initialization step is qualitatively similar to the unsupervised learning step used for training Deep Belief Networks (DBN) (Hinton, Osindero & Teh 2006), where unsupervised pre-training is known to significantly improve the predictive performance of discriminative neural networks (Erhan et al. 2010). The intuition is that pre-training puts us at a region within the parameter space that allows the discriminative learning to rapidly progress. Moreover, since the input is high-dimension, it is harder to overfit the input data versus the low dimensional labels (Hinton, Deng, Yu, Dahl, Mohamed, Jaitly, Senior, Vanhoucke, Nguyen, Sainath et al. 2012).

As demonstrated through the tilt-table example, the proposed approach has the added advantage of having physiological interpretability. Since the features used for prediction are based on the dynamics of the underlying time series, one can link the most predictive features back to the underlying physiology. For instance, tilting is known to disrupt the sympathovagal balance in the direction of increased sympathetic activation (Guzzetti, Piccaluga, Casati, Cerutti, Lombardi, Pagani & Malliani 1988). Notably, modes that were most probable during the tilting events had higher LF/HF ratios, indicating increased sympathetic modulation.

The method discussed in this chapter is directly applicable to outcome prediction in large physiological cohort time series, as described in the previous chapter. Other potential applications may include neural decoding and brain-machine interface (Wu, Black, Mumford, Gao, Bienenstock & Donoghue 2004) and automated speech and hand-writing recognition (Bahl, Brown, De Souza & Mercer 1986). Although here we only discussed a dynamic Bayesian network consisting of discrete latent variables, the marginals-based outcome discriminative approach is similarly applicable to models with mixture of continuous and discrete latent variables, such as the switching Kalman filter (Murphy 1998). Our ongoing and future works involve learning more expressive representations of time series from the inferred marginals. For instance, a convolutional neural network layer may replace the logistic classifier employed here, to extract additional features pertaining to rate of transition among dynamical modes, as well as, features that may represent long-range trends in evolution of the dynamical modes in nonstationary time series.

ACKNOWLEDGMENT

The authors would like to thank Professor George Verghese (MIT-EECS) for his insightful comments, and Dr. Thomas Heldt (MIT-HST) for kindly providing the tilt-table data analyzed in this study. This work was supported in part by the James S. McDonnell foundation postdoctoral grant.

1.5 APPENDIX

Here we present the analytic gradients of the switching VAR model, starting from the regression layer shown in Fig. 1.2, and recursively calculate the the error gradients of the discrete latent switching variables, and ultimately the model parameters Θ.

Hereafter we will use parentheses to index individual states, and we will use brackets to indicate the individual elements of a vector or a matrix. Thus, $A(i)[m,n]$ refers to the m-th row, n-th column element of the matrix of state dynamics for the i-th mode. We will use the symbol $\odot$ to denote the Frobenius inner product of two matrices (or vectors) defined as $A \odot B = \sum_i \sum_j A_{ij} B_{ij}$. Moreover, for a matrix B, indexed by i, j, the colon notation $B(i,:)$ denotes entries ranging over all j values. All the exponentiations involving matrices are element-wise. Finally, $\delta_{m,n}$ denotes a conformable matrix with the (m,n)-th element equal to one and zero elsewhere. Similarly, $\delta_{m,n}^{n,m}$ denotes a conformable matrix with ones at the (m,n)-th and (n,m)-th elements and zero elsewhere.

Gradient of the Regression Layer Parameters

Logistic regression is commonly used for predicting outcomes of categorical variable. For instance, each of the N time-series within a cohort may be associated with an outcome (or label) denoted by $\{O_n^{true}\}_{n=1}^N$). In this work, we use a multinomial regression methods to map the hidden state marginals to the outcome labels of interest at each time series sample. We first provide the analytic gradients of the corresponding error functions with respect to the regression parameters.

