Bayesian Nonparametric Modeling of Epileptic Events

Drausin F. Wulsin

A DISSERTATION in

Bioengineering

presented to the faculties of the University of Pennsylvania
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

2013

Brian Litt	Daniel A. Hammer
Professor of Neurology and Bioengineering	Alfred G. and Meta A. Ennis Professor of Bioengineering and Chemical & Molecular Engineering
Supervisor of Dissertation	Graduate Group Chairperson

Dissertation Committee
Gershon Buchsbaum, Professor of Bioengineering, University of Pennsylvania
Justin A. Blanco, Assistant Professor of Electrical and Computer Engineering, U.S. Naval Academy
Emily B. Fox, Assistant Professor of Statistics, University of Washington
Shane T. Jensen, Associate Professor of Statistics, University of Pennsylvania

To my mother, for her faith in my education.

Acknowledgements

My past four years as a PhD student have been by far the most intellectually challenging and stimulating of my life. I credit much of this success to the network of mentors, peers, family, and friends who have supported me tremendously over this time.

My advisor, Brian Litt, not only provided the support (intellectual, financial, emotional, and otherwise) that has motivated this thesis from the start, but he also trusted me and allowed me complete intellectual freedom. He put my priorities above his own. He suffered my digressions and missteps with the same enthusiasm he did my successes. He welcomed my new ideas with an open, interested mind and helped me focus them toward important, relevant clinical problems. In addition to teaching me much about epilepsy and its clinical treatment, Brian taught me how to step back and ask the big questions, to see where a field is moving, and to understand what details are important and what are not. In running our lab, Brian also taught me how to manage a heterogeneous team of people—all with their own priorities and personalities—and how to keep the group dynamic positive, constructive, and supportive.

I also was fortunate to benefit from the guidance of a few other mentors. Justin Blanco, a senior graduate student in our lab when I entered (now a professor at the U.S. Naval Academy), helped me navigate the early waters of my technical training and research interests. Justin was immensely helpful as a source for new ideas, as a sounding board for my own, and as teacher of many technical and scientific practices hitherto unknown to me. He gave his time generously, especially when carefully reading the drafts of a handful of my papers that never ultimately saw the light of day. I also benefited greatly from Justin's measured perspectives about publishing, scientific skepticism, and the role of research.

A chance conversation with a TA of mine led me to take Shane Jensen's class on Bayesian statistics, a class that fundamentally diverted my intellectual interests and the course of my research. Shane had a knack for explaining the often-confusing aspects of Bayesian analysis in a straightforward manner. He gave his time generously, both during the semester of his course and after. His interest in my own research and technical advice ultimately helped develop the multi-level clustering model we describe in the third chapter of this thesis. Shane helped develop this idea and has been crucial in its maturation, including shepherding two publications through the peer review process.

I also count myself incredibly fortunate to have been able to work with Emily Fox, both

while she was at Penn and now in her position at the University of Washington. After sitting patiently through my research presentation at our first meeting, Emily responded that the work was nice but perhaps a bit ad hoc and suggested I look into the nested Dirichlet process of Rodríguez et al. [103], a model that would become the seed of the multi-level clustering work of this thesis. In our many subsequent discussions, she taught me much about principled model development and testing, providing a seemingly endless number of ideas to explore during this process. These discussions produced the model described in the fourth chapter of this thesis. I also benefited greatly from Emily's high standards for notation and technical language, to say nothing of her incredible attention to detail.

I must also thank Steve Isard who has generously and patiently tutored me in a number of topics—including digital signal processing, numerical analysis, convex optimization, and natural language processing—over the past few years. Steve made accessible topics I never would have been able to tackle on my own. In addition, his vast technical experience over the past fifty years greatly enriched our conversations with a historical perspective unavailable anywhere else. I have relished our one-on-one sessions immensely, knowing such an opportunity is available to very few.

I have also benefited from epileptologists Eric Marsh and Brenda Porter, whose clinical insights and thoughtful discussion played a large role in the third chapter of this thesis. In a world where time is always in short supply, Brenda and Eric were very generous with theirs, especially in manually clustering 193 seizures, twice. Their intellectual honesty and frank representation of their work were also a model in a field where the incentives for spin and overstating results are unfortunately great.

I am grateful to have shared a lab with wonderful peers, all stimulating and interesting in their own way. I especially thank Mark Lippman, Ann Vanleer, and Joost Wagonaar for many helpful conversations about my research and about research in general.

Throughout these four years and for every year before that, my family has provided the love and support that has made this all possible. My parents encouraged my curiosity from the start and have put my best interests ahead of everything, equally when they were together and apart. I am so grateful to have had such a strong foundation in them. And finally, I must thank my fiance Ali, who has kept me sane and balanced over these four years and the five years before that. Throughout the depths and occasional dark days of my PhD, Ali has been my unwavering north star, seeing me safely through it all. I look forward to many future voyages with her as my guide.

ABSTRACT

BAYESIAN NONPARAMETRIC MODELING OF EPILEPTIC EVENTS

Drausin F. Wulsin

Brian Litt

Epilepsy is a common neurological disorder that today plagues over 50 million people worldwide. The 20-40% of patients whose seizures are unable to be controlled with pharmacological treatments commonly receive scalp and intracranial electroencephalogram (EEG) monitoring to determine whether surgical treatment is appropriate. Epileptic events like large, clinical seizures and small, sub-clinical bursts recorded on the EEG are of primary diagnostic interest, but these events—which usually range from a few seconds to a few minutes across tens or hundreds of individual EEG channels—are very complex and high-dimensional. Human epileptologists are well-trained in analyzing individual epileptic events but their ability to generalize across and compare many such events is limited due to the complex, high-dimensional nature of these EEG event recordings. In this work, we develop and apply statistical models for analyzing and understanding large numbers of these events. Our Bayesian nonparametric models naturally incorporate available prior knowledge and uncertainty about these events. While motivated by these epileptic event data, our models generalize to large class of application domains. We first develop and validate

a model for describing seizures that intelligently shares information across the seizures of the same patient and those of other similar patients. We then develop and validate a model for producing a fine-grained parsing of both shorter burst and longer seizure events, allowing for straightforward comparisons between the two. Finally, we apply this later model to large datasets of hundreds of epileptic bursts and seizures, finding that the bursts often display large similarities with the onsets of seizures. These results show the benefit of well-motivated, straightforward Bayesian modeling and the large impact it can have in the quantitative analysis of epileptic events.

Table of Contents

List of Tables

List of Figures

List of Algorithms

Notation Conventions

Notation	Meaning
$\circ$	the Hadamard (element-wise) product
$\oplus$	update operator
$\ominus$	downdate operator
$a \leftarrow b$	store value of b in a
n_{ij}	the element in the j^{th} column of the i^{th} row of matrix $\mathbf{n}$
$n_{i,j}$	the jth element in the vector $\mathbf{n}_i$
$\mathbf{n}_j$, $\mathbf{n}_{*j}$	the j^{th} column of matrix $\mathbf{n}$
$\mathbf{n}_{i*}$	the i^{th} row of matrix $\mathbf{n}$
$\mathbf{m}_{1:i}$	the column vector $(m_1, \ldots, m_i)^{\text{T}}$
$\mathbf{n}_{\cdot j}$	the column-wise sum over the rows of $\mathbf{n}$
$\mathbf{n}_{i\cdot}$	the row-wise sum over the columns of $\mathbf{n}$
$\mathbf{n}^{\text{T}}$	the transpose of matrix $\mathbf{n}$
$\mathbf{e}_j$	the j^{th} column of the identity matrix
$\mathbf{m}^{(\mathbf{j})}$	the sub-vector of $\mathbf{m}$ indexed by the elements of $\mathbf{j}$
$\mathbf{m}^{(-j)}$	the sub-vector of $\mathbf{m}$ of all but the j^{th} element
$\mathbf{n}^{(\mathbf{j},\mathbf{k})}$	the sub-matrix of $\mathbf{n}$ with rows indexed $\mathbf{j}$ and columns by $\mathbf{k}$
δ_θ	a measure with a point-mass at θ
$\delta(i, j)$	the Kronecker delta function
$\mathbb{R}$	the set of all real numbers
$\mathbb{Z}_+$	the set of all non-negative integers
$\mathbb{S}^d_{++}$	the set of $d \times d$ positive-definite matrices
$\{a_k\}$	the set of a_k over all possible values of k
$\lvert \cdot \rvert$	the cardinality of a set
$\mathbf{1}(\cdot)$	the indicator function for a boolean argument
$(\mathbf{a}, \mathbf{b})$	horizontally concatenate row-vectors $\mathbf{a}$ and $\mathbf{b}$
$(\mathbf{c}; \mathbf{d})$	vertically concatenate column-vectors $\mathbf{c}$ and $\mathbf{d}$
$[\mathbf{G} \mid \mathbf{H}]$	horizontally concatenate matrices $\mathbf{G}$ and $\mathbf{H}$

Table 1. General notation

Abbreviation	Meaning
EEG	electroencephalogram
iEEG	intracranial electroencephalogram
EMU	epilepsy monitoring unit
IG	inverse-gamma
IW	inverse-Wishart
HIW	hyper-inverse-Wishart
MCMC	Markov chain Monte Carlo
HMM	hidden Markov model
AR	autoregressive
VAR	vector autoregressive
DP	Dirichlet process
GEM	(distribution of) Griffiths, Engen, and McCloskey
HDP	hierarchical Dirichlet process
NDP	nested Dirichlet process
IBP	Indian buffet process
BP	beta process
BeP	Bernoulli process
MLC-HDP	multilevel clustering hierarchical Dirichlet process
RB	Rao-Blackwellized

Table 2. Common abbreviations used.

Value	Meaning
$\mathbf{x}_{ji}$	the observation vector for channel i of seizure j
$\boldsymbol{\phi}_k$	the parameters associated with observation atom k
$\boldsymbol{\Phi}_k$	the sufficient statistics associated with observation atom k
H	the prior over the channel observation atom parameters $\boldsymbol{\phi}_k$
$\boldsymbol{\pi}_\ell^{(1)}$	seizure cluster ℓ's weights over the channel observation atoms
$\boldsymbol{\beta}^{(1)}$	the global weights over the observation atoms
$\alpha^{(1)}$	DP concentration parameter for weights $\boldsymbol{\pi}_\ell^{(1)}$
$\gamma^{(1)}$	DP concentration parameter for weights $\boldsymbol{\beta}^{(1)}$
$\boldsymbol{\pi}^{(2)}$	patient weights over the seizure clusters
$\boldsymbol{\beta}^{(2)}$	the global patient weights over the seizure clusters
$\alpha^{(2)}$	DP concentration parameter for weights $\boldsymbol{\pi}^{(2)}$
$\gamma^{(2)}$	DP concentration parameter for weights $\boldsymbol{\beta}^{(2)}$
$z_{ji}^{(1)}$	observation atom indicator for channel i of seizure j
$z_j^{(2)}$	seizure cluster indicator for seizure j

Table 3. Notation associated with a single-patient MLC-HDP.

Value	Meaning
$\mathbf{x}_{tji}$	the observation vector for channel i of seizure j from patient t
$\{\mathbf{x}_{-tji}\}$	all the set of all observations *except* $\mathbf{x}_{tji}$
S_{tj}	the set of channels $i = 1, \ldots, N_{tj}$ in seizure j of patient t
$\{S_{-tj}\}$	the set of channels *except* those in S_{tj}
P_t	the set of seizures $j = 1, \ldots, J_t$ of patient t
$\{P_{-t}\}$	the set of all seizures *except* those of P_t
$\boldsymbol{\phi}_k$	the parameters associated with observation atom k
$\boldsymbol{\Phi}_k$	the sufficient statistics associated with observation atom k
H	the prior over the channel observation atom parameters $\boldsymbol{\phi}_k$
$\boldsymbol{\pi}_\ell^{(1)}$	seizure cluster ℓ's weights over the channel observation atoms
$\boldsymbol{\beta}^{(1)}$	the global weights over the observation atoms
$\alpha^{(1)}$	DP concentration parameter for weights $\boldsymbol{\pi}_\ell^{(1)}$
$\gamma^{(1)}$	DP concentration parameter for weights $\boldsymbol{\beta}^{(1)}$
$\boldsymbol{\pi}_l^{(2)}$	patient cluster l's weights over the seizure clusters
$\boldsymbol{\beta}^{(2)}$	the global patient weights over the seizure clusters
$\alpha^{(2)}$	DP concentration parameter for patient weights $\boldsymbol{\pi}_l^{(2)}$
$\gamma^{(2)}$	DP concentration parameter for weights $\boldsymbol{\beta}^{(2)}$
$\boldsymbol{\pi}^{(3)}$	patient population weights over the patient clusters
$\boldsymbol{\beta}^{(3)}$	the global patient population weights over the patient clusters
$\alpha^{(3)}$	DP concentration parameter for weights $\boldsymbol{\pi}^{(3)}$
$\gamma^{(3)}$	DP concentration parameter for weights $\boldsymbol{\beta}^{(3)}$
$z_{tji}^{(1)}$	observation atom indicator for channel i of seizure j for patient t
$z_{tj}^{(2)}$	seizure cluster indicator for seizure j for patient t
$z_t^{(3)}$	the patient cluster indicator for patient t

Table 4. Notation associated with a multi-patient MLC-HDP.

Value	Meaning
r	the order of the channel autoregressive process
$y_t^{(i)}$	the observed scalar value of channel i at time t
$\widetilde{\mathbf{y}}_t^{(i)}$	the vector of r previous observations of channel i from time t
$z_t^{(i)}$	the latent state assigned to channel i at time t
$\epsilon_t^{(i)}$	the innovation of channel i at time t
$\boldsymbol{\pi}_j^{(i)}$	the channel i's feature-constrained transition distribution for state j
Z_t	the latent event state at time t
$\boldsymbol{\phi}_l$	the transition distribution for event state l
$\mathbf{a}_k$	the vector of autoregressive parameters for channel state k
Σ_0	the covariance of the normal prior on channel state parameters $\mathbf{a}_k$
Δ_l	the innovations covariance for event state l
b_0	the degrees of freedom of the HIW prior on Δ_l
D_0	the scale of the HIW prior on Δ_l
$\mathbf{f}^{(i)}$	channel i's binary feature indicator vector
$\boldsymbol{\eta}_j^{(i)}$	the channel i's transition distribution for state j

Table 5. Notation associated with the HIW-spatial BP-AR-HMM.

Chapter 1

Introduction

Epilepsy affects over 50 million people worldwide [1], and the symptoms can be severe for those 20-40% of patients not effectively treated pharmacologically. In addition to the physical and mental risks associated with seizures themselves (e.g., head injuries from falling, memory loss, reduced learning ability), epilepsy can dramatically reduce patient quality of life between seizures, from the often large side effects of anti-epileptic drugs to the incidence of depression to the inability to drive a car, to get and maintain active employment, and to interact socially. Despite over half a century of research into the causes and medical treatment of epilepsy, we still have little idea of many of the underlying cellular and network properties that can give rise to naturally-occurring seizures. This difficulty stems from both the uniqueness of the disorder in each patient and our still-poor understanding of the human brain.

We believe the lack of fundamental insights about the underlying causes and progression of epilepsy in many patients may also be attributed to the sheer amount and heterogeneity of relevant data gathered for each patient, including seizure history, drug treatment history and effects, brain imaging like CT and MR, and electrophysiologic recordings like scalp and intracranial electroencephalogram (EEG). EEG recordings can range from only a few minutes to many weeks, and they require specially trained epileptologists to interpret them. Currently, "EEG reading" is still almost entirely manual, with physicians paging through EEG ten seconds at a time, noting aspects of clinical importance, especially of the recorded seizures.

This work develops models and analysis methods meant to automate some aspects of this EEG interpretation. We aim not to replace physician analysis of EEG but to provide clinicians with a set of useful tools that can simultaneously analyze many epileptic events on the EEG. We believe such automated analyses are at least more consistent and objective than any human can be in the often-large EEG records present for many epilepsy patients. Our methods are designed from the start with the clinician in mind: what analyses and model outputs would be clinically relevant, intuitive, and practical on a large scale?

Quantitative analysis of epileptic EEG activity has a long history. It has existed in varying forms since at least the early 1980s, with advances in computing power since then allowing for ever-more sophisticated techniques on larger datasets. We hope to strike out on a path relatively untrodden by previous work: the development of generative statistical

models of epileptic events. In discriminative models, one attempts to answer a particular question Y given known data X, e.g., "do seizure EEG recordings indicate that this area of brain is responsible for starting seizures?" In generative models, one attempts to describe the observed data X itself, e.g., "what are the different types of brain activity patterns present in these EEG seizure recordings?" In many ways, building generative models is a harder task than building discriminative ones. Generative models require describing the data itself rather than simply using the data to answer a specific question. But we believe such generative models offer a powerful, extensible, general framework for understanding epileptic events. Not only do we believe this framework much closer to the way clinicians interpret and understand epileptic events, but we also believe it can be used to answer many questions—instead of a single question—about epileptic events.

We work within the Bayesian framework for developing these models, as it allows us to explicitly encode prior knowledge and assumptions into the model. Humans use prior knowledge and assumptions in all decision-making. Our models should too. In addition, we believe the Bayesian framework inhibits the creation of ad-hoc, brittle analysis techniques and promotes flexible, modular models that can later be incorporated into larger, more sophisticated models when appropriate.

This thesis is organized into six chapters. In this first chapter, we motivate our work and summarize each of the five subsequent chapters. In Chapter 2, we provide a brief background of epilepsy and its clinical treatment, existing computational models of epilepsy, and light introductions to Bayesian statistical modeling, including the parametric and nonparametric models relevant to this work. In Chapter 3, we develop a Bayesian nonparametric model for describing datasets that cluster on multiple levels and apply it to a dataset of seizure iEEGs from multiple patients. In Chapter 4, we develop a Bayesian nonparametric model for parsing complex events with multiple correlated time series and apply it to datasets of epileptic events like seizures and sub-clinical bursts. In Chapter 5, we use this second model to analyze and begin to understand the relationship between sub-clinical epileptic bursts and clinical seizures present in long-term continuous recordings from dogs with naturally-occurring epilepsy. Finally, in Chapter 6, we summarize and discuss this work, its contributions to the fields of epilepsy and Bayesian nonparametrics, and suggest a general direction of future work. Below, we summarize each of these subsequent chapters in more detail.

■ 1.1 Background

In Chapter 2, we briefly discuss topics relevant to this work.

The clinical treatment of epilepsy We describe epilepsy and its current clinical treatment, including the intracranial electroencephalogram (iEEG) data collected during intracranial electrophysiologic monitoring. We describe the experimental long-term iEEG recordings gathered via an implanted device made by the NeuroVista Corporation. We use these continuous recordings to explore the relationship between sub-clinical epileptic bursts and clinical seizures commonly found in many intracranial recordings. The relationship between

these bursts and seizures is still poorly understood, despite their existence being quite common in the iEEG.

Existing computational models of epilepsy In an effort to situate our work within the broader literature of epilepsy modeling, we briefly summarize some previous approaches to modeling epileptic activity, focusing primarily on research working with EEG recordings, as we do throughout this thesis. We separate this previous work into mechanistic models, spatial models for localizing epileptic activity, and temporal models for describing the evolution of epileptic activity over time.

The Bayesian perspective Throughout the rest of the background chapter, we introduce and briefly describe Bayesian methods and models relevant to this work. Our discussion of this material focuses on the practical aspects necessary to understand the models we develop in Chapters 3 and 4. Readers interested in more thorough, theoretically-minded discussions of much of the same material should see the excellent PhD theses of Sudderth [116] and Fox [38], especially the background sections of each.

We motivate our decision to use the Bayesian framework in constructing our models and describe various prior distributions and the posterior distributions they produce with particular observation models of the data.

Posterior inference We briefly describe methods for Markov chain Monte Carlo (MCMC) posterior inference, including Gibbs sampling and the Metropolis-Hastings algorithm. We address practical aspects of MCMC sampling like burn-in, thinning, and running multiple MCMC chains.

Mixture models As a comparison to the nonparametric models, we describe parametric mixture models in the Bayesian framework, including standard mixture models for datasets with i.i.d. observations and hidden Markov models (HMM) for sequential observations. We give explicit algorithms for calculating the marginal data likelihood of a sequence of observations and also for block-sampling the state sequence from its joint distribution. We also discuss several approaches for determining the number of fixed mixture components to use.

Bayesian nonparametric models Our brief description of Bayesian nonparametrics introduces the models we use as components of the more complex models we develop in Chapters 3 and 4. We discuss the Dirichlet process as well as derivatives like the hierarchical Dirichlet process and the nested Dirichlet process. We then describe a nonparametric version of the HMM, the hierarchical Dirichlet process hidden Markov model. Finally, we motivate and discuss the beta process and its role in inducing sparsity in the feature space of beta process hidden Markov models.

■ 1.2 Clustering Seizures on Multiple Levels

In Chapter 3, we introduce a new model for clustering datasets on multiple levels and describe its application to datasets of seizures from multiple patients.

Model description and posterior inference Our model, which we call the multi-level clustering hierarchical Dirichlet process (MLC-HDP), is motivated by using "clusters of clusters" to describe a dataset, as introduced by the nested Dirichlet process of Rodríguez et al. [103]. A hierarchical Dirichlet process is used at each level of clustering in our model. We first discuss how the model describes the seizures of a single patient. Each channel of each seizure is assumed to be generated from a particular channel type, and each seizure is assumed to be generated from a particular seizure type, represented by a distribution over the possible channel types. Extending this formulation to multiple patients yields a model that clusters over channel types, seizure types, and patient types.

We give the conditional posteriors used during Gibbs sampling as well as explicit algorithms for Markov chain Monte Carlo (MCMC) sampling all model parameters and indicators.

Simulation and seizure iEEG experiments The MLC-HDP is validated and shown to yield superior posterior estimates than the nested Dirichlet process on a set of simulated data. We demonstrate how the hierarchical organization of the channel activity, seizures, and patients implicit in the MLC-HDP produces superior models to flat clustering alternatives like the Dirichlet process. We then show how the MLC-HDP yields seizure clusterings comparable to those of an expert epileptologist and superior to seizure clusterings produced by a simpler, flat Dirichlet process model. Finally, we illustrate how the seizure clusterings can be use to produce a distance metric between seizures that may then be used to visualize similarities between large numbers of seizures among multiple patients.

Model sensitivity analysis We conclude with a number of sensitivity analyses of the MLC-HDP, including the influence of individual patients on estimates of a different patient, the model's sensitivity to fixed priors, and its autocorrelation and convergence behavior under alternate posterior sampling schemes.

■ 1.3 Parsing Epileptic Events in Detail

In Chapter 4, we introduce a new model for parsing complex events comprised of multiple related time series and show its ability to produce detailed and intuitive analysis of iEEG seizures.

Model description and posterior inference Our new model builds on the beta process autoregressive hidden Markov model (BP-AR-HMM), incorporating spatial dependencies between time series using a hyper-inverse Wishart (HIW) prior on the correlations between channels. This model, which we call the HIW-spatial BP-AR-HMM allows for switching between different regimes of spatial relationships. Our model thus allows for asynchronous transitions between the states of individual time series as well as switches between time series relationships over the entire event.

We give the conditional posterior distributions used in MCMC sampling and also provide explicit algorithms for these sampling steps.

Simulation and seizure iEEG experiments We validate our HIW-spatial BP-AR-HMM on simulated data and demonstrate its ability to produce meaningful seizure parsings that agree with clinical judgement. We compare our model with several alternative models and show that accounting for spatial relationships between time series produces better models. In addition, the sparse spatial dependency structure allows us to encode known spatial relationships in time series and scale computations efficiently as the number of time series increases.

Comparing epileptic events of different scales In a proof of principal study, we demonstrate our HIW-spatial BP-AR-HMM's ability to describe a set of 14 sub-clinical epileptic bursts and one seizure from a human epilepsy patient. The iEEG parsings produced by our model indicate that these two classes of events have some similarities, particularly at their onset, but that these similarities end as the seizure onset activity escalates and the bursts finish.

Practical scaling for large datasets Finally, we describe our efforts in scaling the HIW-spatial BP-AR-HMM up to datasets of hundreds of individual epileptic events over thousands of time series. We explore the computational bottlenecks of posterior inference and describe the efficiency gains we achieve through parallelizing computations and using a custom data compression method.

■ 1.4 Exploring the relationship between epileptic bursts and seizures

In Chapter 5, we use a variant of the model developed in Chapter 4 to explore the relationship between sub-clinical epileptic bursts and clinical seizures recorded by devices chronically implanted in dogs with naturally occurring epilepsy.

Quantitative analysis of epileptic events We first describe our methods for epileptic event detection, modeling, and analysis. Events over the entire continuous record are detected by thresholding a simple feature commonly use to indicate epileptic activity. After initial detection, these events are culled in a semi-automated way to reduce artifacts and transient events. These events are modeled using an HIW-spatial autoregressive HMM, which produces event state sequences, among others inferences, for each event. These event state sequences describe channel relationships in each event and how they change over the course of the event. We then describe how these event state sequences are analyzed, producing quantitative comparisons between seizures and sub-clinical bursts.

Seizure and burst similarities We use these quantitative comparisons to show that seizure onsets are most similar to the sub-clinical bursts. We show how these seizure onsets contain the same channel relationships present in bursts and even show that some bursts display event state transition patterns very similar to those of seizure onsets. We conclude our analysis by exploring whether bursts can be used to extract the same types of event state transition information present in seizure onsets and find that despite some similarities, the bursts are ultimately limited in their ability to fully describe the state progressions present in full, clinical seizures.

■ 1.5 Contributions and Discussion

Finally, in Chapter 6, we summarize the work presented in this thesis and discuss its context and role in large-scale analysis of epilepsy data. We recapitulate our contributions to the Bayesian statistics and epilepsy scientific communities and suggest a main avenue for future work.

Chapter 2

Background

2.1 The Clinical Treatment of Epilepsy

Epilepsy is one of the oldest known neurological disorders, though it was not always thought of as such. A Babylonian cunieform tablet from roughly 1050 BCE describes in detail a variety of symptoms of epileptic attacks as well as which demons were supposedly responsible for each of these symptoms [132]. Texts from the Old Testament, from the ancient Egyptians, Chinese, and Indians all describe symptoms of this "falling sickness," usually in the context of demonic possessions or punishments by a god. The etymology of *epilepsy* comes from the Greek $\varepsilon\lambda\alpha\mu\beta\alpha\nu\varepsilon\iota\nu$, meaning "to be seized," "to be taken hold of," "to be attacked," explaining also the origin of *seizure* [33, chap. 1]. In *On the Sacred Disease*, Hippocrates (c. 460-377 BCE) first links the symptoms of epilepsy with a neurological disorder, though the battle between scientific and mystical explanations continued until at least the mid 19th century [33, 121].

Despite the medical, scientific, and technological advances of the last few centuries, our understanding of epilepsy is still quite poor relative to many other neurological diseases. In fact, some argue that epilepsy is not really a disease at all, merely a symptom—the propensity for seizures—caused by both acute neurological injury as well as a group of neurological disorders [33, chap. 1]. The broad space of neurological conditions thought responsible for seizures explains its prevalence in the world today. Epilepsy affects over 50 million people world wide. The World Health Organization estimates 1% of the global burden of disease can be attributed to epilepsy, roughly equivalent to the burdens of lung cancer in men and breast cancer in women [1].

Current clinical treatment of epilepsy generally first involves antiepileptic drugs, which are effective at eliminating seizures in roughly 60-80% of all patients. For those remaining 20-40%, surgical treatment is currently the only major clinical alternative [45]. Surgical treatment generally involves either removing brain tissue or cutting connections between tissues, or both. Figure 2.1 shows diagrams of the human brain. In tissue removal procedures, anything from a specific lesion to specific brain lobes to regions of the cerebral cortex may be removed. Cutting white matter tracts (connections between neurons) ranges from multiple subpial transections—shallow cortical cuts meant to disrupt cortical networks—to full corpus callosotomy, cutting the main bridge between the two hemispheres of the brain. In this work, we primarily consider surgeries for removing (also known as resecting) specific

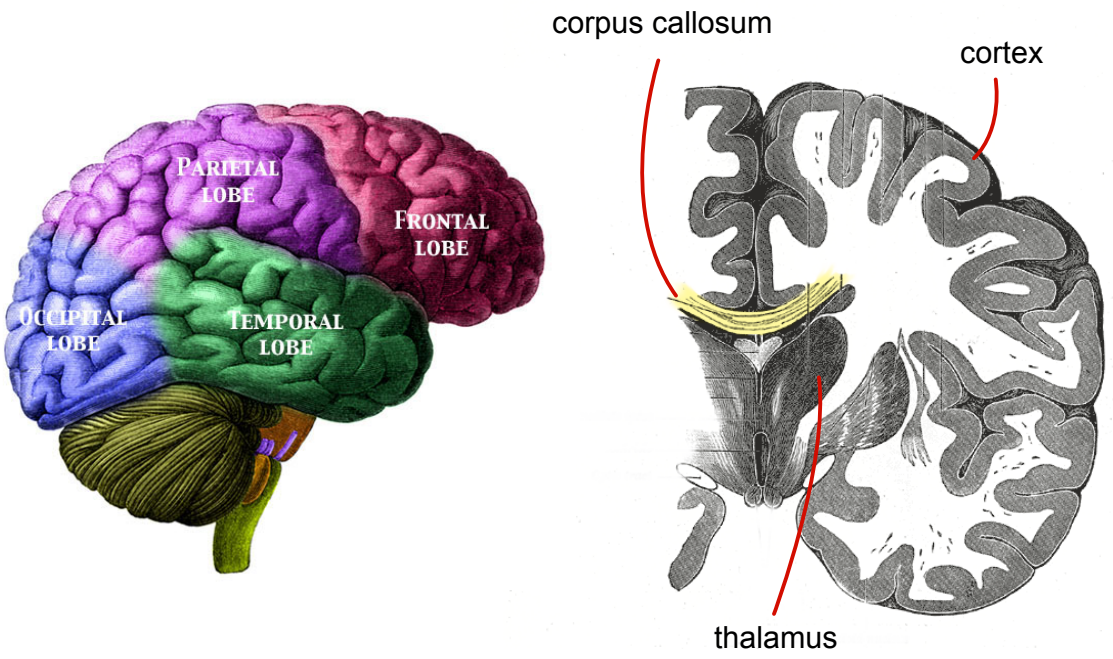

Figure 2.1. Diagrams of the human brain. (**left**) The five main lobes of the brain. (**right**) An axial cross section showing gray and white matter, cortex, thalamus, and the corpus callosum, among other structures.

brain tissue.

This resection rests on the hypothesis that removing the brain areas involved with seizures, or at least their onset, will remove or greatly reduce the occurrence of seizures. Temporal lobe epilepsy, in which seizures occur or begin from sub-cortical structures like the amygdala and hippocampus, generally have high rates of post-surgery seizure freedom [110]. Lesional neocortical epilepsy involves specific areas in the top layers of the cortex with discernable malformation, including tumors, uncharacteristic vasculature, and the irregular cell bodies seen in cortical dysplasia. A patient may have only one or many such lesions. Assuming proper resection of these lesions, many patients achieve seizure freedom from resective surgery. The most difficult patients to treat, and those with the poorest post-surgical outcome are those with nonlesional neocortical epilepsy. In those patients, determining exactly what areas to resect is the main task of epileptologists (neurologists specializing in epilepsy). This decision-making process involves information of many types, including patient seizure history, the seizure semiology (e.g., shaking, eye rolling, abnormal smells, sounds, and sights), imaging such as CT and MR, and both scalp and intracranial electroencephalogram (EEG). In this work, we focus on the analysis of EEG—particularly intracranial EEG (iEEG)—and the insights obtainable from it.

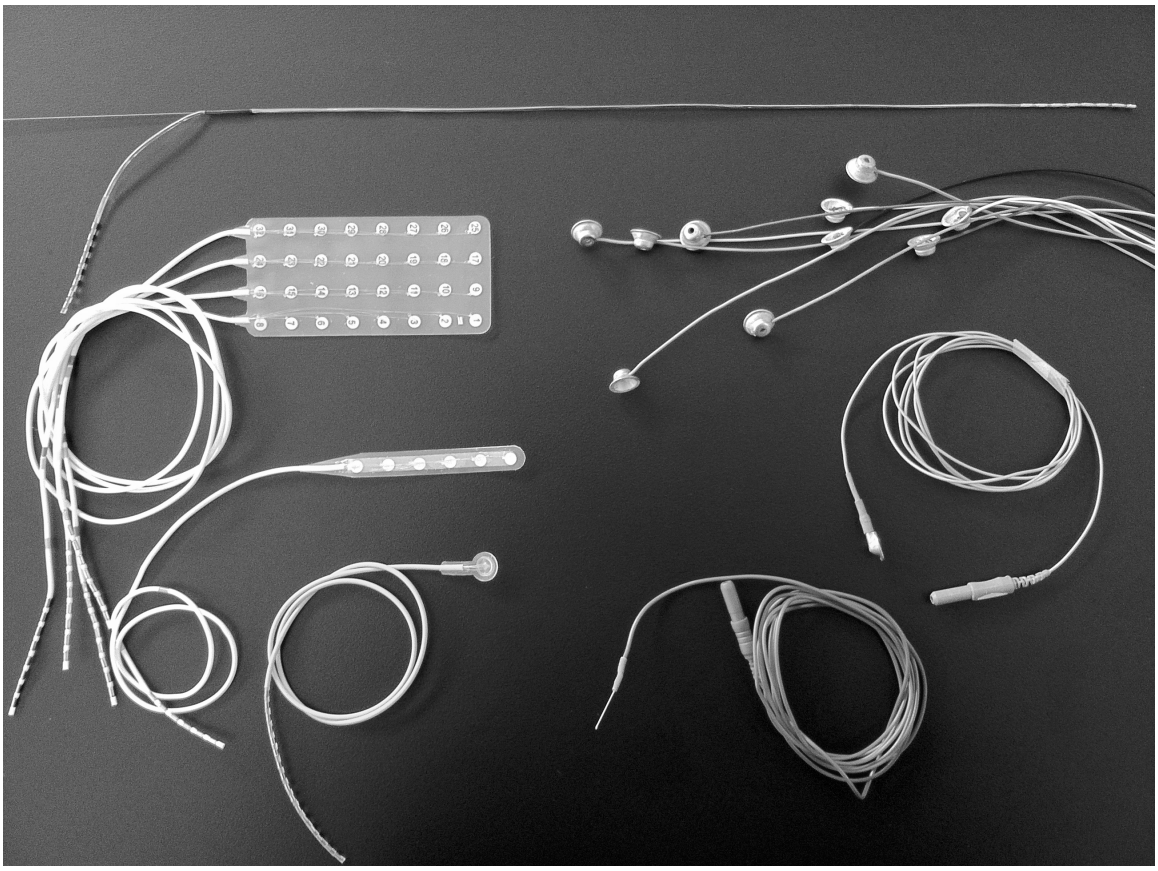

Figure 2.2. A variety of EEG electrodes, including intracranial depth, grid and strips (counter-clockwise from top) as well as scalp EEG electrode and others.

■ 2.1.1 Intracranial Electrophysiologic Monitoring

Electroencephalogram (EEG) involves recording voltage potentials from the scalp or—in the case of intracranial EEG—from the cortical surface or deeper structures like the hippocampus. In the evaluation for epilepsy surgery, patients—especially those with nonlesional neocortical epilepsy—often spend between one and three weeks in an epilepsy monitoring unit (EMU). In the EMU, the antiepileptic medications of these patients are tapered down, allowing the patient to (hopefully) have spontaneous seizures. Epileptologists use the iEEG recorded during these seizures to estimate the epileptogenic zone, the area(s) of the brain necessary to generate seizure activity, which often coindicides with the ictal—or seizure—onset zone, the area(s) of the brain where seizures begin on the EEG [33, chap. 171].

Intracranial EEG is recorded from depth electrodes—implanted to record from deep brain structures—and cortical surface strip and grid electrodes. Figure 2.2 shows a number of such electrodes, and Figure 2.3 shows the surgical implantation of cortical grid and strip

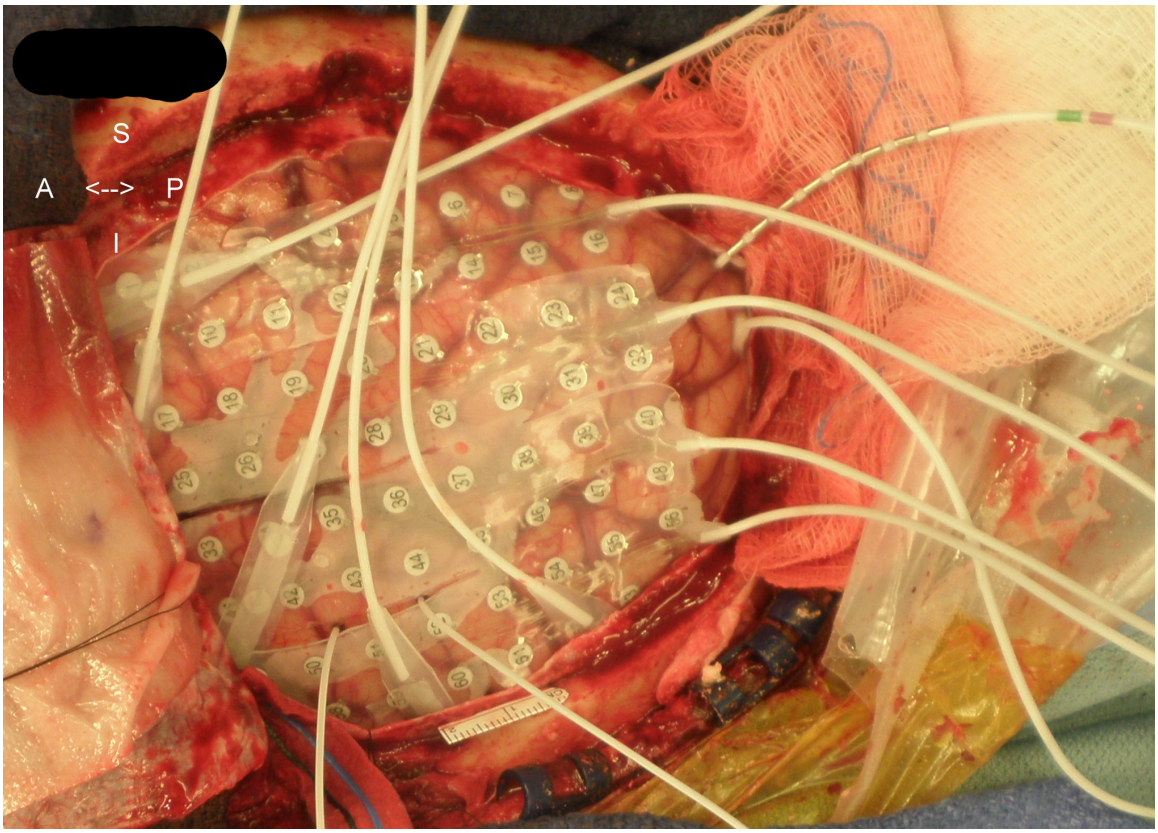

Figure 2.3. Surgical implantation of intracranial electrodes (6 strips, 2 depths, 4 grids).

electrodes. The number and placement of these intracranial electrodes is unique for each patient and depends on the hypothesized epileptogenic and seizure-involved zones. While photographs like that shown in Figure 2.3 and epileptologist notes during the implantation indicate the rough 3D location of each channel relative to the brain, determining the precise location through imaging co-registration techniques is a very challenging task and is thus generally not part of standard clinical practice.

Voltages are recorded continuously from each contact, or channel, on these electrodes at a sampling rate of at least 200 Hz while the patient is in the EMU. Most iEEG waveforms and patterns of clinical interest—spikes, sharp waves, epileptic discharges, bursts, and seizures—fall in the 1-100 Hz frequency range. Recently, interest in higher-frequency transient waveforms of up to 500 Hz [14, 17, 18, 52, 63] has generated interest in recording setups capable of sampling in the kHz range, though such acquisition systems are still experimental and not the standard of clinical care. Currently, epileptologists manually examine and annotate every second of the often weeks of iEEG recorded for each patient.

At the end of this monitoring period, the epileptologists, in consultation with neurosur-

geons, determine whether the patient is likely to benefit from resective surgery, and if so what areas of the brain should be removed. The intracranial electrodes are then removed and the appropriate brain tissue resected, usually during the same surgery.

■ 2.1.2 Electrophysiologic Recordings from Chronic Implants

All human epileptic iEEG recordings come from acute (i.e., temporary) electrode implants during a patient's evaluation in an EMU for resective surgery. But these data almost certainly give an incomplete and distorted view of the brain's "natural" epileptic activity. The brain undergoes large physiologic changes resulting from the electrode implant, both the surgical process of implanting them and the physical pressure they exert on the cortex [75, 91]. The patient's antiepileptic medication doses are often modified many times over the few weeks in the EMU, resulting in seizures that may not be representative of the usual seizures a patient has. Finally, the patients are laying in a hospital bed, a situation that not only affects the patient's physiologic condition—body temperature, blood pressure, sleep cycles—but also is utterly deficient of all environmental factors the patient usually encounters that may or may not influence his epileptic activity.

The NeuroVista Corporation, a startup from Seattle, WA that unfortunately closed recently due to lack of funding, created a device capable of recording 16 channels of iEEG continuously, while a patient is awake and interacting with the natural world in the course of everyday life. These chronic (i.e., long term) electrodes are connected to an acquisition and storage system located in the chest and then telemetered out to an external device that processed them in real time, attempting to predict periods of increased seizure risk.

In this work, we focus on a related set of recordings, taken from dogs with naturally-occurring epilepsy [29]. Dogs are the only other known mammals with naturally occurring seizures, and these seizures are quite similar to those of humans [10, 11, 25]. In addition, dogs have similar rates of drug-resistant medically refractory epilepsy has humans [90, 126, 127]. These similarities make them an ideal model for implantable devices for epilepsy therapy.

■ 2.1.3 Short Bursts of Epileptic Activity

Despite their common occurrence in interictal iEEG, sub-clinical "bursts" of abnormal, presumably epileptic activity—usually 1-3 seconds long—have received scant attention in both the clinical and quantitative epilepsy literature. These bursts are different from other, shorter interictal discharges like local field potential spikes and slow-waves, whose study and relation to seizures has a long history [4, 5]. They are also different from the much faster oscillations of 100-500 Hz mentioned previously. Litt et al. [73] and Traub et al. [124] note that these bursts often occur before seizures and importantly contain prolonged epileptiform discharges, and in a later study Niederhauser et al. [83] develop a detection algorithm for these events. These bursts are termed "sub-clinical" because they do not have any outward clinical manifestations, as full seizures do.

Understanding the relationship between these bursts and seizures motivates the development our model in Chapter 4 and the extensive analysis described in Chapter 5.

■ 2.2 Existing Computational Models of Epilepsy

Development of quantitative approaches for understanding epilepsy has continued to expand over the past few decades thanks to the expansion of EEG (both scalp and intracranial) recordings and the increased prevalence of epilepsy monitoring units, the increases in computational power of standard computers, and the prevailing difficulty of understanding epilepsy as a disease and solving the clinical problems at hand. Much of this quantitative work has focused on seizure detection and prediction, seizure onset localization, seizure dynamics description, and biophysical mechanistic neural models of epilepsy and EEG activity. In this section, we briefly review some of the work focused on modeling epileptic EEG activity. We focus exclusively on the model-based approaches, as they are most relevant to this work and thus avoid vasts areas of valuable discriminative work in areas like seizure, spike, and high frequency oscillation detection, seizure prediction, and seizure onset zone prediction.

Mechanistic models The original Hodgkin-Huxley model of neuronal sodium and potassium ion currents showed that it is possible to simulate realistic action potentials [31, 57]. Since then, much work has been done on creating plausible models of neural activity using networks of physiologically coupled neurons. These approaches generally use various assumptions about brain structure relationships (e.g., feedback connections between deep brain structures like thalamus with surface-level cortex) and local structure connectivity to produce physiologically-plausible EEG activity, including spiking, bursting, and seizures. In the past decade or so, these approaches have moved from modeling small populations of similar neurons to larger, more heterogenous (and thus more realistic) physiologic models.

Inspired by differences in connectivity patterns in the hippocampal CA1 and CA3 regions, Netoff et al. [82] create a small-world neuronal network to understand how network connections and their strength influence transitions between background activity, bursting, and seizing. Wendling et al. [130] explore how the coupling between populations of neurons influences influences the EEG patterns produced during simulation. Kilpatrick and Bressloff [67] use a 2-dimensional spatial neuronal network involving synaptic depression to simulate a wide variety of spatio-temporal EEG patterns.

Incorporating known physiologic connections within cortex and between cortex and thalamic nuclei, Robinson et al. [101] explore how perturbations from the model's steady state reproduce EEG evoked response potentials as well as correlation and coherence measures seen in real signals, and Breakspear et al. [19] use a nonlinear model involving thalamic nuclei and cortical networks to explore the "zones of stability" producing common EEG patterns. Cosandier-Rimele et al. [28] describe this cortico-thalamic relationship by modeling thalamic activity with a dipole source model influencing spatial relationships on the EEG and link it to coupled neuronal populations in cortex that influence temporal dynamics.

Other approaches have explored how particular neurophysiologic variables influence the types of common neuronal activity. Specifically, Stacey and Durand [112] and Stacey et al. [113] suggest that synaptic noise and coupling in hippocampal networks plays a large role in how as an ensemble they produce recognizable burst and high frequency oscillation activity.

Looking toward implantable devices intended to stimulate and suppress seizures, like the NeuroPace RNS device [117], Lopour and Szeri [76] develop cortical model to test charge-neutral electrical stimulation patterns for the seizure suppression.

Source localization The clinical problem of determining brain structures responsible for the initiation of seizures directs much research toward understanding and determining the these structures from the EEG. One well-studied approach involves dipole source localization—solving an inverse problem—from EEG activity. Work by Boon et al. [16] and Van Hese et al. [125] uses dipole modeling to identify seizures with different seizure onset patterns. Merlet and Gotman [78] use a similar approach with interictal spikes. Related work focuses explicitly on localizing frequency-domain EEG activity [2, 6]. Ding et al. [32] use a hybrid approach combining dipole source localization to describe spatial relationships between brain structures and a multivariate autoregressive process to deduce causal relationships from the frequency components of these structures.

Temporal models of epileptic activity Most relevant to our work is research exploring the temporal dynamics and evolution of epileptic activity like seizures. This work often focuses on frequency characteristics of individual channels as well as correlation and coherence measures between channels. Schiff et al. [106] introduce an optimization method for determining the beginning, middle, and end of seizures on multichannel EEG through these features. Chiu et al. [27] describe seizures from single-channel rat hippocampal slices using a hidden Markov model on wavelet-based features that identifies interictal, tonic, chronic, and post-ictal firing patterns. Santaniello et al. [104] also use a hidden Markov model to describe ictal and interictal activity, using features of the singular value decomposition of multichannel EEG. Interestingly, Faul et al. [36] find that a generative Gaussian process model on raw EEG voltage traces actually outperforms other feature-based discriminative classifiers in distinguishing ictal from non-ictal activity.

Eigen-decompositions have also been used to describe and distill the high-dimensional spatio-temporal structure commonly present in epileptic activity. For example, Schiff et al. [105] use the SVD of 1-second time windows to analyze voltage-sensitive dye measurements of different oscillatory wave patterns (planar, spiral, etc) of rat hippocampal slices. Schindler et al. [107] use the eigenvalue spectrum of multichannel EEG correlation structure to describe the dynamics present in a large number of focal-onset seizures. Krystal et al. [70] use the eigen-structure of the time-varying autoregressive model developed by West et al. [131] to decompose the frequency content of single-channel EEG seizures into a few main spectral processes.

■ 2.3 The Bayesian Perspective

The Bayesian perspective on modeling is rooted in the idea that the parameters θ of a model for data X themselves contain uncertainty, and this uncertainty should be openly acknowledged and incorporated into the modeling process. Classical, or frequentist, statistical modeling usually involves determining a specific value for θ, sometimes called a point-

estimate, that is used in describing the data through the probability density $P(X \mid \theta)$. Bayesian modeling involves determining a *distribution* over θ rather than one particular value through application of Bayes' theorem,

$$P(\theta \mid X) = \frac{P(X \mid \theta)P(\theta)}{P(X)}, \tag{2.1}$$

which follows from the conditional probability rule $P(X, \theta) = P(X \mid \theta)P(\theta)$.

In attempting to determine $P(X \mid \theta)$, classical statistical modeling describes the observed data X *given* the known (but really, estimated) parameters θ. In attempting to determine $P(\theta \mid X)$, Bayesian statistical modeling describes the parameters θ *given* the (actually) known data X. We find this philosophical distinction compelling in favor of the Bayesian approach. In addition, we believe the Bayesian approach, in forcing the user to acknowledge and explicitly define the uncertainty present in θ through the prior $P(\theta)$, yields a general, coherent framework for data analysis with a flexibility that allows one to practically construct and use complex models. Readers interested in a more philosophical discussion of the "Bayesian choice" should consult Robert and Casella [100].

In this section, we briefly touch on a number of Bayesian concepts and models relevant to our new work in Chapters 3 and 4. Readers unfamiliar with the material in Sections 2.3.1-2.5 might consult the relevant sections of Bishop [12]. Gelman et al. [47] and MacKay [77] are also good references for much of this material. In Section 2.6, we suggest a number of tutorials and review papers for a more thorough discussion of the Bayesian nonparametric material discussed.

■ 2.3.1 The Prior Distribution

Consider a set of i.i.d. observations $X = \{x_1, \ldots, x_N\}$ and a model with parameters θ for those observations. A prior distribution $p(\theta)$ on the parameters θ allows us to do two things: define the space of possible values that θ can take through the support of the prior and to explicitly encode prior knowledge[1] and assumptions about the values our parameters θ are likely to take. For example, in the case of linear regression, we may believe that the regression coefficients θ should be close to zero. A prior distribution $p(\theta)$ does *not* necessarily imply that θ is truly a random variable or is generated from a particular distribution. Rather, it summarizes the uncertainty and knowledge we have about θ that should be incorporated into the model. For practical purposes often related to conjugacy (discussed in Section 2.3.3) our prior distribution is usually defined by a parametric distribution $p(\theta \mid \lambda)$ given fixed parameters λ. We assume this parametric formulation of the prior throughout this work.

The density $p(x \mid \theta)$ describes the probability of an observation x given the parameters θ. When viewed as a function of x for fixed θ, $p(x \mid \theta)$ is a valid probability distribution—sometimes called the sampling or observation distribution—for x, i.e., it is non-negative over the domain of x and integrates over this domain to one. We can alternately think

[1] Excluding non-informative priors, for now.

of the density $p(x \mid \theta)$ as a function of θ for a fixed observation x. This function of θ is called the *likelihood* function, and its distinction from the distribution of x given a fixed θ is important.

In the Bayesian setting, we are concerned with determining the best *joint* model for our data X and parameters θ, which Bayes's rule shows can be given as the product of the likelihood and the prior,

$$p(X, \theta \mid \lambda) = p(X \mid \theta)p(\theta \mid \lambda). \tag{2.2}$$

■ 2.3.2 The Posterior Distribution

Since the fundamental task of modeling involves determining our model parameters $\theta \in \Theta$, we are usually interested in the *posterior* distribution of only our parameters given the observed data X and our prior parameters λ,

$$\begin{aligned} p(\theta \mid X, \lambda) &= \frac{p(X, \theta \mid \lambda)}{p(X)}, \\ &= \frac{p(X \mid \theta)p(\theta \mid \lambda)}{p(X)}. \end{aligned} \tag{2.3}$$

The marginal likelihood $p(X)$ in the denominator of Equation (2.3) does not depend on θ; it only scales the posterior $p(\theta \mid X, \lambda)$. We thus can (and usually do) write the posterior as *proportional* to simply the product of the likelihood and prior,

$$p(\theta \mid X, \lambda) \propto p(X \mid \theta)p(\theta \mid \lambda). \tag{2.4}$$

Our main task in Bayesian modeling is to obtain the posterior distribution of the parameters θ. Put another way, we are concerned with finding the distribution of θ best explained by our data X rather than finding the value θ that best explains our data, as is the case in maximum likelihood inference.

Sometimes the posterior distribution of our model has a closed-form description that can be derived explicitly from the product of the likelihood and the prior. But in many cases, especially with more complex models, we do not have such a simple description of the posterior distribution and thus must *infer* it by repeatedly sampling from it. We briefly introduce some methods for such posterior inference in Section 2.4.

Posterior predictive distribution In some cases we would like to use our model parameterized by θ to make predictions about a new observation x_+. The *posterior predictive* distribution (sometimes also called simply the predictive distribution) allows us to integrate over the entire space of θ rather than using a single value for it to make inferences about x_+ given the previously-seen observations $X = \{x_1, \ldots, x_N\}$,

$$p(x_+ \mid \{x_1, \ldots, x_N\}, \lambda) = \int_\Theta p(x_+ \mid \theta)p(\theta \mid \{x_1, \ldots, x_N\}, \lambda)d\theta. \tag{2.5}$$

The posterior predictive distribution is especially useful when this integral (a marginalization over the possible values of θ) results in a simple, closed form solution, as is the case

for conjugate priors, discussed in the next section. The posterior predictive distribution is also used in Rao-Blackwellization for collapsing parts of a hierarchical model [23, 46, 74], eliminating inference in intermediate parameters.

Consider predictive distribution in the case where no data is available, often called the *prior predictive* distribution,

$$p(x_+ \mid \lambda) = \int_\Theta p(x_+|\theta)p(\theta \mid \lambda)d\theta. \tag{2.6}$$

Note that this prior predictive distribution is equivalent to the marginal likelihood given the fixed prior parameters λ.

