Parsing Epileptic Events Using a Markov Switching Process for Correlated Time Series

Drausin F. Wulsin WULSIN@SEAS.UPENN.EDU
Dept. of Bioengineering, University of Pennsylvania, Philadelphia, PA USA

Emily B. Fox EBFOX@UW.EDU
Dept. of Statistics, University of Washington, Seattle, WA USA

Brian Litt LITTB@MAIL.MED.UPENN.EDU
Depts. of Neurology and Bioengineering, University of Pennsylvania, Philadelphia, PA USA

Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic "bursts" in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights into the nature and intrinsic dynamics of seizures. A goal of our work is to parse these complex epileptic events into distinct dynamic regimes. A challenge posed by the intracranial EEG (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable numbers of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics. We demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

Proceedings of the 30^{th} International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible enough to describe epileptic events but restrictive enough to distill intelligible information from them. Much of the recent machine learning work in EEG analysis has focused on seizure prediction, (cf., D'Alessandro et al., 2005; Mirowski et al., 2009), an important area of study but one that generally has not focused on parsing the EEG directly, as a human EEG reader would. Such parsings are central for diagnosis and relating various types of abnormal activity. Recent evidence shows that the range of epileptic events extends beyond clinical seizures to include shorter, sub-clinical "bursts" lasting fewer than 10 seconds (Litt et al., 2001). What is the relationship between these shorter bursts and the longer seizures? In this work, we demonstrate that machine learning techniques can have substantial impact in this domain by unpacking how seizures begin, progress, and end.

In particular, we build a Bayesian nonparametric time series model to analyze intracranial electroencephalogram (iEEG) data. We take a modeling approach similar to a physician's in analyzing EEG events: look directly at the evolution of the raw EEG voltage traces. EEG signals exhibit nonstationary behavior during a variety of neurological events, and time-varying autoregressive (AR) processes have been proposed to model single channel data (Krystal et al., 1999). Here we aim to parse the recordings into interpretable regions of activity and thus propose to use autoregressive hidden Markov models (AR-HMMs) to define *locally* stationary processes. In the presence of multiple chan-

nels of simultaneous recordings, as is almost always the case in EEG, we wish to share AR states between the channels while allowing for asynchronous switches. The recent beta process (BP) AR-HMM of (Fox et al., 2009) provides a flexible model of such dynamics: a shared library of infinitely many possible AR states is defined and each time series uses a finite subset of the states. The process encourages sharing of AR states, while allowing for time-series-specific variability.

The BP-AR-HMM assumes independence between time series. In the case of iEEG, this assumption is almost assuredly false. Fig. 1 shows an example of a 4x8 intracranial electrode grid and the residual EEG traces of 16 channels *after* subtracting the predicted value in each channel using a conventional BP-AR-HMM. While the error term in some channels remains low throughout the recording, other channels—especially those spatially adjacent in the electrode grid—have very correlated error traces. We propose to capture correlations between channels by modeling a multivariate innovations process that drives independently evolving channel dynamics. We demonstrate the importance of accounting for this error trace in predicting heldout seizure recordings, making this a crucial modeling step before undertaking large-scale EEG analysis.

To aid in scaling to large electrode grids, we exploit a sparse dependency structure for the multivariate innovations process. In particular, we assume a graph with known vertex structure that encodes conditional independencies in the multivariate innovations process. The graph structure is based on the spatial adjacencies of the iEEG channels, with a few exceptions to make the graphical model fully decomposable. Fig. 1 (bottom left) shows an example of such a graphical model over the channels. Although the relative position of channels in the electrode grid is clear, determining the precise 3D location of each channel is extremely difficult. Unlike in scalp EEG or magentoencephalogram (MEG), which have generally consistent channel positions from patient to patient, iEEG channels vary in number and position for each patient, impeding the use of alternative spatial and multivariate time series modeling techniques.

It is well-known that the correlations between EEG channels usually vary during the beginning, middle, and end of a seizure (Schiff et al., 2005; Schindler et al., 2007). Prado et al. (2006) employ a mixture-of-expert vector autoregressive (VAR) model to describe the different dynamics present in seven channels of scalp EEG. We take a similar approach by allowing for a Markov evolution to an underlying innovations covariance state.

An alternative modeling approach is to treat the channel recordings as a single multivariate time series, perhaps using a switching VAR process as in (Prado et al., 2006). However, such an approach (i) assumes synchronous switches in dynamics between channels, (ii) scales poorly with the number of channels, and (iii) requires identical numbers of channels between patients to share dynamics between event recordings.

We show that our model for correlated time series has better out-of-sample predictions of iEEG data than standard AR- and BP-AR-HMMs and demonstrate the utility of our model in comparing short, subclinical epileptic bursts with longer, clinical seizures. Our inferred parsings of iEEG data concur with key features hand-annotated by clinicians but provide additional insight beyond what can be extracted from a visual read of the data. The importance of our methodology is multifold: (i) the output is interpretable to a practitioner and (ii) the parsings can be used to relate seizure types both within and between patients even with different electrode setups. Enabling such broad-scale automatic analysis, and identifying dynamics unique to sub-clinical seizures, can lead to new insights in epilepsy treatments.

Although we are motivated by the study of seizures from iEEG data, our work is much more broadly applicable in time series analysis. For example, perhaps one has a collection of stocks and wants to model shared dynamics between them while capturing changing correlations. The BP-AR-HMM was applied to the analysis of a collection of motion capture data assuming independence between individuals; our modeling extension could account for coordinated motion with a sparse dependency structure between individuals. Regardless, we find the impact in the neuroscience domain to be quite significant.