The error gradients with respect to the parameters $\beta_{k,j}$ of the multinomial regression and the predictor variables η_t (taken as equal to local smoothed marginals M_t^s) are given by

$$\frac{\partial E}{\partial \beta_{k,j}} = -\sum_{k'=1}^{K} O_{k'}^{true}(\delta_{k,k'} - \mu_k)\eta_j \tag{1.4}$$

$$\frac{\partial E}{\partial \eta_{t,j}} = -\sum_{k=1}^{K}\sum_{k'=1}^{K} O_{t,k'}^{true}(\delta_{k,k'} - \mu_{t,k})\beta_{k,j} \ , \tag{1.5}$$

1.5.1 Gradient of inference operations and hidden state marginals

We start from the filtering step of the switching VAR algorithm and calculate the analytical partial derivatives of each node output(s) with respect to its input(s), as we move forward in time. Next, smoothing of the switching variables is performed and the corresponding analytical gradients are calculated. The back-propagation algorithm starts from the reverse direction (from the output of the smoothed switching variables) and propagates the gradient information backward through the smoothed switching Variables ($t = 1$ to $t = T$), and finally the filtered variables ($t = T-1$ to $t = 1$).

The smoothing step of the switching VAR algorithm for the switching Variables takes the following form (Murphy 1998):

$(M_t^s(i)) = Backward(M_t^f,\ M_{t+1}^s,\ Z)$

$$a_t^s(i,j) = M_t^f(i)Z(i,j)\ , \tag{1.6}$$

$$b_t^s(i,j) = \frac{a_t^s(i,j)}{\sum_{i'} a_t^s(i',j)}\ , \tag{1.7}$$

$$M_t^s(i) = \sum_{j'} b_t^s(i,j')M_{t+1}^s(j')\ , \tag{1.8}$$

for $t = T-1,\cdots,1$. Note, $M_t^s(i) = Prob(S_t = i|y_{1:T})$, with the initial condition $M_T^s = M_T^f$.

Derivatives of the error with respect to the mode proportions η_i are given Eq. (1.5) (in the case of multinomial outcomes). Next, the error is backpropagated through the smoothed switching Variables, as follows:

$$\begin{aligned}\frac{\partial E}{\partial M_1^s(i)} &= \frac{1}{T}\frac{\partial E}{\partial \eta_{\mathbf{t},\mathbf{i}}}\ ,\\ \frac{\partial E}{\partial M_t^s(i)} &= \frac{1}{T}\frac{\partial E}{\partial \eta_{\mathbf{t},\mathbf{i}}} + \sum_{j'}\frac{\partial E}{\partial M_{t-1}^s(j')}b_{t-1}^s(j',i)\ ,\ t = 2\cdots T\end{aligned} \tag{1.9}$$

We also compute the following derivatives:

$$\frac{\partial E}{\partial a_t^s(i,j)} = \sum_{k'}\frac{\partial E}{\partial M_t^s(k')}M_{t+1}^s(j)[\frac{\delta_{k',i}}{\sum_{i'} a_t^s(i',j)} - \frac{a_t^s(k',j)}{(\sum_{i'} a_t^s(i',j))^2}]\ ,\ t = 1\cdots T - \tag{1.10}$$

$L_t(j) = Liklihood(\mathbf{y}_{t-1}, \mathbf{y}_t;\, A(j),\, Q(j))$

$$\mathbf{e}_t = \mathbf{y}_t - A(j)\mathbf{y}_{t-1}\ , \tag{1.11}$$

$$L_t(j) = \mathscr{N}(\mathbf{e}_t;\, 0,\, Q(j))\ . \tag{1.12}$$

With the following derivatives:

$$\frac{\partial L_t}{\partial \mathbf{e}_t} = -L_t Q^{-1}\mathbf{e}_t \tag{1.13}$$

$$\frac{\partial L_t}{\partial A[m,n]} = \frac{\partial L_t}{\partial \mathbf{e}_t} \odot (-\delta_{m,n}\mathbf{y}_{t-1}) \tag{1.14}$$

$$\frac{\partial L_t}{\partial Q} = -\frac{1}{2}L_t(Q^{-1} - Q^{-1}(\mathbf{e}_t\mathbf{e}_t^\intercal)Q^{-1}) \tag{1.15}$$