■ 2.3.3 Conjugate Priors

A prior in the family of distributions $\mathcal{P}$ is conjugate for a likelihood function family $\mathcal{F}$ if the resulting posterior distribution is always of the family $\mathcal{P}$,

$$p(\theta \mid X) \in \mathcal{P} \quad \text{for all} \quad p(\theta) \in \mathcal{P} \quad \text{and} \quad p(X \mid \theta) \in \mathcal{F}. \tag{2.7}$$

Note that if $\mathcal{P}$ is very large (possibly, the space of all functions) all priors are trivially conjugate. When referring to conjugacy, we usually consider only families $\mathcal{F}$ whose functions can be described through parameters like λ. The posterior distribution thus depends only on a related set of parameters that are updated according to the observed data. Simply, conjugate priors are quite straightforward and practical, explaining their widespread use. Of course, a conjugate prior may not always be most appropriate for the modeling task at hand, but in this work we use them almost exclusively, especially since they can usually be made *vague* (or broad) when little prior knowledge is available. Conjugate priors of this form have the intuitive interpretation of defining pseudo-observations through λ, which are then combined with the real observations to yield the parameters that define the posterior.

Exponential family The exponential family of probability distributions have a general form that allows natural conjugacy. For an observation x_i, one parameterization of the exponential family distribution with parameters θ can be given as [47]

$$p(x_i \mid \theta) = f(x_i)g(\theta)\exp\left(\phi(\theta)^{\mathrm{T}}u(x_i)\right), \tag{2.8}$$

where $\phi(\theta)$ and $u(x_i)$ usually have the same dimensionality as θ and $g(\theta)$—sometimes called the inverse partition function—ensures that the distribution integrates to one. The likelihood of N i.i.d. observations $X = \{x_1, \ldots, x_N\}$ is thus

$$p(X \mid \theta) = \prod_{i=1}^{N} f(x_i)g(\theta)\exp\left(\phi(\theta)^{\mathrm{T}}u(x_i)\right) \tag{2.9}$$

$$= \left(\prod_{i=1}^{N} f(x_i)\right) g(\theta)^N \exp\left(\phi(\theta)^{\mathrm{T}} \sum_{i=1}^{N} u(x_i)\right). \tag{2.10}$$

As a function of θ, this joint likelihood can also be written as

$$p(X \mid \theta) \propto g(\theta)^N \exp\left(\phi(\theta)^{\mathrm{T}} t(X)\right), \tag{2.11}$$

for $t(X) = \sum_{i=1}^{N} u(x_i)$.

Sufficient Statistics When the prior density of θ has the same form as the exponential family likelihood function specified in Equation (2.8),

$$p(\theta \mid \lambda) \propto g(\theta)^{\eta} \exp\left(\phi(\theta)^{T} \nu\right), \tag{2.12}$$

for $\lambda = (\eta, \nu)$, we see how the posterior is also a function of the exponential family,

$$\begin{aligned} p(\theta \mid X, \lambda) &\propto p(X \mid \theta) p(\theta \mid \lambda), \\ &\propto g(\theta)^N \exp\left(\phi(\theta)^{\mathrm{T}} t(X)\right) g(\theta)^{\eta} \exp\left(\phi(\theta)^{T} \nu\right), \\ &\propto g(\theta)^{\eta+N} \exp\left(\phi(\theta)^{T} (\nu + t(X))\right). \end{aligned} \tag{2.13}$$

Notice how the data influence the posterior *only* though the value $t(X)$, called the sufficient statistic for θ, since it is sufficient to describe the data's contribution to the posterior. This formulate also reinforces the intuition of the prior parameter λ encoding η pseudo-observations with sufficient statistic ν.

In the following five sections, we consider the posteriors resulting from the multinomial likelihood with a conjugate Dirichlet prior and the normal likelihood with several different conjugate priors. The posteriors are used in the models we introduce in Chapters 3 and 4.

■ 2.3.4 Multinomial Observations

Consider a set of N i.i.d. observations $\{x_1, \ldots, x_N\}$ from a K-dimensional multinomial distribution with probabilities $\boldsymbol{\pi}$,

$$x_i \mid \boldsymbol{\pi} \sim \mathrm{Multi}(\boldsymbol{\pi}), \tag{2.14}$$

(sometimes alternately denoted $x_i \mid \boldsymbol{\pi} \sim \boldsymbol{\pi}$) which have a joint likelihood of

$$\begin{aligned} p(x_1, \ldots, x_N \mid \boldsymbol{\pi}) &\propto \prod_{i=1}^{N} \pi_{x_i}, \\ &\propto \prod_{k=1}^{K} \pi_k^{m_k}, \end{aligned} \tag{2.15}$$

where m_k gives the total number of counts in state k over $x_1, \ldots, x_n$. With a conjugate Dirichlet prior $p(\boldsymbol{\pi} \mid \boldsymbol{\alpha}_0)$, our posterior is

$$\begin{aligned} p(\boldsymbol{\pi} \mid x_1, \ldots, x_N) &\propto p(\boldsymbol{\pi} \mid \boldsymbol{\alpha}_0) p(x_1, \ldots, x_N \mid \boldsymbol{\pi}) \\ &\propto \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} \prod_{k=1}^{K} \pi_k^{m_k} \\ &\propto \prod_{k=1}^{K} \pi_k^{\alpha_k + m_k - 1} \\ &\propto \mathrm{Dir}(\boldsymbol{\alpha}_N), \end{aligned} \tag{2.16}$$

where $\boldsymbol{\alpha}_N = \boldsymbol{\alpha}_0 + \mathbf{m}$.

Posterior prediction The posterior predictive distribution of a new observation x^+ is a multinomial,

$$\begin{aligned} p(x^+ \mid x_1, \ldots, x_N) &\propto \int_{\Delta^{d-1}} p(x^+ | \boldsymbol{\pi}) p(\boldsymbol{\pi}, | x_1, \ldots, x_N) \, d\boldsymbol{\pi} \\ &\propto \mathrm{Multi}\left(\widetilde{\boldsymbol{\alpha}}_N\right) \end{aligned} \tag{2.17}$$

where $\Delta^n = \{(\pi_0, \ldots, \pi_n) \in [0,1]^{n+1} \mid \sum_{i=0}^{n} \pi_i = 1\}$ is the $(n+1)$-dimensional probability simplex and $\widetilde{\boldsymbol{\alpha}}_N = \boldsymbol{\alpha}_N / (\mathbf{1}^T \boldsymbol{\alpha}_N)$.

■ 2.3.5 Normal Observations

Consider the case of N i.i.d. scalar observations $\{x_1, \ldots, x_N\}$, each of which we model using a normal or Gaussian distribution with mean μ and variance σ^2,

$$p(x_i \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right). \tag{2.18}$$

The conjugate prior for this normal likelihood with unknown mean and variance is a joint normal inverse-gamma ($\mathcal{N}$-IG),

$$\begin{aligned} \sigma^2 &\sim \mathrm{IG}(\alpha_0, \beta_0) \\ \mu &\sim \mathcal{N}(\mu_0, \sigma^2 / n_0), \end{aligned} \tag{2.19}$$

with inverse-gamma (IG) prior shape α_0 and scale β_0 and normal prior counts n_0 and mean μ_0. This distribution is denoted $\mathcal{N}$-IG$(n_0, \mu_0, \alpha_0, \beta_0)$. The joint prior density is (omitting the prior parameters in the conditional for notational compactness),

$$\begin{aligned} p(\mu, \sigma^2) &= p(\sigma^2) p(\mu \mid \sigma^2), \\ &= \frac{(\beta_0)^{\alpha_0}}{\Gamma(\alpha_0)} \left(\sigma^2\right)^{-(\alpha_0+1)} \exp\left(-\frac{\beta_0}{\sigma^2}\right) \left(2\pi\sigma^2/n_0\right)^{-1/2} \exp\left(-\frac{1}{2\sigma^2/n_0}(\mu - \mu_0)^2\right), \\ &\propto \left(\sigma^2\right)^{-(\alpha_0+1)} \left(\sigma^2\right)^{-1/2} \exp\left(-\frac{\beta_0}{\sigma^2} - \frac{n_0}{2\sigma^2}(\mu - \mu_0)^2\right). \end{aligned} \tag{2.20}$$

The joint likelihood of the N observations is,

$$\begin{aligned} p(x_1, \ldots, x_N \mid \mu, \sigma^2) &= \prod_{i=1}^{N} p(x_i \mid \mu, \sigma^2) \\ &\propto \prod_{i=1}^{N} \left(\sigma^2\right)^{-1/2} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right) \\ &\propto \left(\sigma^2\right)^{-N/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2\right), \end{aligned} \tag{2.21}$$

which when combined with the prior in Equation (2.20) yields the joint posterior

$$\begin{aligned} p(\mu, \sigma^2 \mid x_1, \ldots, x_N) &\propto p(\mu, \sigma^2) p(x_1, \ldots, x_N \mid \mu, \sigma^2) \\ &\propto \left(\sigma^2\right)^{-(\alpha_0+1)} \left(\sigma^2\right)^{-1/2} \exp\left(-\frac{\beta_0}{\sigma^2} - \frac{n_0}{2\sigma^2}(\mu - \mu_0)^2\right) \cdot \\ &\quad \left(\sigma^2\right)^{-N/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2\right) \end{aligned} \tag{2.22}$$

Completing the square of the quadratic term in the exponential (see Appendix A.1 for the derivation in the multivariate case) yields the posterior distribution $\mathcal{N}$-IG$(n_N, \mu_N, \alpha_N, \beta_N)$ with parameters

$$\begin{aligned} n_N &= n_0 + n, \\ n_N \mu_N &= n_0 \mu_0 + \sum_{i=1}^{N} x_i, \\ \alpha_N &= \alpha_0 + \frac{N}{2}, \\ \beta_N &= \beta_0 + \frac{1}{2}\left(n_0 \mu_0^2 + \sum_{i=1}^{N} x_i^2 - n_N \mu_N^2\right). \end{aligned} \tag{2.23}$$

Note that the inverse-gamma distribution IG(α, β) is equivalent to the scaled inverse-chi-squared distribution $\chi^{-2}(\nu, s^2)$ with $\nu = 2\alpha$ and $s^2 = \beta/\alpha$. See Murphy [81] for a similar analysis using a number of different conjugate priors and parameterizations.

Posterior prediction The posterior predictive distribution for a new observation is a t distribution [47],

$$p(x_+ \mid x_1, \ldots, x_N) \propto t_{2\alpha_n}\left(\mu_N, \frac{n_N + 1}{n_N \alpha_N}\beta_N\right). \tag{2.24}$$

■ 2.3.6 Multivariate Normal Observations

The approach for observations $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ with $\mathbf{x}_i \in \mathbb{R}^d$ is exactly the same as that for scalar observations. We give derivations for the multivariate normal likelihood $p(\mathbf{x}_1, \ldots, \mathbf{x}_N \mid \boldsymbol{\mu}, \Sigma)$ and joint conjugate normal inverse-Wishart prior on $(\boldsymbol{\mu}, \Sigma)$ in Appendix A.1. Briefly, the posterior is

$$p(\boldsymbol{\mu}, \Sigma \mid \mathbf{x}_1, \ldots, \mathbf{x}_N) \propto \\ |\Sigma|^{-(\nu_N + d)/2 - 1} \exp\left(-\frac{1}{2}\mathrm{tr}\left(S_N \Sigma^{-1}\right) - \frac{n_N}{2}\left(\boldsymbol{\mu} - \boldsymbol{\mu}_N\right)^T \Sigma^{-1}\left(\boldsymbol{\mu} - \boldsymbol{\mu}_N\right)\right), \quad (2.25)$$

for normal posterior counts n_N and mean $\boldsymbol{\mu}_N$ and inverse-Wishart posterior degrees of freedom ν_N and scale S_N, whose values are summarized by the posterior parameters

$$\begin{aligned} n_N &= n_0 + N, \\ n_N \boldsymbol{\mu}_N &= n_0 \boldsymbol{\mu}_0 + \sum_{i=1}^{N} \mathbf{x}_i, \\ \nu_N &= \nu_0 + N, \\ S_N &= S_0 + \sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} + n_0 \boldsymbol{\mu}_0 \boldsymbol{\mu}_0^{\mathrm{T}} - n_N \boldsymbol{\mu}_N \boldsymbol{\mu}_N^{\mathrm{T}}. \end{aligned} \quad (2.26)$$

Known mean or covariance Situations where we know (or assume we know) the mean $\boldsymbol{\mu}$ or covariance Σ of our observation distribution are special cases of the situation in which both are unknown. For known mean $\boldsymbol{\mu}$, the posterior for Σ is simply the joint posterior given in Equation 2.25 without the contribution of the normal prior on $\boldsymbol{\mu}$,

$$p(\Sigma \mid \mathbf{x}_1, \ldots, \mathbf{x}_N, \boldsymbol{\mu}) \propto |\Sigma|^{-(\nu_N + d + 1)/2} \exp\left(-\frac{1}{2}\mathrm{tr}\left(S_N \Sigma^{-1}\right)\right), \quad (2.27)$$

for posterior parameters

$$\begin{aligned} \nu_N &= \nu_0 + N, \\ S_N &= S_0 + \sum_{i=1}^{N} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^{\mathrm{T}}. \end{aligned} \quad (2.28)$$

Similarly, in the case of known covariance Σ, the posterior on $\boldsymbol{\mu}$ is

$$p(\boldsymbol{\mu} \mid \mathbf{x}_1, \ldots, \mathbf{x}_N, \Sigma) \propto |\Sigma|^{-1/2} \exp\left(-\frac{n_N}{2}\left(\boldsymbol{\mu} - \boldsymbol{\mu}_N\right)^T \Sigma^{-1}\left(\boldsymbol{\mu} - \boldsymbol{\mu}_N\right)\right), \quad (2.29)$$

for posterior parameters

$$\begin{aligned} n_N &= n_0 + N, \\ n_N \boldsymbol{\mu}_N &= n_0 \boldsymbol{\mu}_0 + \sum_{i=1}^{N} \mathbf{x}_i. \end{aligned} \quad (2.30)$$

■ 2.3.7 The Hyper-Inverse Wishart Prior

In the previous section, we discussed using the inverse-Wishart prior for the unknown covariance parameter Σ, assuming *full covariance*, i.e., a dense covariance matrix in which every element is non-zero. This form of Σ implies that each of the d elements of the observation $\mathbf{x}$ covaries independently with every other element. Describing these dependencies in a Gaussian graphical model $G = (V, E)$ [30], each of the d vertices (together denoted by V) corresponding to an element of $\mathbf{x}$ would have an (undirected) edge connecting it to the $d-1$ other vertices. E denotes the set of edges. In some scenarios, this assumption is reasonable, or we have little knowledge to restrict it in a principled way. In others, however, we may believe the connections in G to be more sparse; not every vertex is connected to every other, giving rise to conditional independencies in G. Dawid and Lauritzen [30] showed that these conditional independencies in G correspond to zero entries in the inverse covariance matrix $\Omega = \Sigma^{-1}$, also known as the precision matrix.

We briefly review a few aspects of undirected graphs relevant to the hyper-inverse Wishart prior [30, 92]. A graph or subgraph $G_p = (V_p, E_p)$ is *complete* if its vertices V_p are fully-connected to each other. A graph G is decomposable if and only if

a) it can be defined by a set of intersecting prime components $\mathcal{P} = \{P_1, \ldots, P_G\}$ and the separators $\mathcal{S} = \{S_2, \ldots, S_G\}$ comprising these intersections,

b) each P_i is nonempty and complete,

c) $S_i = P_i \cup P_h$ for only one $h < i$,

d) each S_i is nonempty and complete.

The prime components $\mathcal{P}$ and separators $\mathcal{S}$ define the graph G.

The hyper-inverse Wishart distribution is defined over the graph G with degrees of freedom b_0 and scale D_0 parameters,

$$\Sigma \sim \mathrm{HIW}_G(b_0, D_0). \tag{2.31}$$

The marginal distribution for the entries of Σ corresponding to each prime component $P \in \mathcal{P}$ and separator $S \in \mathcal{S}$ is inverse-Wishart,

$$\begin{aligned} \Sigma^{(P,P)} &\sim \mathrm{IW}(b_0, D_0^{(P,P)}), \\ \Sigma^{(S,S)} &\sim \mathrm{IW}(b_0, D_0^{(S,S)}), \end{aligned} \tag{2.32}$$

where the superscript $(\cdot)^{(P)}$ indexes the elements of the vector corresponding to the vertices in P and the superscript $(\cdot)^{(P,P)}$ similarly indexes the rows and columns of the sub-matrix. The notation involving S has the parallel meaning. This hyper-Markov distribution implies the density

$$p(\Sigma \mid b_0, D_0) = \frac{\prod_{P \in \mathcal{P}} \mathrm{IW}(\Sigma^{(P,P)} \mid b_0, D_0^{(P,P)})}{\prod_{S \in \mathcal{S}} \mathrm{IW}(\Sigma^{(S,S)} \mid b_0, D_0^{(S,S)})}. \tag{2.33}$$

Observation distribution When the precision matrix $\Omega = \Sigma^{-1}$ is a member of the set $\mathcal{G} \subset \mathbb{S}^{\lceil}_{++}$ comprised all positive-definite matrices with graph structure (i.e., conditional independencies implied by) G, the probability density of an observation $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ for known $\boldsymbol{\mu}$ also decomposes according to the prime components in $\mathcal{P}$ and separators in $\mathcal{S}$,

$$p(\mathbf{x} \mid \boldsymbol{\mu}, \Sigma) = \frac{\prod_{P \in \mathcal{P}} \mathcal{N}(\mathbf{x}^{(P)} \mid \boldsymbol{\mu}^{(P)}, \Sigma^{(P,P)})}{\prod_{S \in \mathcal{S}} \mathcal{N}(\mathbf{x}^{(S)} \mid \boldsymbol{\mu}^{(S)}, \Sigma^{(S,S)})}, \tag{2.34}$$

Posterior distribution Since both the prior in Equation (2.33) and the likelihood in Equation (2.34) similarly factorize over the prime components and separators, the posterior follows by straightforward analogy to the standard IW prior and posterior,

$$p(\Sigma \mid \mathbf{x}_1, \ldots, \mathbf{x}_N, \boldsymbol{\mu}) \propto \mathrm{HIW}_G(b_N, D_N), \tag{2.35}$$

with posterior degrees of freedom and scale parameters,

$$\begin{aligned} b_N &= b_0 + N, \\ D_N &= D_0 + \sum_{i=1}^{N} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^{\mathrm{T}}. \end{aligned} \tag{2.36}$$

Carvalho et al. [22] describe how to sample a value Σ from a hyper-inverse Wishart distribution.

■ 2.4 Posterior Inference

The posteriors for simple models of a few parameters with a conjugate prior are often straightforward to determine and sample from. For example, observations $\{x_1, \ldots, x_N\}$ from normal distribution with unknown mean μ and known variance σ^2 and a conjugate normal prior $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$ lead to a normal posterior,

$$\mu \sim \mathcal{N}\left(\mu_N, \sigma_N^2\right), \tag{2.37}$$

with posterior parameters

$$\begin{aligned} \mu_N &= \sigma_N^2 \left(\frac{1}{\sigma_0^2} \mu_0 + \frac{N}{\sigma^2} \bar{x} \right), \\ \sigma_N^2 &= \left(\frac{1}{\sigma_0^2} + \frac{N}{\sigma^2} \right)^{-1}, \end{aligned} \tag{2.38}$$

where $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$ denotes the observation average. The mean, variance, and mode of $\mathcal{N}(\mu_N, \sigma_N^2)$ are straightforward to calculate, and computational methods exist to sample directly from it.

Often, however, we are interested in more complicated models over a set of d parameters $\theta = (\theta_1, \ldots, \theta_d)$. While we may be able to (and often do) write out the full posterior

distribution over θ, determining basic properties of this posterior and sampling from it are often difficult, if not impossible. Monte Carlo computation techniques allow us to generate samples from the approximate posterior distribution without explicitly sampling from it. When the number of parameters is low, rejection and slice sampling [100] allow simulation from the posterior using samples from another well-defined, easy to simulate distribution like the uniform.

But for posteriors with medium to high complexity (or dimensionality), these posterior computation approaches may not be appropriate. Markov chain Monte Carlo (MCMC) techniques are a very common and powerful class of methods for sampling θ from the approximate posterior when its complexity precludes simpler approaches, e.g., in hierarchical models. MCMC sampling involves sampling a sequence of parameters $(\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(S)})$, where the distribution of $\theta^{(s)}$ depends only on the previous sample $\theta^{(s-1)}$ and is conditionally independent of all other $\theta^{(s'<s-1)}$ (hence the *Markov* nature of the chain). The parameters $\theta^{(s)}$ are sampled from an approximation to the posterior given the value(s) $\theta^{(s-1)}$, with the assumption that this approximation becomes closer and closer to the stationary posterior as s increases.

Below, we briefly introduce several MCMC techniques we employ in this work. Interested readers should consult Brooks et al. [21], Gelman et al. [47], Robert [99], Robert and Casella [100] for expanded, more technical discussions of these and many more topics related to posterior inference.

■ 2.4.1 Gibbs Sampling

As before, consider the case of a joint posterior $p(\theta \mid X, \lambda)$ over multiple parameters $\theta = (\theta_1, \ldots, \theta_d)$ with data X and prior parameters λ. The Gibbs sampler uses the conditional posterior distribution of each $\theta_j^{(s)}$ at iteration s given the other parameters to sample each of the d parameters in turn within one MCMC iteration,

$$\theta_j^{(s)} \sim p\left(\theta_j \mid \{\theta_1^{(s)}, \ldots, \theta_{j-1}^{(s)}, \theta_{j+1}^{(s-1)}, \ldots, \theta_d^{(s-1)}\}, X, \lambda\right). \tag{2.39}$$

Since complex models are often built from simpler components with straightforward posterior inference (e.g., the normal model given in Equation (2.38)), these conditional posterior distributions are often straightforward to sample from. When a parameter θ_j does not depend on all of the other $d-1$ parameters, terms in the full posterior that do not involve θ_j can be absorbed into the proportionality of the posterior, greatly simplifying the sampling of θ_j.

Practical considerations with iterative simulation Iterative Simulation MCMC methods like Gibbs sampling and the more general Metropolis-Hastings algorithm require a few other considerations not necessary when sampling directly from the posterior. First, since the iterations may begin with parameters $\theta^{(1)}$ not representative of the posterior distribution, early samples in the iteration $(\theta^{(2)}, \theta^{(3)}, \ldots)$ may thus be far from the true posterior, indicating that the Markov chain of parameters has not yet converged. It is thus standard practice to discard the first part of the iteration, also known as a (Markov) chain, as *burn-in*. Gelman

et al. [47] suggest discarding the first half of the iterations, though in practice we generally assess convergence of a few parameters and/or heldout data likelihood to determine at what point the chain seems to have reached the stationary posterior and thus at what point we may safely keep the MCMC samples.

While ideally every MCMC chain would fully explore the entire posterior distribution, in reality parameters θ can sometimes become "trapped" in local modes for a long time when the chain exhibits poor mixing. It is thus common practice to start a number of independent MCMC chains with different starting parameters $\theta^{(1)}$ to better ensure that the full posterior is likely to be explored by the samples across all the chains. In addition, Gelman et al. [47] and Brooks and Gelman [20] discuss how monitoring the variance of various scalar parameters within each chain compared to the variance between chains yields insights into when the chains have converged to the true posterior, in which cases the two variances should be very close to each other.

The iterative nature of MCMC algorithms also implies correlation between samples close to each other in the chain. While not in and of itself a problem—since our posterior analyses are usually agnostic to the sample number and the samples are all identically distributed from the posterior—this correlation (specifically, autocorrelation) in effect reduces the number of independent MCMC samples we have. It is thus also standard practice to *thin* the samples, keeping only every kth sample for storage efficiency.

■ 2.4.2 Metropolis-Hastings

In some cases sampling $\theta^{(s)}$ from its conditional posterior given $\theta^{(s-1)}$ is difficult, preventing our use of a Gibbs sampler. The Metropolis-Hastings algorithm is useful in these cases in that it does not depend on the conditional posterior to get the next value $\theta^{(s)}$ but instead samples a new value $\theta^{(*)}$ from a specified proposal distribution $J_s(\theta^{(*)} \mid \theta^{(s-1)})$, where we note that the distribution J_s can optionally depend on s, though in this work we use a fixed $J = J_s$. The proposed value $\theta^{(*)}$ is then *accepted* with a certain probability, given by the *proposal ratio* $\rho(\theta^{(*)} \mid \theta^{(s-1)})$. The proposal ratio is the ratio of the posterior densities of $\theta^{(*)}$ and $\theta^{(s-1)}$ given data X and prior parameters λ, each normalized by the corresponding proposal distribution density,

$$\begin{aligned}\rho(\theta^{(*)} \mid \theta^{(s-1)}) &= \frac{p(\theta^{(*)} \mid X, \lambda)/J_s(\theta^{(*)} \mid \theta^{(s-1)})}{p(\theta^{(s-1)} \mid X, \lambda)/J_s(\theta^{(s-1)} \mid \theta^{(*)})}, \\ &= \frac{p(\theta^{(*)} \mid X, \lambda)}{p(\theta^{(s-1)} \mid X, \lambda)} \frac{J_s(\theta^{(s-1)} \mid \theta^{(*)})}{J_s(\theta^{(*)} \mid \theta^{(s-1)})}. \end{aligned} \tag{2.40}$$

The normalization ratio of the proposal densities accounts for a non-symmetric proposal distribution (i.e., $J_s(\theta^{(*)} \mid \theta^{(s-1)}) \neq J_s(\theta^{(s-1)}) \mid \theta^{(*)})$, as in a gamma distribution), which is sometimes desirable and can increase the speed of the random walk [47]. Such normalization ensures the detailed balance condition where a move from $\theta^{(s-1)}$ to $\theta^{(*)}$ is just as likely as a move from $\theta^{(*)}$ to $\theta^{(s-1)}$, further ensuring that the equilibrium distribution of the Markov chain is equivalent to the target posterior distribution [21]. When the proposal

distribution is symmetric ($J_s(\theta^{(*)} \mid \theta^{(s-1)}) \neq J_s(\theta^{(s-1)}) \mid \theta^{(*)})$), the Metropolis-Hastings proposal ratio reduces to the Metropolis proposal ratio. The proposed value $\theta^{(*)}$ is accepted with probability (w.p.) $\min(1, \rho(\theta^{(*)} \mid \theta^{(s-1)}))$ or otherwise discarded. This update can be written as

$$\theta^{(s)} = \begin{cases} \theta^{(*)} & \text{w.p.} \quad \min(1, \rho(\theta^{(*)} \mid \theta^{(s-1)})), \\ \theta^{(s-1)} & \text{otherwise.} \end{cases} \tag{2.41}$$

We can think of this proposal as always accepting $\theta^{(*)}$ if it has larger normalized posterior density than $\theta^{(s-1)}$ and otherwise accepting it with probability equal to the proposal ratio $\rho(\theta^{(*)} \mid \theta^{(s-1)})$. Gibbs sampling is actually a special case of the Metropolis-Hastings algorithm where the proposal distribution is the conditional posterior and the proposal ratio is one,

$$\begin{aligned} J_s &= p(\theta^{(*)} \mid \theta^{(s-1)}, X, \lambda), \\ \rho(\theta^{(*)} \mid \theta^{(s-1)} &= 1, \end{aligned} \tag{2.42}$$

making the proposal $\theta^{(*)}$ always accepted.

■ 2.5 Mixture Modeling

Mixture modeling is a very common form of unsupervised learning. It is far beyond the scope of this work to provide a full review of this vast area, so we only address here the key concepts required for understanding later portions of this work. Readers new to these concepts may benefit the discussions of mixture modeling given in Gelman et al. [47, chap. 18] and Bishop [12, chap 9] and that of hidden Markov models given Bishop [12, chap. 13].

In this section, we briefly introduce mixture models in the Bayesian setting and then relate these to hidden Markov models for sequential data.

■ 2.5.1 Mixture Models

Observed data often do not adhere to a nicely parameterized distribution like a Gaussian. In this case, we would like a method for describing these data without assuming that they all come from a single parametric distribution. Mixture models, where the density of the data is approximated using a mixture of simpler, often parametric models provide one solution to this problem.

Consider a set of N observations $\{x_1, \ldots, x_N\}$ modeled by K independent model components, each with parameters ϕ_k. The distribution of component k is given by $F(\phi_k)$, and we denote the density of component k at point x_i as $f(x_i \mid \phi_k)$. We assume that each data point x_i has a latent, or hidden, indicator variable z_i that denotes which of the K mixture components describes x_i. We also specify the model components weights $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K) \in [0, 1]^K$, where $\sum_{k=1}^K \pi_k = 1$.

We usually would like to infer component parameters $\{\phi_k\}_{k=1}^K$ and weights $\boldsymbol{\pi}$ from the data, so we place priors (often conjugate) on these parameters. Let H denote the prior

distribution for ϕ_k. The conjugate prior for the weights $\boldsymbol{\pi}$ of a multinomial likelihood is a K-dimensional Dirichlet distribution with parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$. All together, this generative model can be described as

$$\begin{aligned} \boldsymbol{\pi} &\sim \mathrm{Dir}(\boldsymbol{\alpha}), \\ \phi_k &\sim H, & k &= 1, \ldots, K, \\ z_i &\sim \boldsymbol{\pi}, & i &= 1, \ldots, N, \\ x_i &\sim F(\phi_{z_i}), & i &= 1, \ldots, N. \end{aligned} \tag{2.43}$$

The joint distribution of the observations and latent indicator variables given the component parameters and weights is

$$\begin{aligned} p\left(\{(x_i, z_i)\}_{i=1}^N \mid \{\phi_k\}_{k=1}^K, \boldsymbol{\pi}\right) &= \prod_{i=1}^N p(z_i \mid \boldsymbol{\pi}) p(x_i \mid z_i, \{\phi_k\}_{k=1}^K), \\ &= \prod_{i=1}^N \pi_{z_i} f(x_i \mid \phi_{z_i}). \end{aligned} \tag{2.44}$$

Summing over the K mixture components yields the joint marginal distribution for the observations,

$$p\left(\{x_i\}_{i=1}^N \mid \{\phi_k\}_{k=1}^K, \boldsymbol{\pi}\right) = \prod_{i=1}^N \sum_{k=1}^K \pi_k f(x_i \mid \phi_k). \tag{2.45}$$

Conjugate posterior Assuming a conjugate Dirichlet prior on $\boldsymbol{\pi}$, the posterior follows from the product of the multinomial likelihoods on $\{z_i\}_{i=1}^N$ and the Dirichlet prior,

$$p(\boldsymbol{\pi} \mid \{z_i\}_{i=1}^N, \boldsymbol{\alpha}) \propto \mathrm{Dir}(\boldsymbol{\alpha} + \mathbf{m}), \tag{2.46}$$

where $\mathbf{m} = (m_1, \ldots, m_K)$ and $m_k = \sum_{i=1}^N \delta(k, z_i)$ denotes the number observations assigned to component k.

If the prior H on ϕ_k has parameters λ, the posterior of each ϕ_k is proportional to

$$p\left(\phi_k \mid \{(x_i, z_i)\}_{i=1}^N, \lambda\right) \propto p(\phi_k \mid \lambda) \prod_{i \mid z_i = k} f(x_i \mid \phi_k). \tag{2.47}$$

When H is conjugate to F, this posterior distribution is of the same type as H.

■ 2.5.2 Hidden Markov Models

When our observations are sequential, $y_1, \ldots, y_T$, over some integer length T, it may make sense to introduce some sort of sequential relationship between the latent states $\{z_t\}_{t=1}^T$. The hidden Markov model[2] employs a Markovian assumption, where state z_{t+1} is independent

[2]Much of this section closely follows similar discussions in Fox [38].

of states $z_1, \ldots, z_{t-1}$ given z_t. Conditioned on the latent states $\{z_t\}_{t=1}^T$, the observations $\{y_t\}_{t=1}^T$ are independent,

$$y_t \mid z_t \sim F(\phi_{z_t}). \tag{2.48}$$

As in the standard mixture modeling discussed above, ϕ_k denotes the parameter(s) describing component k's particular *emission distribution* $F(\phi_k)$. For $t > 1$, the state z_t evolves according to the transition distribution $\boldsymbol{\pi}_j \in [0,1]^K$ with $\sum_{k=1}^K \pi_{j,k} = 1$ given by the previous state $z_{t-1} = k$,

$$z_t \sim \boldsymbol{\pi}_{z_{t-1}}, \qquad t = 2, \ldots, T, \tag{2.49}$$

and first state z_1 is distributed according to the initial transition distribution $\boldsymbol{\pi}_0$,

$$z_1 \sim \boldsymbol{\pi}_0. \tag{2.50}$$

The joint likelihood on the observations and latent state sequence is

$$p\left(y_{1:T}, z_{1:T}\right) = p\left(z_1\right) p(y_1 \mid z_1) \prod_{t=2}^{T} p\left(z_t \mid z_{t-1}\right) p\left(y_t \mid z_t\right), \tag{2.51}$$

where we have omitted $\boldsymbol{\pi}_0$, $\{\boldsymbol{\pi}_j\}_{j=1}^K$, and $\{\phi_k\}_{k=1}^K$ in the conditionals of $p(z_0)$, $p(z_t \mid z_{t-1})$, and $p(y_t \mid z_t)$, respectively, for notational compactness.

Marginalizing over the state sequence A naive method to calculate the marginal likelihood of the data $y_{1:T}$ involves explicitly summing over the K^T possible values $z_{1:T}$ can take. Of course, this approach quickly becomes computationally intractable as K and T reach even moderate levels. Fortunately, the Markov property of the state sequence $z_{1:T}$ allows us to use a dynamic programming technique called the sum-product algorithm [95] to efficiently calculate $p(y_{1:T})$. We omit $\{\phi_k\}_{k=1}^K$ and $\{\boldsymbol{\pi}_j\}_{j=1}^K$ from the conditionals for notational compactness.

The joint likelihood of the first time point and a state equal to value k is straightforward,

$$p(y_1, z_1 = k) = p(y_1 \mid z_1 = k) p(z_1 = k). \tag{2.52}$$

We get the joint likelihood of the first two time points and the second indicator equal to value ℓ by marginalizing over the values of z_1,

$$\begin{aligned} p(y_{1:2}, z_2 = \ell) &= \sum_{k=1}^{K} p(y_1, z_1 = k) p(y_2 \mid z_2 = \ell) p(z_2 = \ell \mid z_1 = k) \\ p(y_{1:2}, z_2 = \ell) &= p(y_2 \mid z_2 = \ell) \sum_{k=1}^{K} p(y_1, z_1 = k) p(z_2 = \ell \mid z_1 = k) \end{aligned} \tag{2.53}$$

Generalizing to time point t, we have

$$p(y_{1:t}, z_t = \ell) = p(y_t \mid z_t = \ell) \sum_{k=1}^{K} p(y_{1:t-1}, z_{t-1} = k) p(z_t = \ell \mid z_{t-1} = k), \tag{2.54}$$

Algorithm 1 HMM sum-product algorithm for calculating $y_{1:T}$ marginal likelihood

let $\boldsymbol{\pi}_0$ and $\boldsymbol{\pi} = (\boldsymbol{\pi}_1 | \cdots | \boldsymbol{\pi}_K)^{\mathrm{T}}$ be the initial and transition distributions, respectively
let $\mathbf{u}_t \in \mathbb{R}_+^K$ denote the vector of likelihoods $p(y_t \mid z_t = k)$ for each k
let $\boldsymbol{\xi}_t \in \mathbb{R}_+^K$ denote the vector of forward messages $p(y_{1:t-1}, z_{t-1} = k)$ at time t for each k

normalize each $\mathbf{u}_t$ to sum to 1, preventing underflow during computations
for $t = 1, \ldots, T$ **do**
    store the marginal log probability over $\mathbf{u}_t$ in v_t
        $v_t \leftarrow \log\left(\mathbf{1}^{\mathrm{T}}\mathbf{u}_t\right)$
    normalize $\mathbf{u}_t$
        $\widetilde{\mathbf{u}}_t \leftarrow \mathbf{u}_t / \exp(v_t)$
end for

calculate and normalize initial forward messages
    $\boldsymbol{\xi}_1 \leftarrow \widetilde{\mathbf{u}}_1 \circ \boldsymbol{\pi}_0$
    $w_1 \leftarrow \mathbf{1}^{\mathrm{T}}\boldsymbol{\xi}_1$
    $\widetilde{\boldsymbol{\xi}}_1 \leftarrow \boldsymbol{\xi}_1 / w_1$
    $v_1 \leftarrow v_1 + \log(w_1)$

propagate messages forward through each time point
for $t = 2, \ldots, T$ **do**
    transmit messages forward
        $\boldsymbol{\xi}_t \leftarrow \widetilde{\mathbf{u}}_t \circ \left(\boldsymbol{\pi}^{\mathrm{T}}\widetilde{\boldsymbol{\xi}}_{t-1}\right)$
    normalize
        $w_t \leftarrow \mathbf{1}^{\mathrm{T}}\boldsymbol{\xi}_t$
        $\widetilde{\boldsymbol{\xi}}_t \leftarrow \boldsymbol{\xi}_t / w_t$
        $v_t \leftarrow v_t + \log(w_t)$
end for

calculate final marginal log likelihood over state sequences $z_{1:T}$
    $\log p(y_{1:T}) = \sum_{t=1}^{T} v_t$

which recursively depends on the value $p(y_{1:t-1}, z_{t-1} = k)$. This recursive formulation allows for efficient, tractable calculation of the marginal data likelihood $p(y_{1:T})$ via the *forward messages* encoded in $p(y_{1:t}, z_t = k)$ for each $k \in \{1, \ldots, K\}$,

$$p(y_{1:T}) = \sum_{k=1}^{K} p(y_{1:T}, z_T = k). \tag{2.55}$$

Algorithm 1 gives a numerically stable recipe for calculating the (log) marginal likelihood $p(y_{1:T})$.

Sampling the state sequence Consider the conditional likelihood of the last observation y_T given the next-to-last state z_{T-1},

$$p(y_T \mid z_{T-1} = \ell) = \sum_{k=1}^{K} p(y_T \mid z_T = k) p(z_T = k | z_{T-1} = \ell). \tag{2.56}$$

Now consider the conditional likelihood of the last two observations $y_{T-1:T}$ given the third-to-last state z_{T-2},

$$p(y_{T-1:T} \mid z_{T-2} = \ell) = \sum_{k=1}^{K} p(y_{T-1} \mid z_{T-1} = k) p(y_T \mid z_{T-1} = k) p(z_{T-1} = k \mid z_{T-2} = \ell). \tag{2.57}$$

Generalizing to time point t, we have

$$p(y_{t+1:T} \mid z_t = \ell) = \sum_{k=1}^{K} p(y_{t+1} \mid z_{t+1} = k) p(y_{t+2:T} \mid z_{t+1} = k) p(z_{t+1} = k \mid z_t = \ell), \tag{2.58}$$

which depends recursively on the *backward messages* $p(y_{t+2:T} \mid z_{t+1} = k)$ for each $k \in \{1, \ldots, K\}$.

In sampling the state sequence, we may consider sampling the each time point z_t independently or sampling the entire sequence $z_{1:T}$ at once. In this first case of sampling z_t, we desire the posterior distribution of z_t, which we derive using Bayes rule,

$$\begin{aligned} p(y_{1:T}, z_t) &= p(y_{1:t}, z_t) p(y_{t+1:T} \mid z_t), \\ p(z_t \mid y_{1:T}) &\propto p(y_{1:t}, z_t) p(y_{t+1:T} \mid z_t), \end{aligned} \tag{2.59}$$

which is simply the product of the forward messages given in Equation (2.54) and the backward messages given in Equation (2.58) at time t. In the second case of sampling $z_{1:T}$ at once, we desire the joint posterior distribution of $z_{1:T}$, which factors into

$$p(z_{1:T} \mid y_{1:T}) = p(z_T \mid z_{T-1}, y_{1:T}) p(z_{T-1} \mid z_{T-2}, y_{1:T}) \cdots p(z_2 \mid z_1, y_{1:T}) p(z_1 \mid y_{1:T}) \tag{2.60}$$

If we first sample z_1, we can condition on it to then sample z_2 and continue in this fashion until we finish with z_T. The posterior for z_1 is

$$\begin{aligned} p(z_1 \mid y_{1:T}) &\propto p(y_{1:T} \mid z_1) p(z_1) \\ &\propto p(y_{2:T} \mid z_1) p(y_1 \mid z_1) p(z_1), \end{aligned} \tag{2.61}$$

which is easily calculated from the product of the backward messages given in Equation (2.58), the likelihood of observation y_i given each of the K possible values of z_1, and the initial distribution $\boldsymbol{\pi}_0$. The posterior for state z_t with $t > 1$ is similar,

$$p(z_t \mid y_{1:T}) \propto p(y_{t+1:T} \mid z_t) p(y_t \mid z_t) p(z_t \mid z_{t-1}), \tag{2.62}$$

Algorithm 2 HMM sum-product algorithm for block-sampling $z_{1:T}$

let $\boldsymbol{\pi}_0$ and $\boldsymbol{\pi} = (\boldsymbol{\pi}_1 | \cdots | \boldsymbol{\pi}_K)^{\mathrm{T}}$ be the initial and transition distributions, respectively
let $\mathbf{u}_t \in \mathbb{R}_+^K$ denote the vector of likelihoods $p(y_t \mid z_t = k)$ for each k
let $\boldsymbol{\zeta}_t \in \mathbb{R}_+^K$ denote the vector of backward messages $p(y_{t+1:T} \mid z_t = k)$ at t for each k

normalize each $\mathbf{u}_t$ to sum to 1, preventing underflow during computations
for $t = 1, \ldots, T$ **do**
    $\widetilde{\mathbf{u}}_t \leftarrow \mathbf{u}_t / \left(\mathbf{1}^{\mathrm{T}} \mathbf{u}_t\right)$
end for

calculate backward messages over all time points
    $\widetilde{\boldsymbol{\zeta}}_T = \mathbf{1}$
for $t = T-1, \ldots, 1$ **do**
    transmit messages backward
        $\boldsymbol{\tau}_{t+1} \leftarrow \widetilde{\mathbf{u}}_{t+1} \circ \widetilde{\boldsymbol{\zeta}}_{t+1}$
        $\boldsymbol{\zeta}_t \leftarrow \boldsymbol{\pi} \boldsymbol{\tau}_{t+1}$
    normalize
        $\widetilde{\boldsymbol{\zeta}}_t \leftarrow \boldsymbol{\zeta}_t / (\mathbf{1}^T \boldsymbol{\zeta}_t)$
end for
$\boldsymbol{\tau}_1 \leftarrow \widetilde{\mathbf{u}}_1 \circ \widetilde{\boldsymbol{\zeta}}_1$

sample first time point
    $\mathbf{q}_1 \leftarrow \boldsymbol{\pi}_0 \circ \boldsymbol{\tau}_1$
    $\widetilde{\mathbf{q}}_1 \leftarrow \mathbf{q}_1 / (\mathbf{1}^{\mathrm{T}} \mathbf{q})$
    $z_1 \sim \widetilde{\mathbf{q}}_1$

sample other time points
for $t = 2, \ldots, T$ **do**
    $\mathbf{q}_t \leftarrow \boldsymbol{\pi}_{z_{t-1}} \circ \boldsymbol{\tau}_t$
    $\widetilde{\mathbf{q}}_t \leftarrow \mathbf{q}_t / (\mathbf{1}^{\mathrm{T}} \mathbf{q}_t)$
    $z_t \sim \widetilde{\mathbf{q}}_t$
end for

except the last term now depends on the previously sampled value z_{t-1} and its corresponding transition distribution $p(z_t \mid z_{t-1}) = \text{Multi}(\boldsymbol{\pi}_{z_{t-1}})$. In this work, we use the second of these two sampling schemes, *block sampling* the entire state sequence $z_{1:T}$, since it allows for explicit dependence on the previous state indicator and thus in our experience often produces more intuitive full state sequences. Algorithm 2 gives an explicit, numerically stable method for this blocked sampler of $z_{1:T}$.

Emission distribution The emission distribution $F(\cdot)$ defines the type of model used for each component. As with standard mixture models, the choice of $F(\cdot)$ will depend on the data being modeled, though normal and multinomial distributions are common choices. As before, we often use a conjugate prior H to achieve a straightforward posterior for each ϕ_k.

In the standard HMM, observations y_t and y_{t-1} depend on each other only through

their states z_t and z_{t-1}, respectively. Given these states, y_t and y_{t-1} are conditionally independent. In some cases, however, this conditional independence may not be appropriate. Autoregressive (AR) HMMs are one such example where each observation y_t, which for now we assume is scalar, linearly depends on a certain (usually small) number r of observations before it through the AR coefficients $\mathbf{a} = (a_1, \ldots, a_r)$,

$$\begin{aligned} y_t =& \sim \sum_{i=1}^{r} a_r y_{t-i} + \epsilon_t, \\ & \epsilon_t \sim \mathcal{N}(0, \sigma^2), \end{aligned} \tag{2.63}$$

where ϵ_t describes the *innovation* at time t generated by a zero-mean normal with variance σ^2. This formulation, termed an AR(r) process, is also equivalent to a normal model with mean $\mathbf{a}^{\mathrm{T}}\widetilde{\mathbf{y}}_t$, where $\widetilde{\mathbf{y}}_t = (y_{t-1}, \ldots, y_{t-r})$,

$$y_t \mid \widetilde{\mathbf{y}}_t \sim \mathcal{N}(\mathbf{a}^{\mathrm{T}}\widetilde{\mathbf{y}}_t, \sigma^2). \tag{2.64}$$

Fox et al. [40] describe this model for the model general case *vector* autoregressive process (VAR(r)) where each observation $\mathbf{y}_t$ is in $\mathbb{R}^d$. The posterior of the AR(r) parameters for each state k factorizes into two components,

$$p\left(\mathbf{a}_k, \sigma_k^2 \mid \{(y_t, \widetilde{\mathbf{y}}_t)\}_{z_t=k}\right) = p\left(\mathbf{a}_k \mid \sigma_k^2, \{(y_t, \widetilde{\mathbf{y}}_t)\}_{z_t=k}\right) p\left(\sigma_k^2 \mid \{(y_t, \widetilde{\mathbf{y}}_t)\}_{z_t=k}\right). \tag{2.65}$$

We place a conjugate normal inverse-gamma ($\mathcal{N}$-IG) prior on each of our K sets of AR(r) parameters, $(\mathbf{a}_k, \sigma_k^2)$,

$$\begin{aligned} \sigma_k^2 &\sim \mathrm{IG}(\alpha_0, \beta_0), \\ \mathbf{a}_k &\sim \mathcal{N}(\boldsymbol{\mu}_0, \Sigma_0), \end{aligned} \tag{2.66}$$

resulting in the posteriors (see Fox et al. [40] for derivations for the general VAR(r) case)

$$\begin{aligned} p\left(\sigma_k^2 \mid \{(y_t, \widetilde{\mathbf{y}}_t)\}_{z_t=k}\right) &\propto \mathrm{IG}\left(\alpha_k, S_{y|\widetilde{y}}^{(k)}\right), \\ p\left(\mathbf{a}_j \mid \sigma_k^2, \{(y_t, \widetilde{\mathbf{y}}_t)\}_{z_t=k}\right) &\propto \mathcal{N}\left(S_{y\widetilde{y}}^{(k)} S_{\widetilde{y}\widetilde{y}}^{-1(k)}, \sigma_k^2 S_{\widetilde{y}\widetilde{y}}^{(k)}\right), \end{aligned} \tag{2.67}$$

with parameters

$$\begin{aligned} \alpha_k &= n_0 + |\{t \mid z_t = k\}| \\ S_{\widetilde{y}\widetilde{y}}^{(k)} &= \Sigma_0^{-1} + \sum_{t|z_t=k} \widetilde{\mathbf{y}}_t \widetilde{\mathbf{y}}_t^{\mathrm{T}}, \\ S_{y\widetilde{y}}^{(k)} &= \Sigma_0^{-1}\boldsymbol{\mu}_0 + \sum_{t|z_t=k} y_t \widetilde{\mathbf{y}}_t, \\ S_{yy}^{(k)} &= \boldsymbol{\mu}_0^{\mathrm{T}}\Sigma_0^{-1}\boldsymbol{\mu}_0 + \sum_{t|z_t=k} y_t^2, \\ S_{y|\widetilde{y}}^{(k)} &= \beta_0 + S_{yy}^{(k)} - S_{y\widetilde{y}}^{(k)} S_{\widetilde{y}\widetilde{y}}^{-1(k)} S_{y\widetilde{y}}^{\mathrm{T}(k)}. \end{aligned} \tag{2.68}$$

2.5.3 Determining the Number of Components

When working with mixtures of a finite number of components K, determining the appropriate K to use is often of interest. An approach suggested by Tibshirani et al. [123] involves evaluating a *gap statistic* at each $K = 1, 2, \ldots$ until reaches an "elbow," denoting an point of stability in K. This approach does not fit within the Bayesian framework, however, and is also sensitive to the reference distribution used to generate synthetic non-clustered observations. In our experience, determining the appropriate reference distribution is not always straightforward.

An approach suggested by Gelman et al. [47] involves starting with a small K and monitoring the posterior predictive distribution of a test quantity independent of the sufficient statistics used for the model component parameters as K is increased. Another approach suggested by Gelman et al. [47] involves treating K as a hierarchical model parameter and averaging inferences about the observations over the various posterior values of K.

In the next section, we explore Bayesian nonparametric methods that handle this problem in a more elegant way by considering $K = \infty$, though only a finite number of these components will contain any observations (since the observed data is finite).

2.6 Bayesian Nonparametric Models

All of the models discussed above involve finite parameterizations. For example, with Gaussian observations we explicitly specify a mean and covariance; for Gaussian mixture modeling, we explicitly specify K sets of means and covariances. This approach is sometimes well-founded and often practical, but some problems require models with more flexibility. So-called "nonparametric" models—while not truly without parameters—involve fewer assumptions than their parametric equivalents. The models use a potentially infinite number of parameters and thus use priors over these infinite-dimensional parameter spaces. Despite this seemingly abstract formulation, a number of efficient posterior inference schemes exist for these Bayesian nonparametric models. These models, have received a great deal of attention in recent years and have been applied to vast number of problems, including document modeling [cf. 15, 120], gene expression [cf. 9, 64, 102], financial volatility [cf. 43, 53], motion tracking [cf. 116], and speaker diarization [42, cf.], to name but a few.

It is well beyond the scope of this work to recapitulate the vast and growing field of Bayesian nonparametrics. We refer interested readers to a number of helpful tutorials [48, 66, 72, 79, 84, 118, 119, 129] and ourselves often follow the explications given in Fox [38], Sudderth [116]. In the sections below, we address a number of models relevant to our work in subsequent chapters.

2.6.1 Dirichlet Processes

The Dirichlet process, first described by Ferguson [37], can be thought of as a distribution over distributions and plays a role in much of the Bayesian nonparametric literature. Consider a measurable space Θ with a finite partition $A_1, \ldots, A_K$ of that space, where

$\cup_{k=1}^{K} A_k = \Theta$ and $A_j \cap A_k = \emptyset$ for $j \neq k$. Given a base measure H on Θ and a scalar concentration parameter γ, the probability distribution G is a sample from a Dirichlet process $\text{DP}(\gamma, H)$ if it follows a K-dimensional Dirichlet distribution on the partition $A_1, \ldots, A_K$,

$$(G(A_1), \ldots, G(A_K)) \sim \text{Dir}(\gamma H(A_1), \ldots, \gamma H(A_K)). \tag{2.69}$$

Using the expected value of the Dirichlet distribution and Equation (2.69), we see that the expected measure of G over partition M is

$$\mathbb{E}[G(M)] = H(M). \tag{2.70}$$

The base distribution is thus analogous to the mean of the Dirichlet process. The concentration parameter γ can be thought of as a precision parameter, where larger values yield less variation in G from the mean H.

Conjugate posterior Consider an observation $\theta \sim G$ from a Dirichlet process sample $G \sim \text{DP}(\gamma, H)$. This sample falls within one of the K partitions of the space Θ, i.e., $\theta \in A_k$ for some $j \in \{1, \ldots, K\}$. The conjugacy of the Dirichlet distribution to the multinomial implies that the posterior distribution is itself a Dirichlet

$$p(G(A_1), \ldots, G(A_K)) \mid \theta, \gamma, H) \propto \text{Dir}(\gamma H(A_1), \ldots, \gamma H(A_k) + 1, \ldots, \gamma H(A_K)). \tag{2.71}$$

Expanding our data to N observations $\{\theta_1, \ldots, \theta_N\}$, we have a posterior of the form [37]

$$p(G \mid \theta_1, \ldots, \theta_N, \gamma, H) \propto \text{DP}\left(\gamma + N, \frac{1}{\gamma + N}\left(\gamma H + \sum_{i=1}^{N} \delta_{\theta_i}\right)\right), \tag{2.72}$$

where δ_{θ_i} denotes a Dirac unit point mass (also called an atom) at point θ_i in Θ. From Equations (2.70) and (2.72), we see that the expected value of G over a particular region $M \subset \Theta$ is

$$\mathbb{E}[G(M) \mid \theta_1, \ldots, \theta_N, \gamma, H] = \frac{1}{\gamma + N}\left(\gamma H(M) + \sum_{i \mid \theta_i \in M} \delta_{\theta_i}\right). \tag{2.73}$$

Instead of working with the actual observations $\{\theta_i\}_{i=1}^{N}$, we can alternately consider the set of K unique values $\{\phi_k\}_{k=1}^{K}$ present in $\{\theta_i\}_{i=1}^{N}$ (assuming finite γ) and the corresponding empirical frequencies $\{n_k\}_{k=1}^{K}$ of each value ϕ_k,

$$\mathbb{E}[G(M) \mid \theta_1, \ldots, \theta_N, \gamma, H] = \frac{1}{\gamma + N}\left(\gamma H(M) + \sum_{k \mid \phi_k \in M} n_k \delta_{\phi_k}\right). \tag{2.74}$$

As $N \to \infty$, we have $K \to \infty$. Since this expected value is constructed from point masses, it would appear that our DP posterior is a discrete measure with total unit probability. Sethuraman [108] proves this to be the case, leading to an explicit construction of a sample from a DP.