2. A Structured Bayesian Nonparametric Factorial AR-HMM

Observation model Consider an event, a seizure for example, comprised of N univariate time series, which in our case are the voltage measurements of N different EEG electrode channels. We assume that each time series in an event contains T scalar observations, $y_t^{(i)}$. We model $y_t^{(i)}$ as an order r AR-HMM:

$$z_t^{(i)} \sim \boldsymbol{\pi}_{z_{t-1}^{(i)}}, \quad y_t^{(i)} = \mathbf{a}_{z_t^{(i)}}^{\mathrm{T}} \widetilde{\mathbf{y}}_t^{(i)} + \boldsymbol{\epsilon}_t^{(i)}, \tag{1}$$

where $z_t^{(i)}$ denotes the dynamical state of channel i at time t, $\boldsymbol{\pi}_{z_{t-1}^{(i)}}$ the transition distribution given the previous state $z_{t-1}^{(i)}$, $\mathbf{a}_k$ the r AR coefficients associated with channel state k, and $\widetilde{\mathbf{y}}_t^{(i)} = [y_{t-1}^{(i)}, \ldots, y_{t-r}^{(i)}]^{\mathrm{T}}$.

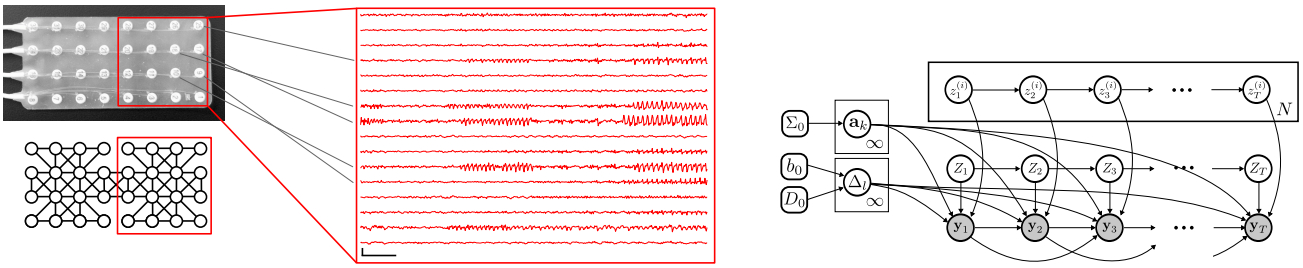

Figure 1. (**top left**) An iEEG grid electrode and (**bottom left**) corresponding graphical model. (**middle**) Residual EEG values *after* subtracting predictions from a BP-AR-HMM assuming independent channels. All EEG scale bars indicate 1 mV vertically and 1 second horizontally. (**right**) Graphical model of the HIW-spatial BP-AR-HMM. Channel states $z_t^{(i)}$ evolve independently for each channel according to feature-constrained transition distributions (omitted for simplicity), and index the AR dynamic parameters $\mathbf{a}_k$ used in generating observation $y_t^{(i)}$. The Markov-evolving event state Z_t indexes the graph-structured covariance Δ_l of the correlated AR innovations resulting in multivariate observations $\mathbf{y}_t = [y_t^{(1)}, \ldots, y_t^{(N)}]^{\mathrm{T}}$ sharing the same conditional independencies.

Importantly, our channels do not evolve independently. We capture the channel correlations via the driving innovations process $\boldsymbol{\epsilon}_t = [\epsilon_t^{(1)}, \ldots, \epsilon_t^{(N)}]^{\mathrm{T}}$. In particular, we assume event-state-specific correlations via

$$Z_t \sim \phi_{Z_{t-1}}, \quad \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \Delta_{Z_t}), \tag{2}$$

where Z_t denotes a Markov-evolving event state, which is distinct from the individual channel states $z_t^{(i)}$, and ϕ_l denotes the event state transition distribution. The flexibility introduced by the event state is particularly important in applications like seizure modeling, where the channels may display one innovation covariance before a seizure (e.g., relatively independent and low-magnitude) but quite a different covariance during a seizure (e.g., correlated, higher magnitude).

Emission parameters To scale to large numbers of electrodes, and to incorporate the physical relationships of the channels, we define a sparse channel dependency structure by introducing a graphical model G and specifying a hyper-inverse Wishart (HIW) prior on Δ_l. The HIW prior (Dawid & Lauritzen, 1993) enforces the hyper-Markov conditions specified by G, leading to conditional independencies in $\boldsymbol{\epsilon}_t$ (and thus in $\mathbf{y}_t$). The AR coefficients $\mathbf{a}_k$ are given a multivariate normal prior. Together, we have

$$\Delta_l \sim \mathrm{HIW}_G(b_0, D_0), \quad \mathbf{a}_k \sim \mathcal{N}(\mathbf{m}, \Sigma_0). \tag{3}$$

Here, b_0 denotes the degrees of freedom and D_0 the scale matrix. We consider $\mathbf{m} = 0$ throughout.

For compactness, we sometimes alternately write

$$\mathbf{y}_t = \mathbf{A}_{\mathbf{z}_t} \widetilde{\mathbf{Y}}_t + \boldsymbol{\epsilon}_t(Z_t), \tag{4}$$

where $\mathbf{y}_t$ is the concatenation of N channel observations at time t and $\mathbf{z}_t$ is the vector of channel states. One can think of this process as a factorial HMM (Ghahramani & Jordan, 1997) since we have $N+1$ independently evolving Markov chains that jointly generate our observation vector $\mathbf{y}_t$. However, here we have a *sparse* dependency structure in how the Markov chains influence a given observation $\mathbf{y}_t$, as induced by the conditional independencies in $\boldsymbol{\epsilon}_t$. See Fig. 1 (right).

Feature constrained channel transition distributions A key goal in modeling the event is to capture shared dynamics across the N related time series (channels). Each channel exhibits some subset of a shared library of AR coefficients $\{\mathbf{a}_k\}$. Let $\mathbf{f}^{(i)}$ be a binary feature vector associated with channel i with $f_k^{(i)} = 1$ indicating that channel i uses the dynamic $\mathbf{a}_k$. The BP-AR-HMM of Fox et al. (2009) provides our sought after framework for defining such a feature model in order to constrain a set of AR-HMM transitions. In particular, through employing a beta process prior (Thibaux & Jordan, 2007), the BP-AR-HMM allows for an infinite library of AR parameters and encourages each time series to use a sparse subset of these parameters with a flexible sharing pattern.