(1.16)

$M_t^f(j) = Forward(M_{t-1}^f, L_t, Z)$

$$a_t^f(j) = L_t(j)M_{t-1}^f(j)Z(i,j)\ , \tag{1.17}$$

$$M_t^f(j) = \frac{a_t^f(j)}{\sum_{j'} a_t^f(j')}. \tag{1.18}$$

For $t = 1, \cdots, T$. Note, $M_t^f(i) = Prob(S_t = i|y_{1:t})$, with the initial condition $M_0^f = \pi$. The partial derivatives are given by

$$\frac{\partial M_t^f(i)}{\partial M_{t-1}^f(j)} = \sum_{k'}[\frac{L_t(i)Z(j,i)\delta_{k',j}}{\sum_{i'} a_t^f(i')} - \frac{L_t(k')Z(j,k')a_t^f(i)}{(\sum_{i'} a_t^f(i'))^2}]\ ,\ t = 1\cdots T. \tag{1.19}$$

$$\frac{\partial M_t^f(k)}{\partial Z(i,j)} = M_{t-1}^f(i)L_t(j)[\frac{\delta_{k,j}}{\sum_{k'} a_t^f(k')} - \frac{a_t(k)}{(\sum_{k'} a_t^f(k'))^2}]. \tag{1.20}$$

$$\frac{\partial M_t^f(k)}{\partial L_t(j)} = \sum_{i'} M_{t-1}^f(i')Z_t^f(i',j)[\frac{\delta_{k,j}}{\sum_{k'} a_t^f(k)} - \frac{a_t^f(k)}{(\sum_{k'} a_t^f(k'))^2}]. \tag{1.21}$$

(1.22)

$$\frac{\partial E}{\partial M_T^f(i)} = \frac{\partial E}{\partial M_T^s(i)}\ ,$$

$$\frac{\partial E}{\partial M_t^f(i)} = \sum_{j'} \frac{\partial E}{\partial a_t^s(i,j')}Z(i,j') + \frac{\partial E}{\partial M_{t+1}^f} \odot \frac{\partial M_{t+1}^f}{\partial M_t^f(i)}\ ,\ t = T-1\cdots 1 \tag{1.23}$$

Gradient with respect to the model parameters

We finally arrive at the error gradients with respect to the model parameters. The derivatives with respect to the Markov switching state transition matrix are given by:

$$\frac{\partial E}{\partial Z(i,j)} = \sum_{t=1}^{T} \frac{\partial E}{\partial M_t^f(j)} \odot \frac{\partial M_t^f(j)}{\partial Z(i,j)} + \sum_{t=1}^{T-1} \frac{\partial E}{\partial a_t^s(i,j)} M_t^f(i) \tag{1.24}$$

For the other model parameters the error gradients are as follows:

$$\frac{\partial E}{\partial A(j)[m,n]} = \sum_{t=1}^{T} \sum_{j'=1}^{J} \frac{\partial E}{\partial M_t^f(j')} \frac{\partial M_t^f(j')}{\partial L_t(j)} \frac{\partial L_t(j)}{\partial A(j)[m,n]}. \tag{1.25}$$

$$\frac{\partial E}{\partial Q(j)[m,n]} = \sum_{t=1}^{T} \sum_{j'=1}^{J} \frac{\partial E}{\partial M_t^f(j')} \frac{\partial M_t^f(j')}{\partial L_t(j)} \frac{\partial L_t(j)}{\partial Q(j)[m,n]}. \tag{1.26}$$