Stick-breaking representation Since a potentially infinite number of values ϕ_k are possible in the DP sample G, we consider a probability mass function $\{\pi_k\}_{k=1}^{\infty}$ over the space of positive integers, where each element k is defined as

$$\begin{aligned} \pi'_k &\sim \text{Beta}(1,\gamma), && k = 1,2,\ldots, \\ \pi_k &= \pi'_k \prod_{\ell=1}^{k-1}(1-\pi'_k), && k = 1,2,\ldots, \\ &= \pi'_k \left(1 - \prod_{\ell=1}^{k-1}\pi_k\right) && k = 1,2,\ldots. \end{aligned} \tag{2.75}$$

This measure is sometimes written $\boldsymbol{\pi} \sim \text{GEM}(\gamma)$ for Griffiths, Engen, and McCloskey [88]. Sethuraman [108] proves that this measure $\boldsymbol{\pi}$ and a corresponding set of infinite samples from H yield an explicit *stick-breaking* construction for a DP sample G,

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}, \qquad \phi_k \sim H \qquad k = 1,2,\ldots. \tag{2.76}$$

The concentration parameter γ controls the mass distribution across the atoms $\{\phi_k\}_{k=1}^{\infty}$. When γ is small, most of the mass is concentrated in the first few atoms, and as γ increases, the mass concentration becomes more and more uniform over the atoms. We can think of this stick-breaking sampling scheme as a special case of the $\pi'_k \sim \text{Beta}(a_k, b_k)$ setting where $a_k = 1$ and $b_k = \gamma$. The related Poisson-Dirichlet process (also known as the Pitman-Yor process) [89] is a two-parameter (a, b) process involving $a_k = a$ and $b_k = b + ka$ that yields heavier-tailed distributions.

Pólya urn predictive distribution Given N observations $\{\theta_i\}_{i=1}^{N}$ from a DP sample G, we consider the predictive distribution for a new sample θ_+ achieved by integrating out the distribution G. Since the N observations take on $K \leq N$ discrete values $\{\phi_k\}_{k=1}^{K}$, Blackwell and MacQueen [13] show that this predictive distribution is

$$p(\theta_+ \mid \theta_1, \ldots, \theta_N, \gamma, H) = \frac{1}{\gamma + N}\left(\gamma H + \sum_{k=1}^{K} n_k \delta_{\phi_k}\right), \tag{2.77}$$

where n_k denotes the number of the N observations with value ϕ_k. This predictive representation allows us to sample a new observation θ_+ without having to explicitly instantiate the DP sample G.

The generative process for a new observation θ_+ can be likened to a Pólya urn model. Consider an urn with N balls, each representing a previous observation, and each colored corresponding to the value ϕ_k that θ_i takes. After each draw from the urn, we replace that ball and add another of the same color. A "new color" ball is drawn with probability proportional to γ normal balls.

Chinese restaurant process Using the predictive distribution in Equation (2.77) for θ_+, we can also consider the predictive distribution of the indicator z_+ given the previous indicators $\{z_i\}_{i=1}^N$, where $\theta_+ = \phi_{z_+}$,

$$p(z_+ \mid z_1, \ldots, z_N, \gamma) = \frac{1}{\gamma + N}\left(\gamma\delta_{K+1} + \sum_{k=1}^{K} n_k \delta_k\right). \tag{2.78}$$

Briefly, the Chinese restaurant process metaphor considers a new customer (θ_+) walking into a Chinese restaurant and choosing one of an infinite number of tables to sit at proportional to the number of people already sitting at each table. Each table k serves a unique a dish ϕ_k. Proportional to γ, the customer selects a new (empty) table at which to sit.

Number of observed values Given N multinomial draws $\{z_1, \ldots, z_N\}$ from the stick-breaking weights $\boldsymbol{\pi} \sim \mathrm{GEM}(\gamma)$, Antoniak [7] showed that the distribution over the number of unique values K in $\{z_1, \ldots, z_N\}$ is

$$p(K \mid N, \gamma) = \frac{\Gamma(\gamma)}{\Gamma(\gamma + N)} s(N, K)\gamma^K, \tag{2.79}$$

where $s(n, m)$ denotes the unsigned Stirling numbers of the first kind. See Stepleton [115] for another proof. Antoniak [7] also notes that this distribution has an expected value of roughly

$$\mathbb{E}[K \mid N, \gamma] \approx \gamma \log\left(\frac{\gamma + N}{\gamma}\right). \tag{2.80}$$

Dirichlet process mixture models Consider again the N observations $\{\theta_i\}_{i=1}^N$ drawn from a Dirichlet process sample G. Recall that since G is discrete with total probability mass one, the samples $\{\theta_i\}_{i=1}^N$ take on one of the $K \leq N$ distinct values $\{\phi_k\}_{k=1}^K$. This formulation entails that the posterior for ϕ_k is *only* influenced by the values $\{\theta_i\}_{i|z_i=j}$ that take value ϕ_k. Thus, a observations with value $\phi_{k'} \neq \phi_k$ but still very close to ϕ_k have just as little influence on the posterior of ϕ_k (i.e., no influence at all) as those with values very far from ϕ_k in Θ. While this property may be useful in some applications, in many it is too restrictive.

The Dirichlet process is quite useful, however, as a prior over a mixture model [7, 34]. In this setting, we assume that an observation x_i is a sample from a distribution $F(\cdot)$ (usually of the exponential family) with parameters θ_i sampled from DP sample G. This hierarchical model can be written as

$$G \sim \mathrm{DP}(\gamma, H), \qquad \theta_i \sim G, \qquad x_i \sim F(\theta_i), \qquad i = 1, \ldots, N. \tag{2.81}$$

Following the Chinese restaurant process described in the previous section, and equivalent indicator variable formulation of this mixture model is

$$\begin{aligned} \boldsymbol{\pi} &\sim \mathrm{GEM}(\gamma), \\ \phi_k &\sim H, & k &= 1, 2, \ldots, \\ z_i &\sim \boldsymbol{\pi}, & i &= 1, \ldots, N, \\ x_i &\sim F(\phi_{z_i}), & i &= 1, \ldots, N. \end{aligned} \tag{2.82}$$

Integrating over the infinite number of clusters yields the density of observation x_i,

$$p\left(x_i \mid \boldsymbol{\pi}, \{\phi_k\}_{k=1}^{\infty}\right) = \sum_{k=1}^{\infty} \pi_k f(x_i \mid \phi_k). \tag{2.83}$$

The stick-breaking prior on $\boldsymbol{\pi}$ controls the complexity of the model, favoring a small number of clusters to describe the observations. As more observations are acquired, more clusters are added.

Finite approximations It is sometimes convenient to consider only a finite number of mixture model components as an approximation to the infinite component model given in Equation (2.82). One such approximation involves using an L-dimensional symmetric Dirichlet prior on $\boldsymbol{\pi}$,

$$\begin{aligned} \boldsymbol{\pi} &\sim \text{Dir}(\gamma/L, \ldots, \gamma/L), \\ G^L &= \sum_{k=1}^{L} \pi_k \delta_{\phi_k}. \end{aligned} \tag{2.84}$$

Ishwaran and Zarepour [61, 62] prove that this *weak-limit* approximation G^L approaches a true DP sample G as $L \to \infty$. A related *truncated* approximation involves simply stopping the stick-breaking sampling procedure after L components. Both of these finite approximations encourage a small number of clusters with new clusters available up until L total clusters.

■ 2.6.2 Hierarchical Dirichlet Processes

Consider a dataset in which observations come from J known groups. For example, these observations could be student test scores with teacher j. We could model these scores using the DP mixture model described in the previous section, but this approach would not make use of the additional information we have available, namely the J different teachers. Perhaps we believe that different teachers are likely to have slightly different distributions of test scores, given that some may be better, more effective teachers than others. Accounting for these different groups motivates Teh et al. [120] to introduce the hierarchical Dirichlet process (HDP).

The HDP involves two levels of DP samples: a global measure $G_0 \sim \text{DP}(\gamma, H)$ and measures $G_j \sim \text{DP}(\alpha, G_0)$ corresponding to each group j of the observations in the dataset,

$$\begin{aligned} G_0 &\sim \text{DP}(\gamma, H), \\ G_j &\sim \text{DP}(\alpha, G_0), && j = 1, \ldots, J \\ \theta_{ji} &\sim G, && j = 1, \ldots, J, \quad i = 1, \ldots, N. \end{aligned} \tag{2.85}$$

Recall that G_0 is a discrete measure with support only at points $\{\phi_k\}_{k=1}^{\infty}$. Under the stick-breaking representation, we can equivalently think of G_j's weights $\boldsymbol{\pi}_j = \{\pi_{j,k}\}_{k=1}^{\infty}$ as a

sample from a Dirichlet process over $\mathbb{Z}_+$ with concentration parameter α and base measure $\boldsymbol{\beta}$. Each child DP sample G_j thus contains the same atoms as G_0 though with different weights,

$$\begin{aligned} G_0 &= \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}, & \boldsymbol{\beta} &\sim \text{GEM}(\gamma) \\ G_j &= \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k}, & \boldsymbol{\pi}_j &\sim \text{DP}(\alpha, \boldsymbol{\beta}), & j &= 1, \ldots, J, \\ \phi_k &\sim H & & & k &= 1, 2, \ldots. \end{aligned} \tag{2.86}$$

A sample $\theta_{ji} \sim G_j$ takes value ϕ_k with probability π_{jk}. Note that the measure G_j is conditionally independent of $G_{j' \neq j}$ given G_0.

Chinese restaurant franchise Teh et al. [120] also provide an alternate formulation that extends the Chinese restaurant process metaphor to what they call the Chinese restaurant franchise, where we imagine a set of J restaurants with the same menu. In this setting, two indicator variables are used: t_{ji} denotes the table customer i is assigned to in restaurant j, and k_{jt} denotes the dish being served by table t in restaurant j. This model is given by

$$\begin{aligned} \boldsymbol{\beta} &\sim \text{GEM}(\gamma), & & & \\ \widetilde{\boldsymbol{\pi}}_j &\sim \text{GEM}(\alpha), & j &= 1, \ldots, J & \\ k_{jt} &\sim \boldsymbol{\beta}, & j &= 1, \ldots, J, & t &= 1, 2, \ldots \\ t_{ji} &\sim \widetilde{\boldsymbol{\pi}}_j & j &= 1, \ldots, J, & i &= 1, \ldots, N. \end{aligned} \tag{2.87}$$

where $\theta_{ji} = \phi_{k_{jt_{ji}}}$. Note that $\widetilde{\boldsymbol{\pi}}_j$ is a distribution over tables, whereas $\boldsymbol{\pi}_j$ in Equation (2.86) is a distribution over dishes with $\pi_{j,k} = \sum_{t|k_{jt}=k} \widetilde{\pi}_{j,t}$. Marginalizing over $\boldsymbol{\beta}$ and $\widetilde{\boldsymbol{\pi}}_j$ yields the predictive distributions for a new customer's assigned table t_{j+} and a new table's selected dish k_{j+} in restaurant j,

$$p\left(t_{j+} \mid t_{j1}, \ldots, t_{jN_j}\right) \propto \alpha \delta_{T_j+1} + \sum_{t=1}^{T_j} n_{jt} \delta_t \tag{2.88}$$

$$p\left(k_{j+} \mid \{\{k_{jt}\}_{t=1}^{T_j}\}_{j=1}^{J}\right) \propto \gamma \delta_{K+1} + \sum_{k=1}^{K} m_{\cdot k} \delta_k, \tag{2.89}$$

where n_{jt} denotes the number of customers at table t in restaurant j, T_j the number of occupied tables in restaurant j, $N_j = \sum_{t=1}^{T_j} n_{jt}$ the total number of customers in restaurant j, m_{jk} the number of tables serving dish k in restaurant j, and $m_{\cdot k}$ the number of tables serving dish k across all the J restaurants.

Hierarchical Dirichlet process mixture models The HDP provides similar infrastructure as the DP for mixture modeling. In this scenario, each sample θ_{ji} represents a set of parameters

of the observation distribution $F(\cdot)$ for x_{ji}. As before, the discrete nature of G_j induces a clustering over the observations in group j since each θ_{ji} takes some value ϕ_k. The atoms $\{\phi_k\}_{k=1}^{\infty}$ are shared between the J groups, but each group retains its own set of mixture weights $\boldsymbol{\pi}_j$. The indicator variable formulation of this model is

$$\begin{aligned} \boldsymbol{\beta} &\sim \text{GEM}(\gamma), \\ \boldsymbol{\pi}_j &\sim \text{DP}(\alpha, \boldsymbol{\beta}), && j = 1, \ldots, J \\ \phi_k &\sim H, && k = 1, 2, \ldots \\ z_{ji} &\sim \boldsymbol{\pi}_j, && j = 1, \ldots, J, \quad i = 1, \ldots, N \\ x_{ji} &\sim F(\phi_{z_{ji}}), && j = 1, \ldots, J, \quad i = 1, \ldots, N. \end{aligned} \tag{2.90}$$

■ 2.6.3 Nested Dirichlet Processes

The nested Dirichlet process (NDP) of Rodríguez et al. [103] slightly modifies the HDP framework for mixture modeling to allow for clustering on both the observations and the group levels. Consider a discrete measure G_ℓ^* over Θ, where

$$G_\ell^* = \sum_{k=1}^{\infty} w_{\ell,k}^* \delta_{\boldsymbol{\phi}_{\ell k}^*}, \qquad \mathbf{w}_\ell^* \sim \text{GEM}(\alpha), \qquad \boldsymbol{\phi}_{\ell k}^* \sim H. \tag{2.91}$$

Note that both the parameter atoms $(\delta_{\boldsymbol{\phi}_{\ell k}^*})_{k=1}^{\infty}$ and the associated weights $\mathbf{w}_\ell$ are unique for each G_ℓ^*. Using these G_ℓ^* measures, we construct another discrete measure Q with probability mass π_ℓ^* at G_ℓ^*,

$$Q = \sum_{\ell=1}^{\infty} \pi_\ell^* \delta_{G_\ell^*}, \qquad \boldsymbol{\pi}^* \sim \text{GEM}(\gamma). \tag{2.92}$$

This formulation implies a mixture of mixtures. Thus, we can think of group j selecting a particular mixture $z_j = \ell$ defined by G_ℓ^*, which is itself a mixture of unique parameter atoms $(\delta_{\boldsymbol{\phi}_{\ell k}^*})_{k=1}^{\infty}$. Given the mixture indicator z_j, group j's observations are modeled exactly as in a DP mixture for the given weights $\mathbf{w}_\ell^*$ and parameter atoms $(\delta_{\boldsymbol{\phi}_{\ell k}^*})_{k=1}^{\infty}$.

It is worthwhile to point out a few important differences between the NDP and the HDP. First, the measures $(G_\ell^*)_{\ell=1}^{\infty}$ of the NDP contain both unique parameter atoms and unique weights, whereas the measures $\{G_j\}_{j=1}^{J}$ of the HDP contain only unique weights, as their parameter atoms are defined by the global measure G_0. Second, in the NDP each group j is assigned to a measure G_ℓ^*, inducing a group-level clustering where zero, one, or more groups may be assigned to each measure G_ℓ^*. This group-level clustering is in contrast to the HDP, where each group i has its own unique measure G_j. Finally, each measure G_ℓ^* in the NDP is truly independent from the others $G_{\ell' \neq \ell}^*$, whereas each measure G_j is only *conditionally* independent from the others $G_{j' \neq j}$ given G_0.

■ 2.6.4 Hierarchical Dirichlet Process Hidden Markov Models

The HDP mixture models discussed in Section 2.6.2 allow for a natural nonparametric extension of the hidden Markov model (HMM) discussed in Section 2.5.2. In the HDP, each

observation x_{ji} has a fixed group j, for example a student's test store for a particular teacher j. In the hierarchical Dirichlet process hidden Markov model (HDP-HMM) described by Teh et al. [120], we let observation x_t's group change from time point to time point through its cluster indicator z_t. Furthermore, we consider an infinite set of groups, i.e., $J = K = \infty$. The $K \times K$ transition distributions matrix given for the finite HMM now becomes a doubly-infinite dimensional matrix.

$$\begin{aligned} \boldsymbol{\beta} &\sim \text{GEM}(\gamma) \\ \boldsymbol{\pi}_j &\sim \text{DP}(\alpha, \boldsymbol{\beta}) && j = 1, 2, \dots \\ \phi_j &\sim H && j = 1, 2, \dots \\ z_t \mid z_{t-1} &\sim \pi_{z_{t-1}} && t = 1, \dots, T \\ x_t \mid z_t &\sim F(\phi_{z_t}) && t = 1, \dots, T \end{aligned} \tag{2.93}$$

The shared global weights $\boldsymbol{\beta}$ encourage the transition weights $\boldsymbol{\pi}_j$ for each state j to be similar.

As in the HMM discussed in Section 2.5.2, the observation distribution $F(\cdot)$ is arbitrary, though for simplicity it is often of the exponential family. Instead of a single parametric distribution, Fox et al. [42] employ a DP mixture of normals to model the complex densities associated with their speaker diarization task. Fox et al. [42] give two different schemes for HDP-HMM posterior inference. Their direct assignment collapsed Gibbs sampler marginalizes the transition distributions $\{\boldsymbol{\pi}_j\}$ as well as the model parameters $\{\phi_j\}$ and works with the predictive distributions for each. Their blocked Gibbs sampler explicitly samples $\{\boldsymbol{\pi}_j\}$ and $\{\phi_j\}$ and uses the weak-limit approximation to the DP.

Encouraging Self-Transitions The HDP-HMM framework given in Equation 2.93 is agnostic to whether the state z_t is the same or different to that of time z_{t-1}. The flexible state space introduced by the HDP allows for possible over-splitting, where observations with quite similar values may be unrealistically assigned to different states. Such over-splitting can have the effect that a sequence of observations $x_1, \dots, x_t$ that *should* be assigned to the same state (i.e., $z_1 = \cdots = z_t$) are in fact assigned to two or more different states, displaying unreasonably fast transition dynamics between these states. One approach to this problem would be to influence the expected number of states through the prior γ in an attempt to reduce this state redundancy. Another, more explicit solution proposed by Fox et al. [42] involves encouraging state self-transitions through an additional "sticky" parameter κ added to the jth component of each distribution $\boldsymbol{\pi}_j$,

$$\boldsymbol{\pi}_j \sim \text{DP}\left(\alpha + \kappa, \frac{1}{\alpha + \kappa}(\alpha\boldsymbol{\beta} + \kappa\delta_j)\right) \tag{2.94}$$

An optional prior is placed on the value $\alpha + \kappa$ to allow for flexibility in this value κ, if desired. This sticky HDP-HMM produces smoother, more intuitive state assignments more free of the fast transitions associated with over-splitting.

■ 2.6.5 Beta Processes

Consider the situation where we have N groups of data (e.g, N separate time series) and K features (or states) used to model our data. It may make sense to restrict the features available to each group, so that the observations in group i can only be assigned to states 1, 2, and 5 but not states 3 and 4, for example. We can represent the availability of feature k to group i using an indicator $f_k^{(i)} = 1$ if it is available or 0 if it is not. The (row) vector of feature availability indicators is thus $\mathbf{f}^{(i)}$, which form a binary matrix F with N rows and K columns when stacked together.

Unless the value of the indicator matrix F is known *a priori*, we would like the ability to model it and thus infer it from the data. The approach to this problem taken by Griffiths and Ghahramani [54] assumes that each feature k is activated with probability ω_k, consistent across the N groups. This formulation implies $f_k^{(i)} \sim \text{Ber}(\omega_k)$. The probability of a given matrix F is thus given by

$$\begin{aligned} p(F \mid \boldsymbol{\omega}) &= \prod_{k=1}^{K} \prod_{i=1}^{N} p(f_k^{(i)} \mid \omega_k), \\ &= \prod_{k=1}^{K} \omega_k^{m_k} (1-\omega_k)^{N-m_k}, \end{aligned} \tag{2.95}$$

where m_k denotes the marginal count (over the N rows) of the groups using feature k. Assuming a conjugate beta prior with parameters $(\alpha/K, 1)$ on each ω_k, we have the generative model

$$\begin{aligned} \omega_k &\sim \text{Beta}(\alpha/K, 1), && k = 1, \ldots, K, \\ f_k^{(i)} &\sim \text{Ber}(\omega_k), && k = 1, \ldots, K, \quad i = 1, \ldots, N. \end{aligned} \tag{2.96}$$

Integrating over the feature probabilities $\boldsymbol{\omega}$ yields the marginal density for F,

$$p(F) = \prod_{k=1}^{K} \frac{\frac{\alpha}{K}\Gamma(m_k + \frac{\alpha}{K})\Gamma(N - m_k + 1)}{\Gamma(N + 1 + \frac{\alpha}{K})}. \tag{2.97}$$

Assuming that the order of the K features is arbitrary, Griffiths and Ghahramani [54] define a set of equivalence classes of F via a left-ordered function (lof) that reorders the columns of F by sorting them ascending (left to right) as if each column represents a binary number, with the first row the most significant bit. Two matrices F and G are said to be lof-equivalent if $lof(F) = lof(G)$. We define the set of matrices, denoted $\{F\}$, as the lof-equivalence class whose members are all the binary matrices lof-equivalent to F. The number of possible values for each column of F is 2^{N-1}. Let K_h for $h = 1, \ldots, 2^{N-1}$ denote the number of columns in F whose decimal equivalent equals h. The cardinality of this set

$\{F\}$ is defined through the multinomial coefficient

$$\begin{aligned} |\{F\}| &= \binom{K}{K_0, \ldots, K_{2^N-1}}, \\ &= \frac{K!}{\prod_{h=1}^{2^N-1} K_h!}. \end{aligned} \tag{2.98}$$

We now can give the probability of the lof-equivalence class of binary matrices $\{F\}$ using Equations (2.97) and (2.98),

$$\begin{aligned} p(\{F\}) &= \sum_{F \in \{F\}} p(F), \\ &= \frac{K!}{\prod_{h=1}^{2^N-1} K_h!} \prod_{k=1}^{K} \frac{\frac{\alpha}{K}\Gamma(m_k + \frac{\alpha}{K})\Gamma(N - m_k + 1)}{\Gamma(N + 1 + \frac{\alpha}{K})}. \end{aligned} \tag{2.99}$$

Indian buffet processes In the nonparametric setting, we are interested in the limit of Equation 2.99 as $K \to \infty$. To do this Griffiths and Ghahramani [54] reorder the first K_+ columns of F such that $m_k > 0$ for $k \leq K^+$ and the rest where $m_k = 0$, allowing them to derive

$$p(\{F\}) = \frac{\alpha^{K_+}}{\prod_{h=1}^{2^N-1} K_h!} \exp\left(-\alpha \sum_{i=1}^{N} \frac{1}{i}\right) \prod_{k=1}^{K_+} \frac{(N - m_k)!(m_k - 1)!}{N!}. \tag{2.100}$$

In the vein of the Chinese restaurant process metaphor, the Indian buffet process (IBP) yields a binary matrix F with the predictive distribution given in Equation 2.100. Consider an Indian restaurant with a lunch buffet with an infinite number of dishes. A new customer enters the restaurant with N current customers. For the K_+ dishes already sampled by the N current customers, the new customer samples each dish k with probability m_k/N. He then also samples Poisson(α/N) new, untried dishes. The lof-equivalence class of matrices $\{F\}$ generated by this IBP has the probability density given in Equation 2.100.

Despite an infinite number of potential dishes, exchangeability entails that the number of dishes chosen by each customer follows a Poisson(α) distribution (since the first customer selects Poisson(α) dishes). This result guarantees that for finite α, the expected value of the number of features active in $\mathbf{f}^{(i)}$ is also finite.

A random measure for the Indian buffet process Thibaux and Jordan [122] show that the *beta process* is the random measure whose predictive distribution yields the IBP. The beta process is an instance of a *completely random measure* [68], entailing that the random measures on disjoint sets of the probability space are mutually independent. Completely random measures can be constructed via a nonhomogenous Poisson process with rate measure η. For a probability space $\Theta \times [0, 1]$, we can define a completely random measure B via the infinite sum of draws (ϕ_k, ω_k),

$$B = \sum_{k=1}^{\infty} \omega_k \delta_{\phi_k}. \tag{2.101}$$

Note that this formulation shows that completely random measures are discrete. For a finite continuous base measure B_0[3] with total mass $B_0(\Theta) = \alpha$ and concentration parameter $c > 0$, we can define a Lévy measure ν over $\Theta \times [0, 1]$ to use as our rate η for the nonhomogenous Poisson process,

$$\nu(d\omega, d\phi) = c\omega^{-1}(1-\omega)^{c-1}d\omega B_0(d\phi) \tag{2.102}$$

Note that this rate is an improper beta distribution over $\Theta \times [0, 1]$. For a discrete base measure B_0 with mass $q_k \in (0, 1)$ at each atom, a sample B necessarily contains each atom ϕ_k with weight ω_k drawn from a beta distribution,

$$\omega_k \sim \text{Beta}(cq_k, c(1-q_k)). \tag{2.103}$$

The sample $B \sim \text{BP}(c, B_0)$ is thus termed a sample from a beta process with concentration parameter $c > 0$ and base measure B_0.

Bernoulli processes Thibaux and Jordan [122] formalize the problem of sampling group i's indicator for feature k, $f_k^{(i)}$, as a *Bernoulli process* and show that the beta process is conjugate to it. The Bernoulli process is parameterized by a base measure B, which can either be continuous or discrete, or the sum of both.

For a discrete base measure B with atoms in $\Theta \times [0, 1]$, which we can also think of as an infinite collection of point masses $\{\delta_{\phi_k}\}$ in Θ, each with probability ω_k, a sample $X^{(i)} \sim \text{BeP}(B)$ from a Bernoulli process contains a subset of these atoms, each with unit mass. One can think of sampling $X^{(i)}$ as considering each feature ϕ_k contained in the discrete measure B and flipping a coin with heads probability ω_k to determine whether $X^{(i)}$ contains feature k or not. The results of these indicator values are represented in the indicator vector $\mathbf{f}^{(i)}$,

$$\begin{aligned} f_k^{(i)} &\sim \text{Ber}(\omega_k), \\ X^{(i)} &= \sum_{k=1}^{\infty} f_k^{(i)} \delta_{\phi_k}. \end{aligned} \tag{2.104}$$

For a continuous base measure B, sampling $X^{(i)}$ involves first sampling the number of atoms L to draw from B and then actually sampling those atoms from $B(\Theta)^{-1}B = \frac{1}{\alpha}B$,

$$\begin{aligned} L &\sim \text{Poisson}(\alpha) \\ \phi_\ell &\sim \frac{1}{\alpha}B, \\ X^{(i)} &= \sum_{\ell=1}^{L} \delta_{\phi_\ell} \end{aligned} \tag{2.105}$$

[3] As Teh and Jordan [119] point out, B_0 is not a probability distribution as it does not necessarily integrate to one.

Conjugate posterior The beta process provides our discrete base measure B for the Bernoulli process, implying the generative scheme

$$\begin{aligned} B &\sim \mathrm{BP}(c, B_0) \\ X^{(i)} &\sim \mathrm{BeP}(B), \qquad i = 1, \ldots, N. \end{aligned} \tag{2.106}$$

Since the beta process is conjugate to the Bernoulli process, the posterior over B given the realizations $\{X^{(1)}, \ldots, X^{(N)}\}$ is itself a beta process,

$$\begin{aligned} p\left(B \mid X^{(1)}, \ldots, X^{(N)}\right) &\propto \mathrm{BP}\left(c + N, \frac{1}{c+N}\left(cB_0 + \sum_{i=1}^{N} X^{(i)}\right)\right), \\ &= \mathrm{BP}\left(c + N, \frac{1}{c+N}\left(cB_0 + \sum_{k=1}^{K_+} m_k \delta_{\phi_k}\right)\right). \end{aligned} \tag{2.107}$$

Note the similarity between this beta process posterior and that of the Dirichlet process (see Equation (2.72)). Assuming B_0 is continuous, the posterior in Equation (2.107) contains a continuous part from the $\frac{c}{c+N}B_0$ term and a discrete part from $\frac{1}{c+N}\sum_{k=1}^{K_+} m_k \delta_{\phi_k}$. Recalling that K_+ denotes the number of unique atoms occurring in $\{X^{(1)}, \ldots, X^{(N)}\}$, the discrete part of the base measure necessarily contains the K_+ atoms, each with posterior weight q_k,

$$q_k = \frac{m_k}{c+N}. \tag{2.108}$$

A sample B from this discrete part of the posterior contains the K_+ atoms each with weight ω_k,

$$\omega_k \sim \mathrm{Beta}((c+N)q_k, (c+N)(1-q_k)), \qquad k = 1, \ldots, K_+. \tag{2.109}$$

The continuous part of the base measure derived from the prior B_0 implies sampling L new atoms not appearing in $\{\phi_k\}_{k=1}^{K_+}$,

$$\begin{aligned} L &\sim \mathrm{Poisson}\left(c\alpha/(c+N)\right), \\ (\phi_\ell, \omega_\ell) &\sim cB_0. \end{aligned} \tag{2.110}$$

The sum of these atoms from the discrete and continuous parts of the base measure yields our posterior sample B,

$$B = \sum_{k=1}^{K_+} \omega_k \delta_{\phi_k} + \sum_{\ell=1}^{L} \omega_\ell \delta_{\phi_\ell} \tag{2.111}$$

Thibaux and Jordan [122] showed that setting $c = 1$ and integrating over B yields the posterior predictive distribution—the Indian buffet process—for $X^{(i)}$.

■ 2.6.6 Beta Process Hidden Markov Models

In Section 2.6.4, we discussed an HMM for a time series $x_{1:T}$. Consider the situation of modeling N time series, each of which we would like to model using its own state sequence. In some scenarios, it may be reasonable and appropriate to assume that every time series has all states available to it, but other times we desire a *sparse* state space available for each time series. Especially in situations with a large number of time series and a large state space, the restricted state space for each time series may provide a more intuitive, compact model for each time series. For example, consider a set of highlight clips from a soccer match and the state space of possible activities in a soccer match: dribbling, passing, shooting, throw-in, penalty received, penalty kicked, celebrating, coach yelling, etc. Each of these clips may only contain a few of these activities, so we would ideally like to model it with only those available rather than the full space of possible actions in a soccer match.

As described by Fox et al. [41, 42], The beta process prior on the Bernoulli process allows for this exact type of state space restriction. In this situation, we have an infinite library of features (states) available, and each time series selects a only finite subset to use. This setting allows for sharing of feature information between time series while also allowing each a unique state space. Each time series's state doubly-infinite dimensional transition distribution matrix thus becomes sparse, with a finite number of non-zero entries. We can describe the features available to each time series i as a sample from a Bernoulli process described above,

$$X^{(i)} \sim \text{BeP}(B), \tag{2.112}$$

where the base measure B is itself a sample from a beta process with concentration parameter $c = 1$ and base measure B_0. The features available in $X^{(i)}$ implies vector of indicators $\mathbf{f}^{(i)}$ describing only a finite number of features available to time series i. The posterior predictive distribution of $\mathbf{f}^{(i)}$ is given the indicators of all the other time series is given by the Indian buffet process.

In Equation (2.94), we described state j's transition distribution for the HDP-HMM as a sample from a Dirichlet process with a base measure over $\mathbb{Z}_+$. In constructing a transition distribution now constrained by the active features available to i, we make use of the fact that a sample from a K-dimensional Dirichlet distribution can also be thought of as the normalized values of K independent gamma samples. Specifically, element of the doubly-infinite transition matrix $\boldsymbol{\eta}^{(i)}$ is a gamma sample with shape γ, with an additional κ mass when it represents a self-transition (i.e., $j = k$),

$$\eta_{jk}^{(i)} \sim \text{Gamma}(\gamma + \kappa\delta(j,k), 1). \tag{2.113}$$

The constrained transition distribution $\boldsymbol{\pi}_j^{(i)}$ is defined as before over $\mathbb{Z}_+$ but only has positive values in the finite number of features given in $\mathbf{f}^{(i)}$,

$$\boldsymbol{\pi}_j^{(i)} = \frac{\boldsymbol{\eta}_j \circ \mathbf{f}^{(i)}}{\sum_{k|f_k^{(i)}=1} \eta_{jk}^{(i)}}. \tag{2.114}$$

The full generative specification of this model, termed the beta process hidden Markov model, is thus

$$\begin{aligned}
B &\sim \mathrm{BP}(1, B_0), \\
X^{(i)} &\sim \mathrm{BeP}(B), \\
\boldsymbol{\eta}_j &\sim \mathrm{Dir}(\gamma, \ldots, \gamma + \kappa, \ldots), && j = 1, 2, \ldots, && i = 1, \ldots, N, \\
z_t^{(i)} \mid \boldsymbol{\eta}^{(i)}, \mathbf{f}^{(i)} &\sim \boldsymbol{\pi}_{z_{t-1}^{(i)}}, && t = 1, \ldots, T^{(i)}, && i = 1, \ldots, N, \\
x_t^{(i)} \mid z_t^{(i)} &\sim F(\phi_{z_t^{(i)}}), && t = 1, \ldots, T^{(i)}, && i = 1, \ldots, N.
\end{aligned} \tag{2.115}$$

Chapter 3

Clustering seizures on multiple levels

In reviewing a patient's iEEG, an epileptologist must determine which areas of the brain should be removed to reduce and hopefully eliminate a patient's seizures [33, chaps. 166-169]. Many factors go into this decision. Epileptologists pay particular attention to which channels they believe are involved in the epileptogenic and ictal onset zones and how activity in these channels spreads to all the channels involved in the seizure. Simplistically, the epileptologists manually cluster the channels into different groups that include onset channels, delayed onset channels, and non-involved channels. On a higher level, they look at all the seizures of a patient to understand the different seizure types a patient tends to display, including those with focal onset (a few very specific channels initiate the seizure), diffuse onset (many or all of the channels involved from the start), and many gradations in between. Understanding the full range of seizure types a patient can manifest is vital in determining how removing particular brain tissue may affect a patient's prognosis.

Finally, from a very high level, the epileptologist considers how likely the patient is to benefit from the extremely invasive resective surgery given the outcomes of previous patients with a similar pattern of seizures. For example, patients with exclusively very focal onset seizures tend to have much higher seizure freedom rates than those with more of a variety of seizure etiologies [33, chap. 167].

In current clinical practice, this analysis process is almost entirely manual and varies greatly depending on the center of treatment and individual training of the epileptologists. The sheer magnitude of the iEEG data makes objective, reproducible analysis difficult for even a single physician. We believe statistical models can help reduce some of the uncertainty in this process and provide objective clinical decision support. On some level, one can think of the clinical analysis process described above as a clustering procedure done on a number of levels: the channel level, the seizure level, and the patient level.

Most existing quantitative approaches to seizure modeling focus on individual seizures since generalizing across seizures and especially across patients is so difficult. Many models aim to understand relationships between the channel activities [cf., 8, 26, 56, 85, 86] in a seizure. Other models focus on the problem of seizure onset detection [cf., 24, 69, 71], a well-studied but still very difficult problem in the epilepsy domain.

Models for a single seizure from one patient are not at all analogous to a physician's representation of the same seizure. That physician has—over the course of his or her

training—seen the iEEGs of hundreds or thousands of seizures from many patients. This experience informs the physician's interpretation of the current seizure of interest. We believe that any model hoping to reasonably represent the iEEG of seizures must also integrate information from many seizures over a diverse patient population. A model that clusters a multi-patient dataset of iEEGs from seizures on the channel level, the seizure level, and the patient level would achieve our desired information sharing across seizures and patients.

Hierarchical Bayesian models provide a ready solution for this problem in that each sub-model in the hierarchy is a blend of local information (e.g., the channel activities in a particular seizure) and also global information (e.g., the other seizures of that patient and even the other seizures of other patients). Nonparametric Bayesian methods are also attractive since they reduce the amount of necessary model selection. Bayesian nonparametric methods often build off the Dirichlet process [37], a discrete probability distribution over distributions. In particular, the hierarchical Dirichlet process of Teh et al. [120] and the nested Dirichlet process of Rodríguez et al. [103] provide ways of sharing information across a hierarchy and clustering on multiple levels of a dataset, respectively, and thus help inspire the model developed in this chapter.

We present our model for the seizures of a single patient in Section 3.1 and a larger version for seizures of multiple patients in Section 3.2. We discuss the posterior inference of these models in Section 3.3. In Section 3.4 we compare our model to the nested Dirichlet process using a simulated dataset and then explore its performance on a dataset of 193 seizures from 10 human epilepsy patients in Section 3.5. We further explore properties of the model through a number of sensitivity analyses presented in Section 3.6. Finally, in Section 3.7 we conclude with a brief discussion and suggest some possible avenues for future work.

■ 3.1 Modeling the Seizures of a Single Patient

Consider the seizure data associated with a single epilepsy patient who has had J seizures, each of which is defined by a collection of N_j recordings from individual iEEG channels. We take these channel recordings over each seizure as our model observations, denoting the observation from the ith channel of seizure j by $\mathbf{x}_{ji}$.

We use an infinite collection $(G_\ell^{(1)})_{\ell=1}^{\infty}$ of discrete measures over a parameter space Θ to model the observations,

$$G_\ell^{(1)} = \sum_{k=1}^{\infty} \pi_{\ell k}^{(1)} \delta_{\phi_k}, \qquad \boldsymbol{\phi}_k \sim H, \tag{3.1}$$

where H denotes the prior distribution on the parameters ϕ_k of each mixture component k. The weights $\boldsymbol{\pi}_\ell^{(1)}$ associated with measure $G_\ell^{(1)}$ are a sample from a Dirichlet process over $\mathbb{Z}_+$,

$$\boldsymbol{\pi}_\ell^{(1)} \sim \mathrm{DP}(\alpha^{(1)}, \boldsymbol{\beta}^{(1)}), \qquad \boldsymbol{\beta}^{(1)} \sim \mathrm{GEM}(\gamma^{(1)}). \tag{3.2}$$

We can equivalently think of each $G_\ell^{(1)}$ as a sample from $\mathrm{DP}(\alpha^{(1)}, G_0^{(1)})$, where $G_0^{(1)}$is itself

a sample from $\mathrm{DP}(\gamma^{(1)}, H)$. As in the NDP, we construct a measure $G^{(2)}$ over the infinite collection of measures $(\delta_{G_\ell^{(1)}})_{\ell=1}^{\infty}$ using another set of weights $\boldsymbol{\pi}^{(2)}$,

$$G^{(2)} = \sum_{\ell=1}^{\infty} \pi_\ell^{(2)} \delta_{G_\ell^{(1)}}, \quad \boldsymbol{\pi}^{(2)} \sim \mathrm{DP}(\alpha^{(2)}, \boldsymbol{\beta}^{(2)}), \quad \boldsymbol{\beta}^{(2)} \sim \mathrm{GEM}(\gamma^{(2)}). \tag{3.3}$$

As with the NDP, this formulation implies a mixture of mixtures. From a generative perspective, each seizure j selects a particular mixture $G_\ell^{(1)}$, indicated by $z_j^{(2)} = \ell$, from $G^{(2)}$ to describe its channel observations. The mixture $G_\ell^{(1)}$ can be though of as the seizure-type cluster to which seizure j is assigned. Conditioned on $z_j^{(2)}$, each channel observation $\mathbf{x}_{ji}$ selects a particular observation model k with parameters $\boldsymbol{\phi}_k$, indicated by $z_{ji}^{(1)} = k$. All together, the full model can be written as,

$$\begin{aligned}
\boldsymbol{\beta}^{(2)} &\sim \mathrm{GEM}(\gamma^{(2)}), \\
\boldsymbol{\pi}^{(2)} &\sim \mathrm{DP}(\alpha^{(2)}, \boldsymbol{\beta}^{(2)}), \\
\boldsymbol{\beta}^{(1)} &\sim \mathrm{GEM}(\gamma^{(1)}), \\
\boldsymbol{\pi}_\ell^{(1)} &\sim \mathrm{DP}(\alpha^{(1)}, \boldsymbol{\beta}^{(1)}), && \ell = 1, 2, \ldots \\
\boldsymbol{\phi}_k &\sim H && k = 1, 2, \ldots \\
z_j^{(2)} &\sim \boldsymbol{\pi}^{(2)}, && j = 1, \ldots, J \\
z_{ji}^{(1)} \mid z_j^{(2)} &\sim \boldsymbol{\pi}_{z_j^{(2)}}^{(1)}, && j = 1, \ldots, J, \quad i = 1, \ldots, N_j \\
\mathbf{x}_{ji} \mid z_{ji}^{(1)} &\sim F(\boldsymbol{\phi}_{z_{ji}^{(1)}}), && j = 1, \ldots, J, \quad i = 1, \ldots, N_j.
\end{aligned} \tag{3.4}$$

We call this model the multilevel clustering hierarchical Dirichlet process (MLC-HDP) and depict it graphically in Figure 3.1 (left). Note that each of the two clustering layers involves its own HDP.

Comparison to similar models For alternatives to our proposed model, consider how the collection of seizures within a patient would be modeled using a DP, an HDP, or an NDP. We use the channel activities in each seizure as the observations for each model. Since the DP contains only one level of clustering, it would consider channel observations of all the seizures at once, without distinguishing between channel observations belonging to one seizure or another. While clustering channel activity alone is indeed a relevant enterprise, it is not as well suited for the more high level clinical analysis of determining similarities (and differences) between seizures, especially for patients with many tens of recorded seizures.

In the HDP, the channels of each seizure j would be grouped together and modeled by a measure G_j unique to each seizure. Since the measures $\{G_j\}_{j=1}^{J}$ are samples from a DP with a global base measure G_0, they share parameter atoms $(\delta_{\boldsymbol{\phi}_k})_{k=1}^{\infty}$ though do not share atom weights $\boldsymbol{\pi}_j$, each of which is unique (and conditionally independent given $\boldsymbol{\beta}$) for each

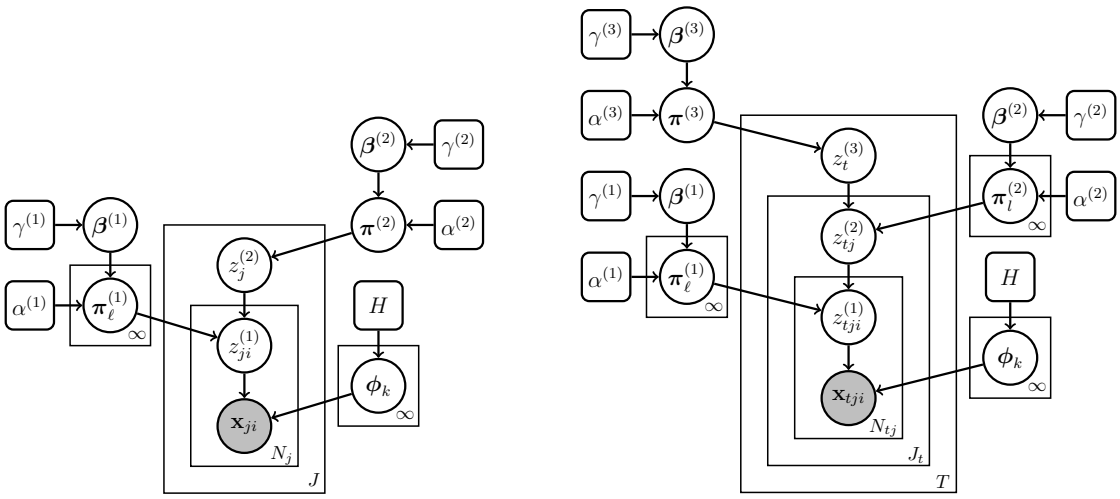

Figure 3.1. (**left**) A graphical model representation of a two-level MLC-HDP. Observed values are given in gray and priors as rounded squares. This two-level model is appropriate for modeling the J seizures of a single epilepsy patient, where each seizure contains N_j channel observations. (**right**) A graphical model representation of a three-level MLC-HDP. This three-level model would be appropriate for modeling the seizures of T different patients, each of whom has J_t seizures, each of which has N_{tj} channel observations.

seizure. Like the DP, the HDP only explicitly clusters the channel observations, requiring downstream analysis if any seizure-specific comparisons are to be made.

Unlike the DP and the HDP, the NDP is indeed a two-level clustering model, where the channel observations are clustered on one level, and the seizures are clustered on another. We believe application of this model is thus more clinically useful than the previous two models because the seizure level cluster indicators $\{z_j^{(2)}\}_{j=1}^J$ can be used to organize the seizures directly without any subsequent analysis. Unfortunately, the NDP's assumption that each seizure-type cluster G_ℓ^* contains its own unique channel-type atoms $(\delta_{\boldsymbol{\phi}_{\ell k}^*})_{k=1}^\infty$ is not realistic for our seizure data. A clinician may deem two seizures to be fundamentally different even if a subset of their channels are behaving quite similarly. We thus would like to share channel-type atoms across seizure type mixtures, something the NDP is not capable of. Furthermore, having unique channel type atoms for each seizure type mixture leads to a large number of total channel-type atoms, making practical computational issues a potential barrier to scaling the model up to a realistic number of seizures (and thus seizure types).

Our MLC-HDP blends the desirable aspects of the HDP and NDP for our seizure iEEG application. It incorporates the multiple levels of clustering introduced by the NDP, but—as in the HDP—the different seizure type mixtures $(G_\ell^{(1)})_{\ell=1}^\infty$ share a common set of channel type atoms $(\delta_{\boldsymbol{\phi}_k})_{k=1}^\infty$, allowing for both across seizure type sharing of channel information and efficient scaling as the number of seizure-types grows. It is important to note that that

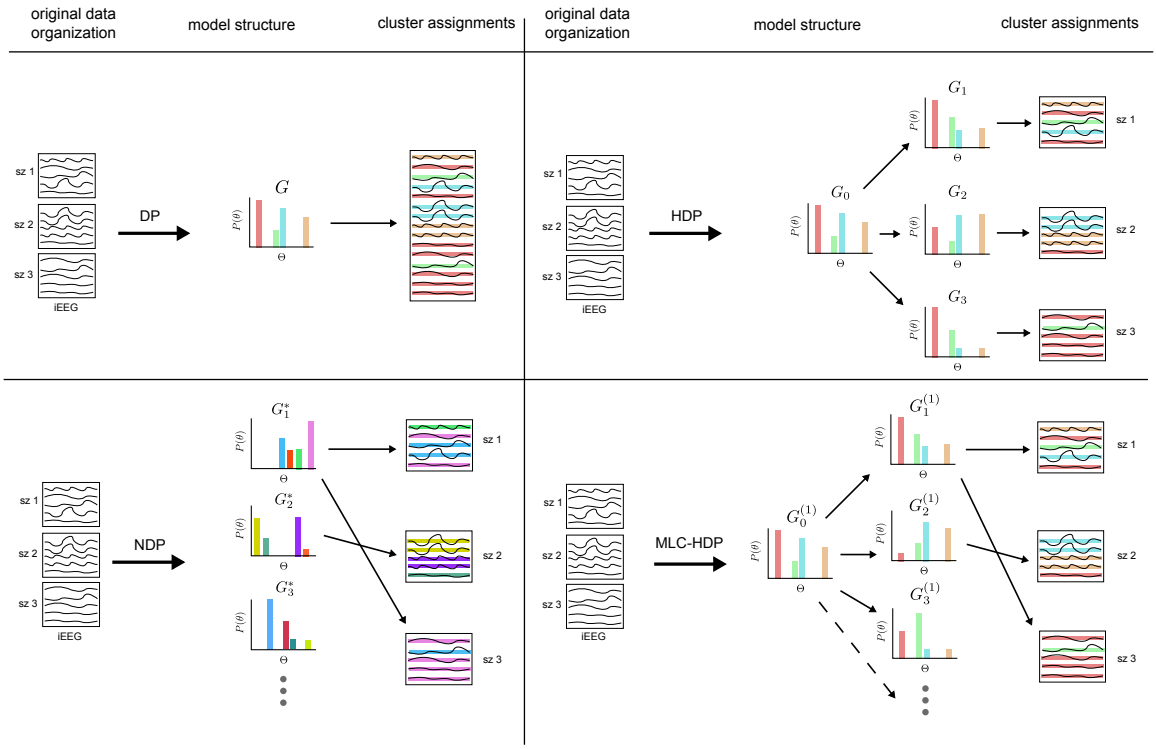

Figure 3.2. A comparison of how four different models—the Dirichlet process (DP), the hierarchical Dirichlet process (HDP), the nested Dirichlet process (NDP), and the multi-level clustering HDP (MLC-HDP)—would be applied to three seizures of a single patient. The DP yields a discrete measure G over the space Θ of channel-type models, producing a clustering over the channels (shown by the color under each channel) that is indiscriminate of the seizure to which each channel belongs. The HDP yields a global discrete measure G_0 and measures $\{G_j\}_{j=1}^3$ for each seizure that share atoms with G_0 but contain unique weights, producing a clustering over channels, which are explicitly grouped by the seizures to which they belong. The NDP yields an infinite collection of independent discrete measures $(G_\ell^*)_{\ell=1}^\infty$ that each contain an infinite collection of unique atoms and weights, yielding a clustering over the seizure types and the channel types. The MLC-HDP also yields an infinite collection of discrete measures $(G_\ell^{(1)})_{\ell=1}^\infty$, but they are DP samples from a global base measure $G_0^{(1)}$ and so share the same infinite collection of atoms while having unique weights. Like the NDP, the MLC-HDP produces a clustering over the seizure types and the channel types.

like the NDP, the MLC-HDP makes the assumption that the channel observations of seizure j are conditionally independent given the seizure type $z_j^{(2)}$. Figure 3.2 depicts the different clustering representations of each of the four models discussed above for a set of seizures from a single patient.

■ 3.2 Modeling the Seizures of Multiple Patients

Instead of modeling only a single epilepsy patient's seizures, as we discussed in the previous section, our real goal is to model a population of T patients, each with J_t seizures and each seizure j of patient t with N_{tj} channel observations. Given the complexity of epilepsy and the unique electrode placement for each patient, any model of seizure activity across patients will necessarily involve a number of simplifying assumptions to make it tractable. First and foremost, we assume that the channels of each seizure and the seizures of each patient are exchangeable. In reality, the spatial relationships between the channels and temporal relationships between the seizures add considerable nuance to this already complex data. Nevertheless, we believe the most important information about a patient's seizures lies in the channel iEEG itself, so it should be the basis for our modeling approach.

Although every patient's seizures are unique, enough similarities exist on the channel and seizure level that we would like to share information between the models for each patient. One approach would hierarchically link our two-level MLC-HDP's observation parameter atoms $(\delta_{\phi_k})_{k=1}^{\infty}$ and weights $\boldsymbol{\pi}_{\ell}^{(1)}$ and the seizure-type weights $\boldsymbol{\pi}^{(2)}$ for each patient. But the clinical aspects of a particular patient's seizures on the iEEG are much more similar to some patients than others, so we would like some way to more *selectively* share information across patients. For example, some patients display diffuse onset seizures, where most of the channels become hyperactive almost simultaneously. In other patients, activity on a small core of channels initiates the seizure and is joined later into the seizure by higher activity in other channels. In still other patients, the seizure is well localized only to a specific brain region, so channels located far from that region have barely any heightened activity during the seizure. We would ideally like to share information mostly between patients with a propensity for similar onset patterns.

Actually, we already achieve this kind of selective information sharing in the two-level MLC-HDP model, where seizure type ℓ incorporates information from any number of similar seizures. We simply need to add another HDP layer that yields a mixture of patient types. Instead of a patient specific measure $G^{(2)}$ over the seizure type measures $(\delta_{G_{\ell}^{(1)}})_{\ell=1}^{\infty}$, as we had in Equation (3.3), we instead define a patient type measure $G_l^{(2)}$,

$$G_l^{(2)} = \sum_{\ell=1}^{\infty} \pi_{l,\ell}^{(2)} \delta_{G_{\ell}^{(1)}}, \quad \boldsymbol{\pi}_l^{(2)} \sim \mathrm{DP}(\alpha^{(2)}, \boldsymbol{\beta}^{(2)}), \quad \boldsymbol{\beta}^{(2)} \sim \mathrm{GEM}(\gamma^{(2)}), \tag{3.5}$$

which we link together with the model $G^{(3)}$ for a population of patients,

$$G^{(3)} = \sum_{l=1}^{\infty} \pi_l^{(3)} \delta_{G_l^{(2)}}, \quad \boldsymbol{\pi}^{(3)} \sim \mathrm{DP}(\alpha^{(3)}, \boldsymbol{\beta}^{(3)}), \quad \boldsymbol{\beta}^{(3)} \sim \mathrm{GEM}(\gamma^{(3)}). \tag{3.6}$$

Note that in Equation (3.5) each patient-type l uses different weights $\boldsymbol{\pi}_l^{(2)}$ over the same seizure type measures $(\delta_{G_{\ell}^{(1)}})_{\ell=1}^{\infty}$, a representation that parallels each seizure type ℓ using different weights $\boldsymbol{\pi}_{\ell}^{(1)}$ over the same channel observation atoms $(\delta_{\phi_k})_{k=1}^{\infty}$ in Equation (3.1).