More formally, in our scenario the feature assignments $f_k^{(i)}$ and their corresponding parameters $\mathbf{a}_k$ are generated by an underlying beta process random measure:

$$B \sim \mathrm{BP}(1, B_0), \; B = \sum_{k=1}^{\infty} \omega_k \delta_{\mathbf{a}_k}, \; f_k^{(i)} \sim \mathrm{Ber}(\omega_k). \tag{5}$$

B defines an infinite collection of feature inclusion probabilities ω_k and AR coefficients $\mathbf{a}_k \in \Omega$, with B_0 a base measure on our parameter space Ω. The resulting feature vectors $\mathbf{f}^{(i)}$ constrain the set of available states $z_t^{(i)}$ can take by constraining the transition distributions, $\boldsymbol{\pi}_j^{(i)}$, to be 0 when $f_k^{(i)} = 0$. In particular, we use $\mathbf{f}^{(i)}$ along with a set of gamma random variables,

to produce the desired transition distribution $\boldsymbol{\pi}_j^{(i)}$ from state j to state k,

$$\eta_{jk}^{(i)} \sim \text{Gamma}(\gamma_c + \delta_{j,k}\kappa_c), \quad \boldsymbol{\pi}_j^{(i)} = \frac{\boldsymbol{\eta}_j^{(i)} \circ \mathbf{f}^{(i)}}{\sum_{k \mid f_k^{(i)}=1} \eta_{jk}^{(i)}}, \tag{6}$$

where $\circ$ denotes the Hadamard (element-wise) product and $\delta_{j,k}$ the Kronecker delta. The positive elements of $\boldsymbol{\pi}_j^{(i)}$ can also be thought of as a sample from a Dirichlet distribution with only $K^{(i)}$ dimensions, where $K^{(i)} = \sum_k f_k^{(i)}$ represents the number of states channel i uses. The parameter κ_c encourages self-transitions, as in the sticky HDP-HMM (Fox et al., 2011a).

Unconstrained event transition distributions We assume a Bayesian nonparametric formulation for the Markov event state process $\{Z_t\}$, as well, by taking $\boldsymbol{\phi}$ to be as in the HDP-HMM (Fox et al., 2011a; Teh et al., 2006). For simplicity, we consider the weak limit approximation (Ishwaran & Zarepour, 2002):

$$\begin{aligned} \boldsymbol{\beta} &\sim \text{Dir}(\gamma_e/L, \ldots, \gamma_e/L), \\ \boldsymbol{\phi}_l &\sim \text{Dir}(\alpha_e \boldsymbol{\beta} + \mathbf{e}_l \kappa_e), \end{aligned} \tag{7}$$

where $\mathbf{e}_l$ is the lth column of identity and L is assumed much greater than the expected number of states. Again, the sticky parameter κ_e promotes self-transitions, reducing state redundancy.

Our resulting structured Bayesian nonparametric factorial HMM is depicted in the graphical model of Fig. 1. We refer to this model as the *HIW-spatial BP-AR-HMM* to denote the dependencies introduced via the innovations process. We note that the infinite factorial HMM of Van Gael et al. (2008) considers a very different structure, allowing for an infinite collection of chains each with a binary state space. The infinite hierarchical HMM (Heller et al., 2009) also considers infinitely many chains with finite state spaces, but with constrained transitions between the chains in a top down fashion. The infinite DBN of Doshi-Velez et al. (2011) considers more general connection structures and arbitrary state spaces. Alternatively, the graph-coupled HMM of Dong et al. (2012) allows graph-structured dependencies in the underlying states of some N Markov chains. Here, we consider a finite set of chains with infinite state spaces that evolve independently and instead capture sparse dependencies in the observations via the innovations driving the AR dynamics.

3. Posterior Computations

Although the components of our model related to the individual channel dynamics are similar to those in the BP-AR-HMM, our posterior computations are significantly different due to the coupling of the Markov chains via the observations $\mathbf{y}_t$. In the BP-AR-HMM, conditioned on the feature assignments, each time series is independent. Here, however, we are faced with a factorial HMM structure and the associated challenges. For example, consider the observation model of Eq. (4). Even conditioned on the event sequence $Z_{1:T}$ and model parameters, we cannot analytically marginalize the state *vector* sequence $\mathbf{z}_{1:T}$ (e.g., via a forward-backward algorithm) since the state space of $\mathbf{z}_t$ is exponentially large. Luckily, the scale of these challenges is mitigated by our underlying graph structure. Conditioned on channel sequences $\{z_{1:T}^{(j)}\}_{j \in \mathbf{i}'}$, we *can* marginalize $z_{1:T}^{(i)}$; because of the graph structure, we need only condition on a *sparse* set of other channels (i.e., neighbors in the graph denoted here by $\mathbf{i}'$). Of course, the dependencies also have to be accounted for in sampling the dynamic parameters $\mathbf{a}_k$.