Removing the Constraints via Parameter Transformations

We can convert the constrained optimization problem to an equivalent unconstrained problem by defining the following transformations:
Let $\bar{Z}(i,j)$ be such that $Z(i,j) = \frac{exp(\bar{Z}(i,j))}{\sum_{j'} exp(\bar{Z}(i,j'))}$, which results in the following gradient for $\bar{Z}$:

$$\frac{\partial E}{\partial \bar{Z}(i,j)} = \frac{\partial E}{\partial \bar{Z}} \odot \frac{\partial Z}{\partial \bar{Z}(i,j)}, \tag{1.27}$$

where

$$\frac{\partial Z(k,l)}{\partial \bar{Z}(i,j)} = \delta_{i,k} Z(i,j)(\delta_{j,l} - Z(k,l)). \tag{1.28}$$

Furhermore, to ensure we optimize over the space of positive semi-definite matrices, we use the Cholesky decomposition representation of the covariance matrices. For instance, in the case of the state-noise covariance matrices, we represent $Q(j) = \Gamma(j)\Gamma(j)^\intercal$, where ($\Gamma(j)$ is a lower diagonal matrix). Then $\frac{\partial E}{\partial \Gamma(j)} = \frac{\partial E}{\partial Q(j)} \frac{\partial Q(j)}{\partial \Gamma(j)}$, given by:

$$\frac{\partial E}{\partial \Gamma(j)[m,n]} = \frac{\partial E}{\partial Q(j)} \odot (\delta_{m,n}\Gamma(j)^\intercal + \Gamma(j)\delta_{m,n}^\intercal), \tag{1.29}$$

and the corresponding gradient vector includes only the lower diagonal elemnts of $\frac{\partial E}{\partial \Gamma(j)}$.

Notes

[1]Lewis Fry Richardson (1881–1953).

Bibliography

Bahl, L., Brown, P., De Souza, P. & Mercer, R. (1986), Maximum mutual information estimation of hidden markov model parameters for speech recognition, *in* 'Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'86.', Vol. 11, IEEE, pp. 49–52.

Bengio, Y., Courville, A. & Vincent, P. (2013), 'Representation learning: A review and new perspectives', *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **35**(8), 1798–1828.

Costa, M., Goldberger, A. L. & Peng, C. K. (2002), 'Multiscale entropy analysis of complex physiologic time series', *Phys Rev Lett* **89**(6), 068102.

Domke, J. (2013), 'Learning graphical model parameters with approximate marginal inference', *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE* **35**(10), 2454–2467.

Eaton, F. & Ghahramani, Z. (2009), 'Choosing a variable to clamp: approximate inference using conditioned belief propagation', *n Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics I* **5**, 145–152.

Erhan, D., Bengio, Y., Courville, A., Manzagol, P., Vincent, P. & Bengio, S. (2010), 'Why does unsupervised pre-training help deep learning?', *The Journal of Machine Learning Research* **11**, 625–660.

Graves, A. (2012), *Supervised sequence labelling with recurrent neural networks*, Vol. 385, Springer.

Graves, A., Fernández, S., Gomez, F. & Schmidhuber, J. (2006), Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks, *in* 'Proceedings of the 23rd international conference on Machine learning', ACM, pp. 369–376.

Guzzetti, S., Piccaluga, E., Casati, R., Cerutti, S., Lombardi, F., Pagani, M. & Malliani, A. (1988), 'Sympathetic predominance an essential hypertension: a study employing spectral analysis of heart rate variability', *Journal of hypertension* **6**(9), 711–717.

Heldt, T., Oefinger, M. B., Hoshiyama, M. & Mark, R. G. (2003), 'Circulatory response to passive and active changes in posture', *Computers in Cardiology* **30**, 263–266. Circulatory response to passive and active changes in posture.