We now have three levels of indicator variables $z_t^{(3)}$, $z_{tj}^{(2)}$, and $z_{tji}^{(1)}$ describing the patient, seizure, and channel type, respectively. All together, this three-level model can be written as,

$$
\begin{aligned}
\boldsymbol{\beta}^{(3)} &\sim \mathrm{GEM}(\gamma^{(3)}), \\
\boldsymbol{\pi}^{(3)} &\sim \mathrm{DP}(\alpha^{(3)}, \boldsymbol{\beta}^{(3)}), \\
\boldsymbol{\beta}^{(2)} &\sim \mathrm{GEM}(\gamma^{(2)}), \\
\boldsymbol{\pi}_l^{(2)} &\sim \mathrm{DP}(\alpha^{(1)}, \boldsymbol{\beta}^{(2)}), && l = 1, 2, \ldots, \\
\boldsymbol{\beta}^{(1)} &\sim \mathrm{GEM}(\gamma^{(1)}), \\
\boldsymbol{\pi}_\ell^{(1)} &\sim \mathrm{DP}(\alpha^{(1)}, \boldsymbol{\beta}^{(1)}), && \ell = 1, 2, \ldots, \\
z_t^{(3)} &\sim \boldsymbol{\pi}^{(3)}, && t = 1, \ldots, T, \\
z_{tj}^{(2)} \mid z_t^{(3)} &\sim \boldsymbol{\pi}_{z_t^{(3)}}^{(2)}, && t = 1, \ldots, T,\ j = 1, \ldots, J, \\
z_{tji}^{(1)} \mid z_{tj}^{(2)} &\sim \boldsymbol{\pi}_{z_{tj}^{(2)}}^{(1)}, && t = 1, \ldots, T,\ j = 1, \ldots, J,\ i = 1, \ldots, N_{jt}, \\
\boldsymbol{\phi}_k &\sim H, && k = 1, 2, \ldots, \\
\mathbf{x}_{ji} \mid z_{tji}^{(1)} &\sim F(\boldsymbol{\phi}_{z_{tji}^{(1)}}) && t = 1, \ldots, T,\ j = 1, \ldots, J,\ i = 1, \ldots, N_{jt}.
\end{aligned}
\tag{3.7}
$$

Though this model involves a number of hyperparameters, we show in Section 3.6 that the model is generally insensitive to the vague priors we place on the α and γ hyperparameters.

In Figure 3.1 (right) we give a graphical model representation of this three-level MLC-HDP. In Figure 3.3, we show the organization of an example seizure dataset across a population of patients under the MLC-HDP's three levels of clustering.

Comparison to similar models While the DP, HDP, and NDP models can also be applied to a multi-patient dataset of seizures, such applications suffer from the same problems discussed previously for those models in the single-patient dataset. The NHDP suggested by Lancelot James in his comment to Rodríguez et al. [103] is a three-level model, but we find it somewhat inappropriate for modeling seizures from multiple patients at once for a few reasons. First, it would assume each patient t has the same number of seizures J in order for the different top-level nested HDPs to be identically distributed. Given that the number of seizures recorded for each patient can vary from only one to over fifty, this assumption is clearly inappropriate for our application. Even if we overlook this assumption, the NHDP would yield clusters on the patient level and on the channel level, but not on the seizure level. While there are conceivably situations in which one might want a very high-level description of the dataset (i.e., the patient clustering) and simultaneously a very low-level description (i.e., the channel clustering), we are more interested in a patient and seizure level clustering to describe a large multi-patient dataset of seizures given that such a dataset would most likely involve tens or even hundreds of thousands of individual channel observations.

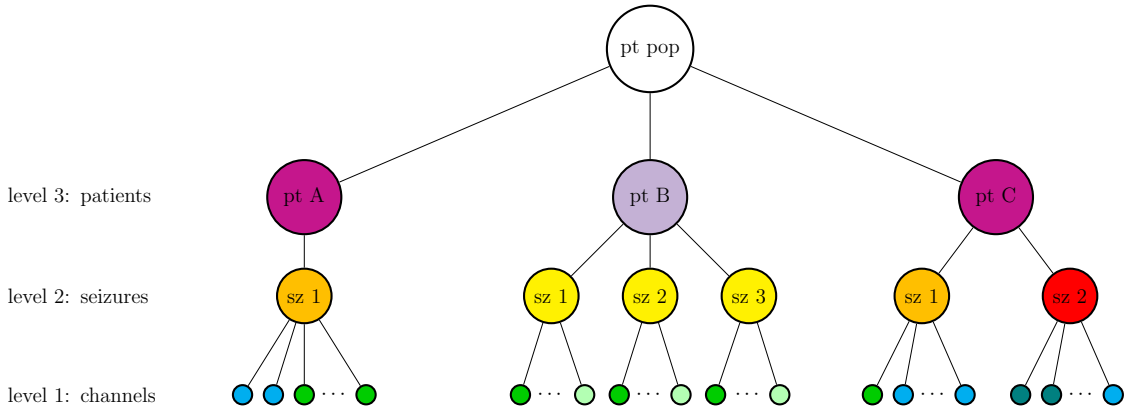

Figure 3.3. A schematic of the epilepsy data under the three-level clustering paradigm of our MLC-HDP model as denoted by the colors at each level. The clusters at each level are shared horizontally across the hierarchy.

■ 3.3 MCMC Posterior Inference

We first describe our choice of observation and prior distributions—$F(\cdot)$ and $H(\cdot)$, respectively—followed by brief descriptions of the Rao-Blackwellization and sufficient statistics used. We then fully describe the steps involved in posterior MCMC.

■ 3.3.1 Observation and Prior Distributions

Throughout the rest of this chapter, we assume the observed data will be the channel features $\mathbf{x}_{tji} \in \mathbb{R}^d$ for channel i's activity in seizure j of patient t. We explicitly describe the features we use in Section 3.5.1. We use a multivariate normal likelihood with diagonal covariance to model these channel features,

$$\mathbf{x}_{tji} \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2), \tag{3.8}$$

where $\boldsymbol{\mu}_k \in \mathbb{R}^d$ and $\boldsymbol{\sigma}_k^2 \in \mathbb{R}_+^d$ are inferred by sharing information (clustering) across channels, seizures, and patients. For convenience, we sometimes represent the set of channel observations occurring in seizure j of patient t as $S_{tj} = \{\mathbf{x}_{tji}\}_{i=1}^{N_{tj}}$ and the set of seizures of patient t as $P_t = \{S_{tj}\}_{j=1}^{J_t}$.

Though the model description of the previous section makes no assumption on the form of the distribution H over the parameter space Θ, we assume that it is conjugate to the observation distribution F. We thus use a normal inverse-Wishart prior $\mathcal{N}\text{-IW}(n_0, \boldsymbol{\mu}_0, \nu_0, I\mathbf{s}_0^2)$ on $(\boldsymbol{\mu}, \boldsymbol{\sigma}_k^2)$ with prior counts n_0, mean $\boldsymbol{\mu}_0$, degrees of freedom ν_0, and (diagonal) scale $I\mathbf{s}_0^2$.

■ 3.3.2 Rao-Blackwellization

Following [120], we use a Rao-Blackwellized sampler (also known as a *collapsed* Gibbs sampler) at each level that marginalizes the level weights $\boldsymbol{\pi}$ and the atom parameters $\boldsymbol{\phi}_k$.

Such marginalization has been shown to improve efficiency of Gibbs samplers [23, 46, 74] by reducing the variance of the sampled value, for us the channel, seizure, and patient type indicators $z^{(1)}_{tji}$, $z^{(2)}_{tj}$, and $z^{(3)}$. In this setting, a finite number $L^{(v)}$ of occupied clusters in level v (of the infinitely many available). We also keep an additional (empty) cluster, which—when filled—prompts the addition of a new, empty cluster. When clusters become empty, they are removed. Thus, at any point in time we have $L^{(v)} + 1$ clusters for level v. For the channel observation level ($v = 1$), we also use the notation K for the number of filled observation-level clusters to be consistent with mixture-modeling convention. We explore properties of the Rao-Blackwellized sampler and the alternatives that do not marginalize level weights and observation model parameters in Section 3.6.3.

■ 3.3.3 Sufficient Statistics

When possible, our implementation of the MLC-HDP parallels that of the HDP by Teh et al. [120]. In addition to the model parameters described in the previous section, we also keep track of two counts variables $\mathbf{n}^{(v)}$ and $\mathbf{m}^{(v)}$ use as sufficient statistics at each level v. At the lowest level, $n^{(1)}_{\ell k}$ represents the number of observations assigned to channel type k across all the seizures in seizure type ℓ. The count $n^{(2)}_{l\ell}$ represents the number of seizures assigned to seizure type ℓ across all the patients assigned to patient type l. The count $n^{(3)}_{l}$ represents the number of patients in the population assigned to patient-type l. The $\mathbf{m}^{(v)}$ auxiliary variables are sampled at each level as described in Equation (3.14), making them matrices at levels 1 and 2 and a vector at level 3. When working with these two variables, we often are interested in a marginal count over one (or both) of the indices, so we substitute a dot ($\cdot$) for the corresponding subscript index.

We let $\mathbf{\Phi}_k$ describe the posterior sufficient statistics for the K non-empty level-1 clusters and let $\mathbf{\Phi}_{K+1}$ describe those for a new cluster. For a normal likelihood, we would have sampled parameter values $\boldsymbol{\phi}_k = (\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k)$ and describe the likelihood of parameters $\boldsymbol{\phi}_\mathbf{k}$ given observation $\mathbf{x}$ as $f_k(\mathbf{x}) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\sigma}^2_k)$, but in our Rao-Blackwellized sampler, we use the posterior predictive likelihood (a multivariate Student-t, see Equation (A.12)) instead, $f_k(\mathbf{x}) = t(\mathbf{x} \mid \mathbf{\Phi}_k)$, where we have overloaded the notation for $f_k(\mathbf{x})$. The function $f_0(\mathbf{x})$ denotes the prior predictive distribution.

■ 3.3.4 Markov Chain Monte Carlo Sampling

The directed graphical model in Figure 3.1 (right) shows the main parameters in the three-level MLC-HDP. It is useful to divide them into three primary sets: the cluster indicators, $z^{(v)}$; the level parameters, $\boldsymbol{\beta}^{(v)}$; and the observation model parameters, $\boldsymbol{\phi}_k$. Optionally, we may also sample the concentration parameters γ and α. Of these steps, sampling the cluster indicators $z^{(v)}$ is usually the most computationally intensive.

One full Gibbs sampler iteration first samples the cluster indicators (starting at the bottom level), then the level parameters and (optionally) the level hyperparameters, and finally the observation parameters if desired. We then remove any clusters that have become empty after the atom indicators sampling. Algorithm 3 summarizes these steps.

Algorithm 3 MLC-HDP master Rao-Blackwellized MCMC sampler

```
for each MCMC sample do
    sample cluster indicators at all levels of model as in Algorithms 4, 5, and 6
    for each level v = 1, ..., 3 do
        add/remove level's clusters if necessary as in Algorithm 7
        sample level weights β^(v) as in Algorithm 8
        (optional) sample hyperparameters γ^(v) and α^(v) as in Algorithm 9
    end for
    (optional) sample obs. level parameters (φ_k)_{k=1}^K (e.g., see Equation (3.16))
end for
```

Sampling cluster indicators At the lowest (observation) level of the model, the cluster indicators are sampled from a multinomial with a probability of selecting cluster k' of

$$p\left(z_{tji}^{(1)} = k' \mid \mathbf{x}_{tji}, \{\mathbf{x}_{-tji}\}, \{z_{-tji}^{(1)}\}, z_{tj}^{(2)}, \alpha^{(1)}, \boldsymbol{\beta}^{(1)}\right) \propto f_{k'}(\mathbf{x}_{tji}) \left(\alpha^{(1)}\beta_{k'}^{(1)} + n_{z_{tj}^{(2)}k'}^{(1)}\right), \tag{3.9}$$

following the Chinese restaurant franchise metaphor for the HDP posterior predictive distribution given by Teh et al. [120, Equation 37]. In this sampler, the sufficient statistics of $\mathbf{x}_{tji}$ are removed from the appropriate sufficient statistics $\boldsymbol{\Phi}_k$ and $n_{\ell k}^{(1)}$ before calculating $f_k(\mathbf{x}_{tji})$. When $k' = K+1$, $f_{k'}(\cdot)$ is the prior predictive distribution $f_0(\cdot)$, and $n_{\ell k'}^{(1)} = 0$. At the seizure level, cluster indicators are sampled similarly

$$p\left(z_{tj}^{(2)} = \ell' \mid S_{tj}, \{S_{-tj}\}, \{z_{-tj}^{(2)}\}, z_t^{(3)}, \alpha^{(2)}, \boldsymbol{\beta}^{(2)}\right) \propto g_{\ell'}(S_{tj}) \left(\alpha^{(2)}\beta_{\ell'}^{(2)} + n_{z_t^{(3)}\ell'}^{(2)}\right), \tag{3.10}$$

where $g_{\ell'}(\cdot)$ is the posterior predictive distribution for seizure cluster ℓ' that results from marginalizing the observation cluster indicator $z_{tj}^{(2)}$ via the multinomial posterior predictive distribution $p(z_{tji}^{(1)} \mid \{z_{-tji}^{(1)}\}_{z_{tj}^{(2)}=\ell'}, \alpha^{(1)}, \boldsymbol{\beta}^{(1)})$ (see Equation (2.17)),

$$g_{\ell'}(S_{tj}) = \prod_{i=1}^{N_{jt}} \left[p\left(z_{tji}^{(1)} = K+1 \mid \{z_{-tji}^{(1)}\}_{z_{tj}^{(2)}=\ell'}, \alpha^{(1)}, \boldsymbol{\beta}^{(1)}\right) f_0(\mathbf{x}_{tji}) + \sum_{k=1}^{K} p\left(z_{tji}^{(1)} = k \mid \{z_{-tji}^{(1)}\}_{z_{tj}^{(2)}=\ell'}, \alpha^{(1)}, \boldsymbol{\beta}^{(1)}\right) f_k(\mathbf{x}_{tji}) \right] \tag{3.11}$$

At the patient level, cluster indicators are also sampled similarly,

$$p\left(z_t^{(3)} = l' \mid P_t, \{P_{-t}\}, \{z_{-t}^{(3)}\}, \alpha^{(3)}, \boldsymbol{\beta}^{(3)}\}\right) \propto h_{l'}(P_t) \left(\alpha^{(3)}\beta_{l'}^{(3)} + n_{l'}^{(3)}\right), \tag{3.12}$$

Algorithm 4 MLC-HDP MCMC sampler for channel cluster indicators

for each patient $t = 1, \ldots, T$ **do**
  store patient t's current cluster index
    $l \leftarrow z_t^{(3)}$
  for each seizure $j = 1, \ldots, J_t$ **do**
    store seizure j's current cluster index,
      $\ell \leftarrow z_{tj}^{(2)}$
    for each channel $i = 1, \ldots, N_{jt}$ **do**
      store channel i's current cluster index,
        $k \leftarrow z_{tji}^{(1)}$
      downdate sufficient statistics
        $\boldsymbol{\Phi}_k \leftarrow \boldsymbol{\Phi}_k \ominus \mathbf{x}_{tji}$,
        $n_{\ell k}^{(1)} \leftarrow n_{\ell k}^{(1)} - 1$
      calculate seizure cluster ℓ's (unnormalized) weights over channel clusters,
        $\widetilde{\boldsymbol{\pi}}_\ell^{(1)} \leftarrow \alpha^{(1)} \left(\beta_1^{(1)}, \ldots, \beta_{K+1}^{(1)}\right) + \left(n_{\ell 1}^{(1)}, \ldots, n_{\ell K+1}^{(1)}\right)$
      sample channel i's cluster indicator,
        $k' \sim \text{Multi}\left(\widetilde{\boldsymbol{\pi}}_\ell^{(1)} \circ (f_1(\mathbf{x}_{tji}), \ldots, f_K(\mathbf{x}_{tji}), f_0(\mathbf{x}_{tji}))\right)$
      update i's channel cluster indicator,
        $z_{tji}^{(1)} \leftarrow k'$
      update sufficient statistics for channel cluster k',
        $\boldsymbol{\Phi}_{k'} \leftarrow \boldsymbol{\Phi}_{k'} \oplus \mathbf{x}_{tji}$
        $n_{\ell k'}^{(1)} \leftarrow n_{\ell k'}^{(1)} + 1$
    end for

    continue to Algorithm 5 for seizure cluster indicator sampling

where $h_{l'}(\cdot)$ is the posterior predictive distribution for patient cluster l' after marginalizing the seizure cluster indicator via the multinomial posterior predictive distribution $p(z_{tj}^{(2)} \mid \{z_{-tj}^{(2)}\}_{z_t^{(3)}=l'}, \alpha^{(2)}, \boldsymbol{\beta}^{(2)})$,

$$h_{l'}(P_t) = \prod_{j=1}^{J_t} \prod_{i=1}^{N_{jt}} \sum_{\ell=1}^{L^{(2)}+1} p\left(z_{tj}^{(2)} = \ell \mid \{z_{-tj}^{(2)}\}_{z_t^{(3)}=l'}, \alpha^{(2)}, \boldsymbol{\beta}^{(2)}\right) \cdot p\left(k = z_{tji}^{(1)} \mid \{z_{-tji}^{(1)}\}_{z_{tj}^{(2)}=\ell'}, \alpha^{(1)}, \boldsymbol{\beta}^{(1)}\right). \quad (3.13)$$

Algorithms 4, 5, and 6 give explicit recipes for sampling these cluster indicators. When sampling from a multinomial, we assume its parameters are automatically normalized to sum to one.

Sampling level parameters As discussed in Section 3.2, each level v's global measure $G_0^{(v)}$ is defined by a set of weights $\boldsymbol{\beta}^{(v)}$ over the measures in the level below (i.e., $(\delta_{\boldsymbol{\phi}_k})_{k=1}^{K+1}$ for $v = 1$, $(\delta_{G_\ell^{(1)}})_{\ell=1}^{L^{(2)}+1}$ for $v = 2$, and $(\delta_{G_l^{(2)}})_{l=1}^{L^{(3)}+1}$ for $v = 3$). We first

Algorithm 5 MLC-HDP MCMC sampler for seizure cluster indicators

continued from Algorithm 4

downdate sufficient statistics

$\mathbf{n}^{(1)}_{\ell *} \leftarrow \mathbf{n}^{(1)}_{\ell *} - \sum_{i=1}^{N_{tj}} \mathbf{e}^{\mathrm{T}}_{z^{(1)}_{tji}}$

$n^{(2)}_{l\ell} \leftarrow n^{(2)}_{l\ell} - 1$

for each seizure cluster $\ell' = 1, \ldots, L^{(2)} + 1$ **do**

calculate seizure cluster ℓ''s weights $\widetilde{\boldsymbol{\pi}}^{(2)}_{\ell'}$ over channel clusters,

$\widehat{\boldsymbol{\pi}}^{(1)}_{\ell'} \leftarrow \alpha^{(1)} \left(\beta^{(1)}_1, \ldots, \beta^{(1)}_{K+1}\right) + \left(n^{(1)}_{\ell' 1}, \ldots, n^{(1)}_{\ell' K+1}\right)$

$\widetilde{\boldsymbol{\pi}}^{(1)}_{\ell'} \leftarrow \widehat{\boldsymbol{\pi}}^{(1)}_{\ell'} / \left(\mathbf{1}^{\mathrm{T}} \widehat{\pi}^{(1)}_{\ell'}\right)$

calculate and store marginal likelihood of seizure j under cluster ℓ',

$g_{\ell'} \leftarrow \prod_{i=1}^{N_{tj}} \left(\widetilde{\pi}^{(1)}_{\ell', K+1} f_0(\mathbf{x}_{tji}) + \sum_{k=1}^{K} \widetilde{\pi}^{(1)}_{\ell', k} f_k(\mathbf{x}_{tji})\right)$

end for

calculate patient cluster l's weights over the seizure clusters,

$\widetilde{\boldsymbol{\pi}}^{(2)}_{l} \leftarrow \alpha^{(2)} \left(\beta^{(2)}_1, \ldots, \beta^{(2)}_{L^{(2)}+1}\right) + \left(n^{(2)}_{l1}, \ldots, n^{(2)}_{lL^{(2)}+1}\right)$

sample seizure j's cluster indicator

$\ell' \sim \mathrm{Multi}\left(\widetilde{\boldsymbol{\pi}}^{(2)}_{l} \circ (g_1, \ldots, g_{L^{(2)}+1})\right)$

update j's seizure cluster indicator,

$z^{(2)}_{tj} \leftarrow \ell'$

update the channel cluster counts for seizure clusters ℓ',

$\mathbf{n}^{(1)}_{\ell' *} \leftarrow \mathbf{n}^{(1)}_{\ell' *} + \sum_{i=1}^{N_{tj}} \mathbf{e}^{\mathrm{T}}_{z^{(1)}_{tji}}$

update patient cluster l's counts for seizure clusters ℓ'

$n^{(2)}_{l\ell'} \leftarrow n^{(2)}_{l\ell'} + 1$

end for

continue to Algorithm 6 for patient cluster indicators sampling

Algorithm 6 MLC-HDP MCMC sampler for patient cluster indicators

continued from Algorithm 5

downdate sufficient statistics

$\mathbf{n}_{\ell *}^{(1)} \leftarrow \mathbf{n}_{\ell *}^{(1)} - \sum_{i=1}^{N_{tj}} \mathbf{e}_{z_{tji}^{(1)}}^{\mathrm{T}}$

$\mathbf{n}_{l *}^{(2)} \leftarrow \mathbf{n}_{l *}^{(2)} - \sum_{j=1}^{J_t} \mathbf{e}_{z_{tj}^{(2)}}^{\mathrm{T}}$

$n_l^{(3)} \leftarrow n_l^{(3)} - 1$

for each seizure cluster $\ell' = 1, \ldots, L^{(2)} + 1$ **do**

calculate and store seizure cluster ℓ''s weights $\widehat{\boldsymbol{\pi}}_{\ell'}^{(2)}$ over channel clusters,

$\widetilde{\boldsymbol{\pi}}_{\ell'}^{(1)} \leftarrow \alpha^{(1)} \left(\beta_1^{(1)}, \ldots, \beta_{K+1}^{(1)}\right) + \left(n_{\ell' 1}^{(1)}, \ldots, n_{\ell' K+1}^{(1)}\right)$

$\widehat{\boldsymbol{\pi}}_{\ell'}^{(1)} \leftarrow \widetilde{\boldsymbol{\pi}}_{\ell'}^{(1)} / \left(\mathbf{1}^{\mathrm{T}} \widetilde{\pi}_{\ell'}^{(1)}\right)$

end for

for each patient cluster $l' = 1, \ldots, L^{(1)} + 1$ **do**

calculate patient cluster l''s weights $\widehat{\boldsymbol{\pi}}_{l'}^{(1)}$ over seizure clusters,

$\widetilde{\boldsymbol{\pi}}_{l'}^{(2)} \leftarrow \alpha^{(2)} \left(\beta_1^{(2)}, \ldots, \beta_{L^{(2)}+1}^{(2)}\right) + \left(n_{l' 1}^{(2)}, \ldots, n_{l' L^{(2)}+1}^{(2)}\right)$

$\widehat{\boldsymbol{\pi}}_{l'}^{(2)} \leftarrow \widetilde{\boldsymbol{\pi}}_{l'}^{(2)} / \left(\mathbf{1}^{\mathrm{T}} \widetilde{\boldsymbol{\pi}}_{l' \cdot}^{(2)}\right)$

calculate marginal likelihood of patient t under cluster l',

$h_{l'} = \prod_{j=1}^{J_t} \prod_{i=1}^{N_{jt}} \sum_{\ell=1}^{L^{(2)}+1} \widehat{\pi}_{l,\ell}^{(2)} \widehat{\pi}_{\ell, z_{tji}^{(3)}}^{(1)}$

end for

update seizure cluster sufficient statistics

$\mathbf{n}_{\ell *}^{(1)} \leftarrow \mathbf{n}_{\ell *}^{(1)} + \sum_{i=1}^{N_{tj}} \mathbf{e}_{z_{tji}^{(1)}}^{\mathrm{T}}$

calculate patient cluster weights,

$\widetilde{\boldsymbol{\pi}}^{(3)} \leftarrow \alpha^{(3)} \left(\beta_1^{(3)}, \ldots, \beta_{L^{(3)}+1}^{(3)}\right) + \left(n_1^{(3)}, \ldots, n_{L^{(3)}+1}^{(3)}\right)$

sample patient t's cluster indicator,

$l' \sim \mathrm{Multi}\left(\widetilde{\boldsymbol{\pi}}^{(3)} \circ (h_1, \ldots, h_{L^{(3)}+1})\right)$

update l's seizure cluster indicator,

$z_t^{(1)} \leftarrow l'$

update the seizure cluster counts for patient clusters l',

$\mathbf{n}_{l' *}^{(2)} \leftarrow \mathbf{n}_{l' *}^{(2)} + \sum_{j=1}^{J_t} \mathbf{e}_{z_{tj}^{(2)}}^{\mathrm{T}}$

update patient cluster counts,

$n_{l'}^{(3)} \leftarrow n_{l'}^{(3)} + 1$

end for

Algorithm 7 MLC-HDP MCMC sampler for adding and removing clusters

let v index the current level

initialize cluster index for this level,
  $r \leftarrow 1$
while cluster index $r \leq L^{(v)} + 1$ **do**
  if $n_{\cdot r}^{(v)} = 0$ **then**
    remove cluster r from level counts,
      $\mathbf{n}^{(v)} \leftarrow \left[\mathbf{n}^{(v)}_{*(1:r-1)} \mid \mathbf{n}^{(v)}_{*(r+1:L^{(v)}+1)}\right]$
    if $v > 1$ **then**
      remove cluster r from the counts in the level below
        $\mathbf{n}^{(v-1)} \leftarrow \left(\mathbf{n}^{(v-1)}_{(1:r-1)*}; \mathbf{n}^{(v-1)}_{(r+1:L^{(v)}+1)*}\right)$
    else
      remove cluster r's model sufficient statistics (recall $K = L^{(1)}$)
        $(\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_{K-1}) \leftarrow (\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_{r-1}, \boldsymbol{\Phi}_{r+1}, \ldots, \boldsymbol{\Phi}_K)$
    end if
    remove cluster r from $\boldsymbol{\beta}^{(v)}$ weights,
      $\rho \leftarrow 1 + \beta_r^{(v)} / \sum_{r'=r+1}^{L^{(v)}+1} \beta_{r'}^{(v)}$
      $\boldsymbol{\beta}^{(v)} \leftarrow \left(\boldsymbol{\beta}_{(1:r-1)}, \rho\boldsymbol{\beta}_{(r+1:L^{(v)}+1)}\right)$
    update number of occupied clusters for this level,
      $L^{(v)} \leftarrow L^{(v)} - 1$
  else if $r = L^{(v)} + 1$ **then** AND $n_{\cdot r}^{(v)} > 0$
    add column of zeros $\mathbf{0} \in \mathbb{Z}^{L^{(v+1)} \times 1}$ to counts for new empty state,
      $\mathbf{n}^{(v)} \leftarrow \left[\mathbf{n}^{(v)} \mid \mathbf{0}\right]$
    if $v > 1$ **then**
      add new space in counts for level $v - 1$
        $\mathbf{n}^{(v-1)} \leftarrow \left(\mathbf{n}^{(v)}; \mathbf{0}^{\mathrm{T}}\right)$
    else
      add new (empty) observation model sufficient statistics
        $(\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_K) \leftarrow (\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_K, \boldsymbol{\Phi}_{K+1})$
    end if
    add another stick-breaking weight element
      $\omega \sim \mathrm{Beta}(1, \gamma^{(v)})$
      $\boldsymbol{\beta}^{(v)} \leftarrow \left(\boldsymbol{\beta}^{(v)}_{(1:L^{(v)})}, \omega\beta^{(v)}_{L^{(v)}+1}, (1-\omega)\beta^{(v)}_{L^{(v)}+1}\right)$
    update number of occupied clusters for this level,
      $L^{(v)} \leftarrow L^{(v)} + 1$
    increment cluster index
      $r \leftarrow r + 1$
  else
    increment cluster index
      $r \leftarrow r + 1$
  end if
end while

Algorithm 8 MLC-HDP MCMC sampler for level weights

let v index the current level

sample and store auxiliary variables for level weight sampling
for $q = 1, \ldots, L^{(v+1)}$ **do**
  for $r = 1, \ldots, L^{(v)}$ **do**
    for $s = 1, \ldots, n_{qr}^{(v)}$ **do**
      $\theta_s \sim \text{Ber}\left(\frac{\alpha^{(v)}\beta_r^{(v)}}{\alpha^{(v)}\beta_r^{(v)}+s}\right)$
    end for
    $m_{qr}^{(v)} \leftarrow \sum_{s=1}^{n_{qr}^{(v)}} \theta_s$
  end for
end for

sample level weights
$\boldsymbol{\beta}^{(v)} \sim \text{Dir}\left(m_{\cdot 1}^{(v)}, \ldots, m_{\cdot L^{(v)}}^{(v)}, \gamma^{(v)}\right)$

sample $\mathbf{m}^{(v)}$ for each $(q, r) \in (\{1, \ldots, L^{(v+1)}\}, \{1, \ldots, L^{(v)}\})$ via the auxiliary variable θ_s,

$$\begin{aligned} \theta_s \mid \{z_*^{(v)}\} &\sim \text{Ber}\left(\frac{\alpha^{(v)}\beta_r^{(v)}}{\alpha^{(v)}\beta_r^{(v)} + s}\right), \qquad \text{for} \quad s = 1, \ldots, n_{qr}^{(v)}, \\ m_{qr}^{(v)} &= \sum_{s=1}^{n_{qr}^{(v)}} \theta_s, \end{aligned} \tag{3.14}$$

and then use it in sampling the level's global weights $\boldsymbol{\beta}^{(v)}$,

$$\boldsymbol{\beta}^{(v)} \sim \text{Dir}\left(m_{\cdot 1}^{(v)}, \ldots, m_{\cdot L^{(v)}}^{(v)}, \gamma^{(v)}\right). \tag{3.15}$$

As previously described, we generally do not directly sample the weights $(\boldsymbol{\pi}_\ell^{(v)})_{\ell=1}^{L^{(v+1)}}$, choosing to instead use the posterior predictive multinomial distribution from $\boldsymbol{\beta}^{(v)}$, $\alpha^{(v)}$, and $n_{qr}^{(v)}$. Algorithm 8 gives an explicit recipe for this sampling for each level $v = 1, \ldots, 3$.

Sampling observation model parameters Each observation model is a multivariate Normal with mean $\boldsymbol{\mu}$ and diagonal covariance $\boldsymbol{\sigma}^2$. As derived in Appendix A.1, the normal inverse-Wishart posterior for observation channel type k is

$$p(\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k \mid \{\mathbf{x}_{tji}\}_{z_{tji}^{(1)}=k}) \propto \mathcal{N}\text{-IW}\left(n_N, \boldsymbol{\mu}_N, \nu_N, S_N\right) \tag{3.16}$$

with parameters

$$\begin{aligned} n_N &= n_0 + |\{\mathbf{x}_{tji} | z_{tji}^{(1)} = k\}|, \\ \boldsymbol{\mu}_N &= \frac{1}{n_N}\left(n_0\boldsymbol{\mu}_0 + \sum_{z_{tji}^{(1)}=k} \mathbf{x}_{tji}\right), \\ \nu_N &= \nu_0 + |\{\mathbf{x}_{tji} | z_{tji}^{(1)} = k\}|, \\ S_N &= I\left(\mathbf{s}_0^2 + n_0\boldsymbol{\mu}_0 \circ \boldsymbol{\mu}_0 + \sum_{z_{tji}^{(1)}=k} \mathbf{x}_{tji} \circ \mathbf{x}_{tji} - n_N\boldsymbol{\mu}_N \circ \boldsymbol{\mu}_N\right), \end{aligned} \tag{3.17}$$

and the posterior predictive distribution for an observation $\mathbf{x}_{tji}$ is

$$p(\mathbf{x}_{tji} \mid \{\mathbf{x}_{-tji}\}_{z_{tji}^{(1)}=k}) \propto t_{\nu_N - d + 1}\left(\boldsymbol{\mu}_N, \frac{n_N + 1}{n_N(\nu_N - d + 1)} S_N\right). \tag{3.18}$$

Sampling level hyperparameters Finally, we have the option to also sample each level v's hyperparameters $\alpha^{(v)}$ and $\gamma^{(v)}$, which can have Gamma(a, b) priors. This step at each level follows exactly that described by Teh et al. [120] and is detailed in Algorithm 9.

3.4 Simulation Experiments

As our model is inspired by the nested Dirichlet process of Rodríguez et al. [103], we first explored some properties of both approaches on simulated data to explore the meaningful differences between the two multi-level clustering models. Note that we omit the DP, HDP, and NHDP models in this comparison because—as previously discussed—they ultimately cannot yield the simultaneous clusterings on the channel, seizure, and patient level that we desire. We use the same simulation data described in Rodríguez et al. [103]: observations are generated from one of four distributions (T1-T4), each of which is a mixture of two to four normal distributions, whose parameters are given in Table 3.1.

Our simulated dataset contained 5 samples from each of the four distributions (T1-T4), where each sample contained 100 observations from that particular distribution. We used the same hyperparameters described in Rodríguez et al. [103], and ran 25 chains for both the MLC-HDP and NDP models, where each chain had a 5000 iteration burn-in and 10-iteration thinning, gathering 10,000 samples total for each model.

We observed that the NDP with inference described as in Rodríguez et al. [103] exhibits high autocorrelation compared to the MLC-HDP in a few investigated common parameters, as shown in the top left and top middle parts of Figure 3.4. Anecdotal results for this simulated dataset indicate that the main cause of this additional autocorrelation is that the weights over the observation level atoms are explicitly sampled only once per iteration in the

Algorithm 9 MLC-HDP MCMC sampler for level hyperparameters

let v index the current label
let the priors for $\gamma^{(v)}$ and $\alpha^{(v)}$ be Gamma$(a_\gamma^{(v)}, b_\gamma^{(v)})$ and Gamma$(a_\alpha^{(v)}, b_\alpha^{(v)})$

sample auxiliary variables for $\gamma^{(v)}$, recalling $\mathbf{m}^{(v)}$ from Line 9 of Algorithm 8
for $r = 1, \ldots, L^{(v)}$ **do**
    $\widehat{m}_r \leftarrow \mathbf{1}\left(m_{\cdot r}^{(v)} > 0\right)$
    $c_r \sim \text{Ber}\left(\frac{m_{\cdot r}^{(v)}}{\gamma^{(v)} + m_{\cdot r}^{(v)}}\right)$
    $d_r \sim \text{Beta}\left(\gamma^{(v)} + 1, m_{\cdot r}^{(v)}\right)$
end for

sample GEM prior for level weights
    $\gamma^{(v)} \sim \text{Gamma}\left(a_\gamma^{(v)} + \widehat{m}_\cdot - c_\cdot, b_\gamma^{(v)} - \sum_{r=1}^{L^{(v)}} \log d_r\right)$

sample auxiliary variables for $\alpha^{(v)}$
for $q = 1, \ldots, L^{(v+1)}$ **do**
    $c_q \sim \text{Ber}\left(\frac{n_{q\cdot}^{(v)}}{\alpha^{(v)} + n_{q\cdot}^{(v)}}\right)$
    $d_q \sim \text{Beta}\left(\alpha^{(v)} + 1, n_{q\cdot}^{(v)}\right)$
end for

sample mass prior for level weights
    $\alpha^{(v)} \sim \text{Gamma}\left(a_\gamma^{(v)} + m_{\cdot\cdot}^{(v)} - c_\cdot, b_\gamma^{(v)} - \sum_{q=1}^{L^{(v+1)}} \log d_q\right)$

	Comp 1			Comp 2			Comp 3			Comp 4		
Dist	w	μ	σ^2	w	μ	σ^2	w	μ	σ^2	w	μ	σ^2
T1	.75	0	1.0	.25	3.0	2.0						
T2	.55	0	1.0	.45	3.0	2.0						
T3	.40	0	1.0	.30	-2.0	2.0	.30	2.0	2.0			
T4	.39	0	1.0	.29	-2.0	2.0	.29	2.0	2.0	.03	10.0	1.0

Table 3.1. Parameters for the true distributions $p_T = \sum_i w_i \mathcal{N}(\mu_i, \sigma_i^2)$ used in the simulation study

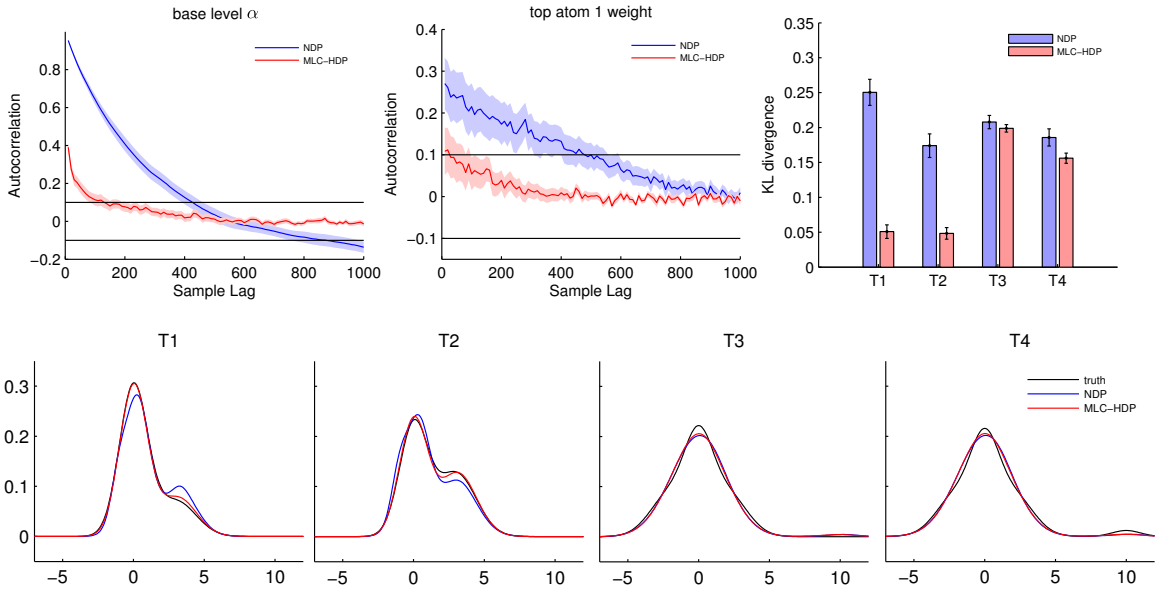

Figure 3.4. Autocorrelation plots for the lower level concentration parameter α (**top left**) and the weight on the first top-level mixture (**top middle**) for the NDP and MLC-HDP models. The dark lines of these plots are the mean over the 25 chains, the shaded region denotes the region of one standard error. (**bottom row**) true and average estimated density functions for each of the four true GMM distributions. (**top right**) the mean Kullback-Leibler (KL) divergence in each of the four distributions for the NDP and MLC-HDP. Error bars bound the region of one standard error.

NDP, whereas they are updated continuously in the Rao-Blackwellized sampling scheme of the MLC-HDP. Results on our iEEG seizure dataset presented in Section 3.6.3 support this hypothesis. Nevertheless, we hesitate to make any strong claims about the autocorrelation inherent in the NDP model versus the MLC-HDP model since autocorrelation is so tied to the particular inference scheme (Rao-Blackwellized, explicitly-sampled weights, adaptive Gibbs sampling, etc) and could potentially be improved by another.

The bottom row of Figure 3.4 shows the MLC-HDP and NDP estimated density functions for each of the four distributions along with the true density functions. The MLC-HDP usually found three top-level groups, and the NDP split evenly between two and three top-level groups, consistent with the result Rodríguez et al. [103] give in their Figure 3 for $J = 20$ and $n = 100$. With this number of observations and samples, both methods were unable to distinguish the T3 and T4 groups, which differ only in a small additional mode at $x = 10$ in the fourth distribution. We use the Kullback-Leibler divergence $D_{\mathrm{KL}}(\text{true} \,||\, \text{estimated})$ [77] for each of the four distributions between the true density functions and the estimated density functions to assess the degree of similarity. These Kullback-Leibler results, shown in the top right of Figure 3.4, illustrate how the estimates of the MLC-HDP model are better than those of the NDP, especially in T1 and T2.

The simulated dataset originally designed by Rodríguez et al. [103] contains common lower level distributions between different top level mixtures. For example T1 and T2

contain data from the same two normal distributions, just in different proportion. The NDP must estimate those base distributions independently for T1 and T2, whereas the MLC-HDP estimates benefit from the data between all the groups.

When attempting to scale up the NDP to a modestly sized ($K = 55$, $L^{(2)} = 35$, and $L^{(3)} = 35$, using the truncated approximations to the DP as in Rodríguez et al. [103]) three-level structure for our iEEG seizure dataset, we found that the total number of observation atoms—more than 67,000—quickly became computationally impractical. The MLC-HDP avoids this pitfall by sharing observation level atoms among the higher level mixtures.

■ 3.5 Human Seizure iEEG Experiments

We compiled a dataset of 193 intracranial EEG seizure records across 10 patients from the Children's Hospital of Philadelphia. These patients display attributes common in epilepsy intracranial EEG: unique electrode placement, large differences in the number of seizures per patient, and differences in the number of usable channels in the seizures of a patient. Table 3.2 describes the number of seizures per patient as well as whether a patient's seizures contain the same number of active (i.e., non-corrupted and without significant artifact) electrodes.

■ 3.5.1 Data Description

We chose to work with iEEG clips of all channels -30 seconds before to +90 seconds after the epileptologist-marked start of the seizure. Instead of working directly with the iEEG channel voltage traces, we extracted a set of simple and intuitive features for each channel: the $\log_{10}$ power in four clinically relevant frequency bands (4-8, 8-13, 13-30, 30-100 Hz) for each channel over the 120 seconds, using a sliding window of 500 ms with 50% overlap. While rich literature exists on EEG features [cf., 3, 50, 97, 111, 114], we chose these features because they closely resemble those we believe actual epileptologists consider when reading EEG. The four frequency features at each of the 479 time points were concatenated into a 1916-dimensional feature vector for each channel's activity during the seizure. We then used principal components analysis (PCA) over all the seizures of all the patients to reduce its dimensionality to 5, retaining 67.7% of the original variance. Figure 3.5 shows an example of the 128 raw iEEG channel traces, the beta band (13-30 Hz) feature for each channel over time, a scree plot for the first 50 principal components of the channel features, and the mean and first five principal components of the channel features. The principal components show how different time points and frequency band powers are emphasized in different principal components. In initial experiments, we found the clustering results to be reasonably similar when using 5, 10, and 20 principal components, an unsurprising result given that the cumulative variance increases only marginally from 5 to 10 to 20 principal components.

In our iEEG seizure experiments, we used Gamma(1,1) priors on $\alpha^{(3)}$, $\alpha^{(2)}$, $\gamma^{(3)}$, and $\gamma^{(2)}$ and Gamma(5,1) priors for $\alpha^{(1)}$ and $\gamma^{(1)}$. The experiments on seizure data described in this section were run before we performed the full sensitivity analysis described in Section 3.6.2,

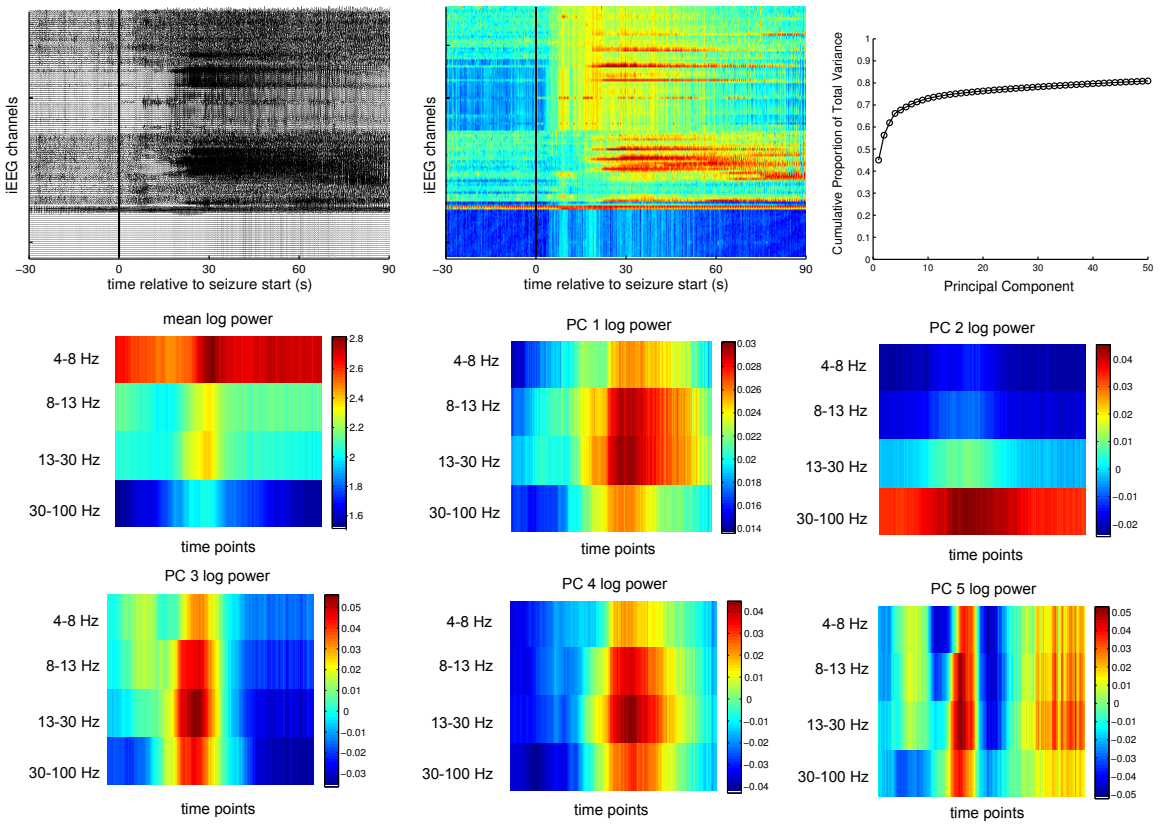

Figure 3.5. (**top left**) An example of 128 iEEG voltage traces over time for a seizure, where the start of the seizure is indicated with a vertical black line. (**top middle**) The $\log_{10}$ power in the clinical beta band (13-30 Hz) for each channel, where red corresponds to larger values and blue to smaller values. (**top right**) A scree plot showing the cumulative proportion of variance of the first 50 principal components of the channel activities. (**middle & bottom rows**) The mean and first five principal components (PC) of the channel features used for each of the four frequency bands and at each of the 479 time points used.

Patient	# Seizures	Same # channels
A	1	yes
B	9	yes
C	4	yes
D	18	no
E	61	no
F	50	no
G	1	yes
H	22	yes
I	13	no
J	14	yes

Table 3.2. The number of recorded seizures for each patient and whether all the seizures of each patient contained the same number of active channels.

which shows the model to be relatively insensitive to the gamma prior used. Given that our feature values are real-valued and continuous, we used a multivariate normal likelihood for the channel observation model with the added constraint that its covariance be diagonal (for computational simplicity). We used a normal inverse-gamma prior for the parameters ϕ_k that has a mean and diagonal covariance equal to those of the channel feature observations. We initialized the model with one patient-type cluster for each patient, 40 seizure-type clusters, and 150 channel-activity observation clusters.

■ 3.5.2 The Advantages of a Hierarchical Model

Since we are aware of no other hierarchical models that share information across seizures and patients in the epilepsy modeling literature, we explored the extent to which this information sharing improves the model for a patient's heldout seizures. We thus were interested in comparing our MLC-HDP model to a flat alternative that is otherwise quite similar. We use the DP because it is the closest flat model to the MLC-HDP. Note that the HDP and NDP models, which also involve hierarchical information sharing albeit in a slightly different fashion than the MLC-HDP (see our discussion of the different models at the end of Section 3.1), would also provide interesting comparisons to the flat DP. But since we believe the MLC-HDP has other features not shared by the HDP (multiple levels of clustering) or the NDP (efficient cluster sharing across the hierarchy), we focus exclusively on comparing the hierarchical MLC-HDP to only the flat DP, paralleling a similar experiment described by Teh et al. [120].

We compared the MLC-HDP trained on a full hierarchy over all patients and seizures to the DP trained with two different datasets: one with only the channel activities from the seizures of patient t and another with channel activities from all patients. We denote the MLC-HDP's modeling scenario as M_3 and the two DP modeling scenarios as M_1 and M_2, respectively. For a given patient t, we created $J_t - 1$ training and testing sets for each of these three modeling scenarios. We summarize these three scenarios below:

M_1 the channel-observations from seizures $1, \ldots, j$ of patient t are used to train a standard DP mixture model;

M_2 the channel-observations from seizures $1, \ldots, j$ of patient t and all the seizures $j' \in \{1, \ldots, J_{t'}\}$ of all the other patients $t' \neq t$ are used to train a standard DP mixture model;

M_3 the same data as M_2 is used but organized in the full patient-seizure-channel hierarchy available in the MLC-HDP model.

To evaluate the three models, similarly to Teh et al. [120], we use the conditional perplexity—whose log is the Shannon entropy [77, 109]—of the future held out seizures $j+1, \ldots, J_t$ given the cluster index for each channel observation

$$PP(S_{tj+1}, \ldots, S_{tJ_t}) = \exp\left(-\frac{1}{J_t - j} \sum_{j'=j+1}^{J_t} \log p\left(S_{tj'} \mid z^{(1)}_{tj'1}, \ldots, z^{(1)}_{tj'N_{tj'}}\right)\right) \tag{3.19}$$

where

$$p\left(S_{tj'} \mid z^{(1)}_{tj'1}, \ldots, z^{(1)}_{tj'N_{tj'}}\right) = \prod_{i=1}^{N_{tj'}} f_{(z^{(1)}_{tj'i})}(\mathbf{x}_{tj'i}) \tag{3.20}$$

Lower perplexity values indicate better models. We ran 25 chains, each yielding 200 samples after a 500-iteration burn in and 20-iteration spacing.

The left side of Figure 3.6 shows the results of this experiment for patient B. Corresponding plots for other patients are similar. Early on, the DP model with only patient B's seizures (M_1) suffers from a low amount of training data, though as more and more seizures are added to the training set, its model improves considerably. The DP model incorporating training data from all other patients (M_2) demonstrates much better performance than M_1 when the amount of training data from patient B is small but becomes slightly worse than M_1 after the first five seizures of patient B are included in the training set. Though it has the same training data as M_2, the hierarchical MLC-HDP (M_3) model has consistently lower perplexity (better performance) than both M_1 and M_2 models across every number of patient B training seizures. These results show the value of a hierarchical model for our iEEG seizure application. This organization allows the model to intelligently blend local data (e.g., the seizures for a particular patient) with global data (e.g., the seizures from all other patients). Such hierarchical organization is particularly appropriate for a dataset like ours where the number of seizures varies widely between patients.

■ 3.5.3 Automated and Manual Seizure Clustering

To assess how well the MLC-HDP clusters the 193 seizures from 10 patients, we had two board-certified epileptologist manually cluster the seizures for each patient. The physicians were given deidentified printouts of each seizure and also had access to them in iEEG viewing software. Clustering one patient at a time, the physicians were instructed to cluster

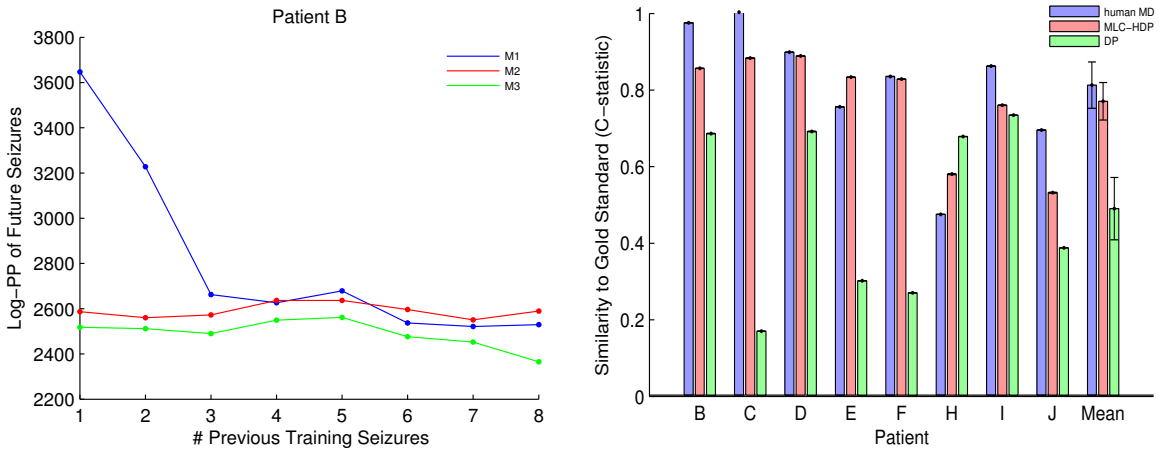

Figure 3.6. (**left**) The mean log-perplexity of patient B's future seizures in each of the three models: M_1, a DP with training data from previous seizures from the patient; M_2, a DP using training data from previous seizures from the patient as well as all the seizures from all the other patients; M_3, an MLC-HDP model with its full patient-seizure-channel hierarchy and clustering. (**right**) The seizure clustering similarity between the clusters found by a DP, an MLC-HDP, and physician expert to another expert's clustering for each patient individually and the patient average. Patients A and G are excluded because they only had one seizure. Errorbars bound the region of one standard error.

according to whatever clinical aspects they felt were most important. This task is inherently subjective and uncertain, so we used two physicians to quantitatively assess the inter-expert uncertainty in this task. These two physicians have trained and worked with each other for over ten years, so their markings should be as close as two separate human markers can be. For our subsequent analysis, we arbitrarily chose one of the physicians as the "gold standard." Our results do not change substantially when the markings of the other physician are used as the gold standard instead.