Algorithm 1 Outline of one MCMC iteration

for channels $i = 1, \ldots, N$ **do**
  sample active features,
    $\mathbf{f}^{(i)} \mid \mathbf{y}_{1:T}, \mathbf{z}_{1:T}^{(\mathbf{i}')}, Z_{1:T}, \mathbf{f}^{(-i)}, \boldsymbol{\eta}^{(i)}, \{\mathbf{a}_k\}, \{\Delta_l\}$
  sample state sequences,
    $z_{1:T}^{(i)} \mid \mathbf{y}_{1:T}, \mathbf{z}_{1:T}^{(\mathbf{i}')}, Z_{1:T}, \mathbf{f}^{(i)}, \boldsymbol{\eta}^{(i)}, \{\mathbf{a}_k\}, \{\Delta_l\}$
  sample transition parameters,
    $\boldsymbol{\eta}^{(i)} \mid z_{1:T}^{(i)}, \mathbf{f}_k^{(i)}$
end for
for active features $k \in \{k \mid \sum_i f_k^{(i)} > 0\}$ **do**
  sample AR coefficients,
    $\mathbf{a}_k \mid \mathbf{y}_{1:T}, \mathbf{z}_{1:T}, Z_{1:T}, \{\mathbf{a}_{k'}\}_{k' \neq k}, \{\Delta_l\}$
end for
sample event state sequence,
  $Z_{1:T} \mid \mathbf{y}_{1:T}, \mathbf{z}_{1:T}, \boldsymbol{\phi}, \{\mathbf{a}_k\}, \{\Delta_l\}$
sample event transition parameters,
  $\boldsymbol{\phi} \mid Z_{1:T}, \boldsymbol{\beta} \quad \boldsymbol{\beta} \mid Z_{1:T}$
for event states $l = 1, \ldots, L$ **do**
  sample innovation covariance,
    $\Delta_l \mid \mathbf{y}_{1:T}, \mathbf{z}_{1:T}, Z_{1:T}, \{\mathbf{a}_k\}$
end for

Algorithm 1 provides a high level overview of the steps involved in one iteration of MCMC sampling, with more details below and complete derivations provided in Supplement B. For brevity, we omit the hyperparameters from the conditioning set throughout.

Individual channel variables We harness the fact that we can compute the marginal conditional likelihood given $\mathbf{f}^{(i)}$ and the neighborhood set of other channels $\mathbf{z}_{1:T}^{(\mathbf{i}')}$ in order to block sample $\{\mathbf{f}^{(i)}, z_{1:T}^{(i)}\}$. That is, we first sample $\mathbf{f}^{(i)}$ marginalizing $z_{1:T}^{(i)}$ and then

sample $z_{1:T}^{(i)}$ given the sampled $\mathbf{f}^{(i)}$. Sampling the active features $\mathbf{f}^{(i)}$ for channel i follows as in Fox et al. (2009), using the Indian buffet process (IBP) (Griffiths & Ghahramani, 2005) predictive representation associated with the beta process, but using a likelihood term that conditions on neighboring channel state sequences $\mathbf{z}_{1:T}^{(\mathbf{i}')}$ and observations $\mathbf{y}_{1:T}^{(\mathbf{i}')}$. We additionally condition on the event state sequence $Z_{1:T}$ to define the sequence of distributions on the innovations. Generically, this yields (omitting the dependency on $\boldsymbol{\eta}^{(i)}, \{\mathbf{a}_k\}, \{\Delta_l\}$)

$$p(f_k^{(i)} \mid \mathbf{F}^{-ik}, y_{1:T}^{(i)}, \mathbf{y}_{1:T}^{(\mathbf{i}')}, \mathbf{z}_{1:T}^{(\mathbf{i}')}, Z_{1:T}) \propto p(f_k^{(i)} \mid \mathbf{F}^{-ik}) p(y_{1:T}^{(i)} \mid \mathbf{y}_{1:T}^{(\mathbf{i}')}, \mathbf{z}_{1:T}^{(\mathbf{i}')}, Z_{1:T}, \mathbf{f}^{(i)}). \quad (8)$$

Here, the first term is given by the IBP prior and the second term is the marginal conditional likelihood as derived in Supplement A. The quantity $\mathbf{F}^{-ik}$ defines the indicators for features other than k for time series other than i. Supplement B contains details on the feature sampling of Fox et al. (2009).

Conditioned on $\mathbf{f}^{(i)}$, we block sample the state sequence $z_{1:T}^{(i)}$ by first calculating backward messages $\boldsymbol{\psi}_t$ for $t = 1, \ldots, T$ and then forward sampling (again omitting dependency on $\boldsymbol{\eta}^{(i)}, \{\mathbf{a}_k\}, \{\Delta_l\}$),

$$z_t^{(i)} \mid y_{t:T}^{(i)}, \mathbf{y}_{t:T}^{(\mathbf{i}')}, z_{t-1}^{(i)}, \mathbf{z}_{t:T}^{(\mathbf{i}')}, Z_{1:T}, \mathbf{f}^{(i)} \sim \boldsymbol{\pi}_{z_{t-1}^{(i)}}^{(i)} \circ \mathbf{u}_t \circ \boldsymbol{\psi}_t, \quad (9)$$

where $\mathbf{u}_t$ denotes the likelihoods under each of the available $K^{(i)}$ channel states. Recall that $\boldsymbol{\pi}^{(i)}$ is a function of $\mathbf{f}^{(i)}$ and $\boldsymbol{\eta}^{(i)}$. Details are in Supplement B.

For sampling the transition parameters $\boldsymbol{\eta}^{(i)}$, we follow the correction described by Hughes et al. (2012, Supplement) and sample from the posterior given by

$$p(\eta_{jk}^{(i)} \mid z_{1:T}^{(i)}, f_k^{(i)}) \propto \frac{(\eta_{jk}^{(i)})^{n_{jk}^{(i)} + \gamma_c + \delta_{j,k}\kappa_c - 1} e^{\eta_{jk}^{(i)}}}{\sum_{k' \mid f_k^{(i)}=1} \eta_{jk'}^{(i)}}, \quad (10)$$

where $n_{jk}^{(i)}$ denotes the number of times channel i transitions from state j to state k. We sample $\boldsymbol{\eta}_j^{(i)} = C_j^{(i)} \bar{\boldsymbol{\eta}}_j^{(i)}$ from its posterior via two auxiliary variables,

$$\begin{aligned} \bar{\boldsymbol{\eta}}_j^{(i)} &\sim \mathrm{Dir}(\gamma_c + \mathbf{e}_j \kappa_c + \mathbf{n}_j^{(i)}) \\ C_j^{(i)} &\sim \mathrm{Gamma}(K\gamma_c + \kappa_c, 1), \end{aligned} \quad (11)$$

where $\mathbf{n}_j$ gives the transition counts from state j.