Heskes, T. & Zoeter, O. (2002), Expectation propagation for approximate inference in dynamic bayesian networks, *in* 'Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence', Morgan Kaufmann Publishers Inc., pp. 216–223.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N. et al. (2012), 'Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups', *Signal Processing Magazine, IEEE* **29**(6), 82–97.

Hinton, G., Osindero, S. & Teh, Y. (2006), 'A fast learning algorithm for deep belief nets', *Neural computation* **18**(7), 1527–1554.

Ivanov, P. C., Rosenblum, M. G., Peng, C. K., Mietus, J., Havlin, S., Stanley, H. E. & Goldberger, A. L. (1996), 'Scaling behaviour of heartbeat intervals obtained by wavelet-based time-series analysis.', *Nature* **383**(6598), 323–327.

Kim, M. & Pavlovic, V. (2009), 'Discriminative learning for dynamic state prediction', *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **31**(10), 1847–1861.

Lafferty, J., McCallum, A. & Pereira, F. C. (2001), 'Conditional random fields: Probabilistic models for segmenting and labeling sequence data'.

Lasko, T. A., Denny, J. C. & Levy, M. A. (2013), 'Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data', *PloS one* **8**(6), e66341.

Lehman, L., Adams, R., Mayaud, L., Moody, G., Malhotra, A., Mark, R. & Nemati, S. (2014), 'A physiological time series dynamics-based approach topatient monitoring and outcome prediction', *Biomedical and Health Informatics, IEEE Journal of* **PP**(99), 1–1.

Marlin, B. M., Kale, D. C., Khemani, R. G. & Wetzel, R. C. (2012), Unsupervised pattern discovery in electronic health care data using probabilistic clustering models, *in* 'Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium', ACM, pp. 389–398.

McCallum, A., Freitag, D. & Pereira, F. C. N. (2000), Maximum entropy markov models for information extraction and segmentation, *in* 'Proceedings of the Seventeenth International Conference on Machine Learning', ICML '00, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 591–598.

Memisevic, R. (2006), An introduction to structured discriminative learning, Technical report, Citeseer.

Murphy, K. P. (1998), 'Switching kalman filter', *Compaq Cambridge Research Laboratory, Tech. Rep. 98-10.* . Cambridge, MA.

Nemati, S., Edwards, B. A., Sands, S. A., Berger, P. J., Wellman, A., Verghese, G. C., Malhotra, A. & Butler, J. P. (2011), 'Model-based characterization of ventilatory stability using spontaneous breathing', *J Appl Physiol* **111**(1), 55–67. eng.

Nemati, S., Lehman, L.-w. H., Adams, R. P. & Malhotra, A. (2012), Discovering shared cardiovascular dynamics within a patient cohort, *in* 'Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE', IEEE, pp. 6526–6529.

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1988), 'Learning representations by back-propagating errors', *Cognitive modeling* .

Stein, P. K., Domitrovich, P. P., Huikuri, H. V., Kleiger, R. E. & , C. I. (2005), 'Traditional and nonlinear heart rate variability are each independently associated with mortality after myocardial infarction.', *J Cardiovasc Electrophysiol* **16**(1), 13–20.

Stoyanov, V., Ropson, A. & Eisner, J. (2011), Empirical risk minimization of graphical model parameters given approximate inference, decoding, and model structure, *in* 'Proceedings of AISTATS'.

Sutskever, I. (2013), Training recurrent neural networks, PhD thesis, University of Toronto.

Woodland, P. & Povey, D. (2000), Large scale discriminative training for speech recognition, *in* 'ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)'.

Wu, W., Black, M. J., Mumford, D., Gao, Y., Bienenstock, E. & Donoghue, J. P. (2004), 'Modeling and decoding motor cortical activity using a switching kalman filter', *Biomedical Engineering, IEEE Transactions on* **51**(6), 933–942.

Author index

Peterson, K., 42
Tranah, D.A., 42
Young, P.D.F., 42

Subject index

diffraction, 42
force
  hydrodynamic, 42
  interactive, 42