In addition to the MLC-HDP and physician seizure clusterings, we also desired a baseline seizure clustering from another model. Of the related models we have previously considered (DP, HDP, NDP, NHDP), the NDP is the only other one that naturally yields an explicit seizure clustering when channel activities are used as the observations. Unfortunately, the NDP involves assumptions and computational burdens that are impractical for this problem, as we discuss in Section 3.1. For the baseline seizure clustering, we thus decided to work with a parameterization of a seizure that is agnostic to the number of channel recordings in that seizure. Such a setting allows us to straightforwardly compare seizures with a single patient and across multiple patients. It also allows us to cluster seizures using a standard DP. We believe this method for producing baseline seizure clusterings is the closest reasonable alternative to those produced by our MLC-HDP.

For the seizure parameterization, we worked with the six features of Schiff et al. [106] since we believe they capture the most important dynamics of a seizure, namely, the synchronization of different areas of the brain and their frequency characteristics. These features were calculated using the same sliding window as our frequency-band features for

individual channels and are described in Appendix B. As with the channel features, these seizure features were concatenated across time windows and reduced to the top 20 principal components, retaining 72.3% of the variance.

We use the c-statistic of Rand [96] to determine the similarity between each seizure clustering and the gold standard because it elegantly handles different numbers of clusters and labels between two differ clusterings. The similarity c between two clusterings Y and Y' of N objects is given by

$$c(Y, Y') = 1 - \frac{1}{\binom{N}{2}} \left(\frac{1}{2} \sum_i \left(\sum_j n_{ij} \right)^2 + \frac{1}{2} \sum_j \left(\sum_i n_{ij} \right)^2 - \sum_i \sum_j n_{ij}^2 \right) \tag{3.21}$$

where n_{ij} is the number of points simultaneously in the ith cluster of Y and the jth cluster of Y'.

We clustered the seizures from each patient separately and also all together using the MLC-HDP and DP and found that both models yield superior clustering performance when all the seizures are clustered at once. Clusterings over all patients are thus used for both models in subsequent discussion. This finding it intuitive: when the human experts cluster each patient's seizures separately, they do not forget about the many thousands of seizures they have seen before this task, whereas the MLC-HDP and DP models have only the seizures we give them.

The right side of Figure 3.6 compares the MLC-HDP and DP clustering performance to that of the other physician (i.e., not the gold standard). The physicians usually agree best with each other, as one would expect, but the MLC-HDP's clusterings are often close to those of the physician. Patient H shows an interesting case in that the DP's clusterings were closer to one physician expert that than other expert's (and the MLC-HDP's) were. The two physicians disagreed most on this patient. The seizures of patient H are extremely similar, and it seems that the models and the experts disagreed on the best way to split them up. The gold-standard expert and the DP had fewer clusters in the seizures of patient H, whereas the MLC-HDP and the other doctor split them more.

We believe the main difference between the MLC-HDP and DP clusterings comes from the differences in how each method represents a seizure. In the seizure features used by the DP, the activity of a few channels can be washed out by that of all the channels. Since the MLC-HDP explicitly models the activity of each individual channel, a few important channels can more easily sway the entire model of the seizure. Both physicians indicated that they followed the clinical practice of defining a seizure in large part by the activity of a few "leading" channels. These results incline us to believe that any attempt to model seizures must begin with modeling the activity of individual channels and build from there. We also believe that the absence of such methods until now explains the almost non-existence of seizure clustering within and between patients in the epilepsy literature.

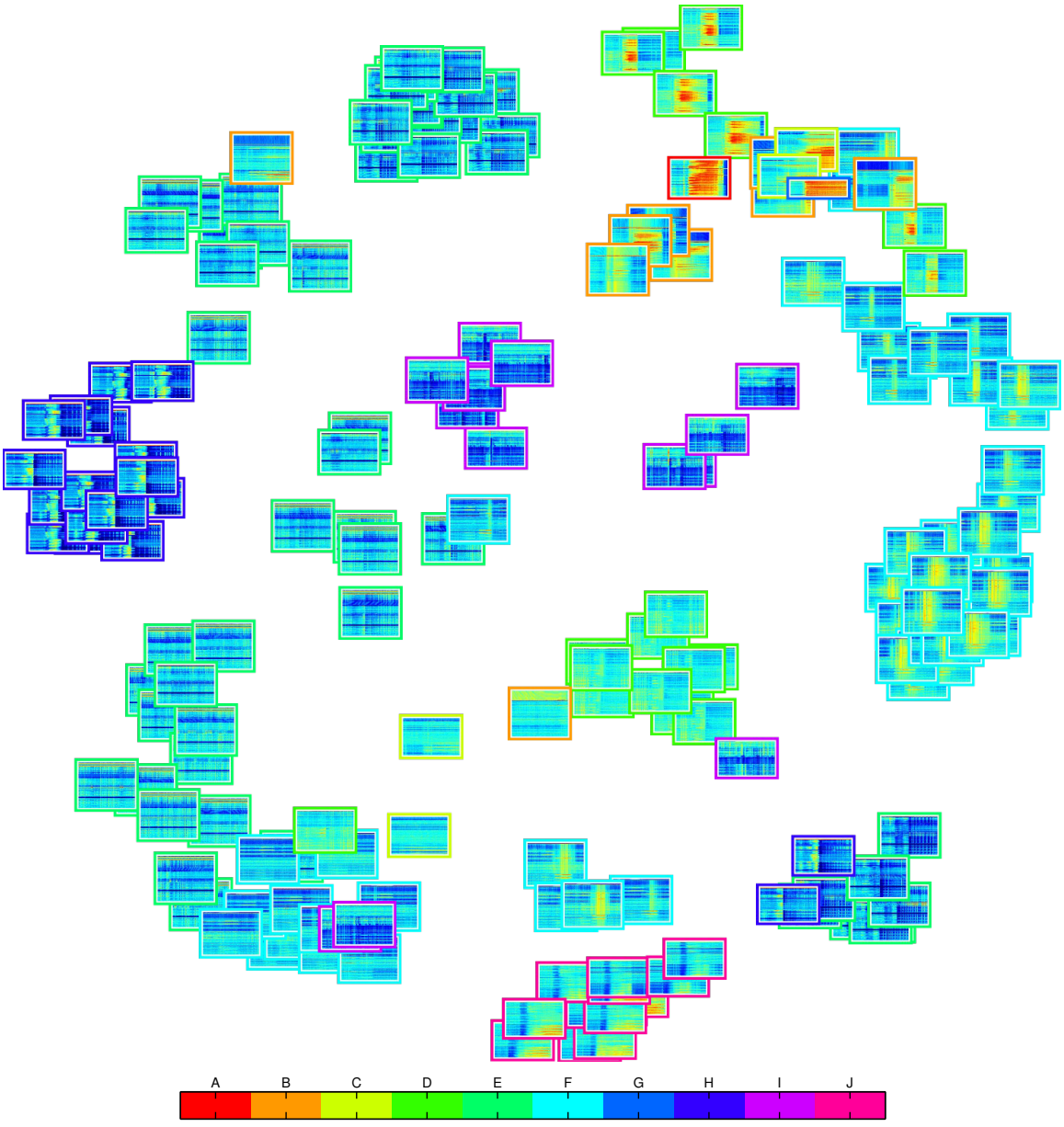

Figure 3.7. A 2D representation of the similarities between seizures across all patients in the MLC-HDP model, where seizure images closer together indicate more similarity. Image positions are slightly jittered to make more seizures visible. Each seizure image depicts the 13-30 Hz feature (blue: low intensity, red: high intensity) across all the channels (rows) at each time point (columns). The image border colors denote to which patient they belong. The channel order of the seizures is consistent within a patient but not so between patients. Note that seizures from different patients can have quite similar dynamics.

3.5.4 Seizure Similarity Across Patients

We can use the seizure clusterings over the 5000 MCMC samples (200 samples from each of the 25 chains) to derive a similarity metric between seizures. We can similarly derive similarity metrics between the channels and seizures. Similarity metrics that generalize between patients have received scant attention in the epilepsy literature. This metric is simply the posterior probability with which two seizures occur in the same cluster. We use the posterior probability of the two seizures *not* clustering together as a distance metric. We then use least squares multidimensional scaling [55, pg. 502] to find a 2-dimensional projection of the seizures, where seizures closer together in the 2D space are more similar.

The seizure images corresponding to the 193 seizures in our dataset are plotted in Figure 3.7. We note that the seizures of the same patients (which share the same color outline) are often situated near each other, though they can also vary greatly. This 2D representation also allows us to find seizures of different patients that are similar to each other. For example, patients A, B, C, D, and H all have seizures similar in spatio-temporal morphology and close in the 2D representation (top right area of Figure 3.7). Such a task would be considerably more difficult if one had to wade though all the 193 seizures individually or even try to look at them in the aggregate. We hope that methods such as this will be helpful in the future of organizing and mining much larger intracranial EEG datasets with hundreds of patients and thousands of seizures.

3.6 Model Sensitivity Analysis

To further explore the MLC-HDP model, we ran experiments to understand practical aspects of the model's behavior with our iEEG seizure dataset. Unless otherwise specified, all priors and other parameters were kept the same as those used in Section 3.5.

3.6.1 The Influence of Individual Patients

One of the motivating ideas behind using a hierarchical model to model a collection of patients instead of a flat model is that it "evens the playing field" between patients relative to the number of seizures each has. Alternatively, in a flat model, patients with many seizures might overly-influence the model, whereas patients with just a few seizures may influence it very little.

To test the extent to which this skewing (or lack thereof) occurs in the MLC-HDP model, we examined the perplexity of all of patient B's heldout seizures under a set of models, each of which had another patient removed from the training dataset. Ideally, every patient should improve the model relative to patient B's seizures, and thus when each patient is removed, the model should worsen (and the perplexity increase). We ran 25 chains, taking 200 samples from each at 20-iteration thinning after 500 burn-in iterations.

The left side of Figure 3.8 shows the results of this experiment for both hierarchical MLC-HDP and flat DP models (with the same architecture as the M_3 and M_2 models, respectively, in the experiments of Section 3.5.2). In addition to underscoring that the

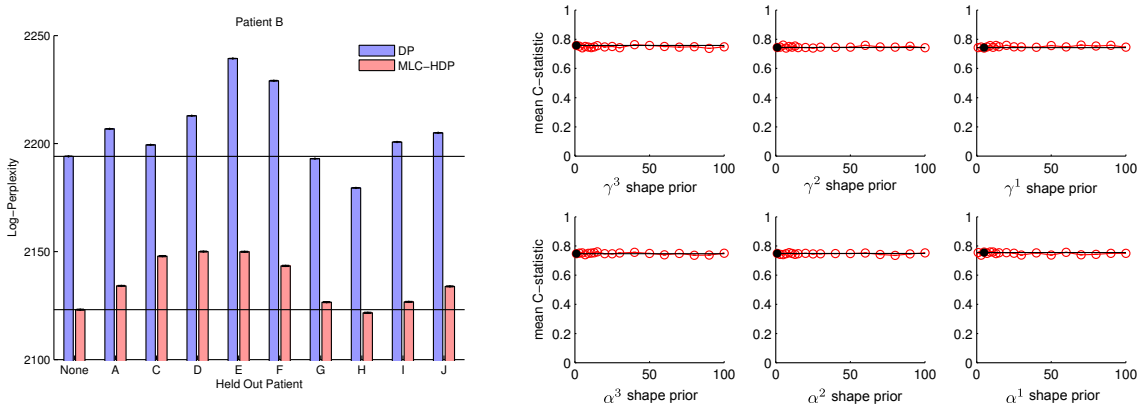

Figure 3.8. (**left**) The log-perplexity of held-out seizures from patient B when also removing each of the other patients (or none) from the training set for the MLC-HDP and DP models. The horizontal black line denotes the baseline perplexity with all the other patients in the training set. Lower perplexities indicate better models. Errorbars (barely visible) denote the region of one standard error. (**right**) The average seizure clustering performance (c-statistic) of the MLC-HDP model when the concentration parameter Gamma prior's shape parameter for one of the α or γ concentration parameters at one of three levels is varied over values between 1 and 100. The standard shape value used in the rest of this work is shown as a black dot, and its associated clustering performance is shown with a black line.

MLC-HDP simply yields superior models to the DP (since the perplexities are lower), we also note that the MLC-HDP's range of variation from the case where none of the patients were held out is smaller. We interpret this smaller variation as a sign that the MLC-HDP model is influenced less, and also more regularly, by removing one patient from the dataset. Furthermore, the patients with relatively few numbers of seizures (A, C, and G) have an influence more balanced with the patients with relatively many seizures (D, E, and F) in the MLC-HDP. Finally, we note that the detrimental effect of patient H's seizures on the model is much smaller (and almost negligible) in the MLC-HDP than the DP model. Recall that patient H was also the patient where the two physician experts disagreed about seizure clustering more than the MLC-HDP and the DP models did with either. These seizures of patient H do have a different form than those of most other patients, so it is not entirely surprising that they have led to outlying results in some of our experiments.

■ 3.6.2 Prior Sensitivity

Throughout this paper, we have intentionally used vague priors since we have very little basis for anything stronger. We generally set the observation model prior for the base distribution from the data but have no such convenient method for the gamma priors on the α and γ concentration parameters at each of the three model levels. We thus varied the gamma prior shape parameter between 1 and 100 for each of the six concentration parameters in turn (and keeping the other five fixed at the values given in Section 3.5.2) to assess how sensitive the global model performance was to these priors. We used the average patient clustering performance (relative to the same gold-standard markings used

in Section 3.5.3) for our metric of global model performance. We ran 10 chains, taking 100 samples per chain at 20-iteration spacing after 500-iteration burn-in, for 18 different shape values for each of the 6 concentration parameters. For each MCMC sample and each of the eight patients, we calculate the c-statistic between the MLC-HDP model's seizure clustering and the physician gold-standard clustering of that patient as described in Equation (3.21). These c-statistics are averaged across patients and MCMC samples to yield a single value for the given gamma prior shape value for the particular concentration parameter.

The right side of Figure 3.8 shows these results for each of the six concentration parameters. We notice that varying the prior shape parameter over much larger values than those normally used has very little effect on the global clustering performance. We also ran a permutation test on gold-standard physician seizure clusterings to obtain a null distribution for the mean c-statistic across the patients. Over 100,000 permutations, we found the 95% confidence interval of the mean c-statistic null-distribution to be $[0.48, 0.64]$, well below the values around 0.8 yielded by the MLC-HDP.

To fully understand the how the prior distributions on our $\gamma^{(v)}$ and $\alpha^{(v)}$ hyperparameters affects the clustering, we simulated clusterings at each level from the priors over the same range of gamma prior shape parameters used in Section 3.6.2 of the paper. This simulation used the same Rao-Blackwellized sampler described in the paper, resulting in the clusterings at each level being independent. Thus, changing a parameter on one level has no influence on the prior clustering of a different level. We clustered the same number of channels, seizures, and patients at each level as existed in our original seizure dataset (476, 193, and 10, respectively) and gathered 1000 samples for each of the 18 Gamma prior shape values for each of six concentration parameters.

Figure 3.9 shows the number of clusters yielded at each level of the model. As expected, the number of clusters increases as the gamma prior shape parameter increases. The prior distribution in the number of clusters at each level for the prior values used in our seizure experiments is similarly reasonable.

Figure 3.10 shows the probability of clustering two different channels, seizures, and patients at each of their respective levels. As the number of clusters increases with increasing gamma prior shape, the probability of any two items being clustered together goes down, as we would expect. These distributions show broad support over the $[0, 1]$ domain, especially for the seizure and patient clusters.

Though the prior distributions do change depending on these hyperparameter values, we see in Section 3.6.2 of the main text that posterior inference is not sensitive to differing hyperparameter values.

■ 3.6.3 Alternative Sampling Schemes

As discussed previously in Section 3.3, throughout this work we use a collapsed Rao-Blackwellized (RB) sampler when possible, following Teh et al. [120] and previous literature [23, 46, 74]. Nevertheless, we were interested in the properties of different sampling schemes for our application and so tested convergence and autocorrelation properties for four different combinations of observation and higher-level Gibbs samplers. For the observation

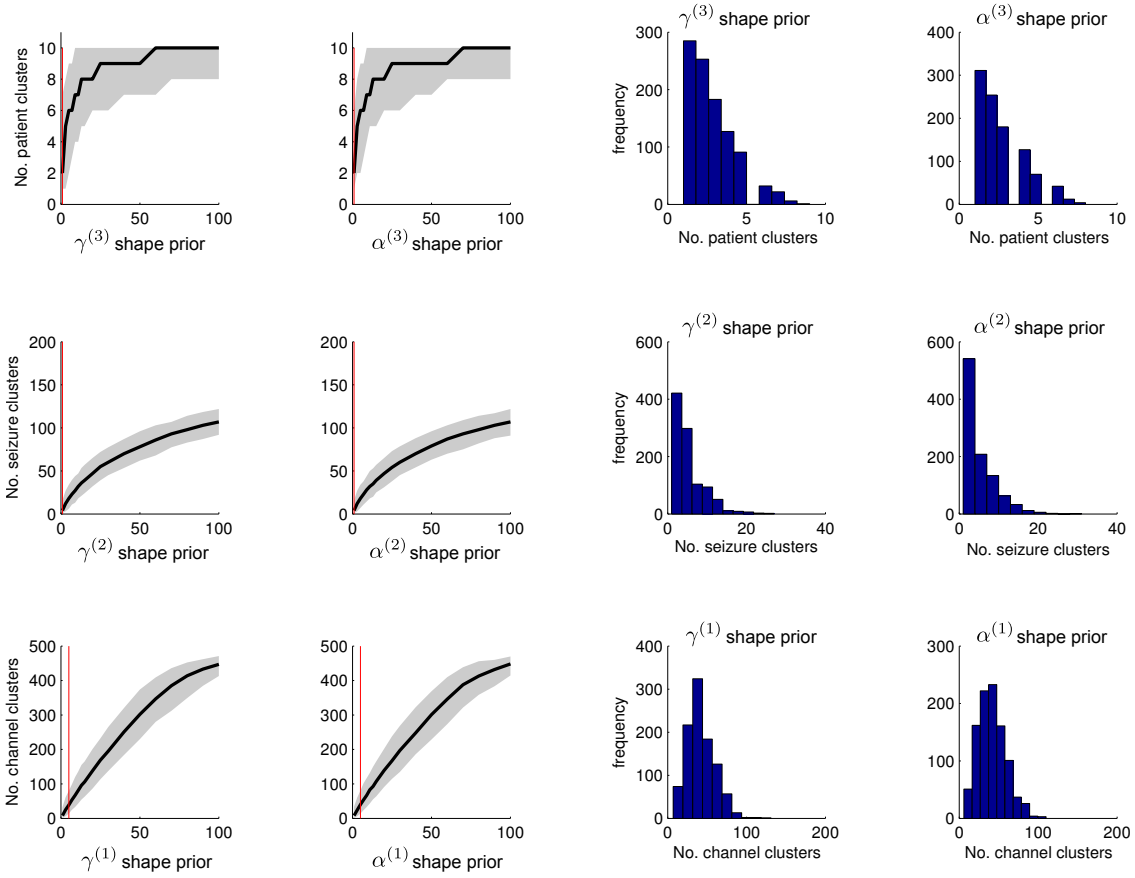

Figure 3.9. (**first two columns**) The number of clusters produced at each level of the hierarchy in prior simulations over a range of gamma prior shape values for the two concentration parameters at each level. The black lines denote the distribution median and the gray area the 95% confidence region. The red line denotes the actual gamma prior shape value used for our seizure experiments. (**last two columns**) Histograms of the number of clusters produced at each level of the hierarchy in prior simulations for the gamma prior shape values used in our seizure experiments (i.e., the cross section distribution at the red lines in the first two columns of the figure).

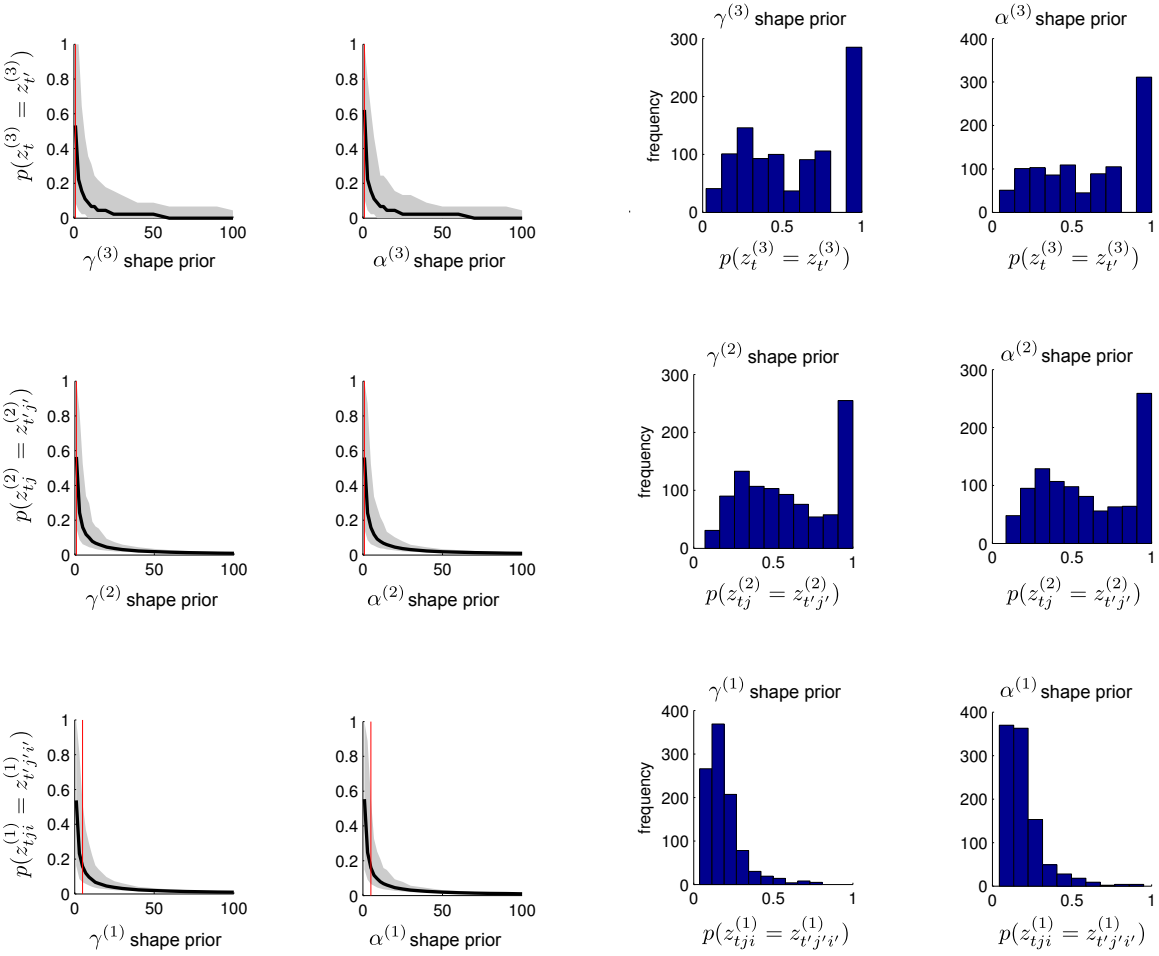

Figure 3.10. (**first two columns**) The probability of any two channels ($tji \neq t'j'i'$), seizures ($tj \neq t'j'$), or patients ($t' \neq t'$) being clustered together over a range of gamma prior shape values for the two concentration parameters. The black lines denote the distribution median and the gray area the 95% confidence region. The red line denotes the actual gamma prior shape value used for our seizure experiments. (**last two columns**) Histograms of the probability of two channels, seizures, or patients being clustered together produced at each level of the hierarchy in prior simulations for the gamma prior shape values used in our seizure experiments (i.e., the cross section distribution at the red lines in the first two columns of the figure).

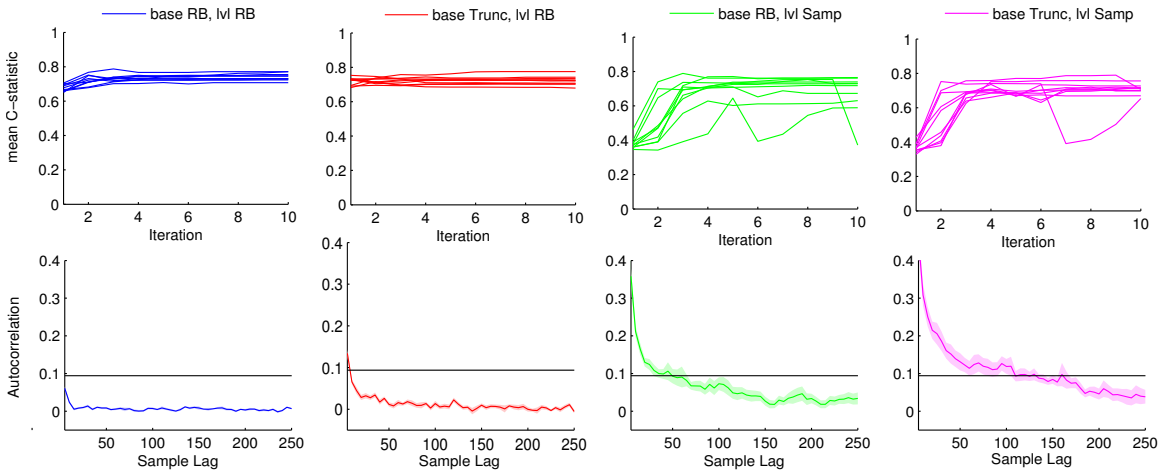

Figure 3.11. A comparison between collapsed Rao-Blackwelled (RB) and explicit (Trunc or Samp) samplers for the observation atom parameters $\boldsymbol{\phi}_k$ and the level atom weights $\boldsymbol{\pi}$. Each column represents one of the four possible combinations of base- and level-samplers. The top row shows the convergence of average seizure clustering performance for 10 chains in the first 10 MCMC iterations. The bottom row shows the average autocorrelation of the observation level concentration parameters $\alpha^{(1)}$ over 250 iteration lags, where the upper confidence bound is shown as a black line.

atoms, we compared the collapsed RB sampler to a common truncated version [60, 62, 116], where we approximate the infinite number of atoms with a (large) finite number whose parameters we explicitly sample. This truncated sampler is used by Rodríguez et al. [103] in their NDP model. For the higher levels, we either use the collapsed RB sampler or the explicitly sample the weight parameters from the Dirichlet posterior (see Equation (2.16)). For each of these four sampler combinations, we investigated the convergence of the average seizure clustering performance and the autocorrelation of the lowest concentration parameter $\alpha^{(1)}$.

Figure 3.11 shows the results from these experiments. We first note that explicitly sampling the higher-level weight parameters (shown in the third and fourth columns of Figure 3.11) seems to have the most pernicious effect on both convergence and autocorrelation. At the higher levels of the model, where patient and seizure counts are relatively small (10 patients and 193 seizures versus 21,476 individual channel-activity observations), the Dirichlet posterior for the level weights has relatively low counts in each of the components, and so the variability of the sampled weights is quite high. This high variability is most likely the cause of the slower convergence and higher autocorrelation results.

We next note that the differences between the two observation atom samplers is much smaller than those of the two level weight samplers. Interestingly, the observation-level truncated sampler seems to have slightly better convergence properties than the RB sampler. This result is similar to that of an analogous experiment in Fox et al. [42]. In that paper, the authors argue that sampling the cluster indicators all at once instead of one at a time (as in the RB sampler) improves mixing and thus convergence. In our seizure application, this

also seems to be the case, though the differences are small. Our results also indicate that slight improvement in convergence of the truncated sampler comes at the cost of slightly increased autocorrelation. Anecdotally, we have found that in smaller (simulated) datasets, the higher autocorrelation is more pronounced.

■ 3.7 Discussion and Future Work

In the past decade, epilepsy monitoring units and treatment centers have become more systematic in storing and documenting the EEG, both intracranial and scalp, acquired from patients. In our lab alone, we have records from over a hundred patients between the Hospital of the University of Pennsylvania and the Children's Hospital of Pennsylvania. As information sharing and collaboration between treatment centers continues to improve, datasets of hundreds of patients will become commonplace. Our ability to process and understand this vast volume of EEG data has not kept pace with our ability to collect it. While human EEG readers can reasonably explore and analyze a single patient's record, doing so over hundreds of patients and making comparisons between them quickly becomes an impossible task.

The work we have presented in this chapter is an initial attempt at creating a principled framework through which to explore, compare, and understand the records from many patients at the same time. It tries, albeit crudely, to model epileptic events (seizures or otherwise) in a way similar to how we believe clinicians do: start by describing the channel activity in each event, use that channel activity of each event as the basis for comparing events between each other, and describe a patient by the types of events he tends to have. We have shown that this approach can both produce seizure clusterings similar to those of physicians and also allows for intuitive visualizations depicting the similarities between large numbers of seizures.

The multi-level modeling approach we have taken here allows for a number of extensions, both on the modeling and analysis fronts. We discuss a few below.

Alternate observation models At the lowest level of the model, we assume in this chapter that the channel observations can be represented by a low-dimensional projection in feature space. Specifically, we use a five-dimensional Gaussian with diagonal covariance. This assumption not only forces the events described to have the same time length, but it also is brittle with respect to the time alignment of the events within this fixed time window (which for us was -30 seconds to +90 seconds around the seizure onset). Using an explicitly temporal model, like a Markov switching model, would allow us to capture much more of the fine-grained details of each event. In fact, it is possible that much or all of the model we discuss in the next chapter could be integrated at the lower level(s) of the MLC-HDP, allowing for both very global and very local comparisons of epileptic events between patients.

Incorporating patient meta-data While we focus exclusively on patient iEEG in this work, the clinical record is actually far richer. In addition to complex CT and MR imaging, we also have raw-text clinical notes that include patient seizure and medication history, often

video from the EMU, clinical judgements about epilepsy type and origin (lesional versus non-lesional, temporal lobe versus neocortical), seizure onset zone judgements, and seizure freedom rates for those who ultimately underwent resective surgery. While some of these data are more objective than others, all add significant nuance and richness unable to be completely captured by a patient's EEG alone. Incorporating some of these data at the patient level of our model would bring it ever closer to approximating the judgements of human clinicians.

Visualization of channel and seizure inferences The channel clusterings produced by the MLC-HDP could be viewed in their 2-dimensional configuration in the electrode grids and strips or (even better), in the 3-dimensional configuration in context with the patient's actual brain. Such channel cluster visualizations would facilitate their use in every-day clinical decision making by non-technical clinicians. Specifically, we expect these visualizations would provide useful decision support when epileptologists determine the physiologic regions of seizure onset and spread.

Visualizing seizure-level clusterings in a timeline would clarify the temporal similarities present in the seizure when clinicians review an entire patient's iEEG record. A clustering timeline such as this would allow clinicians to see how a patient's epilepsy changes over the period of days or weeks. Such decision support would give clinicians a starting point, with which to either agree or disagree, for their own clinical judgements.

Chapter 4

Parsing Epileptic Events in Detail

Despite over three decades of quantitative analysis, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy and the paucity of quantitative tools flexible enough to describe epileptic events but restrictive enough to distill intelligible information from them. While most of the machine learning and statistical modeling research in epilepsy has focused on discriminative problems of clinical importance like seizure prediction, seizure onset zone detection, or isolated event detection, generative statistical models of epileptic events like seizures have received almost no attention. Mechanistic models of epilepsy have a long history in the quantitative epilepsy literature, but these tend to be principle-driven rather than data-driven and thus of less use for answering questions from a given set of iEEG recordings. This absence of generative models most likely stems from the fact that training discriminative models is hard enough, so there seems little hope for the more difficult task of constructing generative models of epilepsy. Our still-poor understanding of how epilepsy develops and how seizures occur seems to diminish the prospects for generative models of epilepsy.

Though generative models of epilepsy as a disease may still be beyond our reach, we believe generative models of specific epileptic events like seizures are feasible thanks in part to much recent work in Bayesian modeling of complex time-series. The Bayesian setting allows us to construct models of complex events like seizures from much simpler models like autoregressive (AR) processes, Markov switching processes, and Bernoulli processes. We build a model meant to mimic in many ways what a human pays attention to when reading EEG: changes in the activity of individual channels and changes in the relationships between the channels. Not only does this modeling approach yield intuitive channel and event parsings, it also forces the model to interact with the data at a very low level, describing the original, raw iEEG data in the parlance of the clinicians who work with it most closely. We believe this approach yields models powerful enough to answer many different clinical questions rather than just one.

Epileptic events like seizures are by definition non-stationary. The activity within each iEEG channel and relationships between the channels changes over the course of the event. Time-varying AR processes have been proposed for single- [70] and multi-channel [93] EEG data, but we take a slightly different modeling approach that models a single channel's activity as switching between a set of locally stationary AR processes via the AR-HMM

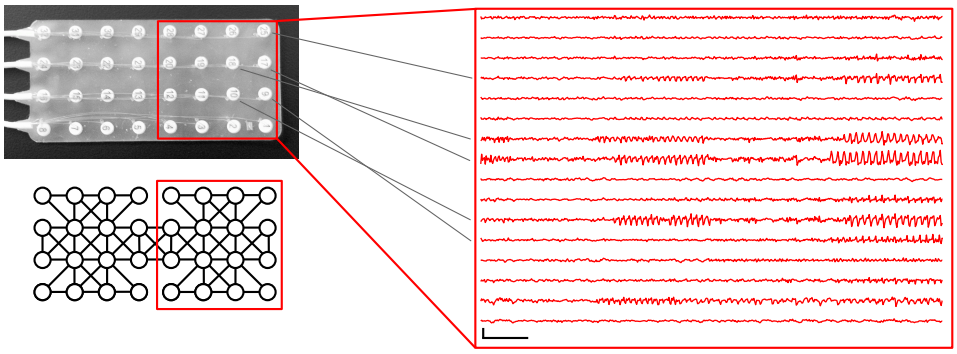

Figure 4.1. (**top left**) An iEEG grid electrode and (**bottom left**) corresponding graphical model encoding channel spatial adjacencies. (**right**) Residual EEG values *after* subtracting predictions from a BP-AR-HMM assuming independent channels. Scale bars indicate 1 mV vertically and 1 second horizontally.

(see Sections 2.5.2 and 2.6.4). As our iEEG almost always contains recordings from many channels at once, we wish to share the AR states between the channels while still allowing asynchronous state switching in each. (Synchronous state switching would imply a single vector-autoregressive process instead of the collection of separate scalar autoregressive processes we assume.) The beta process (BP) AR-HMM described in Section 2.6.5 allows for such sharing: a common library of infinitely many possible AR states is defined, though each time series uses only a finite subset of these states. This formulation 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 unrealistic. Figure 4.1 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. Accounting for these channel correlations incorporates one of the key iEEG features clinicians look for—the change in relationships between the channels—when parsing a seizure; it also improves our model on heldout seizures, as we show in Section 4.5.1.

Our innovations process involves a covariance term that encodes the spatial relationships between the iEEG channels. A straightforward approach would be to assume full covariance between the channels, implying that each channel may covary independently with every other channel. At the spatial scale of clinical iEEG, which has contacts roughly 1 cm apart from each other, we believe that full channel connectivity is not physiologically plausible (or at least not necessary). We thus restrict the spatial relationships available to our model by assuming a known graph structure encoding spatial adjacencies like that shown in the bottom left of Figure 4.1. While the 3-dimensional coordinates of channels are very difficult

to determine, the 2-dimensional relationships on the electrode grids or strips are trivial. As part of a Gaussian graphical model, adjacency graphs such as these encode conditional independencies in the multivariate innovations process across all the channels. The spatial graphs we employ encode channel adjacencies, with a few exceptions to make the graph fully decomposable.

It is well-known that the correlations between EEG channels generally vary during the beginning, middle, and end of a seizure [106, 107]. Prado et al. [94] employ a mixture-of-expert vector autoregressive models to describe the different dynamics present in seven channels of scalp EEG. We take a similar approach by allowing for Markov evolution for an underlying innovations covariance state.

We describe our model in detail in Section 4.1 and its posterior inference in Section 4.2. In Section 4.3 we give results from a simple experiment on simulated data. We explore seizure parsings produced by our model in Section 4.4 and compare it to other similar models in Section 4.5. We use the model in an initial exploration of the relationships between a set of sub-clinical epileptic bursts and a full clinical seizure in Section 4.6. In Section 4.7 we explore a number of practical considerations necessary for scaling our model and analysis to large datasets like that discussed in the next chapter. We conclude in Section 4.8 with a brief discussion of possible future directions for this work.

■ 4.1 A Markov Switching Process for Correlated Time Series

Observation model Consider an event with N univariate time series of length T. This event could be a seizure, where each time series is one of the iEEG voltage-recording channels. To avoid confusion between the N univariate time series which themselves make up a single T-dimensional event time series, we refer to the univariate time series as channels and multivariate time series as the event. We denote the scalar value for each channel i at each (discrete) time point t as $y_t^{(i)}$ and model it using an r-order AR-HMM (see Sections 2.5.2 and 2.6.4),

$$\begin{aligned} z_t^{(i)} &\sim \boldsymbol{\pi}^{(i)}_{z_{t-1}^{(i)}}, \\ y_t^{(i)} &= \mathbf{a}^{\mathrm{T}}_{z_t^{(i)}} \widetilde{\mathbf{y}}_t^{(i)} + \epsilon_t^{(i)}. \end{aligned} \tag{4.1}$$

Recall that in the AR-HMM, a channel i is modeled by a set of K independent dynamical states, each of which is an autoregressive (AR) model with parameters $\mathbf{a}_k$. We denote channel i's dynamical state at time t as $z_t^{(i)}$. Thus, $\boldsymbol{\pi}^{(i)}_{z_{t-1}^{(i)}}$ is the transition distribution given the state $z_{t-1}^{(i)}$ at the previous time point. We denote the vector of r previous observations as $\widetilde{\mathbf{y}}_t^{(i)} = \left(y_{t-1}^{(i)}, \ldots, y_{t-r}^{(i)}\right)^{\mathrm{T}}$.

Instead of evolving independently as in Fox et al. [42], the innovations $\boldsymbol{\epsilon}_t = \left(\epsilon_t^{(1)}, \ldots, \epsilon_t^{(N)}\right)^{\mathrm{T}}$

capture channel relationships through an event-state-specific correlations process,

$$\begin{aligned} Z_t &\sim \boldsymbol{\phi}_{Z_{t-1}}, \\ \boldsymbol{\epsilon}_t &\sim \mathcal{N}(\mathbf{0}, \Delta_{Z_t}), \end{aligned} \tag{4.2}$$

where Z_t denotes a Markov-evolving event state distinct from the individual channel states $\{z_t^{(i)}\}$. Similar to $\boldsymbol{\pi}^{(i)}_{z^{(i)}_{t-1}}$ for channel i, $\boldsymbol{\phi}_{Z_{t-1}}$ denotes the event's transition distribution given its state Z_{t-1} at the previous time point. Each event state covariance $\boldsymbol{\phi}_l$ describes a different set of channel relationships. Note that this formulation decouples the autoregressive predictions (influenced by the individual channel states) from the innovations (influenced by the event state). Not only does this decoupling allow for modeling a very broad range of activity. This flexibility 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 As in the AR-HMM, we place a multivariate normal prior on the AR coefficients,

$$\mathbf{a}_k \sim \mathcal{N}(\mathbf{m}_0, \Sigma_0), \tag{4.3}$$

with mean $\mathbf{m}_0$ and covariance Σ_0. Throughout this work, we let $\mathbf{m}_0 = \mathbf{0}$.

We define a sparse channel dependency structure on Δ_l by introducing a graphical model G and specifying a hyper-inverse Wishart (HIW) prior,

$$\Delta_l \sim \text{HIW}_G(b_0, D_0), \tag{4.4}$$

where b_0 denotes the degrees of freedom and D_0 the scale. As discussed in Section 2.3.7, the HIW prior enforces hyper-Markov conditions specified by G, leading to conditional independencies in $\boldsymbol{\epsilon}_t$ and thus also in $\mathbf{y}_t$. In addition to encoding the spatial proximities of iEEG electrodes, the HIW prior also yields a sparse precision matrix Δ_l^{-1}, allowing for more efficient scaling to the large numbers of channels commonly present in iEEG.

For compactness, we sometimes alternately write

$$\mathbf{y}_t = \mathbf{A}_{\mathbf{z}_t} \widetilde{\mathbf{Y}}_t + \boldsymbol{\epsilon}_t(Z_t), \tag{4.5}$$

where $\mathbf{y}_t$ is the concatenation of N channel observations at time t and $\mathbf{z}_t$ is the vector of concatenated channel states. One can think of this process as a factorial HMM [49] 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 $\boldsymbol{\epsilon}_t$ derived from the graphical model G of the channel relationships, as shown in Figure 4.1 (bottom left).

Feature constrained channel transition distributions We assume that each channel i exhibits some subset of a shared library of AR states with 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 described by $\mathbf{a}_k$. The BP-AR-HMM of Fox et al. [44] described in Section 2.6.6 provides this framework for defining this feature model to constrain the AR-HMM transitions. Through the beta process prior discussed Section 2.6.5, the BP-AR-HMM allows for an infinite library of AR states but encourages each channel to only use a sparse subset of them.

Following the formulation of the BP-AR-HMM, the feature assignments $f_k^{(i)}$ and their corresponding parameters $\mathbf{a}_k$ are generated by a beta process random measure and the conjugate Bernoulli process (BeP),

$$\begin{aligned} B &\sim \mathrm{BP}(1, B_0), \\ X^{(i)} &\sim \mathrm{BeP}(B), \end{aligned} \tag{4.6}$$

with base measure B_0 over our parameter space $\Theta \times [0,1]$, where $\Theta = \mathbb{R}^r$ for our r-order autoregressive parameters $\mathbf{a}_k$. The discrete measures B and $X^{(i)}$ can be represented as the infinite sum of point mass atoms ($\mathbf{a}_k \in \Theta, \omega_k \in [0,1]^K$) and $f_k^{(i)} \in \{0,1\}$, respectively,

$$\begin{aligned} B &= \sum_{k=1}^{\infty} \omega_k \delta_{\mathbf{a}_k}, \\ X^{(i)} &= \sum_{k=1}^{\infty} f_k^{(i)} \delta_{\mathbf{a}_k}, \end{aligned} \tag{4.7}$$

implying $f_k^{(i)} \sim \mathrm{Ber}(\omega_k)$. The resulting feature vectors $\mathbf{f}^{(i)}$ constrain the set of available states $z_t^{(i)}$ can take by constraining each transition distributions, $\boldsymbol{\pi}_j^{(i)}$, to be 0 when $f_k^{(i)} = 0$.

We use $\mathbf{f}^{(i)}$ along with a set of gamma random variables to produce the desired channel transition distribution $\boldsymbol{\pi}_j^{(i)}$ of state j,

$$\eta_{jk}^{(i)} \sim \mathrm{Gamma}(\gamma_c + \kappa_c \delta(j,k)) \tag{4.8}$$

$$\boldsymbol{\pi}_j^{(i)} = \frac{\boldsymbol{\eta}_j^{(i)} \circ \mathbf{f}^{(i)}}{\sum_{k | f_k^{(i)}=1} \eta_{jk}^{(i)}}. \tag{4.9}$$

The positive elements of $\boldsymbol{\pi}_j^{(i)}$ can also be thought of as a sample $\widetilde{\boldsymbol{\pi}}_j^{(i)}$ from a finite Dirichlet distribution with only $K^{(i)}$ dimensions, where $K^{(i)} = \sum_k f_k^{(i)}$ represents the number of states channel i uses. For convenience, we sometimes denote the set of transition distributions $\{\boldsymbol{\eta}_j^{(i)}\}_j$ as $\boldsymbol{\eta}^{(i)}$. As in the sticky HDP-HMM of Fox et al. [42], the parameter κ_c encourages self-transitions (i.e., state j at time $t-1$ to state j at time t).

Unconstrained event transition distributions For simplicity, we follow the standard HDP-HMM as described in Section 2.6.4 and do not constrain the event state transition distribution $\boldsymbol{\phi}_l$. For additional simplicity, we consider the weak limit approximation to the Dirichlet process [62] that involves a finite L number of dimensions,

$$\begin{aligned} \boldsymbol{\beta} &\sim \mathrm{Dir}(\gamma_{\mathrm{e}}/L, \ldots, \gamma_{\mathrm{e}}/L), \\ \boldsymbol{\phi}_l &\sim \mathrm{Dir}(\alpha_{\mathrm{e}}\boldsymbol{\beta} + \kappa_{\mathrm{e}}\mathbf{e}_l). \end{aligned} \tag{4.10}$$

Again, the sticky parameter κ_{e} promotes self-transitions, reducing state redundancy.

We term this model the HIW-spatial BP-AR-HMM, though the correlations in the channels i induced by the graph G need not necessarily be the spatial correlations that they are for our seizure modeling application. The complete model formulation can be specified as

$$\begin{aligned} B &\sim \mathrm{BP}(1, B_0), \\ X^{(i)} &\sim \mathrm{BeP}(B) && i = 1, \ldots, N \\ \boldsymbol{\eta}_j^{(i)} &\sim \mathrm{Dir}(\gamma, \ldots, \gamma + \kappa_c, \ldots), && j = 1, 2, \ldots,\ i = 1, \ldots, N, \\ \mathbf{a}_k &\sim \mathcal{N}(\mathbf{m}_0, \Sigma_0), \\ \boldsymbol{\beta} &\sim \mathrm{Dir}(\gamma_e/L, \ldots, \gamma_e/L), \\ \boldsymbol{\phi}_l &\sim \mathrm{Dir}(\alpha_e\boldsymbol{\beta} + \kappa_e\mathbf{e}_l), && l = 1, \ldots, L, \\ \Delta_l &\sim \mathrm{HIW}_G(b_0, D_0), \\ Z_t &\sim \boldsymbol{\phi}_{Z_{t-1}}, && t = 1, \ldots, T \\ \boldsymbol{\epsilon}_t &\sim \mathcal{N}(\mathbf{0}, \Delta_{Z_t}), && t = 1, \ldots, T \\ z_t^{(i)} &\sim \boldsymbol{\pi}_{z_{t-1}^{(i)}} && t = 1, \ldots, T,\ i = 1, \ldots, N, \\ y_t^{(i)} &= \tilde{\mathbf{y}}_t^{(i)} \mathbf{a}_{z_t^{(i)}}^{\mathrm{T}} + \epsilon_t^{(i)}, && t = 1, \ldots, T,\ i = 1, \ldots, N \end{aligned} \tag{4.11}$$

We depict this model in the directed acyclic graphs shown in Figure 4.2. Note that while we formally consider a model of only single event for notational simplicity, our formulation scales straightforwardly to multiple independent events. In this case, each event e of length T_e would have its own state sequence $Z_{1:T_e}^{(e)}$ and transition distributions $\boldsymbol{\phi}^{(e)}$.

■ 4.2 MCMC Posterior Inference

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. Yet the underlying graph structure of the channel dependencies mitigates the scale of these challenges. Conditioned on channel sequences $\{\mathbf{z}_{1:T}^{(\mathbf{i}')}\}$, we *can* marginalize $z_{1:T}^{(i)}$; because of the graph

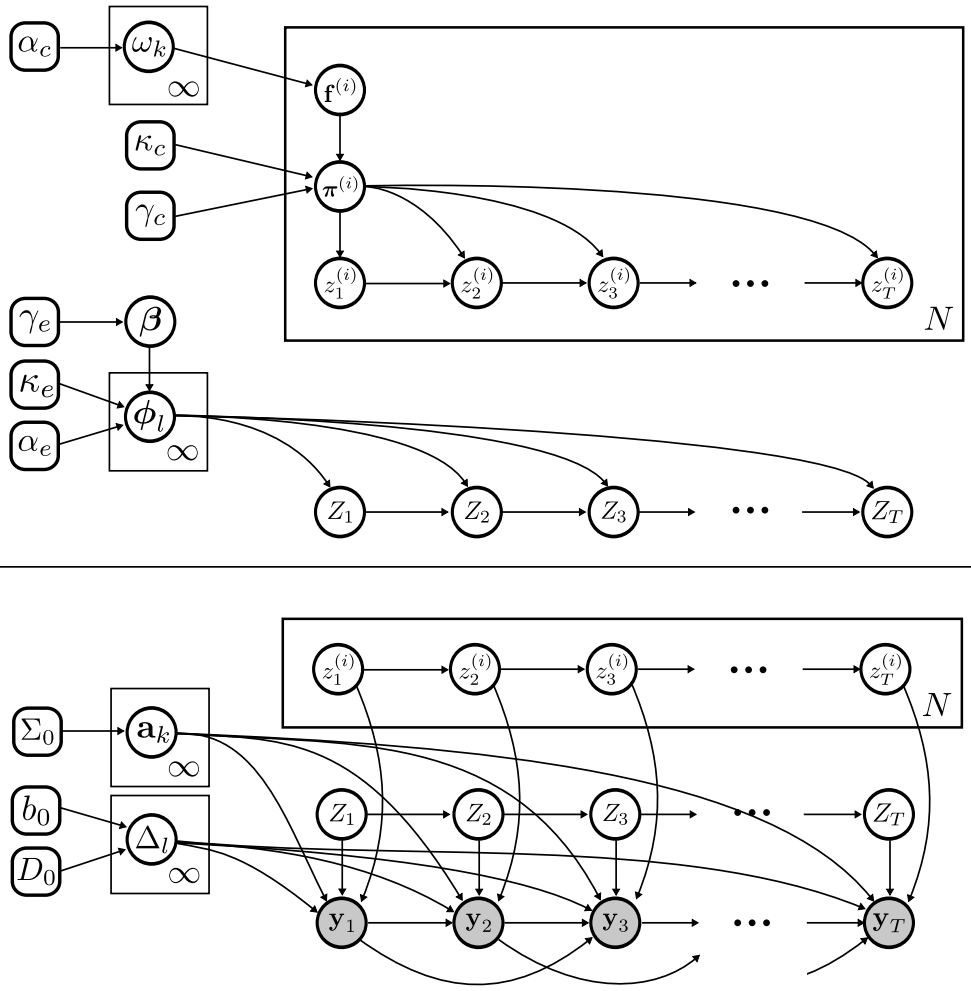

Figure 4.2. Graphical models of the lower and upper portions of the HIW-spatial BP-AR-HMM shown separately for simplicity. (**top**) The event state Z_t evolves independently of each channel i's state $z_t^{(i)}$ according to transition distributions $\{\boldsymbol{\phi}_l\}$, which are coupled by global transition distribution $\boldsymbol{\beta}$. The channel i feature indicators $\mathbf{f}^{(i)}$ are samples from a Bernoulli process with weights $\{\omega_k\}$, and they constrain the channel transition distributions $\boldsymbol{\pi}^{(i)}$. (**bottom**) 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.

Algorithm 10 HIW-spatial BP-AR-HMM master MCMC sampler

for each MCMC iteration **do**
  get a random permutation of the channel indices,
    $\mathbf{h} \in {}^{n}\mathrm{P}_{n}$
  for each channel $i \in \mathbf{h}$ **do**
    sample feature indicators $\mathbf{f}^{(i)}$ as in Algorithms 11 and 12
    sample state sequence $z_{1:T}^{(i)}$ as in Algorithm 2 using the feature constrained finite
      transition distribution $\widetilde{\boldsymbol{\pi}}^{(i)} \in \mathbb{R}^{K^{(i)} \times K^{(i)}}$ and likelihoods matrix $\mathbf{u}^{(i)} \in \mathbb{R}_{+}^{(K^{(i)}+1) \times T}$
    sample state transition parameters $\boldsymbol{\eta}^{(i)}$ as in Algorithm 13
  end for
  sample event states sequence $Z_{1:T}$ as in Algorithm 2 using the transition distribution
    $\boldsymbol{\phi} \in \mathbb{R}^{L \times L}$ and likelihoods matrix $\mathbf{v} \in \mathbb{R}^{L \times T}$
  sample event state transition parameters $\boldsymbol{\phi}$ as in Algorithm 14
  sample observation model parameters $(\{\mathbf{a}_k\}, \{\Delta_l\})$ as in Algorithm 15
  (sample hyperparameters as given in Equations (4.33), (4.34), (4.35), (4.36), and (4.38))
end for

structure, we need only condition on a *sparse* set of other channels $\mathbf{i}'$ (i.e., neighbors in the graph).

On a high level, each MCMC iteration proceeds through sampling individual channel states, individual events states, observation model parameters, channel dynamics model parameters, event dynamics model parameters, and hyperparameters. Algorithm 10 summarizes these steps and links to explicit algorithms for each step, all of which are described in detail below. In the notation below, we omit hyperparameters in the conditionals for compactness.

4.2.1 Individual Channel Variables

Sampling the variables associated with the individual channel i involves first sampling active features $\mathbf{f}^{(i)}$ (while marginalizing $z_{1:T}^{(i)}$), then conditioning on these feature assignments $\mathbf{f}^{(i)}$ to block sample the state sequence $z_{1:T}^{(i)}$, and finally sampling the transition distribution $\boldsymbol{\pi}^{(i)}$ given the feature indicators $\mathbf{f}^{(i)}$ and state sequence $z_{1:T}^{(i)}$.