AR coefficients Each observation $\mathbf{y}_t$ is generated based on a *vector* of AR parameters $[\mathbf{a}_{z_t^{(1)}}, \cdots, \mathbf{a}_{z_t^{(N)}}]$. Thus, sampling $\mathbf{a}_k$ involves conditioning on $\{\mathbf{a}_{k'}\}_{k' \neq k}$ and disentangling the contribution of $\mathbf{a}_k$ on each $\mathbf{y}_t$. As derived in Supplement B,

$$\mathbf{a}_k \mid \mathbf{y}_{1:T}, \mathbf{z}_{1:T}, Z_{1:T}, \{\mathbf{a}_{k'}\}_{k' \neq k}, \{\Delta_l\} \sim \mathcal{N}(\mathbf{a}_k; \boldsymbol{\mu}_k, \Sigma_k), \quad (12)$$

where

$$\Sigma_k^{-1} = \Sigma_0^{-1} + \sum_{t=1}^{T} \bar{\mathbf{Y}}_t^{(\mathbf{k}^+)} \Delta_{Z_t}^{-(\mathbf{k}^+, \mathbf{k}^+)} \bar{\mathbf{Y}}_t^{T(\mathbf{k}^+)},$$

$$\Sigma_k^{-1} \boldsymbol{\mu}_k = \sum_{t=1}^{T} \bar{\mathbf{Y}}_t^{(\mathbf{k}^+)} \left(\Delta_{Z_t}^{-(\mathbf{k}^+, \mathbf{k}^+)} \mathbf{y}_t^{(\mathbf{k}^+)} + \Delta_{Z_t}^{-(\mathbf{k}^+, \mathbf{k}^-)} \boldsymbol{\epsilon}_t^{(\mathbf{k}^-)} \right).$$

The vectors $\mathbf{k}^+$ and $\mathbf{k}^-$ denote the indices of channels assigned and not assigned to state k at time t, respectively. We use these to index into the rows and columns of the vectors $\boldsymbol{\epsilon}_t$, $\mathbf{y}_t$, and matrix Δ_{Z_t}. Each column of matrix $\bar{\mathbf{Y}}_t^{(\mathbf{k}^+)}$ is the previous r observations for one of the channels assigned to state k at time t.

Event variables Conditioned on the channel state sequences $\mathbf{z}_{1:T}$ and AR coefficients $\{\mathbf{a}_k\}$, we can compute an innovations sequence as $\boldsymbol{\epsilon}_t = \mathbf{y}_t - \mathbf{A}_{\mathbf{z}_t} \widetilde{\mathbf{Y}}_t$. These innovations are the observations of the HMM of Eq. (2). Conditioned on the truncated HDP-HMM event transition distributions $\boldsymbol{\phi}$ and emission parameters $\{\Delta_l\}$, we can use a standard backward filtering forward sampling scheme to block sample $Z_{1:T}$.

In sampling the event transition distributions $\boldsymbol{\phi}$, we recall the L weak limit approximation of Eq. (7) and first sample the parent transition distribution $\boldsymbol{\beta}$ as described in Supplement B and then sample each $\boldsymbol{\phi}_l$ from its Dirichlet posterior,

$$\boldsymbol{\phi}_l \sim \mathrm{Dir}(\alpha_e \boldsymbol{\beta} + \mathbf{e}_l \kappa_e + \mathbf{n}_l), \quad (13)$$

where $\mathbf{n}_l$ is a vector of transition counts of $Z_{1:T}$ from state l to the L different states.

Event state parameters Finally, we sample the innovation covariance Δ_l for each event state l from its HIW posterior: $\Delta_l \sim \mathrm{HIW}_G(b_l, D_l)$, with

$$b_l = b_0 + |\{t \mid Z_t = l\}|, \quad D_l = D_0 + \sum_{t \mid Z_t = l} \boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_t^T. \quad (14)$$

Hyperparameters The prior and conditional posteriors of the hyperparameters γ_c, κ_c, α_e, κ_e, γ_e, and $\alpha_c = B_0(\Omega)$ are provided in Supplement B.

4. Experiments

Simulated data We first simulated six channels of toy data from five different first-order AR processes and three different event innovation covariances to confirm that our model was able to accurately estimate the true channel and event state sequences, the true

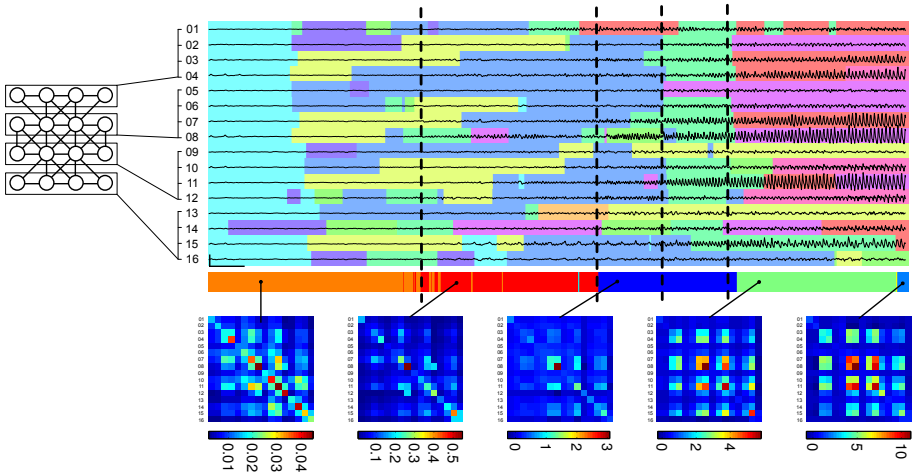

Figure 2. The graph used for a 16 channel iEEG electrode and corresponding traces over 25 seconds of a seizure onset with colors indicating the inferred channel states. The event states are shown below along with the associated innovation covariances. Vertical dashed lines indicate the EEG transition times marked independently by an epileptologist.

channel state AR coefficients, and the true event state innovation covariances. We found our model[1] able to estimate all of these parameters remarkably well and describe the simulation parameters, experimental details, and results in Supplement C.

Analyzing seizures We tested the HIW-spatial BP-AR-HMM on two similar seizures (events) from a patient of the Children's Hospital of Pennsylvania. These seizures were chosen because qualitatively they displayed a variety of dynamics throughout the beginning, middle, and end of the seizure and thus are ideal for exploring the extent to which our HIW-spatial BP-AR-HMM can parse a set of rich neurophysiologic signals. We used the 90 seconds of data after the clinically-determined starts of each seizure from 16 channels, whose spatial layout in the electrode grid is shown in Fig. 2 along with the graph encoding our conditional independence assumptions. The data were low-pass filtered and downsampled from 200 to 50 Hz, preserving the clinically important signals but reducing the computational burden of the analysis. The data was also scaled to have 99% of values within [-10, 10] for numerical reasons. We examined an order 5 HIW-spatial BP-AR-HMM and ran 10 MCMC chains for 6000 iterations, discarding 1000 samples as burn-in. Details are in Supplement D.

The HIW-spatial BP-AR-HMM inferred state sequences for the sample corresponding to a minimum expected Hamming distance criterion are shown in Fig. 2. The results were analyzed by a board-certified epileptologist (B.L.). He agreed with the model's judgement in identifying the subtle changes from the background dynamic (cyan) initially present in all channels. Furthermore, the model's grouping of spatially-proximate channels into similar state transition patterns (e.g., channels 03, 07, 11, 15) was clinically intuitive and consistent with his own reading of the raw EEG. Using only the raw EEG and prior to disclosing our results, he identified roughly six points in the duration of the seizure where the dynamic fundamentally changes. The three main event state transitions shown in Fig. 2 occurred almost exactly at the same time as three of his own marked transitions. These event states allow for a more global summary of the dynamics of the seizure and provide an important addition to the channel state sequences of the standard (non-spatial) and our HIW-spatial BP-AR-HMMs.

In Supplement E, we show how the event state parsing of the same seizure's offset is similarly clinically intuitive and distinguishes subtle transitions in the dynamics: strong correlations in the spikings of a few channels to a more widespread correlation structure and synchronized discharge pattern. The automatic identification of brief intervals of synchronized spiking makes it easy for a clinician to calculate changes in the inter-spike interval, a quantity of clinical importance. While interpreting these state sequences and covariances from the model, it is important to keep in mind that they are ultimately estimates of a system whose parsing even highly-trained physicians disagree upon. Nevertheless, the prospect of a reproducible, objective, and automated method for parsing such complex, multichannel events that closely mirrors those of practicing clinicians opens the possibility of large-scale analysis of hundreds or even thousands of such events.

We compared our HIW-spatial BP-AR-HMM to a full-

[1]Matlab code for this and other models in this paper is available at `www.seas.upenn.edu/~wulsin`

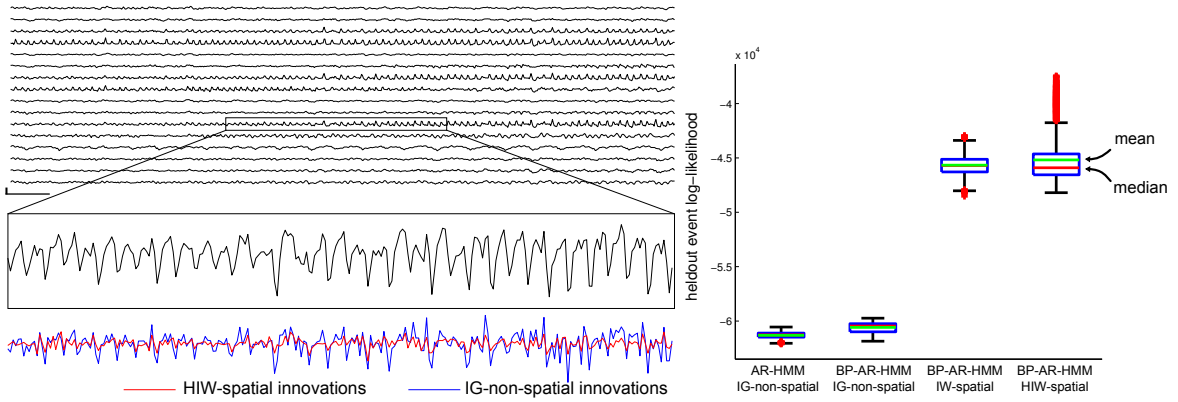

Figure 3. (**left**) An example 16-channel clip of iEEG with the middle section of one channel zoomed in and innovations from a non-spatial MNIW prior and a spatial N-HIW prior BP-AR-HMM shown below. (**right**) Boxplots of the heldout event log-likelihoods from the two non-spatial and two spatial models with mean and median posterior likelihood given in green and red lines. Boxes denote the middle 50% prediction interval.