Channel marginal likelihood Let $\mathbf{i}' \subseteq \{1, \ldots, N\}$ index the neighboring channels in the graph upon which channel i is conditioned. The conditional likelihood of observation $y_t^{(i)}$ under AR model k given the neighboring observations $\mathbf{y}_t^{(\mathbf{i}')}$ at time t is

$$p\left(y_t^{(i)} \mid \widetilde{\mathbf{y}}_t^{(i)}, \mathbf{y}_t^{(\mathbf{i}')}, z_t^{(i)} = k, \mathbf{z}_t^{(\mathbf{i}')}, Z_t, \{\mathbf{a}_k\}, \{\Delta_l\}\right) \propto \mathcal{N}\left(\widetilde{\mu}_t, \widetilde{\sigma}_t^2\right) \tag{4.12}$$

for

$$\begin{aligned} \widetilde{\mu}_t &= \mathbf{a}_k^T \widetilde{\mathbf{y}}_t^{(i)} + \Delta_{Z_t}^{(i,\mathbf{i}')} \Delta_{Z_t}^{-1(\mathbf{i}',\mathbf{i}')} \left(\mathbf{y}_t^{(\mathbf{i}')} - \mathbf{A}_{\mathbf{z}^{(\mathbf{i}')}} \widetilde{\mathbf{Y}}_t^{(\mathbf{i}')}\right), \\ \widetilde{\sigma}_t^2 &= \Delta_{Z_t}^{(i,i)} - \Delta_{Z_t}^{(i,\mathbf{i}')} \Delta_{Z_t}^{-1(\mathbf{i}',\mathbf{i}')} \Delta_{Z_t}^{(\mathbf{i}',i)}, \end{aligned} \tag{4.13}$$

which follows from the conditional distribution of the multivariate normal [47, pg. 579]. Using the sum-product algorithm to marginalize over the exponentially many state sequences $z_{1:T}^{(i)}$ as described in Section 2.5.2 and Algorithm 1, we can calculate the channel marginal likelihood,

$$p\left(y_{1:T}^{(i)} \mid \mathbf{y}_{1:T}^{(\mathbf{i}')}, \mathbf{z}_{1:T}^{(\mathbf{i}')}, Z_{1:T}, \mathbf{f}^{(i)}, \boldsymbol{\eta}^{(i)}, \{\mathbf{a}_k\}, \{\Delta_l\}\right), \tag{4.14}$$

of channel i's observations over all $t = 1, \ldots, T$ given the observations $\mathbf{y}_{1:T}^{(\mathbf{i}')}$ and the assigned states $\mathbf{z}_{1:T}^{(\mathbf{i}')}$ of neighboring channels $\mathbf{i}'$ and given the event state sequence $Z_{1:T}$. As previously discussed, taking the non-zero elements of the infinite-dimensional transition distributions $\boldsymbol{\pi}^{(i)}$, derived from $\mathbf{f}^{(i)}$ and $\boldsymbol{\eta}^{(i)}$ as in Equation (4.9), yields a set of $K^{(i)}$-dimensional active feature transition distributions $\widetilde{\boldsymbol{\pi}} = \{\widetilde{\boldsymbol{\pi}}_j\}$, reducing this marginalization to a series of matrix-vector products.

Sampling active features, $\mathbf{f}^{(i)}$ We briefly describe the active feature sampling scheme given in detail by Fox et al. [41]. Recall that for our HIW-spatial BP-AR-HMM, we need to condition on neighboring channel state sequences $\mathbf{z}_{1:T}^{(\mathbf{i}')}$ and event state sequences $Z_{1:T}$. Sampling the feature indicators $\mathbf{f}^{(i)}$ for channel i via the Indian buffet process (IBP) involves considering those features shared by other channels and those unique to channel i. Let $K_+ = \sum_{k=1}^{K} \mathbf{1}\left(f_k^{(1)} \vee \cdots \vee f_k^{(N)}\right)$ denote the total number of active features used by at least one of the channels. We consider the set of shared features across channels not including those specific to channel i as $\mathcal{S}^{(-i)} \subseteq \{1, \ldots, K_+\}$ and the set of unique features for channel i as $\mathcal{U}^{(i)} \subseteq \{1, \ldots, K_+\}/\mathcal{S}^{(-i)}$.

Shared features The posterior for each shared feature $k \in \mathcal{S}^{(-i)}$ for channel i is given by

$$p\left(f_k^{(i)} \mid y_{1:T}^{(i)}, \mathbf{y}_{1:T}^{(\mathbf{i}')}, \mathbf{z}_{1:T}^{(\mathbf{i}')}, Z_{1:T}, \mathbf{f}_k^{(-i)}, \mathbf{f}_{k' \neq k}^{(i)}, \boldsymbol{\eta}^{(i)}, \{\mathbf{a}_k\}, \{\Delta_l\}\right) \propto p\left(f_k^{(i)} \mid \mathbf{f}_k^{(-i)}\right) p\left(y_{1:T}^{(i)} \mid \mathbf{y}_{1:T}^{(\mathbf{i}')}, \mathbf{z}_{1:T}^{(\mathbf{i}')}, Z_{1:T}, \mathbf{f}_{k' \neq k}^{(i)}, f_k^{(i)}, \boldsymbol{\eta}^{(i)}, \{\mathbf{a}_k\}, \{\Delta_l\}\right). \tag{4.15}$$

Recalling the form of the IBP posterior predictive distribution from Section 2.6.5, we have $p\left(f_k^{(i)} = 1 \mid \mathbf{f}_k^{(-i)}\right) = m_k^{(-i)}/N$, where $m_k^{(-i)}$ denotes the number of other channels that use feature k. We use this posterior to formulate a Metropolis-Hastings proposal that flips the current indicator value $f_k^{(i)}$ to its complement $\bar{f}_k^{(i)}$ with probability $\rho(\bar{f}_k^{(i)} \mid f_k^{(i)})$,

$$f_k^{(i)} = \begin{cases} \bar{f}_k^{(i)}, & \text{w.p.} \quad \rho(\bar{f}_k^{(i)} \mid f_k^{(i)}), \\ f_k^{(i)}, & \text{otherwise}, \end{cases} \tag{4.16}$$

where

$$\rho(\bar{f}_k^{(i)} \mid f_k^{(i)}) = \min\left(\frac{p\left(\bar{f}_k^{(i)} \mid y_{1:T}^{(i)}, \mathbf{y}_{1:T}^{(\mathbf{i}')}, \mathbf{z}_{1:T}^{(\mathbf{i}')}, Z_{1:T}, \mathbf{f}_k^{(-i)}, \mathbf{f}_{k' \neq k}^{(i)}, \boldsymbol{\eta}^{(i)}, \{\mathbf{a}_k\}, \{\Delta_l\},\right)}{p\left(f_k^{(i)} \mid y_{1:T}^{(i)}, \mathbf{y}_{1:T}^{(\mathbf{i}')}, \mathbf{z}_{1:T}^{(\mathbf{i}')}, Z_{1:T}, \mathbf{f}_k^{(-i)}, \mathbf{f}_{k' \neq k}^{(i)}, \boldsymbol{\eta}^{(i)}, \{\mathbf{a}_k\}, \{\Delta_l\},\right)}, 1\right). \tag{4.17}$$

Algorithm 11 gives an explicit recipe for sampling these shared features for each channel.

Algorithm 11 HIW-spatial BP-AR-HMM MCMC sampler for shared channel features

let i denote a particular channel index
sample a new channel state from prior for potential feature birth
  $\mathbf{a'}_+ \sim \mathcal{N}(\mathbf{0}, \Sigma_0)$
store likelihoods under the K existing channel states for each time point as in Equation (4.12)
  $u_{t,k}^{(i)} \leftarrow p\left(y_t^{(i)} \mid \widetilde{\mathbf{y}}_t^{(i)}, \mathbf{y}_t^{(\mathbf{i'})}, z_t^{(i)} = k, \mathbf{z}_t^{(\mathbf{i'})}, Z_t, \{\mathbf{a}_k\}, \{\Delta_l\}\right) \qquad t = 1, \ldots, T, \quad k = 1, \ldots, K$
store likelihoods under the potential birth feature
  $u_{t,K+1}^{(i)} \leftarrow p\left(y_t^{(i)} \mid \widetilde{\mathbf{y}}_t^{(i)}, \mathbf{y}_t^{(\mathbf{i'})}, z_t^{(i)} = K+1, \mathbf{z}_t^{(\mathbf{i'})}, Z_t, \{\mathbf{a}_k\}, \mathbf{a'}_+, \{\Delta_l\}\right) \qquad t = 1, \ldots, T$
store the current marginal likelihood given the feature indicators $\mathbf{f}^{(i)}$ as in Equation (4.14)
  $\ell_{\mathbf{f}^{(i)}}^{(i)} \leftarrow p\left(y_{1:T}^{(i)} \mid \mathbf{u}^{(i)}, \mathbf{f}^{(i)}, \boldsymbol{\eta}^{(i)}\right)$

get set of shared features
  $\mathcal{S}^{(-i)} = \{k \mid f_k^{(i')} = 1,\ i' = 1, \ldots, i-1, i+1, \ldots, N,\ k = 1, \ldots, K_+\}$

for each shared feature $k \in \mathcal{S}^{(-i)}$ **do**
    propose new feature indicator $\bar{\mathbf{f}}^{(i)}$ by flipping kth indicator to its complement
      $\bar{\mathbf{f}}^{(i)} \leftarrow \left(f_1^{(i)}, \ldots, f_{k-1}^{(i)}, \neg f_k^{(i)}, f_{k+1}^{(i)}, \ldots, f_K^{(i)}\right)$
    store the marginal likelihood under proposed feature indicators $\bar{\mathbf{f}}^{(i)}$ as in Equation (4.14)
      $\ell_{\bar{\mathbf{f}}^{(i)}}^{(i)} \leftarrow p\left(y_{1:T}^{(i)} \mid \mathbf{u}^{(i)}, \bar{\mathbf{f}}^{(i)}, \boldsymbol{\eta}^{(i)}\right)$
    calculate proposal ratio $\rho(\bar{f}^{(i)} \mid f^{(i)})$ for accepting $\bar{f}^{(i)}$ as in Equation (4.17)
      $\rho(\bar{f}^{(i)} \mid f^{(i)}) \leftarrow \min\left(\frac{\ell_{\bar{\mathbf{f}}^{(i)}}}{\ell_{\mathbf{f}^{(i)}}} \left(\frac{m_k^{(-i)}}{N - m_k^{(i)}}\right)^{1 - 2f_k^{(i)}}, 1\right)$
    flip feature indicator $f_k^{(i)}$ with probability $\rho(\bar{f}^{(i)} \mid f^{(i)})$
      $a \sim \mathrm{Ber}(\rho(\bar{f}^{(i)} \mid f^{(i)}))$
    if $a = 1$ **then**
      $f_k^{(i)} \leftarrow \bar{f}_k^{(i)}$
      $\ell_{\mathbf{f}^{(i)}}^{(i)} \leftarrow \ell_{\bar{\mathbf{f}}^{(i)}}^{(i)}$
    end if
end for

continue to Algorithm 12 for unique feature sampling

Unique features We either propose a new feature or remove a unique feature for channel i using a birth and death reversible jump MCMC sampler [21, 51, 98] (see Fox et al. [44] for details relevant to the BP-AR-HMM). We denote the number of unique features for channel i as $n^{(i)} = |\mathcal{U}^{(i)}|$. We define the vector of shared feature indicators as $\mathbf{f}_-^{(i)} = \mathbf{f}^{(i)}_{k'|k' \in \mathcal{S}^{(-i)}}$ and that for unique feature indicators as $\mathbf{f}_+^{(i)} = \mathbf{f}^{(i)}_{k'|k' \in \mathcal{U}^{(i)}}$, which together $\left(\mathbf{f}_-^{(i)}, \mathbf{f}_+^{(i)}\right)$ define the full feature indicator vector $\mathbf{f}^{(i)}$ for channel i. Similarly, $\mathbf{a}_+^{(i)}$ and $\boldsymbol{\eta}_+^{(i)}$ describe the model dynamics and transition parameters associated with these unique features. We propose a new unique feature vector $\mathbf{f'}_+$ and corresponding model dynamics $\mathbf{a'}_+$ and transition parameters $\boldsymbol{\eta'}_+$ (sampled from their priors in the case of feature birth) with a proposal distribution of

$$p\left(\mathbf{f}'_+, \mathbf{a}'_+, \boldsymbol{\eta}'_+ \mid \mathbf{f}_+^{(i)}, \{\mathbf{a}_k\}_+^{(i)}, \boldsymbol{\eta}_+^{(i)}\right) = p\left(\mathbf{f}'_+ \mid \mathbf{f}_+^{(i)}\right) p\left(\mathbf{a'}_+ \mid \mathbf{f}'_+, \mathbf{f}_+^{(i)}, \{\mathbf{a}_k\}_+^{(i)}\right) p\left(\boldsymbol{\eta'}_+ \mid \mathbf{f}'_+, \mathbf{f}_+^{(i)}, \boldsymbol{\eta}_+^{(i)}\right). \quad (4.18)$$

A new unique feature is proposed with probability 0.5 and each existing unique feature is removed with probability $0.5/n^{(i)}$. This proposal is accepted with probability

$$\rho\left(\mathbf{f}'_+, \mathbf{a}'_+, \boldsymbol{\eta}'_+ \mid \mathbf{f}_+^{(i)}, \{\mathbf{a}_k\}_+^{(i)}, \boldsymbol{\eta}_+^{(i)}\right) = \min\left(\frac{p\left(y_{1:T}^{(i)} \mid \mathbf{y}_{1:T}^{(\mathbf{i}')}, \mathbf{z}_{1:T}^{(\mathbf{i}')}, [\mathbf{f}_-^{(i)}\ \mathbf{f}'_+], \boldsymbol{\eta}^{(i)}, \boldsymbol{\eta}'_+, \{\mathbf{a}_k\}, \{\Delta_l\}\right)}{p\left(y_{1:T}^{(i)} \mid \mathbf{y}_{1:T}^{(\mathbf{i}')}, \mathbf{z}_{1:T}^{(\mathbf{i}')}, [\mathbf{f}_-^{(i)}\ \mathbf{f}_+^{(i)}], \boldsymbol{\eta}^{(i)}, \{\mathbf{a}_k\}, \{\Delta_l\}\right)} \cdot \frac{\text{Poisson}\left(n'_i \mid \alpha_c/N\right)}{\text{Poisson}\left(n_i \mid \alpha_c/N\right)} \frac{p\left(\mathbf{f}_+^{(i)} \mid \mathbf{f}'_+\right)}{p\left(\mathbf{f}'_+ \mid \mathbf{f}_+^{(i)}\right)}, 1\right). \quad (4.19)$$

Algorithm 12 gives an explicit recipe for sampling these unique features for each channel.

Chanel state sequence, $z_{1:T}^{(i)}$ We sample the state sequence for all the time points of channel i, given that channel's feature-constrained transition distributions $\boldsymbol{\pi}^{(i)}$, the state parameters $\{\mathbf{a}_k\}$, the observations $y_{1:T}^{(i)}$, and the neighboring observations $\mathbf{y}_{1:T}^{(\mathbf{i}')}$ and current states $\mathbf{z}_{1:T}^{(\mathbf{i}')}$. The joint probability of the state sequence $z_{1:T}^{(i)}$ is given by

$$p\left(z_{1:T}^{(i)} \mid y_{1:T}^{(i)}, \mathbf{y}_{1:T}^{(\mathbf{i}')}, \mathbf{z}_{1:T}^{(\mathbf{i}')}, \mathbf{f}^{(i)}, \boldsymbol{\eta}^{(i)}, \{\mathbf{a}_k\}, \{\Delta_l\}\right) = p\left(z_1^{(i)} \mid y_1^{(i)}, \mathbf{y}_1^{(\mathbf{i}')}, \mathbf{z}_1^{(\mathbf{i}')}, \mathbf{f}^{(i)}, \boldsymbol{\eta}^{(i)}, \{\mathbf{a}_k\}, \{\Delta_l\}\right) \cdot \prod_{t=2}^{T} p\left(z_t^{(i)} \mid y_{t:T}^{(i)}, \mathbf{y}_{t:T}^{(\mathbf{i}')}, z_{t-1}^{(i)}, \mathbf{z}_{t:T}^{(\mathbf{i}')}, \mathbf{f}^{(i)}, \boldsymbol{\eta}^{(i)}, \{\mathbf{a}_k\}, \{\Delta_l\}\right). \quad (4.20)$$

Algorithm 12 HIW-spatial BP-AR-HMM MCMC sampler for unique channel features

continued from Algorithm 11

get set of unique features for channel i

$\mathcal{U}^{(i)} = \{k \mid f_k^{(i)} = 1, k \notin \mathcal{S}^{(-i)}\}$

$n^{(i)} = |\mathcal{U}^{(i)}|$

initialize proposed unique feature indicators to current indicators

$\mathbf{f'}_+ \leftarrow \mathbf{f}_+^{(i)}$

propose a birth ($a = 1$) or death ($a = 0$) move

$a \sim \text{Ber}(0.5)$

if $a = 1$ **then**

$f'_{+_{n^{(i)}+1}} = 1$

$n'^{(i)} \leftarrow n^{(i)} + 1$

else

$k' \sim \text{Multi}(1/n^{(i)}, \ldots, 1/n^{(i)})$

$f'_{+_{k'}} = 0$

end if

store the marginal likelihood under proposed feature indicators $(\mathbf{f}_-, \mathbf{f'}_+)$ as in Equation (4.14)

$\ell^{(i)}_{\mathbf{f}'^{(i)}} \leftarrow p\left(y^{(i)}_{1:T} \mid \mathbf{u}^{(i)}, (\mathbf{f}_-, \mathbf{f'}_+), \boldsymbol{\eta}^{(i)}\right)$

calculate proposal ratio $\rho(\mathbf{f'}_+, \mathbf{a'}_+, \boldsymbol{\eta'}_+ \mid \mathbf{f}_+^{(i)}, \{\mathbf{a}_k\}_+^{(i)}, \boldsymbol{\eta}_+^{(i)})$ as in Equation 4.19

$$\rho(\mathbf{f'}_+, \mathbf{a'}_+, \boldsymbol{\eta'}_+ \mid \mathbf{f}_+^{(i)}, \{\mathbf{a}_k\}_+^{(i)}, \boldsymbol{\eta}_+^{(i)}) \leftarrow \min\left(\frac{\ell^{(i)}_{\mathbf{f}'^{(i)}}}{\ell^{(i)}_{\mathbf{f}^{(i)}}} \frac{\text{Poisson}(n'^{(i)} \mid \alpha_c/N)}{\text{Poisson}(n^{(i)} \mid \alpha_c/N)} \left(\frac{0.5/n'^{(i)}}{0.5}\right)^{2a-1}, 1\right)$$

accept proposed unique features $\mathbf{f}_+$ with probability $\rho(\mathbf{f'}_+, \mathbf{a'}_+, \boldsymbol{\eta'}_+ \mid \mathbf{f}_+^{(i)}, \{\mathbf{a}_k\}_+^{(i)}, \boldsymbol{\eta}_+^{(i)})$

$b \sim \text{Ber}(\rho(\mathbf{f'}_+, \mathbf{a'}_+, \boldsymbol{\eta'}_+ \mid \mathbf{f}_+^{(i)}, \{\mathbf{a}_k\}_+^{(i)}, \boldsymbol{\eta}_+^{(i)}))$

if $b = 1$ **then**

update unique feature indicators and current marginal likelihood

$\mathbf{f}_+^{(i)} \leftarrow \mathbf{f'}_+$

$\ell^{(i)}_{\mathbf{f}^{(i)}} \leftarrow \ell^{(i)}_{\mathbf{f}'^{(i)}}$

if $a = 1$ **then**

add new feature to set of existing features

$\{\mathbf{a}_k\} \leftarrow \{\mathbf{a}_k\} \cup \mathbf{a'}_+$

make space for new birth state in all feature indicators and transition distributions

for $i' = 1, \ldots, N$ **do**

if $i' \neq i$ **then**

$\mathbf{f}^{(i')} \leftarrow (\mathbf{f}^{(i')}, 0)$

end if

sample new $\{\boldsymbol{\eta}_j^{(i')}\}$ for each j as in Equation (4.23)

end for

end if

end if

Algorithm 13 HIW-spatial BP-AR-HMM MCMC sampler for channel state trans. params.

for $i \in \{1, \ldots, N\}$ **do**
  count the number of times channel i transitions from state j to state k
    $n_{jk}^{(i)} \leftarrow 0$ $\quad j = 1, \ldots, K \quad k = 1, \ldots, K$
  for $t \in 2, \ldots, T$ **do**
    $n^{(i)}_{z_{t-1}^{(i)} z_t^{(i)}} \leftarrow n^{(i)}_{z_{t-1}^{(i)} z_t^{(i)}} + 1$
  end for
  sample each transition distribution using the transition counts column vector $\mathbf{n}_j^{(i)} = \mathbf{n}_{j*}^{\mathrm{T}(i)}$
  for $l \in \{1, \ldots, K\}$ **do**
    $\bar{\boldsymbol{\eta}}_j \sim \mathrm{Dir}\left(\gamma_c + \kappa_c \mathbf{e}_j + \mathbf{n}_j^{(i)}\right)$
    $C_l^{(i)} \sim \mathrm{Gamma}(K\gamma_c + \kappa_c, 1)$
    $\boldsymbol{\eta}_j^{(i)} = C_j^{(i)} \bar{\boldsymbol{\eta}}_j^{(i)}$
  end for
end for

Again following the sum-product algorithm (see Section 2.5.2 and Algorithm 2), at each time point t we sample a state whose conditional probability is given by

$$p\left(z_t^{(i)} \mid y_{t:T}^{(i)}, \mathbf{y}_{t:T}^{(\mathbf{i}')}, z_{t-1}^{(i)}, Z_{1:T}, \mathbf{z}_{t:T}^{(\mathbf{i}')}, \mathbf{f}^{(i)}, \boldsymbol{\eta}^{(i)}, \{\mathbf{a}_k\}, \{\Delta_l\}\right) \propto \mathrm{Multi}\left(\widetilde{\boldsymbol{\pi}}^{(i)}_{z_{t-1}^{(i)}} \circ \mathbf{u}_t^{(i)} \circ \boldsymbol{\psi}_t\right), \tag{4.21}$$

where $\widetilde{\boldsymbol{\pi}}^{(i)}_{z_{t-1}^{(i)}}$ is the transition distribution given the assigned state at $t-1$, $\mathbf{u}_t^{(i)} \in \mathbb{R}^{K^{(i)}}$ is the vector of likelihoods under each possible state at time t (as in Equation (4.12)), and $\boldsymbol{\psi}_t \in \mathbb{R}^{K^{(i)}}$ is the vector of backwards messages from time point $t+1$ to t.

Channel transition parameters, $\boldsymbol{\eta}^{(i)}$ Following the correction described by Hughes et al. [58, Supplement], the posterior for the transition variable $\eta_{jk}^{(i)}$ is given by

$$p(\eta_{jk}^{(i)} \mid z_{1:T}^{(i)}, f_k^{(i)}) \propto \frac{(\eta_{jk}^{(i)})^{n_{jk}^{(i)} + \gamma_c + \kappa_c \delta(j,k) - 1} e^{\eta_{jk}^{(i)}}}{\sum_{k' | f_{k'}^{(i)} = 1} \eta_{jk'}^{(i)}}, \tag{4.22}$$

where $n_{jk}^{(i)}$ denotes the number of times channel i transitions from state j to state k. We can sample from this posterior via two auxiliary variables,

$$\begin{aligned} \bar{\boldsymbol{\eta}}_j^{(i)} &\sim \mathrm{Dir}(\gamma_c + \kappa_c \mathbf{e}_j + \mathbf{n}_j^{(i)}) \\ C_j^{(i)} &\sim \mathrm{Gamma}(K\gamma_c + \kappa_c, 1) \\ \boldsymbol{\eta}_j^{(i)} &= C_j^{(i)} \bar{\boldsymbol{\eta}}_j^{(i)}. \end{aligned} \tag{4.23}$$

Algorithm 13 gives an explicit recipe for this transition variable sampling.

■ 4.2.2 Event State

Since we model the event state process with a (truncated approximation to the) HDP-HMM, inference is more straightforward than with the channel states. We block sample the event state sequence $Z_{1:T}$ and then sample the event state transition distributions $\boldsymbol{\phi}$.

Event marginal likelihood Let $\mathbf{z}_t$ denote the vector of N channel states at time t. Since the space of $\mathbf{z}_t$ is exponentially large, we cannot integrate it out to compute the marginal conditional likelihood of the data given the event state sequence $Z_{1:T}$ (and model parameters). Instead, we consider the conditional likelihood of an observation at time t given channel states $\mathbf{z}_t$ and event state $Z_t = l$,

$$p(\mathbf{y}_t \mid \widetilde{\mathbf{Y}}_t, \mathbf{z}_t, Z_t = l, \{\mathbf{a}_k\}, \Delta_l) \propto \mathcal{N}(\mathbf{A}_{\mathbf{z}_t} \widetilde{\mathbf{Y}}_t, \Delta_l). \tag{4.24}$$

Recalling Equation (4.5), we see that this conditional likelihood of $\mathbf{y}_t$ is equivalent to a zero-mean multivariate normal model on the channel innovations $\boldsymbol{\epsilon}_t$,

$$p(\boldsymbol{\epsilon}_t \mid Z_t = l, \Delta_l) \propto \mathcal{N}(\mathbf{0}, \Delta_l).$$

As with the channel marginal likelihoods, we use the sum-product algorithm to marginalize over the possible event state sequences $Z_{1:T}$ as described in Section 2.5.2 and Algorithm 1, yielding a likelihood conditional on the channel states $\mathbf{z}_t$ and autoregressive parameters $\{\mathbf{a}_k\}$, in addition to the event transition distribution $\boldsymbol{\phi}$ and event state covariances $\{\Delta_l\}$,

$$p\left(\mathbf{y}_{1:T} \mid \mathbf{z}_{1:T}, \boldsymbol{\phi}, \{\mathbf{a}_k\}, \{\Delta_l\}\right). \tag{4.25}$$

Event state sequence, $Z_{1:T}$ The mechanics of sampling the event state sequence $Z_{1:T}$ directly parallel those of sampling the individual channel state sequences $z^{(i)}_{1:T}$. The joint probability of the event state sequence is given by

$$\begin{aligned} p(Z_{1:T} \mid \mathbf{y}_{1:T}, \mathbf{z}_{1:T}, \boldsymbol{\phi}, \{\mathbf{a}_k\}, \{\Delta_l\}) \propto \\ p(Z_1 \mid \mathbf{y}_1, \mathbf{z}_1, \boldsymbol{\phi}, \{\mathbf{a}_k\}, \{\Delta_l\}) \prod_{t=2}^{T} p(Z_t \mid \mathbf{y}_{t:T}, \mathbf{z}_{t:T}, Z_{t-1}, \boldsymbol{\phi}, \{\mathbf{a}_k\}, \{\Delta_l\}). \end{aligned} \tag{4.26}$$

We again use the sum-product algorithm (see Algorithm 2) to block sample each event state whose conditional probability distribution over the L states is given by

$$p\left(Z_t \mid \widetilde{\mathbf{Y}}_t, \mathbf{z}_t, \boldsymbol{\phi}, \{\mathbf{a}_k\}, \{\Delta_l\}\right) \propto \text{Multi}\left(\boldsymbol{\phi}_{Z_{t-1}} \circ \mathbf{v}_t \circ \boldsymbol{\psi}_t\right), \tag{4.27}$$

where $\boldsymbol{\phi}_{Z_{t-1}}$ is the transition distribution given the assigned state at $t-1$, $\mathbf{v}_t \in \mathbb{R}^L$ is the vector of likelihoods under each of the L possible states at time t (as in Equation (4.24)), and $\boldsymbol{\psi}_t \in \mathbb{R}^L$ is again the vector of backwards messages from time point $t+1$ to t.

Algorithm 14 HIW-spatial BP-AR-HMM MCMC sampler for event state trans. params.

count the number of times the event transitions from state l to state m
  $n_{lm} \leftarrow 0$ $\quad l = 1, \ldots, L \quad m = 1, \ldots, L$
for $t \in \{2, \ldots, T\}$ **do**
  $n_{Z_{t-1} Z_t} \leftarrow n_{Z_{t-1} Z_t} + 1$
end for

sample each transition distribution using the transition counts column vector $\mathbf{n}_l = \mathbf{n}_{l*}^{\mathrm{T}}$
for $l \in \{1, \ldots, L\}$ **do**
  $\boldsymbol{\phi}_l \sim \mathrm{Dir}\left(\gamma_e + \kappa_e \mathbf{e}_l + \mathbf{n}_l\right)$
end for

Event transition parameters, $\boldsymbol{\phi}$ The Dirichlet posterior for the event state l's transition distribution $\boldsymbol{\phi}_l$ simply involves the transition counts $\mathbf{n}_l$ from event state l to all L states,

$$p\left(\boldsymbol{\phi}_l \mid Z_{1:T}, \boldsymbol{\beta}\right) \propto \mathrm{Dir}(\alpha_e \boldsymbol{\beta} + \mathbf{e}_l \kappa_e + \mathbf{n}_l), \tag{4.28}$$

for global weights $\boldsymbol{\beta}$, concentration parameter α_e, and self-transition parameter κ_e. Algorithm 14 gives an explicit recipe for this transition distribution sampling.

Global transition parameters, $\boldsymbol{\beta}$ The Dirichlet posterior of the global transition distribution $\boldsymbol{\beta}$ involves the auxiliary variables $(\bar{m}_{\cdot 1}, \ldots, \bar{m}_{\cdot L})$,

$$p(\boldsymbol{\beta} \mid Z_{1:T}) \propto \mathrm{Dir}(\gamma_e/L + \bar{m}_{\cdot 1}, \ldots, \gamma_e/L + \bar{m}_{\cdot L}), \tag{4.29}$$

where these auxiliary variables are defined as

$$\begin{aligned}
\bar{m}_{ll'} &= \begin{cases} m_{ll'}, & l \neq l' \\ m_{ll} - w_l, & l = l' \end{cases} \\
m_{ll'} &= \sum_{r=1}^{n_{ll'}} \theta_r \\
\theta_r &\sim \mathrm{Ber}\left(\frac{\alpha_e \beta_l + \kappa_e \delta(l, l')}{\alpha_e \beta_l + \delta(l, l') + r}\right), \\
w_l &\sim \mathrm{Bin}\left(m_{ll'}, \frac{\rho_e}{\rho_e + \beta_l(1 - \rho_e)}\right).
\end{aligned} \tag{4.30}$$

See Fox [38, Appendix A] for full derivations. Note that this formulation is similar to that in Equations (3.14) and (3.15) but with the additional "override" variables w_l associated with the self-transition parameter κ_e.

■ 4.2.3 Observation Model Parameters

We sample channel state AR coefficients $\{\mathbf{a}_k\}$ and the event state covariances $\{\Delta_l\}$ from their posteriors using a standard Gibbs sampler. Algorithm 15 gives an explicit recipe for sampling these observation model parameters.

Algorithm 15 HIW-spatial BP-AR-HMM MCMC sampler for obs. model params.

calculate channel innovations at each time point for each channel
for $t \in \{1, \dots, T\}$ **do**
  for $i \in \{1, \dots, N\}$ **do**
    $\epsilon_t^{(i)} = y_t^{(i)} - \mathbf{a}_{z_t^{(i)}}^{\mathrm{T}} \widetilde{\mathbf{y}}_t^{(i)}$
  end for
end for

initialize sufficient statistics for the K current channel states
  $S^k_{\bar{\mathbf{Y}}\Delta\bar{\mathbf{Y}}^{\mathrm{T}}} \leftarrow \mathbf{0}^{r\times r}$ $k = 1, \dots, K$
  $S^k_{\bar{\mathbf{Y}}\Delta\mathbf{y}} \leftarrow \mathbf{0}^{r\times 1}$ $k = 1, \dots, K$
collect sufficient statistics $S^k_{\bar{\mathbf{Y}}\Delta\bar{\mathbf{Y}}^{\mathrm{T}}}$ and $S^k_{\bar{\mathbf{Y}}\Delta\bar{\mathbf{y}}}$ for the K current channel states
for $t \in \{1, \dots, T\}$ **do**
  for $k \in \{1, \dots, K\}$ **do**
    get the indices of the channels assigned ($\mathbf{k}^+$) and not assigned ($\mathbf{k}^-$) to state k at t
      $\mathbf{k}^+ \leftarrow \{i \mid z_t^{(i)} = k,\ i = 1, \dots, N\}$
      $\mathbf{k}^- \leftarrow \{i \mid z_t^{(i)} \neq k,\ i = 1, \dots, N\}$
    if $|\mathbf{k}^+| > 0$ **then**
      $\bar{\mathbf{Y}}_t^{(\mathbf{k}^+)} \leftarrow \left[\widetilde{\mathbf{y}}_t^{(k_1^+)} \mid \cdots \mid \widetilde{\mathbf{y}}_t^{(k_{|\mathbf{k}^+|}^+)}\right]$
      $S^k_{\bar{\mathbf{Y}}\Delta\bar{\mathbf{Y}}^{\mathrm{T}}} \leftarrow S^k_{\bar{\mathbf{Y}}_t\Delta\bar{\mathbf{Y}}_t^{\mathrm{T}}} + \bar{\mathbf{Y}}_t^{(\mathbf{k}^+)} \Delta_{Z_t}^{-1(\mathbf{k}^+,\mathbf{k}^+)} \bar{\mathbf{Y}}_t^{\mathrm{T}(\mathbf{k}^+)}$
      $S^k_{\bar{\mathbf{Y}}\Delta\mathbf{y}} \leftarrow S^k_{\bar{\mathbf{Y}}\Delta\mathbf{y}} + \bar{\mathbf{Y}}_t^{(\mathbf{k}^+)} \left(\Delta_{Z_t}^{-1(\mathbf{k}^+,\mathbf{k}^+)} \mathbf{y}_t^{(\mathbf{k}^+)} + \Delta_{Z_t}^{-1(\mathbf{k}^+,\mathbf{k}^-)} \boldsymbol{\epsilon}_t^{(\mathbf{k}^-)}\right)$
    end if
  end for
end for

sample each channel state k's AR coefficients as in Equation (4.31)
for $k \in \{1, \dots, K\}$ **do**
  $\Sigma_k \leftarrow \left(\Sigma_0^{-1} + S^k_{\bar{\mathbf{Y}}\Delta\bar{\mathbf{Y}}^{\mathrm{T}}}\right)^{-1}$
  $\mathbf{a}_k \sim \mathcal{N}\left(\Sigma_k S^k_{\bar{\mathbf{Y}}\Delta\mathbf{y}}, \Sigma_k\right)$
end for

collect sufficient statistics $S^l_{\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}}}$ for the L event states
  $S^l_{\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}}} \leftarrow \mathbf{0}^{N\times N}$ $l = 1, \dots, L$
  $S^l_n \leftarrow 0$ $l = 1, \dots, L$
for $t \in \{1, \dots, T\}$ **do**
  $S^{Z_t}_{\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}}} \leftarrow S^{Z_t}_{\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}}} + \boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t^{\mathrm{T}}$
  $S^{Z_t}_n \leftarrow S^{Z_t}_n + 1$
end for

sample each event state l's covariance as in Equation (4.32)
for $l \in \{1, \dots, L\}$ **do**
  $\Delta_l \sim \mathrm{HIW}_G(b_0 + S^l_n, D_0 + S^l_{\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}}})$
end for

AR coefficients, $\{\mathbf{a}_k\}$ Each observation $\mathbf{y}_t$ is generated based on a *matrix* of AR parameters $[\mathbf{a}_{z_t^{(1)}} \mid \cdots \mid \mathbf{a}_{z_t^{(N)}}]$. 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 Appendix A.2, the posterior for $\mathbf{a}_k$ is a multivariate normal

$$p(\mathbf{a}_k \mid \mathbf{y}_{1:T}, \mathbf{z}_{1:T}, Z_{1:T}, \{\mathbf{a}_{k'}\}_{k' \neq k}, \{\Delta_l\}) \propto \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k), \tag{4.31}$$

where

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

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

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 innovation covariances, $\{\Delta_l\}$ The posterior of the event state l depends on the number of time points assigned to state l and the outer products of the innovations $\boldsymbol{\epsilon}_t = \mathbf{y}_t - \mathbf{A}_{\mathbf{z}_t} \widetilde{\mathbf{Y}}_t$,

$$p(\Delta_l \mid \mathbf{y}_{1:T}, \mathbf{z}_{1:T}, Z_{1:T}, \{\mathbf{a}_k\}) \propto \mathrm{HIW}_G(b_l, D_l), \tag{4.32}$$

where

$$b_l = b_0 + |\{t \mid Z_t = l,\ t = 1, \ldots, T\}|,$$
$$D_l = D_0 + \sum_{t \mid Z_t = l} \boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_t^{\mathrm{T}},$$

Details on how to efficiently sample from a HIW distribution are provided in [22].

■ 4.2.4 Hyperparameters

Below we give brief descriptions for the MCMC sampling of the hyperparameters in our model. Full derivations are given in Fox [38, Section 5.2.3, Appendix C].

Channel dynamics model hyperparameters, γ_c, κ_c We use Metropolis-Hastings steps to propose a new value γ_c' from gamma distributions with fixed variance $\sigma_{\gamma_c}^2$ and accept with probability $\min(r(\gamma_c' \mid \gamma_c), 1)$,

$$\begin{aligned} r(\gamma_c' \mid \gamma_c) &= \frac{p(\{\boldsymbol{\pi}^{(i)}\} \mid \gamma_c', \kappa, \{\mathbf{f}^{(i)}\}) p(\gamma_c' \mid \gamma_c^2/\sigma_{\gamma_c}^2, \gamma_c/\sigma_{\gamma_c}^2) p(\gamma_c \mid \gamma_c', \sigma_{\gamma_c}^2)}{p(\{\boldsymbol{\pi}^{(i)}\} \mid \gamma_c, \kappa, \{\mathbf{f}^{(i)}\}) p(\gamma_c \mid \gamma_c^2/\sigma_{\gamma_c}^2, \gamma_c/\sigma_{\gamma_c}^2) p(\gamma_c' \mid \gamma_c, \sigma_{\gamma_c}^2)} \\ &= \frac{p(\{\boldsymbol{\pi}^{(i)}\} \mid \gamma_c', \kappa, \{\mathbf{f})^{(i)}\}}{p(\{\boldsymbol{\pi}^{(i)}\} \mid \gamma_c, \kappa, \{\mathbf{f})^{(i)}\}} \frac{\Gamma(\nu) \gamma_c^{\nu' - \nu - a}}{\Gamma(\nu') \gamma_c^{\nu - \nu' - a}} \exp\left(-b(\gamma_c' - \gamma_c) \sigma_{\gamma_c}^{2(\nu - \nu')} \right), \end{aligned} \tag{4.33}$$

where $\nu = \gamma_c^2/\sigma_{\gamma_c}^2$, $\nu' = {\gamma_c'}^2/\sigma_{\gamma_c}^2$, and we have a Gamma$(a, b)$ prior on γ_c. Recall that the transition parameters $\{\boldsymbol{\pi}^{(i)}\}$ are independent over i, and thus their likelihoods multiply. The proposal and acceptance ratio for κ_c is similar.

Channel active features model hyperparameter α_c We place a Gamma$(a_{\alpha_c}, b_{\alpha_c})$ prior on α_c, which implies a gamma posterior of the form

$$p(\alpha_c \mid \{\mathbf{f}^{(i)}\}) \propto \text{Gamma}(a_{\alpha_c} + K_+, b_{\alpha_c} + \sum_{i=1}^{N}(1/i)), \tag{4.34}$$

where $K_+ = \sum_{k=1}^{K} \mathbf{1}\left(f_k^{(1)} \vee \cdots \vee f_k^{(N)}\right)$ denotes the number of channel states activated in at least one of the channels.

Event dynamics model hyperparameters, γ_e**,** α_e**,** κ_e**,** ρ_e Instead of sampling α_e and κ_e independently, we an additional parameter $\rho_e = \kappa_e/(\alpha_e + \kappa_e)$ and sample $(\alpha_e + \kappa_e)$ and ρ_e, which is simpler than sampling α_e and κ_e independently.

($\boldsymbol{\alpha_e + \kappa_e}$**)** With a Gamma$(a_{\alpha_e+\kappa_e}, b_{\alpha_e+\kappa_e})$ prior on $(\alpha_e+\kappa_e)$, we use auxiliary variables $\{r_l\}_{l=1}^L$ and $\{s_l\}_{l=1}^L$ to define the posterior,

$$p\left(\alpha_e + \kappa_e \mid Z_{1:T}\right) \propto \text{Gamma}\left(a_{\alpha_e+\kappa_e} + \bar{m}_{\cdot\cdot} - \sum_{l=1}^{L} s_l, b_{\alpha_e+\kappa_e} - \sum_{l=1}^{L} \log r_l\right), \tag{4.35}$$

where $\bar{m}_{\cdot\cdot} = \sum_{l,l'=1}^{L} \bar{m}_{ll'}$ is the sum over auxiliary variables $\bar{m}_{ll'}$ defined in Eq. (4.30), and the auxiliary variables r_l and s_l are sampled as

$$\begin{aligned} r_l &\sim \text{Beta}(\alpha + \kappa + 1, n_{l\cdot}), \\ s_l &\sim \text{Ber}(n_{l\cdot}/(n_{l\cdot} + \alpha + \kappa)). \end{aligned}$$

$\boldsymbol{\rho_e}$ With a Beta(c_{ρ_e}, d_{ρ_e}) prior on ρ_e, we use auxiliary variables $\{w_{l\cdot}\}_{l=1}^L$ to define the posterior,

$$p(\rho_e \mid \bar{\mathbf{m}}, \boldsymbol{\beta}) \propto \text{Beta}\left(c_{\rho_e} + \sum_{l=1}^{L} w_{l\cdot}, d_{\rho_e} + \bar{m}_{\cdot\cdot} - \sum_{l=1}^{L} w_{l\cdot}\right). \tag{4.36}$$

For $w_{ls} \sim \text{Ber}(\rho)$ over $s = 1, \ldots, m_{ll}$, the posterior of the auxiliary variable $w_{l\cdot}$ is

$$p(w_{l\cdot} \mid \bar{m}_{ll}, \beta_l) \propto \text{Bin}(\bar{m}_{ll}, \rho_e + \beta_l(1 - \rho_e)) \tag{4.37}$$

$\boldsymbol{\gamma_e}$ With a Gamma$(a_{\gamma_e}, b_{\gamma_e})$ prior on γ_e, we use auxiliary variables v and q to define the posterior,

$$p(\gamma_e \mid \mathbf{m}) \propto \text{Gamma}\left(a_{\gamma_e} - q + \sum_{l=1}^{L} \mathbf{1}(\bar{m}_{\cdot l} > 0), b_{\gamma_e} - \log v\right). \tag{4.38}$$

The auxiliary variables are sampled via

$$\begin{aligned} v &\sim \text{Beta}(\gamma_e + 1, \bar{m}_{\cdot\cdot}), \\ q &\sim \text{Ber}(\bar{m}_{\cdot\cdot}/(\gamma_e + \bar{m}_{\cdot\cdot})). \end{aligned}$$

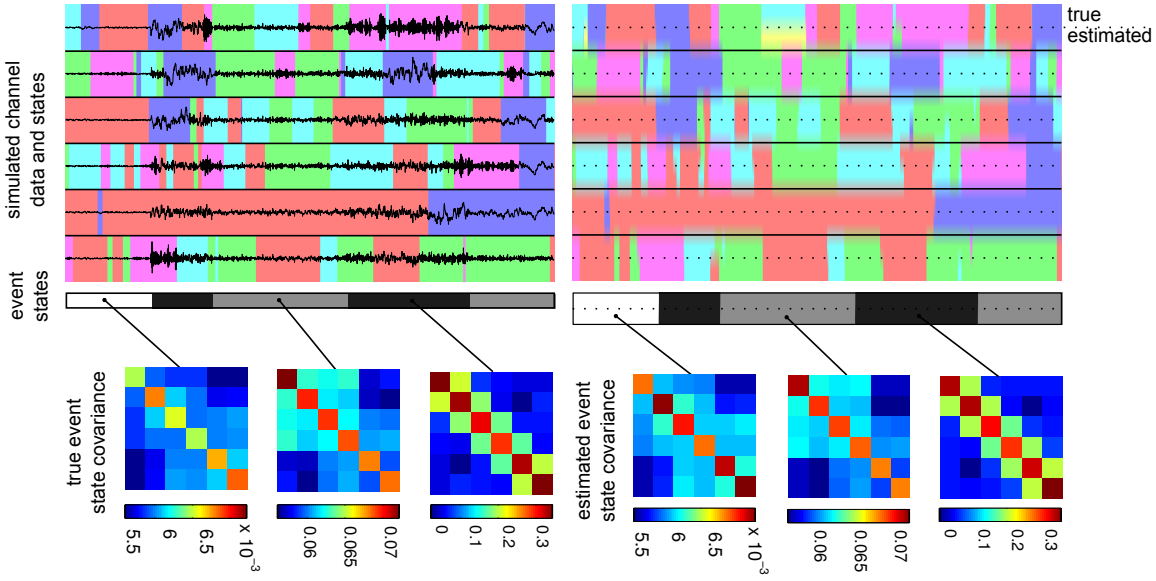

Figure 4.3. (**top left**) The six simulated channel time series overlaid on the five true channel states, denoted by different colors; the three true event states are shown in grayscale in the bar below. (**top right**) The true and estimated channel (color) and event (grayscale) states shown below for comparison after 6000 MCMC iterations. The true (**bottom left**) and estimated (**bottom right**) event state innovation covariances.

■ 4.3 Simulation Experiments

To initially explore some characteristics of our HIW-spatial model, we tested it on a small simulated dataset.

Data We simulated data from six time series in a 2x3 spatial arrangement, with vertices connecting all adjacent nodes (i.e., two cliques of 4 nodes each). We generated 2000 scalar observations using an first-order AR process with five channel states—AR coefficients linearly spaced between -0.9 and 0.9—and three event states with covariances shown in the bottom left of Figure 4.3. Channel and event state transition matrices were set to 0.99 and 0.9, respectively, for a self-transition and uniform between the other states. We generated channel feature indicators using $\alpha_c = 10$.

Results We ran the MCMC sampler for 6000 iterations, taking 500 samples after 1000-iteration burn-in and 10-sample thinning. Figure 4.3 shows the generated data and its true states along with the inferred states and event state covariances for a representative posterior sample. The event state matching is almost perfect, and the channel state matching is quite good, though we see that the sampler added an additional (yellow) state in the middle of the first time series when it should have assigned that section to the cyan state. The scale and structure of the estimated event state covariances match the true covariances quite well. Furthermore, Table 4.1 shows how the posterior estimates of the channel state AR

channel state	true $\mathbf{a}_k$	post. $\mathbf{a}_k$ mean	post. $\mathbf{a}_k$ 95% interval
1	-0.900	-0.906	[-0.917, -0.896]
2	-0.450	-0.456	[-0.474, -0.436]
3	0	-0.009	[-0.038, 0.020]
4	0.450	0.445	[0.425, 0.466]
5	0.900	0.902	[0.890, 0.913]

Table 4.1. The true and estimated values for the channel state coefficients in the simulated dataset.

coefficients also center well around the true values. While of course this simulated event is much simpler than a seizure, it confirms that our model is able to accurately estimate channel and event states as well as model parameters.

4.4 Parsing a Seizure

We tested the HIW-spatial BP-AR-HMM on two similar seizures (events) from a patient of the Children's Hospital of Pennsylvania.

4.4.1 Data and Methodology

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 Figure 4.4 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 posterior inference. The data was also scaled to have 99% of values within [-10, 10] for numerical reasons.

We examined a 5th-order HIW-spatial BP-AR-HMM and ran 10 MCMC chains for 6000 iterations, discarding 1000 samples as burn-in and using 10-sample thinning.

Determining a representative MCMC sample We often would like to display the channel and event states assigned over the time points of an event. But our MCMC sampler yields many sets of such states that are difficult to summarize in a single figure. We thus follow the method used by Fox et al. [39, 41, 42] to determine a particular sample with a representative event state sequence. This procedure involves calculating the pair-wise Hamming distances between the states of all MCMC samples. Since state indices are not necessarily consistent from one MCMC iteration to the next (especially when explicit state relabeling is involved as with the beta process), we first optimally relabel one set of states $Z^r_{1:T}$ to the other $Z^s_{1:T}$ using the Munkres assignment algorithm [80] on a $K^s \times K^r$ distance matrix D based on a

proximity matrix P between the K^s unique states in $Z^s_{1:T}$ and the K^r unique states in $Z^r_{1:T}$,

$$\begin{aligned} P_{k^s k^r} &= \sum_{t=1}^{T} \mathbf{1}(Z_t^s = k^s \wedge Z_t^r = k^r), & k^s = 1, \ldots, K^s, \quad k^r = 1, \ldots, K^r, \\ D_{k^s k^r} &= 1 - D_{k^s k^r} / D_{\cdot\cdot} & k^s = 1, \ldots, K^s, \quad k^r = 1, \ldots, K^r. \end{aligned} \tag{4.39}$$

This optimal state assignment is performed for each pair of MCMC samples $(s, r) \in (\{1, \ldots, S\}, \{1, \ldots, S\})$, and then the Hamming distance is calculated between the states $Z^s_{1:T}$ and the relabeled states $\widetilde{Z}^r_{1:T}$. We use the sample s^* with the minimum average Hamming distance to the other states as our representative sample,

$$s^* = \arg\min_{s=1,\ldots,S} \frac{1}{S-1} \sum_{r \neq s} \sum_{t=1}^{T} \mathbf{1}\left(Z_t^s \neq \widetilde{Z}_t^r\right). \tag{4.40}$$

Graphically summarizing event dynamics The channel and event states sequence visualization can deliver a detailed parsing of a particular seizure, but we may also be interested in higher-level summaries of such epileptic events given that we intend to apply this model to hundreds, if not thousands, of individual events so that we can better compare them. We thus desire a way of summarizing such events in a way that makes it easy to summarize the high-level event dynamics.

We use a finite automata-like visualization on the event state dynamics as one of the many possible approaches to this summarization problem. We describe the transitions between event states by a state transition diagram similar to those used to describe finite state automata. In this diagram, we denote transitions between two particular event states by a directed edge whose width is proportional to the number of transitions between the two event states. We do not show self-transitions (since most of transitions in the event state sequence are self transitions) and instead convey the prominence of each event state by the size of each event state, with an area proportional to the number of times it occurs in the event state sequence. These state transition diagrams are produced by an automated process. Finally, below or next to each event state, we depict the partial correlation coefficients between nodes in our sparse graph describing spatial channel relationships. The partial correlation coefficient between two channels i and j represents their correlation given their other connected channels in the graph G. For precision matrix Ω, the partial correlation coefficient between i and j is given by

$$r_{ij} = -\frac{\Omega_{ij}}{\sqrt{\Omega_{ii}\Omega_{jj}}}. \tag{4.41}$$

Since our hyper-inverse Wishart prior implies that the precision matrix has a zero entry for two channels i and j not connected in G, they have zero partial correlation given their neighbors and thus have no connecting edge in our correlation visualization. We show positive partial correlations in red and negative in blue, where the thickness of the edge denotes the magnitude of the correlation. We use the event state sequence and covariances given by the representative MCMC sample determined as previously described for these visualizations.

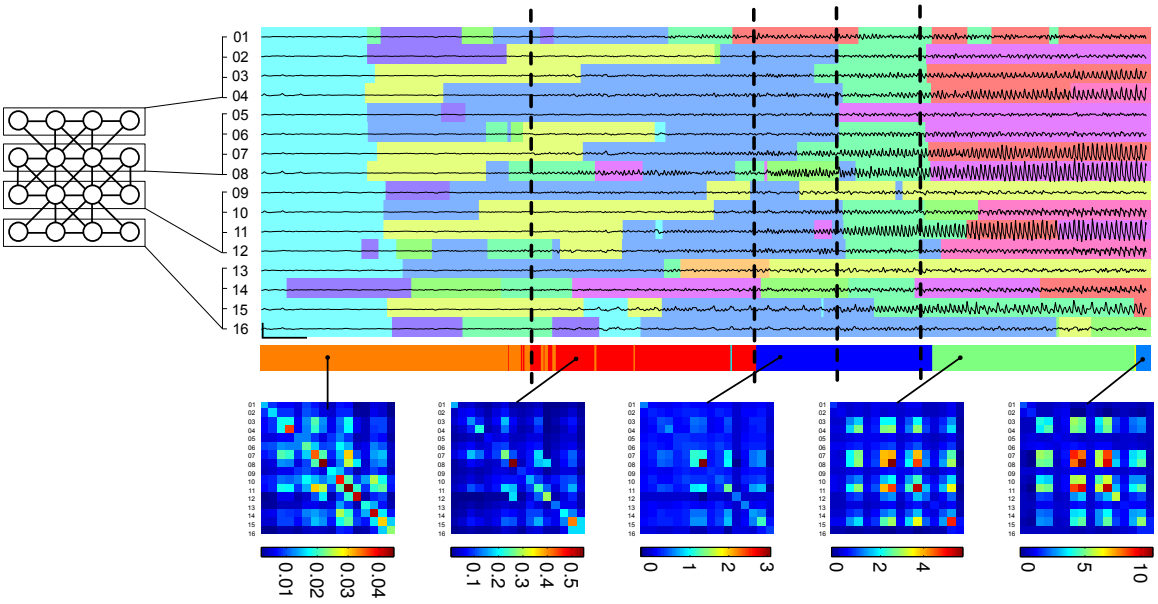

Figure 4.4. The graph used for a 16 channel iEEG electrode and corresponding traces over 25 seconds of a seizure onset with colors indicating the inferred channel states. The event states are shown below along with the associated innovation covariances. Vertical dashed lines indicate the EEG transition times marked independently by an epileptologist. Vertical and horizontal scale bars denote 1 mV and 1 second, respectively.