covariance model with an IW prior on Δ_l (IW-spatial). We additionally compared it to non-spatial alternatives where channels evolve independently: the BP-AR-HMM of Fox et al. (2009) and an AR-HMM without the feature-based modeling provided by the beta process (Fox et al., 2011b). Both of these models use inverse gamma priors on the individual channel innovation variances. Fig. 3 (left) shows how conditioning on the innovations of neighboring channels in the HIW-spatial model improves the prediction of an individual channel, as seen by its reduced innovation trace relative to the IG-non-spatial model. The quantitative benefits of accounting for these correlations are seen in our predictions of heldout events, as depicted in Fig. 3 (right). We infer a set of AR coefficients $\{\mathbf{a}_k\}$ and event covariances $\{\Delta_l\}$ on one seizure, and then compute the heldout log-likelihood on a separate seizure, constraining it to only select among the inferred AR and event states. We can analytically marginalize the heldout event state sequence $Z_{1:T}$ but perform a Monte Carlo integration over the feature vectors $\mathbf{f}^{(i)}$ and channel states $\mathbf{z}_{1:T}$ using our MCMC sampler[2]. Fig. 3 (right) compares the heldout log-likelihoods for the IG-non-spatial and (H)IW-spatial models listed above, collected over 5000 samples across 10 chains, each with a 1000-sample burn in and 10-sample thinning. As expected, the HIW-spatial model has significantly larger predictive power than the non-spatial models. Though hard to see due to the large spatial/non-spatial difference, the BP-based model also improves on the standard non-feature-based AR-HMM. Performance is also at least as competitive as a full-covariance model (IW-spatial) but most importantly has significant computational gains based on the graph structure. The model scales linearly with the number of events, and the conditional independencies introduced by using a hyper-inverse Wishart prior allow the matrix operations to grow linearly with the number of channels rather than quadratically, as they do in the IW-spatial model. In Supplement F, we give results from a similar comparison on a much larger dataset of 50 events.

Comparing epileptic bursts and seizures We applied the HIW-spatial BP-AR-HMM to six channels of iEEG over 15 events from one patient. These events comprise 14 short sub-clinical epileptic bursts of roughly five to eight seconds and a final, 2-3 minute clinical seizure. Our hypothesis was that the sub-clinical bursts display initiation dynamics similar to those of a full, clinical seizure and thus contain information about the seizure-generation process.

The events were automatically extracted from the patient's continuous iEEG record by taking sections of iEEG whose median line-length feature (Esteller et al., 2001) crossed a preset threshold, also including 10 seconds before and after each event. The iEEG was preprocessed in the same way as in the previous section. The six channels studied came from a depth electrode implanted in the left temporal lobe of the patient's brain. We ran our MCMC sampler on the 15 events ($N = 15 \cdot 6$ with disconnected channel graphs between events) and selected a representative sample as in (Fox et al., 2011a). The hyperparameter settings, number of MCMC iterations, chains, and thinning was as in the previous experiment.

Fig. 4 compares two of the 14 sub-clinical bursts and the onset of the single seizure. We have aligned the three events relative to the beginnings of the red event state common to all three, which we take roughly as

[2]For each original MCMC sample, a secondary chain is run fixing all but $z_t^{(i)}$, Z_t, $f^{(i)}$, $\boldsymbol{\eta}^{(i)}$, $\boldsymbol{\phi}$. We approximate $p(\mathbf{y}_{1:T} \mid \boldsymbol{\phi}, \{\mathbf{a}_k\}, \{\Delta_l\})$ by averaging the secondary chain's closed-form $p(\mathbf{y}_{1:T} \mid \mathbf{z}_{1:T}, \boldsymbol{\phi}, \{\mathbf{a}_k\}, \{\Delta_l\})$. See Eq. (7) of the Supplementary Materials.

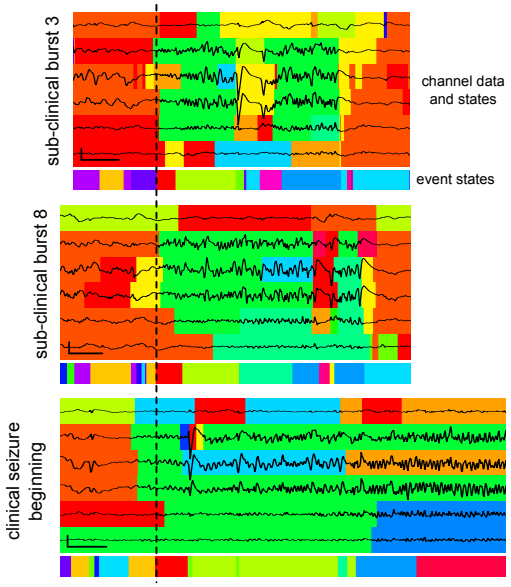

Figure 4. 6 iEEG traces from two sub-clinical bursts and onset of the single seizure with colors indicating inferred channel and event states. The dashed lines indicates the start of the red state in the three events.

the unequivocal start of the epileptic activity. The individual states of the four middle channels are also all green throughout most of the red event state. It is interesting to note that at this time the fifth channel's activity in all three events is much lower than those of the three channels above it, yet it is still assigned to the green state and continues in that state along with the other three channels as the event state switches from the red to the lime green state in all three events. While clinical opinions can vary widely in EEG reading, a physician would most likely not consider this segment of the fifth channel similar to the other three, as our model consistently does. But on a relative voltage axis, the segments actually look quite similar. In a sense, the fifth channel has the same dynamics as the other three but just with smaller magnitude. This kind of relationship is difficult for the human EEG readers to identify and shows how models such as ours are capable of providing new perspectives not readily apparent to a human reader. Additionally, we note the commonalities in event state transitions.

The similarities mentioned above, among others, suggest some relationship between these two different classes of epileptic events. However, all bursts make a notable departure: a large one-second depolarization in the middle three channels, highlighted at the end by the magenta event state and followed shortly thereafter by the end of the event. Neither the states assigned by our model nor the iEEG itself indicates that dynamic present in the clinical seizure. This difference leads us to posit that perhaps these sub-clinical bursts are a kind of false-start seizure, with similar onset patterns but a disrupting discharge that prevents the event from escalating to a full-blown seizure. Validation of such a hypothesis through a more comprehensive study would greatly improve our basic-science understanding of seizures and epileptogenesis.

5. Discussion

We presented a modeling framework for automating the parsing of EEG data, especially in the challenging scenario of multiple recordings taken from patients with variable numbers of channels, as is common in iEEG data. Our framework builds on the BP-AR-HMM, enabling learning a shared dictionary of AR dynamics that are asynchronously switched between by the individual channels. In contrast to the BP-AR-HMM, our model captures correlations between the time series, which we demonstrated is crucial in fitting heldout seizure data. We harness the spatial structure of the channels to define a set of conditional independencies that both improve out-of-sample predictions and reduce the computational burden, allowing scalability to large electrode grids. We additionally introduce a Markov event state to capture the time-varying correlations. We showed how this event state further improved the clinical interpretability of our model.

In addition to providing clinically intuitive parsings of the onset and offset of a seizure, we demonstrated how our event and channel state estimates facilitate comparisons between sub-clinical epileptic bursts and clinical seizures, suggesting new clinical hypotheses about their relationship. Clearly, validating such speculations necessitates testing on more epileptic events from a large class of patients. We have delved into a clinical analysis of the iEEG to illustrate how our model brings a quantitative structure to these highly complex multi-channel events. We see the model and its parsing capabilities as a data exploration tool that will help clinicians make sense of the vast quantities of iEEG data collected from epilepsy patients.

While we focus on modeling epileptic EEG events in this work, our model is more generally applicable to multiple correlated time series, especially in scenarios where there are multiple recordings with variable numbers of time series (e.g., motion capture sensors, financial data streams, etc.).

Acknowledgements

This work is supported in part by AFOSR Grant FA9550-12-1-0453 and DARPA Grant FA9550-12-1-0406 negotiated by AFOSR, NINDS RO1-NS041811, RO1-NS48598, and U24-NS063930-03, and the Mirowski Discovery Fund for Epilepsy Research.

References

D'Alessandro, M., Vachtsevanos, G., Esteller, R., Echauz, J., Cranstoun, S., Worrell, G, Parish, L, and Litt, B. A multi-feature and multi-channel univariate selection process for seizure prediction. *Clinical Neurophysiology*, 116:506–516, 2005.

Dawid, A. P. and Lauritzen, S. L. Hyper Markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics*, 21(3):1272–1317, 1993.

Dong, W., Pentland, A., and Heller, K. A. Graph-Coupled HMMs for Modeling the Spread of Infection. In *Proceedings of the Twenty-Eighth Conference Conference on Uncertainty in Artificial Intelligence*, 2012.

Doshi-Velez, F., Wingate, D., Tenenbaum, J., and Roy, N. Infinite dynamic bayesian networks. In *Proceedings of the 28th International Conference on Machine Learning*, 2011.

Esteller, R., Echauz, J., Tcheng, T., Litt, B., and Pless, B. Line length: an efficient feature for seizure onset detection. In *Proceedings of the 23rd EMBS Conference*, 2001.

Fox, E. B., Sudderth, E. B, Jordan, M. I., and Willsky, A. S. Sharing features among dynamical systems with beta processes. *Advances in Neural Information Processing Systems*, 22, 2009.

Fox, E. B., Sudderth, E. B., Jordan, M. I., and Willsky, A. S. A sticky HDP-HMM with application to speaker diarization. *The Annals of Applied Statistics*, 5(2A): 1020–1056, 2011a.

Fox, E. B., Sudderth, E. B., Jordan, M. I., and Willsky, A. S. Bayesian Nonparametric Inference of Switching Dynamic Linear Models. *IEEE Transactions on Signal Processing*, 59(4):1569–1585, April 2011b.

Ghahramani, Z. and Jordan, M. I. Factorial hidden Markov models. *Machine learning*, 31, 1997.

Griffiths, T. L and Ghahramani, Z. Infinite Latent Feature Models and the Indian Buffet Process Infinite. *Gatsby Computational Neuroscience Unit, Technical Report #2005-001*, 2005.

Heller, K., Teh, Y. W., and Gorur, D. The Infinite Hierarchical Hidden Markov Model. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, 2009.

Hughes, M., Fox, E., and Sudderth, E. Effective Split-Merge Monte Carlo Methods for Nonparametric Models of Sequential Data. In *Advances in Neural Information Processing Systems*, 2012.

Ishwaran, H. and Zarepour, M. Exact and approximate sum representations for the Dirichlet process. *Canadian Journal of Statistics*, 30(2):269–283, 2002.

Krystal, A. D., Prado, R., and West, M. New methods of time series analysis of non-stationary EEG data: eigenstructure decompositions of time varying autoregressions. *Clinical Neurophysiology*, 110:2197–2206, 1999.

Litt, B., Esteller, R., Echauz, J., D'Alessandro, M., Shor, R., Henry, T., Pennell, P., Epstein, C., Bakay, R., Dichter, M., and Vachtsevanos, G. Epileptic seizures may begin hours in advance of clinical onset: a report of five patients. *Neuron*, 30(1):51–64, April 2001.

Mirowski, P., Madhavan, D., Lecun, Y., and Kuzniecky, R. Classification of patterns of EEG synchronization for seizure prediction. *Clinical Neurophysiology*, 120(11): 1927–1940, 2009.

Prado, R., Molina, F., and Huerta, G. Multivariate time series modeling and classification via hierarchical VAR mixtures. *Computational Statistics & Data Analysis*, 51 (3):1445–1462, December 2006.

Schiff, S. J., Sauer, T., Kumar, R., and Weinstein, S. L. Neuronal spatiotemporal pattern discrimination: the dynamical evolution of seizures. *NeuroImage*, 28(4): 1043–1055, December 2005.

Schindler, K., Leung, H., Elger, C. E., and Lehnertz, K. Assessing seizure dynamics by analysing the correlation structure of multichannel intracranial EEG. *Brain*, 130: 65–77, January 2007.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

Thibaux, R. and Jordan, M. I. Hierarchical beta processes and the Indian buffet process. In *Proceedings of the Tenth Conference on Artificial Intelligence and Statistics*, 2007.

Van Gael, J., Saatci, Y., Teh, Y. W., and Ghahramani, Z. Beam sampling for the infinite hidden Markov model. In *Proceedings of the 25th International Conference on Machine Learning*, 2008.