■ 4.4.2 Seizure Analysis

Exploring onset dynamics The HIW-spatial BP-AR-HMM inferred state sequences for the sample corresponding to a minimum expected Hamming distance criterion are shown in Figure 4.4. The results were analyzed by a board-certified epileptologist who agreed with the model's judgement in identifying the subtle changes from the background dynamic (cyan) initially present in all channels. 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 independently identified roughly six points in the duration of the seizure where the dynamic fundamentally changes. The three main event state transitions shown in Figure 4.4 occurred almost exactly at the same time as three of his own marked transitions. The other two transitions he marked did not occur during the onset. 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 the summary diagram shown in Figure 4.5 corresponding to the same data and event states shown in Figure 4.4, we see the onset dynamics from a high-level perspective. The event states 4 and 11 (orange and red, respectively, in Figure 4.4) are quite similar except

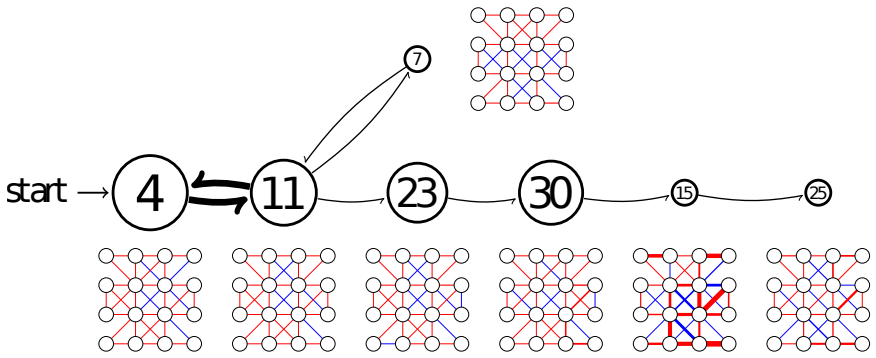

Figure 4.5. A finite automata-like diagram summarizing the seizure onset event state dynamics shown in Figure 4.4. Each vertex corresponds to an event state and has an area proportional to the number of time points assigned to that state. Transition arrow widths are proportional to the number of transitions from one state to another. The undirected graphical models depicting the partial correlation coefficients implied by the event state are shown beneath or next to them. Red edges denote positive correlation and blue edges negative. The edge width is proportional to the correlation magnitude.

for a slight negative partial correlation between channels 3 and 6 and a slightly increased correlation between channels 8 and 12 in event state 11. An inspection of the iEEG shown in Figure 4.4 shows how channels 8 and 12 become more active in the transition from event state 8 to 11. As the seizure transitions to event state 23, we see the partial correlation between channels 8 and 12 become stronger in magnitude but negative. In addition, positive partial correlations between channels 15 and 11, 11 and 7, and 7 and 8 increase. In event state 30, even stronger partial correlations exist between channels 15 and 16, 15 and 11, 11 and 8, and 7 and 8.

The summary of the seizure onset shown in Figure 4.5 parallels the type of EEG reading epileptologists perform manually when examining a patient's seizures, where they describe the channel relationships and how they change over the onset of the seizure. As we will see in Section 5.2.3, summary diagrams such as this also provide an intuitive and straightforward visualization for comparing a number of events to each other simultaneously.

Exploring offset dynamics Figure 4.6 shows the event state parsing at the offset of the same seizure whose onset is shown in Figure 4.4. Though the channel states are intuitive for this offset, we have shown only the event states to illustrate how well the model is capable of distinguishing subtle transitions in the event dynamics like that from the first half to the second half of the offset. In this parsing of the transition, we see how the seizure moves from 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.

Figure 4.7 shows a diagram of the event state dynamics of the same seizure offset. Notice how changes in relationships of channels 7, 8, 11, and 12 while subtle in the iEEG become

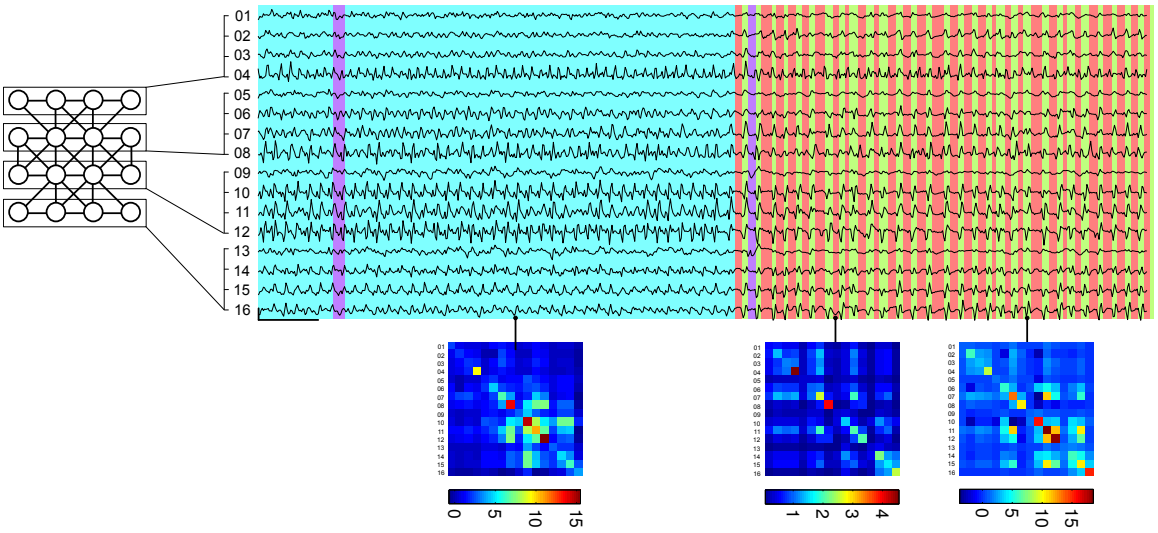

Figure 4.6. A representative sample showing the event state parsing (copied across all EEG channels) of a seizure by the HIW-spatial BP-AR-HMM model.

much more apparent in the partial correlations shown in the undirected graphs of event states 14 and 28.

Clinical relevance 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, we believe that the event states directly describe the activity of particular clinical interest.

In modeling the correlations between channels, the event states give insight into how different physiologic areas of the brain interact over the course of a seizure. In the clinical workup for resective brain surgery, these event states could help define and specifically quantify the full range of ways in which neurophysiologic regions initiate seizures and how others are recruited over the numerous seizures of a patient.

The ultimate clinical aim of this work, however, involves understanding the relationship between epileptic bursts and seizures. Because the event state aspect of our model involves the Markov assumption, the intrinsic length of the event has little bearing on the states assigned to particular time points. Thus, these event states allow us to straightforwardly compare the neurophysiologic relationship dynamics in short bursts (often less than two seconds long) to those in much longer seizures (on the order of two minutes long).

■ 4.5 Model Comparison

While our HIW-spatial BP-AR-HMM model clearly provides useful clinical epileptic event parsing, we are also interested in how it compares statistically to other similar models.

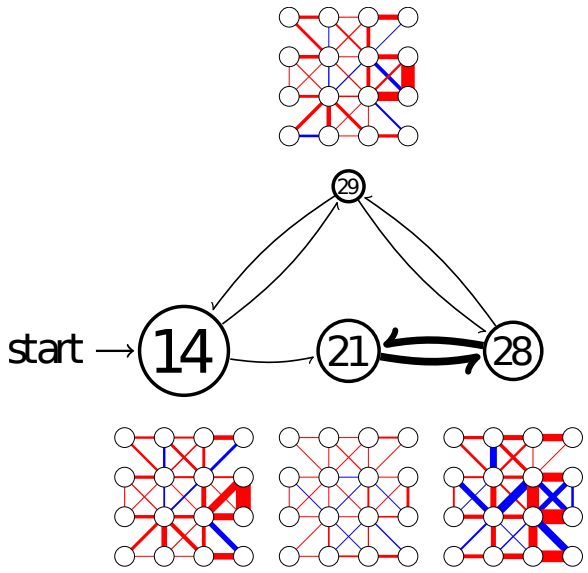

Figure 4.7. A diagram summarizing the event state dynamics of the seizure offset depicted in Figure 4.6.

■ 4.5.1 The Advantages of a Spatial Model

We explored the extent to which the spatial information encoded in the HIW prior improve the model on heldout data.

Data and methodology We compared our HIW-spatial BP-AR-HMM to a full-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. [41] and an AR-HMM without the feature-based modeling provided by the beta process [43]. Both of these models use inverse gamma (IG) priors on the individual channel innovation variances. 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. MCMC samples were collected over 5000 samples across 10 chains, each with a 1000-sample burn in and 10-sample thinning.

We 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. For each original MCMC sample, a secondary chain is run fixing all but $z_t^{(i)}$, Z_t, $\mathbf{f}^{(i)}$, $\boldsymbol{\eta}^{(i)}$, and $\boldsymbol{\phi}$. We approximate $p(\mathbf{y}_{1:T} \mid \boldsymbol{\phi}, \{\mathbf{a}_k\}, \{\Delta_l\})$ by averaging the secondary chain's closed-form $p(\mathbf{y}_{1:T} \mid \mathbf{z}_{1:T}, \boldsymbol{\phi}, \{\mathbf{a}_k\}, \{\Delta_l\})$ given in Equation (4.25).

Results Figure 4.8 (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

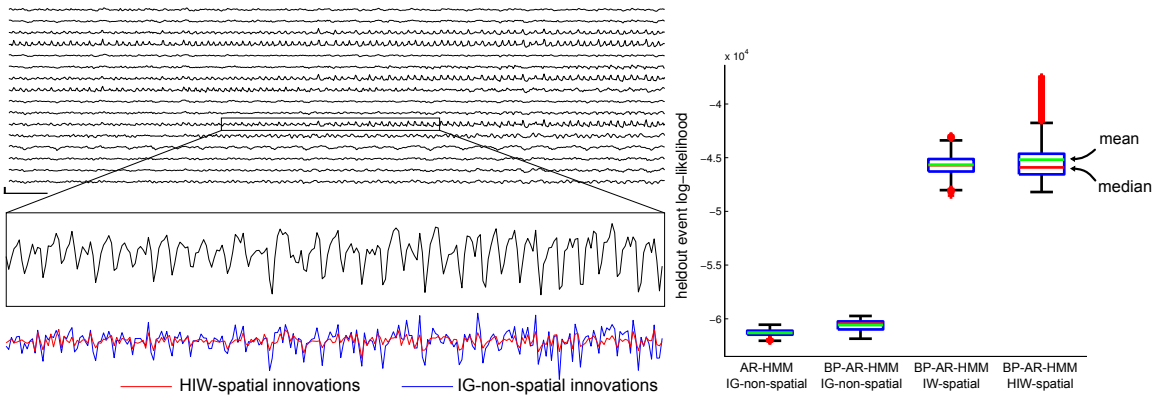

Figure 4.8. (**left**) An example 16-channel clip of iEEG with the middle section of one channel zoomed in and innovations from a non-spatial $\mathcal{N}$-IG prior and a spatial $\mathcal{N}$-HIW prior BP-AR-HMM shown below. (**right**) Boxplots of the heldout event log-likelihoods from the two non-spatial and two spatial models with mean and median posterior likelihood given in green and red lines. Boxes denote the middle 50% prediction interval.

in Figure 4.8 (right), which compares the heldout log-likelihoods for the IG-non-spatial and (H)IW-spatial models listed above. 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).

■ 4.5.2 The Advantages of Sparse Spatial Dependencies

The sparse dependencies among channels induced by the hyper-inverse Wishart prior allow the model to scale efficiently to the large number of channels present in iEEG. Specifically, the matrix operations associated with computing the conditional channel likelihood given by Equation (4.12) scale roughly linearly in the number of channels versus quadratically as they do with the full-covariance IW prior.

Data and Methodology To quantify this complexity difference, we performed an experiment comparing the IW and HIW priors. We ran an IW-spatial BP-AR-HMM and a HIW-spatial BP-AR-HMM on five datasets of 8, 16, 32, 64, and 96 channels from the same dataset used in Section 4.4.1. The IW prior entailed a fully-connected channel graph, and the HIW-prior used the spatial proximity connections similar to those shown on the left of Figure 4.4. All priors and other parameters of these two sets of models were kept the same, except the spatial dependencies of the IW and HIW priors on Δ_l. We ran the two models on each of the five datasets for at least 1000 MCMC iterations, using a profiler to tabulate the time spent in each step of the MCMC iteration.

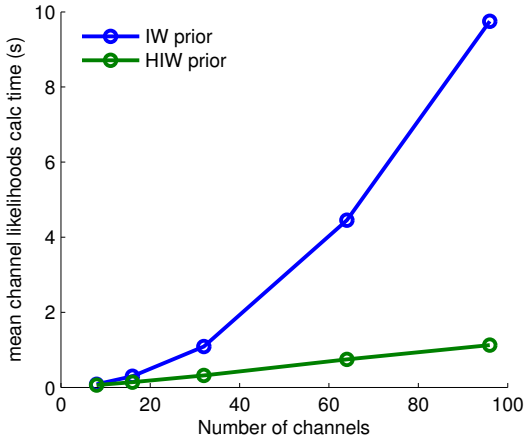

Figure 4.9. The average time per MCMC iteration required to calculate all of the channel likelihoods under each AR model at each time point.

Results Figure 4.9 shows the average time required to calculate the channel likelihoods at each time point under each AR channel state (see Equation (4.12)), which are used both for calculating the marginal likelihood (over all the state sequences $z_{1:T}^{(i)}$) used in active feature sampling as well as sampling the state sequences $z_{1:T}^{(i)}$. We see how the computation time increases quadratically in the IW prior and linearly in the HIW prior as the number of channels increases.

Anecdotally, we also found that the IW prior experiments—especially those with larger numbers of channels—tended to occasionally have numerical underflow problems associated with the inverse term $\Delta_{Z_t}^{-1(\mathbf{i}',\mathbf{i}')}$ in the conditional channel likelihood calculation. This underflow in the IW prior model calculations is not surprising since the matrices inverted are of dimension $N-1$ (for N channels), whereas in the HIW prior, the sparse spatial dependencies make these matrices no larger than eight-by-eight.

■ 4.6 Comparing Epileptic Events of Different Scales

As previously discussed, our HIW-spatial BP-AR-HMM has the capability of comparing events with very different time scales. We applied it 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.

Data and methodology The events were automatically extracted from the patient's continuous iEEG record by taking sections of iEEG whose median line-length feature [35] crossed a preset threshold, also including 10 seconds before and after each event. The iEEG was

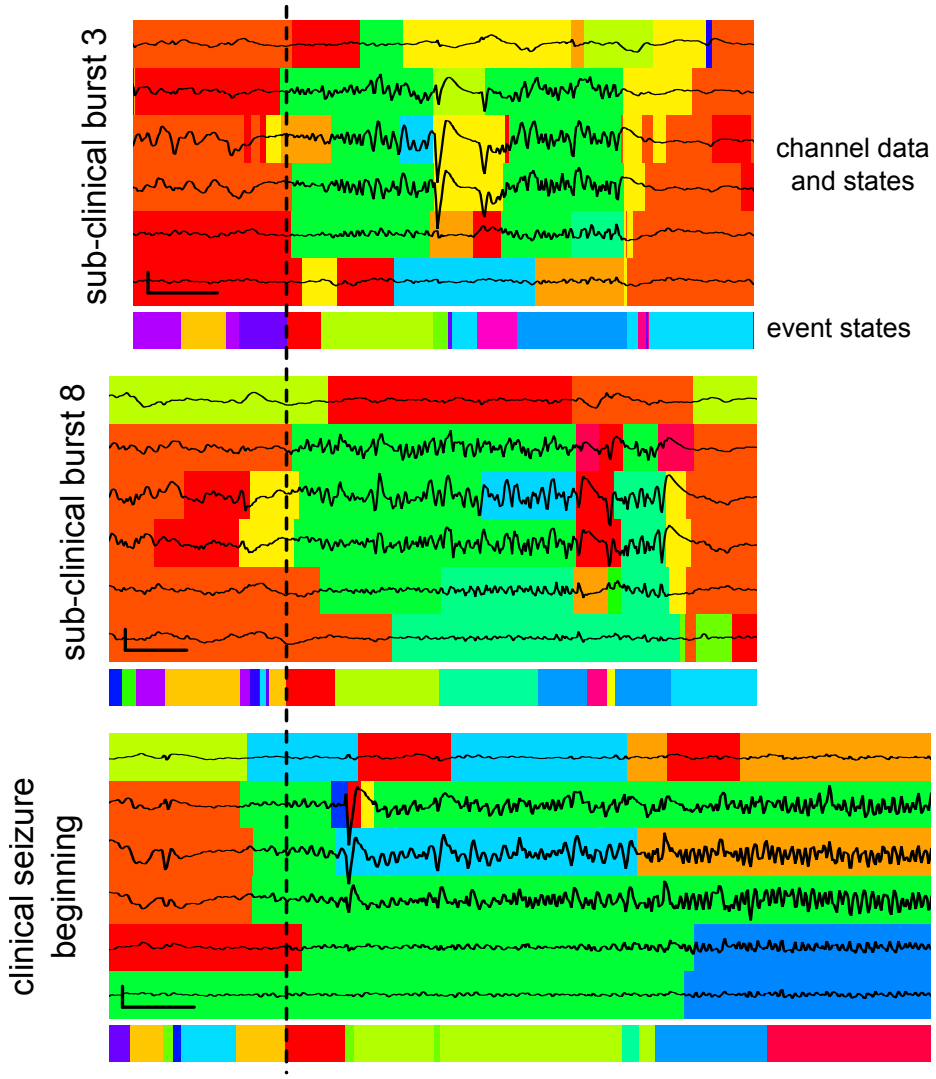

Figure 4.10. 6 iEEG traces from two sub-clinical bursts and onset of the single seizure with colors indicating inferred channel and event states. The dashed lines indicates the start of the red state in the three events. Vertical and horizontal scale bars denote 1mv and 1 second, respectively.

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 using the minimum average Hamming distance method previous described in Section 4.4.1. The hyperparameter settings, number of MCMC iterations, chains, and thinning was as in the previous experiment.

Results Figure 4.10 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 treat as the 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 similarities 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 from the seizure: 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.

In the next chapter, we use our model to explore this hypothesis more deeply in an analysis of hundreds of such events.

■ 4.7 Practical Scaling for Large Datasets

After describing the benefits of our HIW-spatial BP-AR-HMM in modeling seizures, both statistical (compared to other, similar models) and qualitative (in delivering intuitive parsing of the channel and event dynamics), we cannot honestly overlook the practical computational difficulties that accompany these benefits. Simply, posterior inference of our model is slow, prohibitively so for the large datasets of epileptic bursts and seizures we desire to examine. In fact, an informal calculation indicated that it would take roughly three months of constant running on our available computers for this model to yield a reasonable number of posterior MCMC samples on one of our large epileptic event datasets.

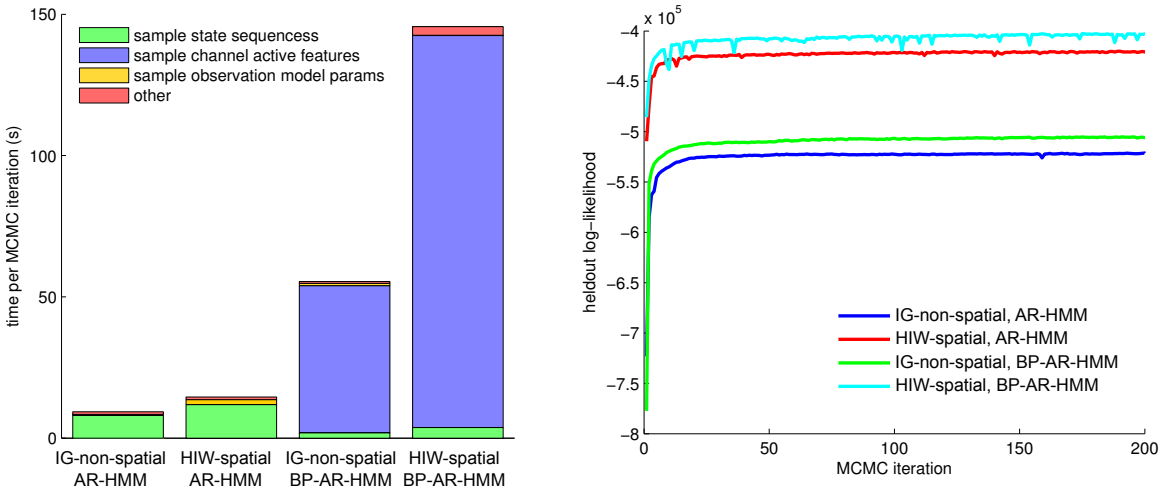

Figure 4.11. (**left**) The average time spent for each MCMC iteration and a stacked breakdown of the most time-consuming steps of each iteration. (**right**) The log-likelihood of the heldout set for each of the four models over the first 200 MCMC iterations.

We should qualify that in this section, we focus exclusively on our particular goal of modeling epileptic activity from 16 channels of iEEG. The model may behave differently and have different computational properties on other datasets. We must also note that our "large" dataset of hundreds of epileptic events, with a total size in memory of a few hundred megabytes, is hardly exemplary "Big Data" standards, where dataset are often too large to fit entirely in memory. Throughout this work, we assume that our entire dataset and the current parameters of the model can easily fit in the available computer memory.

In the rest of this section, we describe various practical problems encountered and our solutions to them when scaling our HIW-spatial BP-AR-HMM to run on hundreds of epileptic events.

■ 4.7.1 Computational Bottlenecks

In determining how to speed up posterior inference in our model, we first examined the time spent in each step of an MCMC iteration and found that sampling the channel active feature $\mathbf{f}^{(i)}$, the channel states $\mathbf{z}_{1:T}^{(i)}$ and event states $Z_{1:T}$, and the observation model parameters $\{\mathbf{a}_k\}$ and $\{\Delta_l\}$ comprised the majority of the time spent in each iteration. Note of course that now we are working with a dataset many events rather than the simpler case of a single event as used (for notational simplicity) in Section 4.1. We thus ran a small experiment comparing the time for each of these steps in four similar models: an IG-non-spatial AR-HMM, a HIW-spatial AR-HMM, an IG-non-spatial BP-AR-HMM, and a HIW-spatial BP-AR-HMM. The AR-HMM models used an HDP-HMM involving a DP with a fixed number of states, as described in Fox et al. [42].

Data and methodology We used a dataset of 50 epileptic events, gathered at random from the 777 events in the dataset for dog 002 described in Section 5.1.1. The first 25 events were used for training the model, and the second 25 were heldout for evaluatation. Since the beta process models use a Metropolist-Hastings step for sampling the κ self-transition parameter versus the Gibbs sampling step used in the standard AR-HMM models, we fixed the κ hyperparameters to the same values across all four models. We ran 1000 MCMC iterations for each model, using a profiler during this MCMC computation to determine the time spent in each step of the full MCMC iteration.

Results Figure 4.11 shows the profiling and heldout likelihood results from each of the four models. Two results in particular are striking. First, the active feature sampling associated with the beta process takes a large amount of time compared to the other sampling steps of the AR-HMM (which of course do not have an active feature sampling step). Second the HIW-spatial components of both the AR-HMM and BP-AR-HMM yield improvements in modeling the heldout events. While the beta process and its constraints on the available channel features improves both the IG-non-spatial and the HIW-spatial, these improvement are not as great as that yielded by the HIW-spatial over the IG-non-spatial model. Furthermore, given the true scale of our epileptic events dataset (with hundreds of events), we do not find that the at least five-fold increase in computational burden implied by the beta process models has a proportional payoff in modeling improvement. Since in the previous section we showed the varied benefits of parsing and event into both channel states *and* event states, we believe the HIW-spatial AR-HMM offers the best balance of computational and modeling performance for large event datasets.

In addition, the sparse set of channel features selected by each channel is of less interest in and of itself than in other applications since our main focus in this application is on the event state sequence rather than what each particular channel within the state is doing. We focus less on the inferred channel state sequences as well, mostly because there are so many more of them to consider in a single event. The loss of the active feature indicators for each channel is thus only a slight loss in modeling power rather than a loss in a parameter of primary interest.

■ 4.7.2 Parallelizing Computations

An additional computation advantage of the AR-HMM models is that sampling their channel and event states can be parallelized, allowing us to take advantage of the multi-core machines commonly available. While Matlab (Mathworks, Natick, MA) does have innate multi-threading capability in many of its matrix operations, we desired a way to run all the computations involved with sampling channel and event states on an explicit thread. Since most of the computationally intensive parts of our code were already implemented in C++ using the Eigen matrix library, we were able to take advantage of native C/C++ POSIX threads to spawn new threads and assign various computations to each. In machines without hyperthreading [59], we have found that we achieve almost no performance improvement in spawning more threads than available CPU cores and often find a slight

loss in performance, due most likely to the additional overhead and inefficiencies involved in spreading computation over a number of cores. Since we usually run our computations on 4- or 8-core machines, we usually use 4 or 8 threads.

The naive way of divvying up which events are processed by which threads would be to just to assign an roughly equal number of events to each thread. But this strategy ignores the length of each event, which in our datasets of seizure and short bursts, can vary greatly from event to event. Since the amount of computation for each event is usually directly related to its length, this naive divvying would most likely lead to imbalances in how much computation each thread must perform in a given function call. We thus assign events to threads one at a time, assigning each event to the thread with the current fewest number of total time points to process.

This simple event divvying procedure ensures that no thread processes many more or many fewer time points than any of the others. Note, however, that this procedure does not necessarily produce an assignment that optimally balances to the total number of time points among the threads. A slightly more sophisticated method for producing such an optimal assignment would involve sorting the events by increasing time length and then sequentially assigning a block of events to each thread. Nevertheless, anecdotal observation of the activity of the multiple CPU cores using our simpler assignment procedure indicated that the threads were balanced reasonably well (i.e., they all finish processing their events within a second or two of each other).

While we would naively expect our computations to speed up by the factor of our parallelization, at some point the parallelization overhead and the non-parallel parts of the code degrade this efficiency. We explored the degree to which an HIW-spatial AR-HMM MCMC iteration is sped up with varying degrees of parallelization.

Data and methodology Using the full dataset of 777 events from dog 002, we collected 1000 MCMC samples from over chains using 1, 2, 4, or 8 parallel threads whenever possible. As before, we ran these computations inside a profiler to determine the total time spent in each step of the MCMC iteration.

Results Figure 4.12 (left) shows the breakdown of a HIW-spatial AR-HMM MCMC iteration for each of the four parallelizations. While the computational time clearly decreases as the number of threads increases, the relationship seems not to be completely linear, showing how the overhead and other possible inefficiencies associated with multithreading can degrade performance. Nevertheless, this roughly 4.5x speedup between a single thread and eight threads is vital for finishing MCMC computations on our large datasets in a reasonable amount of time.

■ 4.7.3 MCMC Sample Compression

Each HIW-spatial AR-HMM MCMC sample contains a number of different values of interest: observation model parameters, hyperparameter values (if they are sampled rather than fixed), various statistics of interest like event likelihoods, event and channel transition distributions, and the assigned channel and event state sequences. In a large-scale experiment

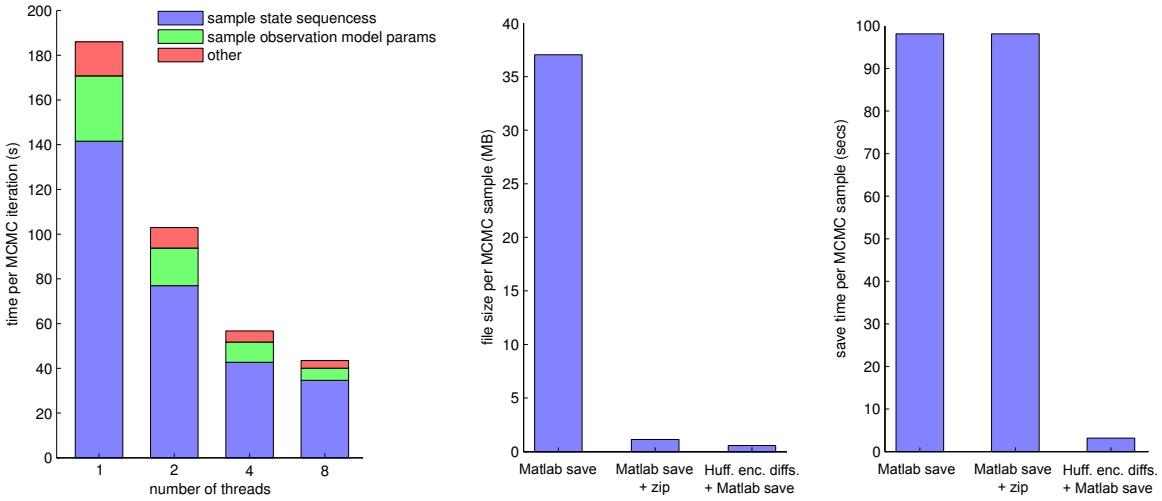

Figure 4.12. (**left**) The breakdown of time spent in two of the main steps of a HIW-spatial AR-HMM MCMC iteration for different numbers of parallel threads. (**middle**) The size of the saved model parameters and data states from an single MCMC sample saved using three different saving schemes: saving a Matlab struct directly to a `.mat` file, saving a Matlab struct to a `.mat` file and then zip-compressing at the end of MCMC sampling, and compressing the data states using Huffman encoded differences in Matlab and then saving the struct to a `.mat` file. (**right**) The average time required to save an MCMC sample's parameters and data states using each of the three saving schemes.

of perhaps 10 MCMC chains, each of which involves perhaps 1000 samples, the amount of memory and hard disk space required to store these values quickly becomes large. For example, in one of our datasets of 902 events, 100 MCMC samples occupy roughly 7.5 GB of memory and would be stored using roughly 3.5 GB of disk space (a smaller value thanks to Matlab's compression routines built in to its `.mat` file save functionality). As experiments scale up, we found this memory usage and disk storage became burdensome. We were able to mitigate it to a degree by not saving parameters of secondary interest like the channel and event transition distributions, but the state sequences—of primary interest for our downstream analysis—comprised the majority of this space.

The increased size of the samples affects two aspects of the experiment: the amount of time required to periodically save results to disk, and the size of those files once saved to disk. Of course, we could reduce the cost in time of periodically saving MCMC samples by saving them less often, but this less frequent saving comes with the cost of more risk of lost samples should experiments (or the machines on which they run) fail. Though Matlab natively performs some degree of lossless compression when saving data to a `.mat` file, we found the disk space used when running multiple chains for many hundreds (or more) of iterations quickly becomes unwieldy.

Fortunately, the channel and event state sequences are constrained in value (by the truncation level, often less than 50) and also usually contain a great deal of structure, in

that long sequences of time points are often assigned to the same state. Both of these properties make the state sequences ideal for lossless compression, which would not only ameliorate the large size of the data stored on disk but also reduce the amount of time require to save that data to disk.

The field of data compression is vast and rich, but we focus on one classic, straightforward approach for compressing our state sequences. Huffman coding is a type of prefix code satisfying the source coding theorem, which states that the expected length $L(C, X)$ (in bits) of code words in a prefix code C for the data X is,

$$H(X) \leq L(C, X) \leq H(X) + 1, \tag{4.42}$$

where $H(X)$ is the Shannon entropy of the data X [77, chap. 5]. For example, let us assume that we have sequences of states that take one of 50 possible values. Assuming a worse-case scenario where each value is equally probable, the average code length would be $\log_2 50 = 5.6$ bits, which is smaller than the single byte (8 bits) required natively to store each value in the best-case scenario. In reality for our channel and event state sequences, some states are assigned much more often than others (some of which may not even be assigned at all), a property that further reduces the entropy of the states, increasing our possible compression ratio. Finally, we can take advantage of first-order structure in the state sequences to even further reduce the entropy of our states. Instead of encoding the states themselves, we encode the *transitions* in the states, which are very often self-transitions since adjacent time points are often assigned to the same state (even more so because of the self-transition parameter κ in the HDP-HMM models). In fact, anecdotal analysis of our state sequences indicated that roughly 98% of state transitions were self-transitions, or a state difference of zero. Since we originally store the channel and event state sequences as 16-bit unsigned integers, we can achieve significant compression savings via Huffman encoded differences (HED).

Our state sequence compression scheme involves simply HED for the channel and event states and storing those encoded (and compressed) values in a Matlab struct with the rest of the (uncompressed) MCMC sample values. We then save this struct to a standard `.mat` file using Matlab, taking advantage of another level of compression through Matlab's internal save routines. For simplicity, we perform the compression for each MCMC sample independently.

Data and methodology To explore the extent of which our compression scheme reduces both the size of each MCMC sample on hard disk and the time required to save it to disk, we compared our HED compression on only the state sequences (leaving the other parameters of the MCMC sample intact) to two other approaches: a naive Matlab save of the MCMC samples without any explicit compression on our part, and the same naive Matlab save followed by file compression using the Unix operating system's zip utility. For our experiments, we worked with a Matlab struct array of 100 MCMC samples, each containing the various values of interest (including the channel and event state sequences) to be saved. These data were taken from one of the files produced by our modeling of 902

events for dog 005, as discussed in the next chapter. We examined the average storage size (on disk) per MCMC sample and also the average time required for each save to disk. Note that we do not include the zip-compression time in this calculation, since it will usually only be executed once, after all the MCMC samples have been saved to the particular file.

Results Figure 4.12 (middle) shows the average size (on disk) of a single MCMC sample's data when saving using each of the three saving schemes, with the average size in memory at 74.5 MB. Matlab's internal compression achieves a roughly 2x compression ratio, but this is far less than that achieved by the other two methods. The Matlab save followed by zip-compression reduces the size to roughly 1.1 MB per sample, and our Huffman encoded differences approach yields an even better 55 KB per sample. It is interesting to note that this 136x compression ratio is far larger than what we can achieve by naive Huffman encoded differences (which can only hope to achieve no better than a 16x ratio from the 16-bit integers we store the state sequences in). Clearly, Matlab is able to compress the Huffman encoded differences quite effectively.

The right side of Figure 4.12 shows the average time per save of the MCMC samples under each of the three saving schemes. Since the zip compression works only after Matlab has naively saved all of the uncompressed data, its save time is just as slow as the that without any compression. In contrast, compressing the state sequences using HED prior to saving them through Matlab greatly reduces the save time since much less data is being saved.

These results show how using a simple encoding method like Huffman coding can dramatically reduce both the ultimate disk space required for the MCMC samples and also the time required to save those samples.

■ 4.8 Discussion and Future Work

As the amount and complexity of collected EEG data continues to expand, it becomes less and less feasible to analyze all of this data manually, using the valuable time of highly-trained physicians. The past two decades have seen a great expansion of automated EEG analysis methods. And yet, very few of these automated methods are flexible enough to fully emulate the manual physician readings. If automated methods are to be trusted on a large scale over much data, they must yield intuitive analyses that mesh well with existing clinical practices.

In this chapter, we develop a model—the HIW-spatial BP-AR-HMM—capable of producing EEG parsings we believe are intuitive and similar to the manual parsings produced by epileptologists. Our model builds the beta process autoregressive hidden Markov model, a Markov switching process that describes the dynamics of a collection of independent time series, in our case, the individual iEEG channels of a particular epileptic event. Our contribution involves linking these channels together using a Gaussian graphical model, where channel activity influences the activity of spatially-proximate channels. We show that this spatial model improves predictions on heldout data and also yields clinically intuitive parsings. We apply this model to a dataset of 14 sub-clinical bursts and one clinical seizure

and show that the bursts contain individual channel and across-channel dynamics similar to those present in the seizure onset. Finally, we describe our strategies for scaling up our model to efficiently process hundreds of individual events over thousands of time series.

Despite these initial scaling efforts, we believe the computational complexity of posterior inference of our HIW-spatial BP-AR-HMM and AR-HMM can be optimized even further, especially as our datasets continue to grow in size.

State sequence downsampling Downsampling the channel and event state sequences could be an inelegant though perhaps effective method for reducing the computational burden of the state sequence sampling. Though sampling frequencies of 200 Hz and higher are probably necessary for capturing the important aspects of the raw iEEG data, it is quite possible that the state sequence resolution need not be that high. Instead of assigning a state for each time point, we could assign a state to perhaps five or ten time points, where the likelihood of the block of time points under the particular state is simply the product of the likelihoods of each of the time points within that block. This state sequence downsampling would reduce the matrix-vector products associated with marginalizing and sampling the state sequences. This approach might first be tried with the channel state sequences, since far more of them must be sampled in each MCMC iteration than the event state sequences and since they are currently of secondary interest in our downstream analysis, so a little loss of state sequence resolution could perhaps be tolerated.

Variational methods A perhaps complementary approach would involve using variational methods [65, 77, 128] like those described in Ghahramani and Jordan [49] for a factorial HMM, which our HIW-spatial (BP-)AR-HMM actually is. Variational methods have used in a wide variety of Bayesian models where traditional Gibbs sampling may be slow to converge and mix. These methods deterministically optimize a variational approximation to the posterior distribution and may perhaps improve posterior sample mixing, allowing us to perhaps reduce the number of MCMC iterations between collected samples.

Chapter 5

Exploring the relationship between epileptic bursts and seizures

Epileptic bursts have long been noticed on the iEEGs of patients in the epilepsy monitoring unit (EMU). The relationship of these bursts to clinical seizures is poorly understood, however, since they have no clinical manifestation. These sub-clinical bursts, while undetectable by the patient and human doctors, represent a class of epileptic events far more frequent than seizures. How are these bursts related to seizures? In this chapter, we begin to answer that question by describing these sub-clinical bursts alongside the full clinical seizures using the HIW-spatial AR-HMM presented in Chapter 4.

We examine the continuous iEEG recordings from three dogs with chronic implants. Dogs are the only other animal known to have spontaneously occurring epilepsy, so those with epilepsy offer the closest possible comparison to the human disease. These dogs were implanted as part of a trial for a chronically implanted seizure warning system, developed by the NeuroVista corporation. In this system, iEEG is collected continuously by an internal device and telemetered out to an external processing unit that attempts to predict when the dog is at heightened risk of having a seizure. Figure 5.1 shows a diagram describing this recording setup. These continuous data, which for some dogs span over year, offer a far more complete picture of a patient's epilepsy than the recordings collected over a few weeks in the EMU.

In the EMU, patients are still recovering from the invasive electrode implantation surgery, which greatly disrupts the physiological state of the brain and so possibly also the seizures and other activity recorded during that time. In addition, the modulation of the patient's anitepileptic medication (in an effort to allow more seizures to occur) while in the EMU probably also influences the seizures recorded during that time. Chronic recordings from long-term implants offer a picture of a patient's epilepsy less muddled by these distortions and thus present an ideal setting for examining the relationship between sub-clinical bursts and clinical seizures.

We examined the recordings from three dogs that were implanted as part of an early trial by the NeuroVista corporation for their seizure warning device. For consistency with other work involving these dogs, we refer to them by the implant number assigned by NeuroVista. The three dogs we study (implant numbers 002, 004, and 005) were recorded

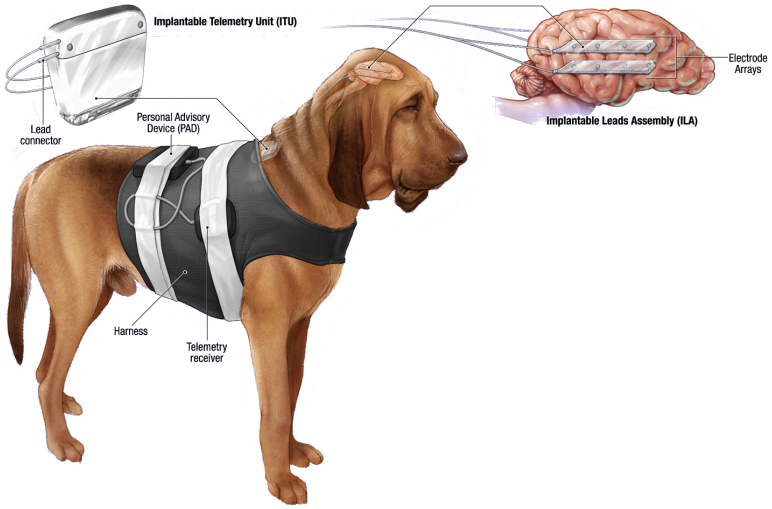

Figure 5.1. NeuroVista canine continuous iEEG monitoring setup. Two four-channel strips are implanted on each hemisphere, and the data is recorded and stored by the Implantable Telemetry Unit (ITU). The ITU broadcasts seizure warnings to the Personal Advisory Device (PAD). (Figure adapted with permission from Patterson et al. [87])

for 475, 329, and 45 days. These dogs had a 37, 14, and 91 clinical seizures. Every one of these seizures was marked independently by two board-certified epileptologists who then came to a consensus about the exact temporal location of every seizure onset. The subclinical bursts were far too numerous to manually mark, so we use an automated detection scheme for them that is described in the next section.

Dog 005 died in status epilepticus after only 45 days. In status epilepticus, a patient seizes almost continuously, a situation associated with many adverse physiologic changes that can cause severe, permanent neurological damage if not treated immediately [33, chap. 64]. In some cases, as in dog 005, it is fatal. While the data from this dog do not offer as long-term a picture of its epilepsy as the data in the other dogs, they do document how the physiologic changes occurring in status epilepticus are reflected in the iEEG.

In our study of the bursts and seizures recorded in these three dogs, we hypothesized that subsets of the bursts would contain some limited similarities with seizures, especially with the seizure onsets. This relationship was found in the 14 bursts and a single seizure described previously in Section 4.6. We believed that these similarities between the bursts and seizure onset would hold up to a point, at which the seizure progresses into a full clinical event, whereas the bursts end or "reset" in some way.

Figure 5.2 shows a high-level summary of the processing steps involved in our quantitative analysis. Epileptic burst events are detected from the continuous iEEG record and then culled to eliminate artifacts and other events not of interest to our problem. These events are preprocessed before they are fed into the main modeling phase of the analysis,

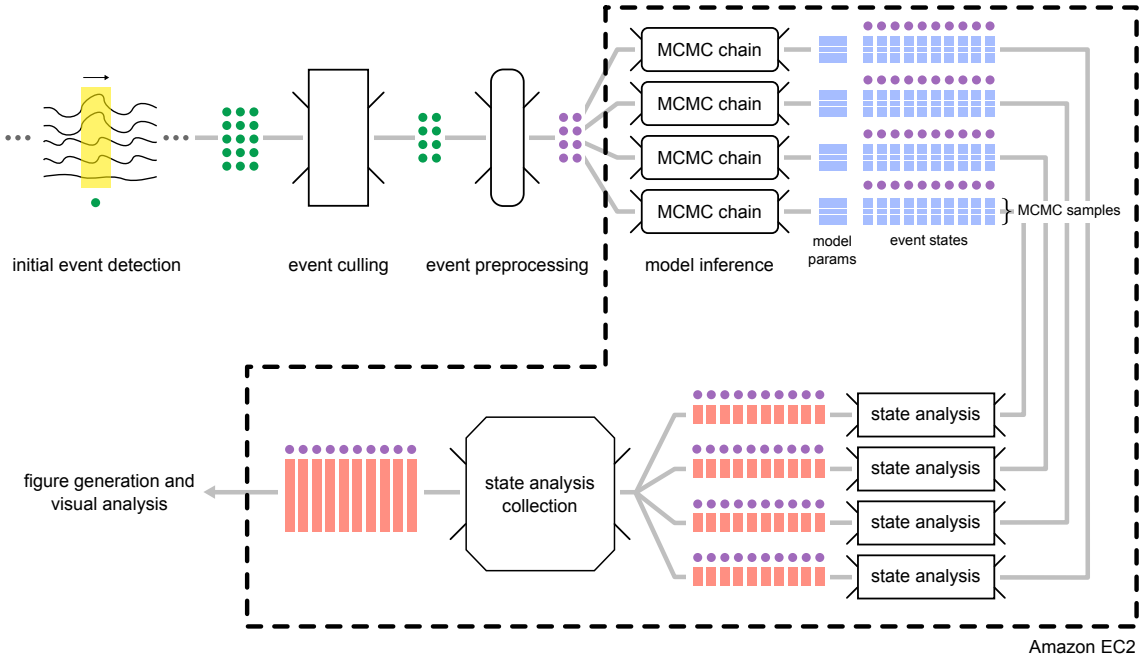

Figure 5.2. A high level description of epileptic event analysis methodology. (Clockwise from the top left) A feature calculated within a sliding window is used in conjunction with a threshold for initial event detection (green dots). These events are further culled in a semi-manual process aimed at eliminating false-positive epileptic events. The final set of events are preprocessed (downsampled and scaled), yielding the dataset of events (purple dots) used in the modeling phase. These data are uploaded to a cluster of Amazon EC2 instances for the modeling and quantitative analysis. A number of parallel MCMC chains of the HIW-spatial AR-HMM model are run, generating a number of MCMC samples, each of which contain (among others) model parameters (e.g., channel state AR coefficients, event state covariances) and event states corresponding to each epileptic event. The event states from these MCMC samples are then analyzed in parallel, yielding co-states and Hamming distances for each seizure event, which are then collected to yield max co-states and a representative sample across all MCMC samples. The results of this analysis are then used for qualitative event analysis.

which involves posterior MCMC inference over a number of parallel chains using the HIW-spatial AR-HMM model described in Section 4.7.1. The resulting posterior MCMC samples are then analysed and collected from the separate chains before we can use them in our qualitative analysis.

We describe these processing steps in detail in Section 5.1. In Section 5.2, we present the results of our seizure and burst analysis, examining which aspects of seizures tend to be most similar to the bursts and also exploring the subtle channel dynamics present in both bursts and seizures. We describe how our identification of the bursts most similar to seizures and compare the event state progressions between a seizure and its fifteen most similar bursts. Finally in this section, we investigate whether the bursts most similar to each seizure occur with any temporal regularity, e.g., just before the seizure occurs. In Section 5.3 we investigate the extent to which the physiological relationships involved in seizures can be seen from looking at only the bursts. Finally, in Section 5.4 we suggest a number of avenues to explore in future work.

■ 5.1 Quantitative Analysis of Epileptic Events

Modeling the epileptic bursts and seizures involves more than just running MCMC on a dataset. In this section, we describe in detail all aspects of event detection, dataset preparation, modeling, and subsequent quantitative analysis.

■ 5.1.1 Event Detection

The models discussed in Chapter 4 are capable of integrating data to up to a few million total time points (on 16 channels) for a reasonable runtime of no more than a few weeks. But the continuous records for the three dogs contain data from 475, 329, and 45 days, which correspond to roughly 16.4, 11.4, and 1.5 billion time points, respectively, for the 400 Hz sampling frequency, far more than we can hope to model and analyze with our existing techniques. We instead broke this long continuous record into a set of discrete epileptic events, since non-background epileptic activity is of most interest. Given the length of the continuous records, we developed a simple initial detection paradigm to reduce the scale of the data used in subsequent analysis.

Collecting candidate events For our initial event detection, we calculated the average line length feature [35] across all 16 channels within a 2-second moving window, yielding a scalar value that we used as a rough proxy for the degree of epileptic activity within that window. Time periods with average line length above a pre-defined threshold were saved as candidate events. These time periods were saved with an additional 30 seconds of padding at both ends of the above-threshold time period. For each dog, we started this simple event detection method for a few different threshold values before deciding on the value to use for the entire continuous data record. When in doubt, we chose a higher (less permissive) threshold with the intent that our pool of candidate events to be as "pure" (i.e., as few non-epileptic events and/or artifacts) as possible. This process was unique for each dog.

subject	recording length	initial detections	final detections	
			seizures	non-seizures
dog 002	475.7 days	1,846 (354.1 hrs)	37 (32.3 mins)	740 (30.9 mins)
dog 004	329.9 days	16,026 (1,320.7 hrs)	14 (15.2 mins)	758 (52.0 mins)
dog 005	45.8 days	6437 (4650.6 hrs)	91 (112.3 mins)	811 (65.7 mins)

Table 5.1. A summary for each of the three dogs examined of the total recording length, the number of initial event detections, and the number of final seizure and non-seizure detections after further culling.

Manual inspection and culling After this initial event detection, we further reduced the pool of events through a combination of quantitative and qualitative criteria. We always kept events containing the seizures previously marked by expert epileptologists. We excluded candidate events with maximum average line length feature values above a certain (manually-defined) value in an attempt to remove artifact candidate events, which often displayed large-amplitude, high-frequency noise simultaneously on all channels. Under the assumption that most of the non-seizure events would be short epileptic bursts on the order of 2-10 seconds, we eliminated any candidate events with above-threshold durations longer than a predefined time (usually 30 seconds). Similarly, we removed those candidate events with above-threshold durations less than a fixed length (usually, 500 ms) in an attempt to capture events with more than a single spike or slow-wave discharge. We also eliminated any candidate event containing missing data (as the continuous record occasionally does).

In one of the dogs (004), we noticed that three of the channels sometimes had quite noisy, high-amplitude recordings, due possibly to poor contact between the metal electrode contact and the cortical surface of the brain. This noise led to a very large number of false-positive candidate event detections. We thus recalculated the average moving-window line length feature on all the channels *except* those we had determined to be noisy and were able to exclude a large number of false-positive candidate events this way. These three channels were omitted for all the events of dog 004 in all subsequent analysis.

As the parameters for these candidate event elimination criteria are fairly subjective, we used a semi-manual approach in setting them. For each set of parameters, we quickly scanned through a number (usually a majority) of the resulting events to ensure that they contained as few false-positive detections as possible. When in doubt, we again erred on the side of having less permissive parameters, since we believed it more important that our resulting events dataset miss a few true positives rather than contain true-negatives. We chose to favor specificity over sensitivity. Events passing all of these criteria were saved with time padding (1 second for non-seizures, 10 seconds for seizures) before and after the above-threshold time points.

Table 5.1 gives a summary of the original data and event detections length for each of the three examined dogs. Notice that this process has reduced the length of the data from tens or hundreds of days in the original continuous recording to just a few hours worth of data in the final set of events. The magnitude of this degree of data reduction stems both from the sparsity of epileptic activity even in quite diseased dogs and also our stringent criteria

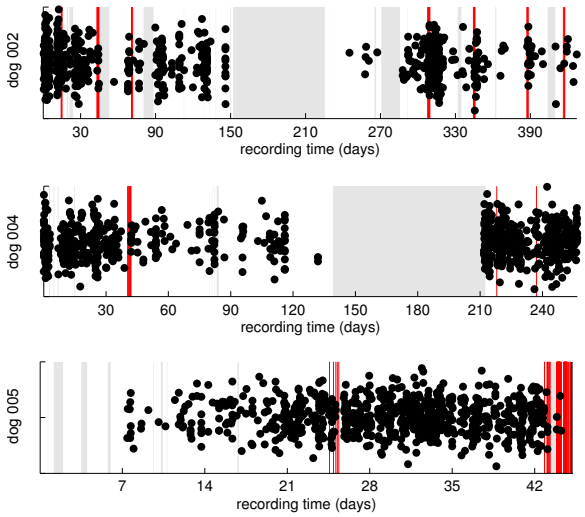

Figure 5.3. Timelines of the seizures (red bars) and sub-clinical bursts (dots jittered vertically for display) for each dog over the span of the continuous recording. Gray periods in the recording denote times of no available data. Though hard to see, most of the seizures occur in groups spaced a few hours from each other. The last 73 days of dog 004's record are omitted because only events excluded during the culling (and no seizures) occurred during that time frame.

for an event's inclusion in the final set for each dog. While this manual inspection process is inherently subjective and also inefficient, we felt these drawbacks were outweighed by the advantage of producing a dataset with very few false-positive detections for our downstream modeling and analysis. Furthermore, the manual component of this process usually only lasted a few hours.

Figure 5.3 shows timelines of the final set of epileptic events used for each dog. The final eighty or so seizures of dog 005 occurred within a few days of each other while the dog was in status epilepticus, a state of severe and repeated seizures associated with many negative physiologic changes, including greatly increased morbidity and mortality [33, chap. 64]. Dog 005 died while in status epilepticus. We note a few interesting characteristics of these status epilepticus seizures toward the end of Section 5.2.1.

Preprocessing After the sets of epileptic events were finalized, we low-pass filtered and downsampled the event iEEG from 400 Hz to 200 Hz, preserving what we believe to be the most clinically relevant aspects of the records (sub-100 Hz signals) while also reducing the modeling computational burden. We also rescaled the iEEG voltages to have a 99% confidence interval of roughly [-10, 10] (for numerical reasons) as in Section 4.4.1.

■ 5.1.2 Modeling

After epileptic event detection and preprocessing described in the previous section, we performed posterior inference on the events for each dog individually. We ran 10 chains

of the HIW-spatial AR-HMM for each of the three datasets, using the first 2500 MCMC iterations as burn-in and then taking 750 MCMC samples from each chain with 10-sample thinning, resulting in a 7500 total MCMC samples for each dog across the 10 chains.

MCMC using Amazon Web Services Unfortunately, the magnitudes of the dog datasets were too great for us to run MCMC on our local computational resources if we wanted to take advantage of the parallelization available in the AR-HMM (see Section 4.7.2) and also run multiple MCMC chains. We thus performed these MCMC computations on Amazon Web Services' Elastic Compute Cloud (EC2).

Our code is implemented in Matlab (with portions optimized in `C++`). Unfortunately, Matlab's current licensing requirements make it difficult to run native Matlab (and thus native Matlab programs) on generic machines such as those available on EC2. To do so would require installing a Matlab license manager on each machine and linking it to an existing license account, a task that while technically possible is prohibitively complicated for our purposes. Matlab offers an alternative to its native interpreted runtime environment, however, in the form of the Matlab compiled runtime environment (MCR). In this setting, existing Matlab code—including all necessary libraries and compiled MEX (`C` or `C++`) functions—can be compiled into either a `C` library or standalone application. This compiled code can then be run on any machine with Matlab's freely-available MCR installed.

Amazon Web Services[1] allow a user to start up an EC2 instance from an existing Amazon machine image (AMI), eliminating the need for the user to worry about installing an operating system and many other standard tools every time he needs to boot a new machine. Furthermore, the user can customize existing AMIs—making additional configurations and installing additional software as desired—and save to a new AMI. After creating this custom AMI once, the user can start an arbitrary number of EC2 instances starting with this machine image. We used this custom AMI capability to create an AMI with MCR and a few other additional utilities.

We then used the (free) third-party StarCluster[2] software to easily spawn and link together a number of EC2 instances into a single cluster running Sun Grid Engine (SGE). We used an EC2 m1.small instance type for the head node and either c1.xlarge or m2.xlarge spot instances for the worker nodes (one worker per MCMC chain).

Each MCMC chain was started as an SGE job, where particular parameters (priors, data file locations, etc) were specified in a unique external text file for each SGE job.

■ 5.1.3 Event State Analyses

Each MCMC sample contains a number of different values: channel state AR coefficients, event state covariances, state sequences for each channel and each event, transitions distribution parameters for each channel and each event, etc. All of these values have relevance and meaning for our epileptic event application, but we focus mostly on the event state sequences since they provide the most direct way to compare our two classes of epileptic

[1] See http://aws.amazon.com/documentation/ for extensive documentation.
[2] http://star.mit.edu/cluster/

events, bursts and seizures.

Event co-states One way to compare two particular events involves examining which time points in each tend to be assigned to the same event state over the range of MCMC samples. We can conclude that time points in two different events frequently assigned to the same event state display similar channel relationships. Consider two events indexed by e_1 and e_2 with time lengths $T^{(e_1)}$ and $T^{(e_2)}$ and event state sequences $Z^{(e_1,s)}_{1:T^{(e_1)}}$ and $Z^{(e_2,s)}_{1:T^{(e_2)}}$, respectively, for a particular MCMC sample s. Over S MCMC samples, the number of times $C^{(e_1,e_2)}_{t_1,t_2}$ a particular time point t_1 in event e_1 is assigned to the same event state as a particular time point t_2 in event e_2 is given by

$$C^{(e_1,e_2)}_{t_1,t_2} = \left|\{s \mid Z^{(e_1,s)}_{t_1} = Z^{(e_2,s)}_{t_2}, s = 1, \ldots, S\}\right|. \tag{5.1}$$

We often call these frequencies between two events the *co-states* between e_1 and e_2. A challenge, however, is that the space required to store the co-states $C^{(e_1,e_2)} \in \mathbb{Z}_+^{T^{(e_1)} \times T^{(e_2)}}$ is relatively large. Consider the case of storing the event co-states between all the time-points of all the events. For E events and total event time points $T = \sum_{e=1}^{E} T^{(e)}$, an upper-triangular matrix of these co-states would have $T(T-1)/2$ (potentially) non-zero entries. For dog 002, $T \approx 760,000$ (the smallest of the three dogs), which entails roughly 289 billion non-zero entries. Storing these values as 16-bit integers would require roughly 578 GB of disk space, probably more than is practical. In addition, calculating this number of values would take an impractical length of time without implementing the embarrassingly parallel task on a graphical processing unit chip or in some sort of large-scale Map/Reduce framework. For practical time constraints, we were not interested in pursuing these implementation options.

We thus chose to limit our event co-state analysis in a few ways. First, we only calculate co-states between each seizure and each other non-seizure (i.e., the seizure/seizure and non-seizure/non-seizure co-states are not calculated). Since we are most interested in comparing seizures to non-seizure bursts, this reduction is sensible. In addition, we use Huffman encoding [77, chap. 5] to compress the co-state counts (achieving a compression ratio of roughly 5x).

Co-state frequencies between each seizure and each burst were calculated separately (in parallel) for each chain on a cluster of EC2 instances as described in the previous section. The co-state counts for each chain were subsequently added together to yield counts across all chains.

While visualizing the matrix $C^{(e_1,e_2)}$ of co-states between two events is trivial, we often desire a method for summarizing this information in a way that can be displayed under the EEG of both events or summarized for one seizure across all the other bursts, for example. We achieve this co-state summarization by storing the maximum frequencies across the rows

and columns of a co-states matrix $C^{(e_1,e_2)}$,

$$\begin{aligned} m_{t_1}^{(e_1,e_2)} &= \max_{t_2=1,\ldots,T^{(e_2)}} \left(C_{t_1,t_2}^{(e_1,e_2)} \right), \\ m_{t_2}^{(e_2,e_1)} &= \max_{t_1=1,\ldots,T^{(e_1)}} \left(C_{t_1,t_2}^{(e_1,e_2)} \right). \end{aligned} \tag{5.2}$$

The vector $\mathbf{m}^{(e_1,e_2)}$ gives the maximum co-state frequency in each time point of event e_1 over all the time points of event e_2, and vice versa for $\mathbf{m}^{(e_2,e_1)}$. We can also interpret $m_{t_1}^{(e_1,e_2)}$ as the maximum posterior probability that event e_1's time point t_1 is clustered with least one of e_2's time points. The interpretation for $m_{t_2}^{(e_2,e_1)}$ is similar.

Representative MCMC Samples As previously described in Section 4.4.1, it is often useful to display a single set of event states rather than the states over all the MCMC samples. Using the Munkres assignment algorithm and minimum Hamming distance method described in Section 4.4.1, we determined a representative MCMC sample for each seizure individually. The Munkres assignment algorithm and Hamming distance calculations were performed for each chain in parallel on the same cluster of EC2 instances previously described. We then stitched together these Hamming distances into a single $S \times S$ upper-triangular matrix (for S total MCMC samples).

■ 5.2 Seizure and Burst Similarities

The primary clinical goal of this work is to better understand the relationship between shorter epileptic bursts and longer seizures. The maximum event co-states and representative event states produced by the quantitative analysis procedure described in the previous section give us one way to explore this relationship.

■ 5.2.1 Periods in Seizures of Most Pronounced Burst Similarity

We first examine what time points with each seizure overlap most often with those in the bursts. Recall that the maximum event co-states $\mathbf{m}^{(e,b)}$ in a seizure e for a particular burst b give the maximum (posterior) probability that each time point of the seizure e is clustered with at least one of the time points of the burst b. Averaging over the B bursts, we get an average maximum probability that each time point of seizure e clusters with at least one of the time points of a burst,

$$\bar{\mathbf{m}}^{(e)} = \frac{1}{B} \sum_{b} \mathbf{m}^{(e,b)}, \tag{5.3}$$

where b indexes each non-seizure bursts. This probability allows us to visualize which time points of a particular seizure are more likely to cluster with those of the bursts.

Examining individual seizures for each dog Figure 5.4 shows these time point averages $\bar{\mathbf{m}}^{(e)}$ for a representative seizure from each dog. We first notice that all three seizure onsets are the most similar to the bursts. It is interesting, though perhaps not surprising, to notice

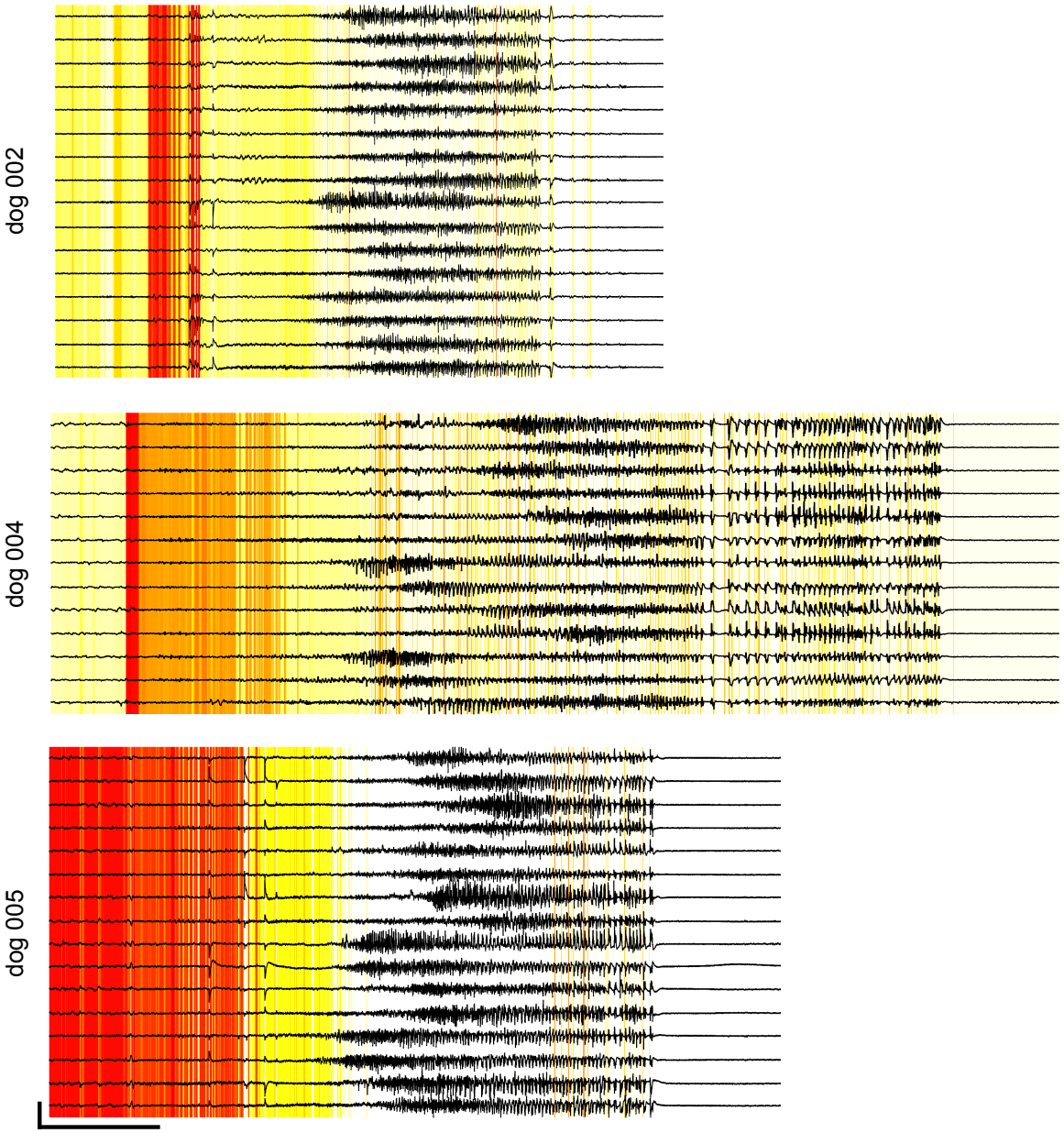

Figure 5.4. The average maximum event co-states $\bar{\mathbf{m}}^{(e)}$ for a representative seizure for each dog. For each time point of seizure e, we average (across the bursts) the maximum event co-states $\mathbf{m}^{(e,b)}$ between the seizure e and every burst b. These maximum event co-state averages are displayed in color under the corresponding EEG time points and convey the average similarity between the time points of the seizure and those of the bursts. White indicates low values and red high values. The horizontal scale bar denotes 10 seconds and the vertical 1 mV.

that the higher magnitude activity in the height of the seizure has relatively little similarity with the bursts. In dogs 002 and 004, it seems that the state(s) of the initial discharges are most present in the bursts, whereas in dog 005 the entire onset is very similar. These results, especially those of the initiating discharges of dogs 002 and 004, reach a similar finding as our analysis in Section 4.6, where the onset patterns of the sub-clinical bursts and the clinical seizures progressed through similar event states.

It is worth noting that the average maximum probability metric $\bar{\mathbf{m}}^{(e)}$ displayed under the EEG in Figure 5.4 is unable to distinguish between a few bursts with high maximum probability and many more bursts with lower maximum probability. To further examine how these maximum probabilities vary over the bursts, we show them stacked in the rank-order of the burst occurrence (in time) in Figure 5.5. The average maximum probabilities depicted in Figure 5.4 are simply the column-wise average of those shown in Figure 5.5. Note how temporally (and thus rank-order) proximate bursts often display quite similar maximum co-state probabilities. The bursts in dogs 002 and 004 display much more heterogeneous probabilities than those in dog 005. In all three dogs, we see how the seizure onset region contains the highest similarity with the bursts, though this large similarity is sometimes (in the case of dogs 002 and 004) present in only a subset of the bursts.

Examining all seizures for each dog Figure 5.6 shows the average maximum co-states across all the seizures for each dog, where each row contains the same information underlying the EEG shown in Figure 5.4. This visualization allows us to see how the average maximum co-states change over the seizures. Of particular interest, note how the seizures within each group (denoted by the horizontal black lines) tend to display similar average maximum co-states.

In dog 002, the first two groups of seizures have little onset similarity with the bursts, though the later groups all display strong onset similarities. The high-amplitude seizure activity is in general not very similar to the bursts, though very discrete periods of the offsets tend to display strong similarities. The flickers of similarity occasionally present at the end of the seizures occur at discrete, low-amplitude post-ictal discharges.

In dog 004, 14 of the 16 recorded seizures occurred within the period of just a few days. These all tend to display strong onset similarities with a subset of the bursts, generally those bursts that occur before the large data gap shown in Figure 5.3. The two seizures occurring much later in the record contain similarities across more of the bursts. As in dog 002, all of the seizures in dog 004 contains patterns of very brief but very strong similarity at seizure offset.

Dog 005 is a particularly interesting case in that it contained two main groups of seizures, the second of which occurred while the dog was in status epilepticus. The five seizures in the first group all contain very strong onset similarities and a few brief periods of offset similarity across almost all the bursts, as in dogs 002 and 004. The second group of seizures offer a fascinating perspective of how the physiologic changes associated with status epilepticus (SE) include changes in iEEG seizure dynamics.

In the early stages of SE (phase I), cerebral blood flow and oxygen keep up with the significant metabolic demand associated with epileptic seizures. But as these seizures continue

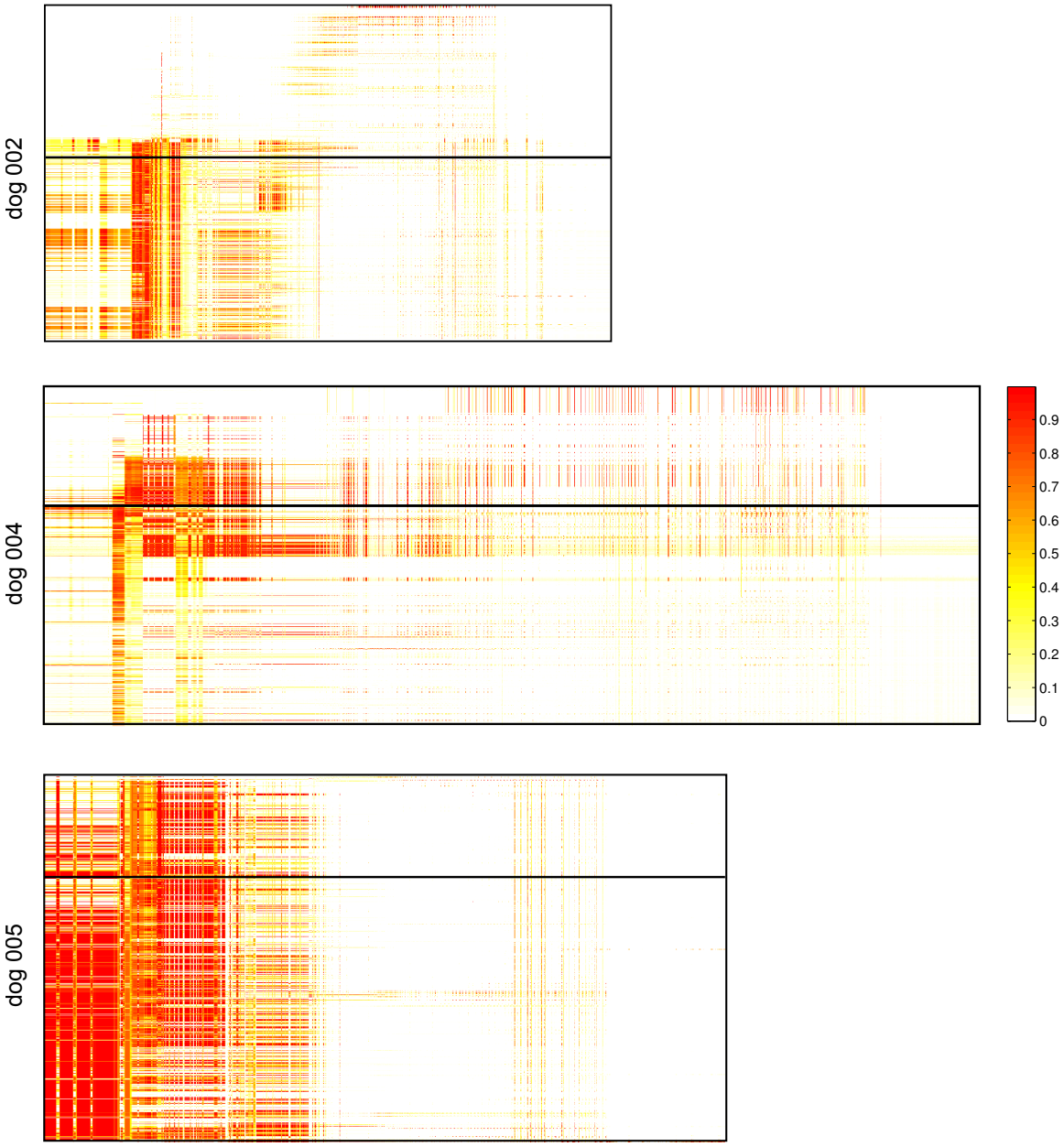

Figure 5.5. The maximum event co-state probabilities $\mathbf{m}^{(e,b)}$ across all the bursts b (rows) for the representative seizures e shown in Figure 5.4. The columns of each matrix denote the time points of the seizure. The horizontal black line denotes the rank position of the seizure within the bursts. The color scale denotes the maximum posterior probability of each seizure time point clustering with at least one of the time points in each burst. The scale is the same across the three dogs.

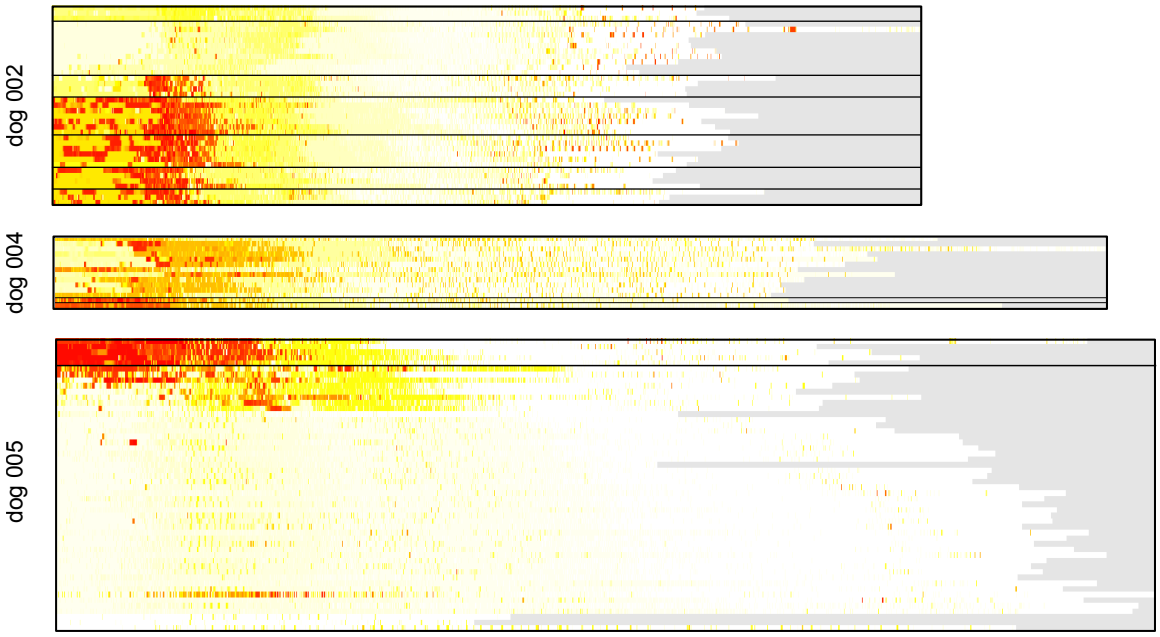

Figure 5.6. The average maximum event co-states $\bar{\mathbf{m}}^{(e,b)}$ for all the seizures of each dog. The maximum co-states of each seizure (rows) are stacked down from the top in the rank order in which they occurred. Horizontal black lines denote groups of seizures occurring with fewer than 24 hours between each seizure. See Figure 5.3 for an alternate visualization of the groups of seizures. Each dog's seizures are shown left-aligned by the start of the event. Since the seizures were not necessarily the same length, time points beyond the end of a seizure event are shown in gray. Only the first 53 seizures of dog 005 are shown for brevity.

into phase II SE, cerebral glucose and oxygen levels drop, with the previous hypermetabolism transitioning into a hypometabolism and ischemia. These changes in metabolic rate occur simultaneously with reductions in cerebral perfusion, increases in intracranial pressure, and increased blood-brain barrier permeability [33, pgs. 743-744]. Thought it is hard to determine definitively, the second group of dog 005's seizures seem to parallel these other physiologic changes from phase I to phase II SE. In particular the first seven seizures in the second group display tapering degrees of onset similarity with the bursts. Roughly 14 hours after the seventh of these seizures, another string of seizures with very little burst similarity (onset or otherwise) begins, ending (presumably) in the dog's death. This transition from seizures in phase I SE with burst similarities to seizures in phase II SE without the similarities allows us to see how the brain veers farther and farther away from baseline epileptic activity as SE progresses.

■ 5.2.2 Identifying Subtle Relationships Common to Seizures and Bursts

Having shown in the previous section that seizures often display particular similarity to bursts in their onsets, we now turn to a more focused analysis of these similarities. Which

time points and/or event states of the seizure onset are most similar to those of the bursts? How confident can we be that the event state sequence assignments are actually intuitive and illuminating? We begin to address these questions by examining a particular seizure from dog 002. In the fine-grain analysis we undertake in this section and the next, it is expedient to focus on only a particular representative seizure, though the analysis methods employed of course generalize to any seizure. Furthermore, for computational simplicity and explicative clarity, we use the event state sequences and model parameters from a single representative MCMC sample, though the analysis could also be generalized across the MCMC posterior samples.

Periods most similar between a seizure onset and bursts We begin by examining the five seconds at the start of the seizure's onset. The EEG of the full seizure and the zoomed-in five seconds of onset are shown in Figure 5.7. We compare this seizure onset to three bursts with time points having a high probability of being clustered with the time points in the seizure onset. We show the full onset event state sequence in the second row of Figure 5.7, and in the bottom three rows we highlight only those time points $\mathcal{T}^{(e)}$ and $\mathcal{T}^{(b)}$ of the seizure e and burst b, respectively, with a reasonably high probability ($p > 0.75$) of being clustered with at least one of the time points of the other event,

$$\begin{aligned} \mathcal{T}^{(e)} &= \{t \mid m_t^{(e,b)} > \rho, t = 1, \ldots, T^{(e)}\}, \\ \mathcal{T}^{(b)} &= \{t \mid m_t^{(b,e)} > \rho, t = 1, \ldots, T^{(b)}\}, \end{aligned} \tag{5.4}$$

where $\rho = 0.75$. We notice that some bursts like that shown in the third row have event state sequences that are quite similar to the seizure onset. That burst progresses from the red to the green to the cyan event states, similar in many ways to the onset's event state progression. It is also interesting to note where the burst and seizure onset differ. For example, a brief orange event state occurs twice in the third-row burst (which is actually most likely parts of two bursts) that denotes a synchronized discharge. These brief periods in the bursts are clustered with a similar-looking discharge in the seizure onset. In addition, the event states leading up to this discharge (red, then green, then cyan) seem to be a subset of those leading up to the first discharge in the seizure onset. But after this single discharge, the seizure then progresses into a set a much higher-amplitude discharges than those demarcated by the orange state in the burst. It is almost as if the burst displays the dynamics of a mini seizure onset but then "resets" and does not transition to higher-amplitude discharges and a full seizure.

The burst compared to the seizure onset in the fourth row of the figure, shows how some of the detected "bursts" may actually contain few or no changes in event state. In the event detection phase of our experimentation, we doubted these bursts with static EEG (which mostly occurred in the first two months of the dog 002's recording) would have any relevance to seizures but included them anyway to avoid undue manual selection bias. Interestingly, the static EEG in the burst is actually quite similar to periods in the seizure onset—shown in cyan—leading up to the discrete discharges. It is almost as if the channels in the burst are stuck in an intermediate state involved in the onset of an eventual seizure.

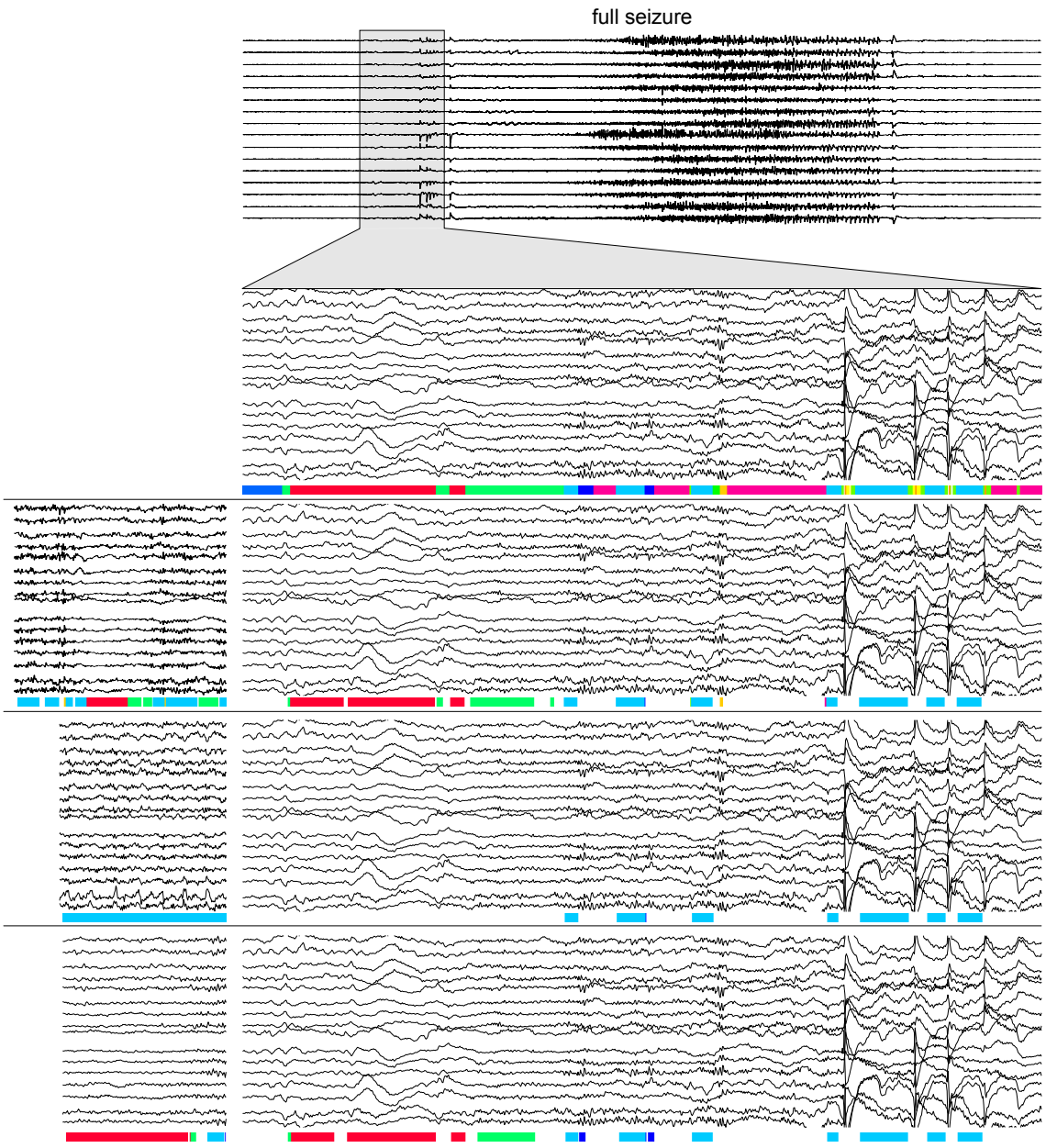

Figure 5.7. A comparison of the onset from a representative seizure of dog 002 to three bursts. The full seizure and onset period examined are shown at the top of the figure. The expanded onset EEG and event states (shown by different colors across all the channels) are shown directly below the full seizure. The three bottom rows show the seizure onset EEG and event states compared to those of three bursts similar to the seizure. Event states in the bursts and the seizure are shown only for those time points $\mathcal{T}^{(b)}$ and $\mathcal{T}^{(e)}$, respectively, with a high probability ($p > 0.75$) of clustering with at least one of the time points of the other event.

The burst shown in the bottom row of Figure 5.7 in displays perhaps a middle ground between the other two bursts. While the EEG dynamics do evolve over the course of the burst, they do not lead to as pronounced a discharge as the burst in the third row. Similar "ramping up" dynamics seem to occur several times in the onset before they progress to higher-amplitude discharges.

Confirming event state similarities between a seizure and the bursts As much of the quantitative analysis in this chapter hinges upon the event states, we wanted to confirm that they actually denote time points with common dynamics across all events, both seizures and bursts. While the periods of EEG assigned to the same event states in Figure 5.7 do qualitatively seem similar, we wanted like to quantitatively support this claim. One way to accomplish this validation involves calculating the channel covariances of the time points over all the events for any event state of interest.

We thus calculated the covariances at time points assigned to four states (red, green, cyan, orange) shown in Figure 5.9 for the same seizure but now across all the bursts of dog 002. For each event state l, we calculated these covariances across the time points $\mathcal{T}_l^{(e)}$ of the single seizure e and also across the time points $\mathcal{T}_l^{(\mathcal{B})}$ of all bursts $\mathcal{B}$,

$$\begin{aligned} \mathcal{T}_l^{(e)} &= \left\{ t \mid z_t^{(e)} = l, t = 1, \ldots, T^{(e)} \right\}, \\ \mathcal{T}_l^{(\mathcal{B})} &= \left\{ (b,t) \mid z_t^{(b)} = l, t = 1, \ldots, T^{(b)}, b \in \mathcal{B} \right\}. \end{aligned} \tag{5.5}$$

We calculated these covariances using original EEG data values $\mathbf{y}_t$,

$$\begin{aligned} \widehat{\Sigma}_{l,\mathbf{y}_t}^{(e)} &= \frac{1}{\left|\mathcal{T}_l^{(e)}\right| - 1} \sum_{t \in \mathcal{T}_l^{(e)}} \mathbf{y}_t^{(e)} \mathbf{y}_t^{\mathrm{T}(e)}, \\ \widehat{\Sigma}_{l,\mathbf{y}_t}^{(\mathcal{B})} &= \frac{1}{\left|\mathcal{T}_l^{(\mathcal{B})}\right| - 1} \sum_{(t,b) \in \mathcal{T}_l^{(\mathcal{B})}} \mathbf{y}_t^{(b)} \mathbf{y}_t^{\mathrm{T}(b)}, \end{aligned} \tag{5.6}$$

as well as the innovation values $\boldsymbol{\epsilon}_t = \mathbf{y}_t - \mathbf{A}_{\mathbf{z}_t} \widetilde{\mathbf{Y}}_t$, calculated by subtracting off the assigned channel state AR prediction from the original data values,

$$\begin{aligned} \widehat{\Sigma}_{l,\boldsymbol{\epsilon}_t}^{(e)} &= \frac{1}{\left|\mathcal{T}_l^{(e)}\right| - 1} \sum_{t \in \mathcal{T}_l^{(e)}} \boldsymbol{\epsilon}_t^{(e)} \boldsymbol{\epsilon}_t^{\mathrm{T}(e)}, \\ \widehat{\Sigma}_{l,\boldsymbol{\epsilon}_t}^{(\mathcal{B})} &= \frac{1}{\left|\mathcal{T}_l^{(\mathcal{B})}\right| - 1} \sum_{(t,b) \in \mathcal{T}_l^{(\mathcal{B})}} \boldsymbol{\epsilon}_t^{(b)} \boldsymbol{\epsilon}_t^{\mathrm{T}(b)}. \end{aligned} \tag{5.7}$$

Figure 5.8 shows these four event states covariances: the original EEG data values for the seizure, the original EEG data values for all the bursts, the innovation values for the seizure, and the innovation values for all the bursts. This figure shows how modeling the covariances of the *innovations* is crucial for effectively finding periods with similar relationships in the bursts and in the seizures.

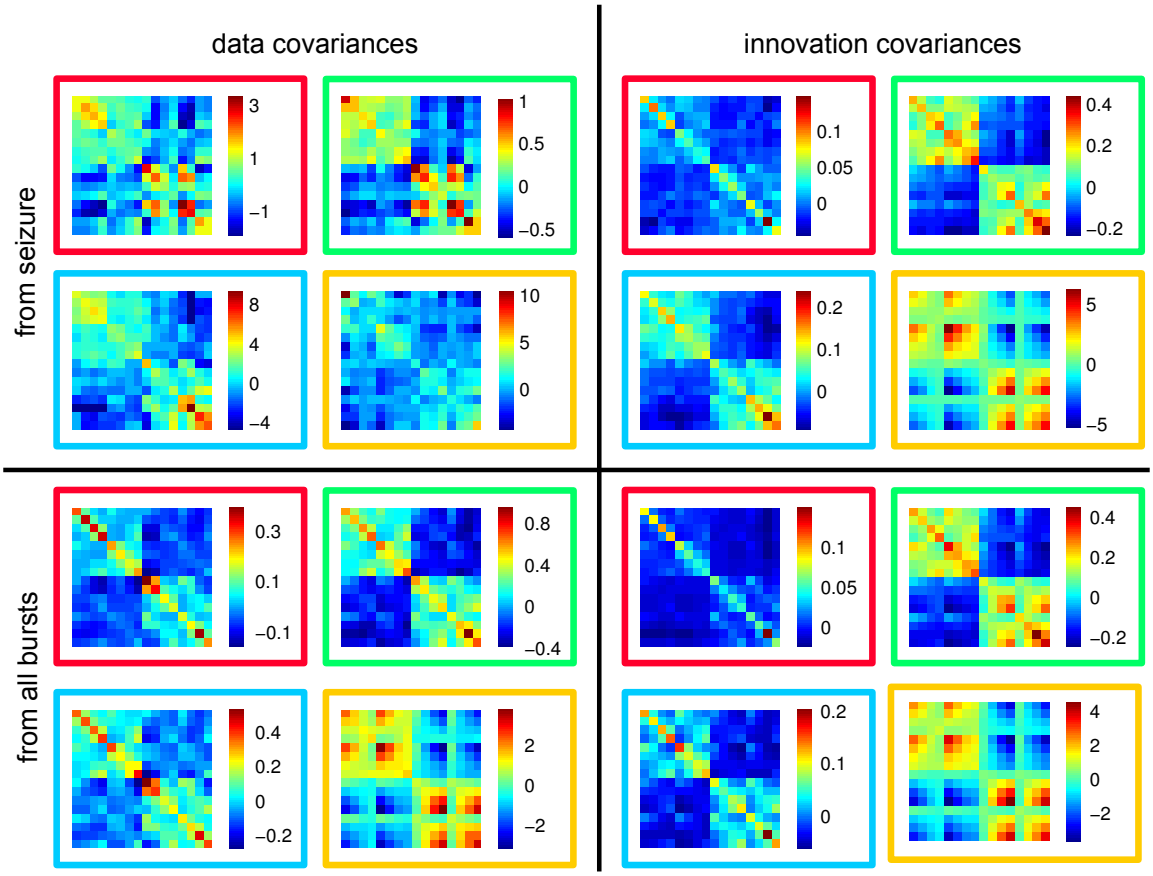

Figure 5.8. A comparison of covariances corresponding to four main event states (red, green, cyan, orange) shown in Figure 5.7 calculated in four different ways. In the upper left quadrant, the covariances $\widehat{\Sigma}^{(e)}_{l,\mathbf{y}_t}$ for the *EEG data* from the *single seizure* are shown using only EEG data in the time points assigned to each of the four event states in their respective covariances. The bottom left quadrant shows the covariances $\widehat{\Sigma}^{(\mathcal{B})}_{l,\mathbf{y}_t}$ for the *EEG data* from *all bursts* for each event state. In the top right quadrant, the covariances $\widehat{\Sigma}^{(e)}_{l,\boldsymbol{\epsilon}_t}$ of the *innovations* (of the channel state AR predictions) from the *single seizure* are shown for the values corresponding to each event state. The bottom right quadrant shows the covariances $\widehat{\Sigma}^{(\mathcal{B})}_{l,\boldsymbol{\epsilon}_t}$ of the *innovations* from *all bursts* for each event state.

Let us first consider the covariances $\Sigma^{(e)}_{l,\mathbf{y}_t}$ and $\Sigma^{(\mathcal{B})}_{l,\mathbf{y}_t}$ calculated on the original data values from the seizure and from all the bursts, respectively. We notice that the covariances of each of the four event states are quite different between the seizure and the bursts. Not only do they have different correlation structure (shown by the patterns within the covariance matrix), but they also have quite different magnitudes (as seen by the scale bars). The similarities we believe exist between the channel relationships are not apparent when examining only the original data values. The channel relationships in these four states are obscured by the correlations and magnitude of the individual channel dynamics. But after subtracting off these individual channel dynamics and working with the channel innovations, we get much clearer picture of how similar channel relationships are indeed present across the seizure and bursts, both in correlation structure and in magnitude. This necessity of modeling the *innovation* covariance rather than that of the original EEG signal is intuitive since one some level bursts are inherently different—especially in voltage magnitude—than seizures, and so any method that wants to compare the two classes of events must account for this difference.

■ 5.2.3 Finding Bursts Most Similar to Seizures

In the previous section, we showed how some bursts can contain state transition similar to those found in a seizure onset. In Figure 5.7 we examine a few bursts with high average maximum event co-states with the seizure, but we also wanted to examine bursts with similar event state *transitions* rather than just the event states themselves. Consider again two events indexed by e_1 and e_2 with event state sequences $Z^{(e_1)}_{1:T^{(e_1)}}$ and $Z^{(e_2)}_{1:T^{(e_2)}}$. From a state sequence $Z^{(e_1)}_{1:T^{(e_1)}}$, we can construct a state transition matrix $\mathbf{n}^{(e_1)} \in \mathbb{Z}_+^{L\times L}$ (for L event states) with elements n_{jl} that give the number of times the sequence transitions from state j to state l,

$$n^{(e_1)}_{jl} = \left|\{t \mid Z^{(e_1)}_{t-1} = j, Z^{(e_1)}_t = l, t = 1, \ldots, T^{(e_1)}\}\right|. \tag{5.8}$$

Note that this transition counts matrix $\mathbf{n}^{(e_1)}$ is involved in posterior sampling of the event state transition parameters $\boldsymbol{\phi}^{(e_1)}$ as in Equation (4.28). We could alternately use the transition parameters $\boldsymbol{\phi}^{(e_1)}$ to describe the state transitions but choose $\mathbf{n}^{(e_1)}$ in this analysis for its simplicity. We thus compare the event state transitions in events e_1 and e_2 through their transition counts matrices $\mathbf{n}^{(e_1)}$ and $\mathbf{n}^{(e_2)}$.

If these two events are a seizure and burst, time lengths $T^{(e_1)}$ and $T^{(e_2)}$ are probably on a quite different scale, so it makes little sense to compare the actual values in the elements of each matrix. Instead, we work with a $\{0, 1\}$ truncation of the counts that indicates only whether the state sequence $Z^{(e_1)}_{1:T^{(e_1)}}$ *ever* transitions from state j to state l,

$$\widetilde{n}^{(e_1)}_{jl} = \max(n^{(e_1)}_{jl}, 1). \tag{5.9}$$

The number of non-zero elements in the intersection of $\mathbf{n}^{(e_1)}$ and $\mathbf{n}^{(e_2)}$ yields a proximity

measure that captures the similarity between the transitions of events e_1 and e_2,

$$P_{\widetilde{\mathbf{n}}}^{(e_1,e_2)} = \left|\{(j,l) \mid n_{lj}^{(e_1)} > 0 \wedge n_{lj}^{(e_2)} > 0\}\right| = \left|\{(j,l) \mid \widetilde{n}_{lj}^{(e_1)} = 1 \wedge \widetilde{n}_{lj}^{(e_2)} = 1\}\right|. \tag{5.10}$$

A close reading of a seizure onset and the bursts most similar to it Figure 5.9 shows the first fifteen seconds of the onset from the same seizure explored in the previous section along with its event state sequence for the (same) representative MCMC sample. In addition, we show the EEG and event states of the fifteen bursts most similar to this seizure onset period (rather than the whole seizure). Note that ten of these burst events actually contain multiple discharges that have been grouped together by our event detection algorithm because they occur so close to each other.

Using the EEG alone, it is difficult to compare the channel relationships (and how they change) between the bursts and the seizure onset. The event state sequences facilitate this comparison somewhat. In many of the bursts, though, we notice the transition from the red (29) state to the green (12) state to the cyan (16) state occurs, as it does in the seizure onset. Not surprisingly, the similarity metric described in Equation (5.10) has found bursts similar to that the third row of Figure 5.7.

Nevertheless, it is difficult in this visualization to progress much beyond the transitions in these three states to a more comprehensive state transition comparison. We thus appeal to the finite automata-like diagrams described in Section 4.4.1 to give an alternate picture of the event state transitions in the seizure onset and the bursts. Figure 5.10 shows these diagrams for the seizure onset and each of the bursts shown in Figure 5.9. For compactness, we have excluded the least frequent event states together comprising less than 5% of the time points across the seizure onset and all the bursts. While these diagrams are complex, a few key aspects of the seizure onset and burst transitions stand out, all of which are confirmed upon reexamination of the event state sequences shown in Figure 5.9. First, state 29 (red) is the most common state in almost all of the bursts, followed by (in no particular order) by states 12 (green), 16 (cyan), and 18 (lighter blue). These four states all appear at the beginning of the seizure onset. These states often transition between each other in the bursts. Second, most of these bursts transition from state 16 to state 4 (orange), often by way of state 20 (darker blue). This later transition pattern is similar to one in the seizure onset. The time spent in states 4 and 20 is usually brief. Third, at this point, the bursts either end or seem to "reset," transitioning briefly to state 27 or perhaps another state not shown and then eventually back to state 29. This ending or resetting in the bursts is now different from the dynamic in the seizure onset, where states 4 and 20 transition to state 27, which occupies a large number of the time points and transitions a number of times between itself and state 23 (purple). This difference between the transition patterns of the bursts and seizure onset seems to be the distinguishing feature of the two. For some physiologic reason, the bursts transition quickly through state 27 (or not at all) while the seizure onset stays in that state longer before moving to a new state (23) that then marks its metamorphosis into a full-blown seizure.

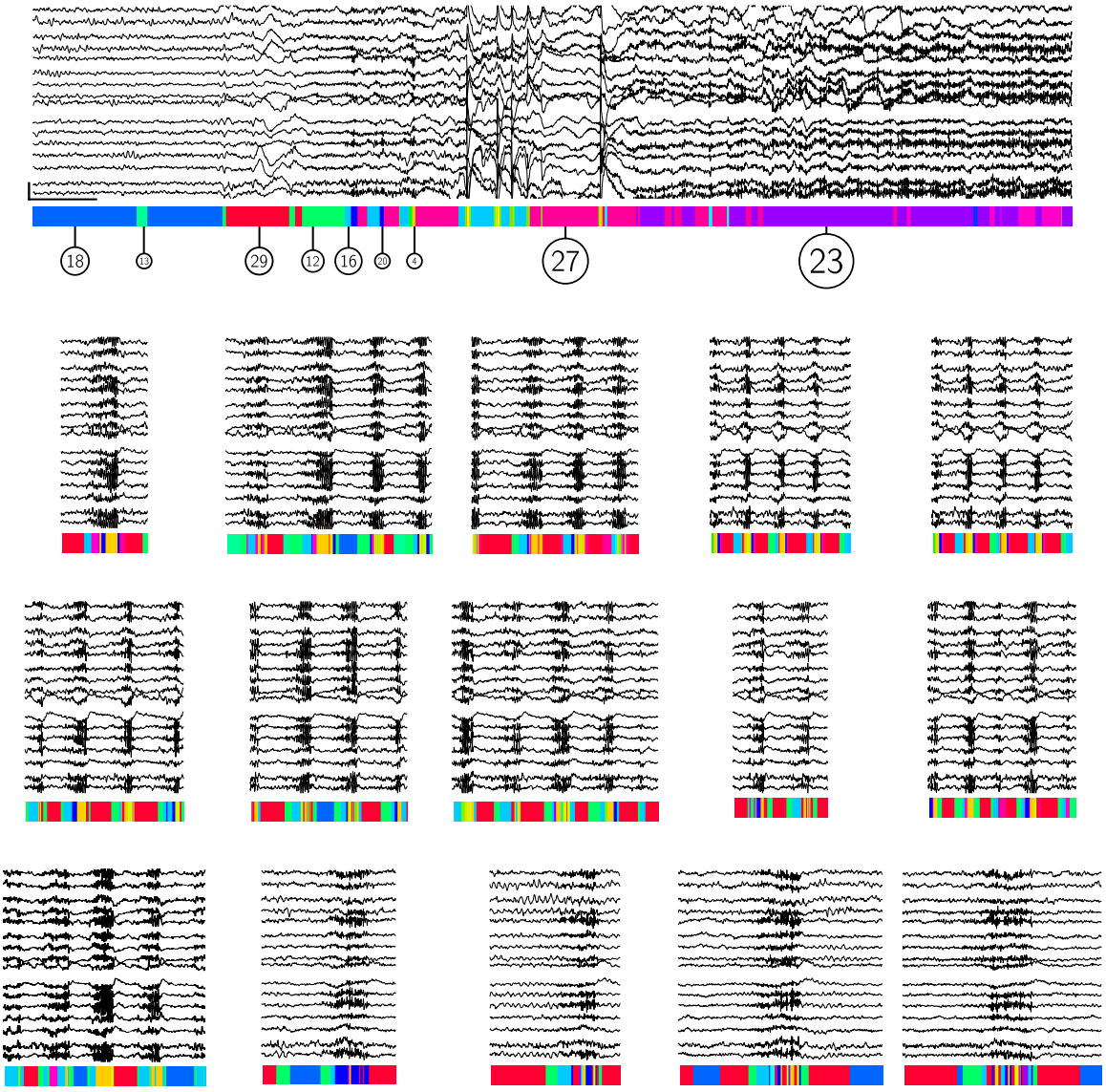

Figure 5.9. (**top row**) The EEG of a seizure onset and its assigned event state sequence. The colors denoting different event states are also labeled for reference in Figure 5.10, where the area of the labels is proportional to the number of time points assigned to that state. (**bottom three rows**) The EEG and event states for the fifteen bursts with event state transition matrices most similar to that of the seizure onset. The event states were taken from a representative MCMC sample for the seizure event states. Vertical and horizontal scale bars denote 250 μV and 1 second, respectively.

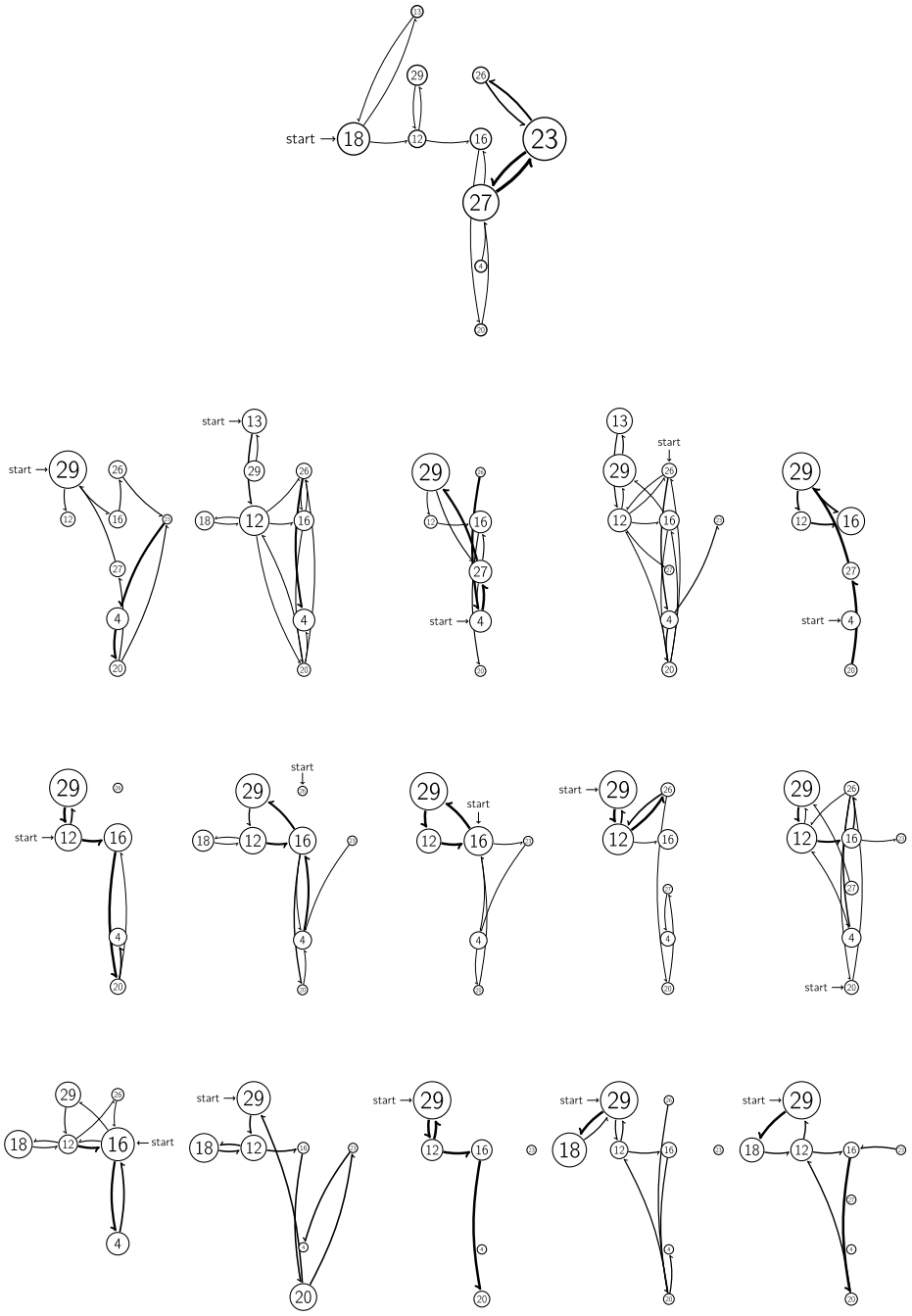

Figure 5.10. The event state transition diagrams for the seizure onset (**top row**) and bursts (**bottom three rows**) shown in Figure 5.9. Event states accounting for less than 5% of the total time points are not shown for compactness.

Due to the diversity of seizures within dog 002 and especially between the three dogs, we hesitate to draw any global conclusions about particular event state transition patterns based on our detailed analysis of this single seizure and fifteen similar bursts. Though we have examined other seizures of dog 002 and found comparable relationships (both the similarities and differences in event state transitions) between the seizure onset and the bursts, it would be impractical to do this type of "close reading" for the 142 seizures across the three dogs. Rather, we believe such analyses of individual seizure onsets lend support to our broad claim that the sub-clinical bursts display channel relationships in many ways quite similar to those of the seizure onset.

Examining burst similarity over time Given that our analysis indicated that bursts contain many similarities with seizure onsets, we were interested in examining whether the most similar bursts have any temporal relationship to seizures. For example, perhaps the most similar bursts occur just before the seizure. If there is some sort of temporal relationship, perhaps these bursts could be used in a seizure prediction setting, where the bursts indicate a heightened risk of seizures. To explore this potential relationship, we calculated the similarities of all the bursts to each individual seizure of each dog. We used two similarity measures between each burst b and each seizure e: the average over the maximum event co-state frequency $\mathbf{m}^{(b,e)}$ defined in Equation (5.2),

$$P_{\mathbf{m}}^{(e,b)} = \frac{1}{T^{(b)}} \sum_{t=1}^{T^{(b)}} \mathbf{m}_t^{(b,e)}, \tag{5.11}$$

and the similarity $P_{\tilde{\mathbf{n}}}^{(e,b)}$ (see Equation (5.10)) between the event state transition matrices.

Figures 5.11, 5.12, and 5.13 show similarities of all the bursts for a number of representative seizures from each of the three dogs, respectively, as calculated by the two similarity measures. While some patterns emerge from these figures—like the roughly equal similarity of all of dog 002's bursts after the first few months to all the subsequent seizures (see also Figure 5.4)—no striking temporal relationships seem consistent across events to the seizures of each dog individually. In general, these bursts seem equally likely both before and after the seizures, exhibiting no strong temporal localization that would be amenable to seizure prediction methods.

Of course, other burst detection methods and other similarity metrics than what we could have used may possibly lead to such a temporal relationship between the bursts and the seizures. Our event detection method was aimed at finding events similar to seizures rather than events that might be predictive of seizures. Instead of detecting events with a channel *average* feature above a threshold, one might detect events with a channel *maximum* feature above a threshold. Furthermore, the events that are predictive of seizures (if such events even exist) may not actually be similar to seizures, as we assume they would be in our analysis. Thus we can only say with confidence that the burst similarities shown in Figures 5.11, 5.12, and 5.13 do not convey any kind of special temporal relationship between the bursts we have detected and the seizures.

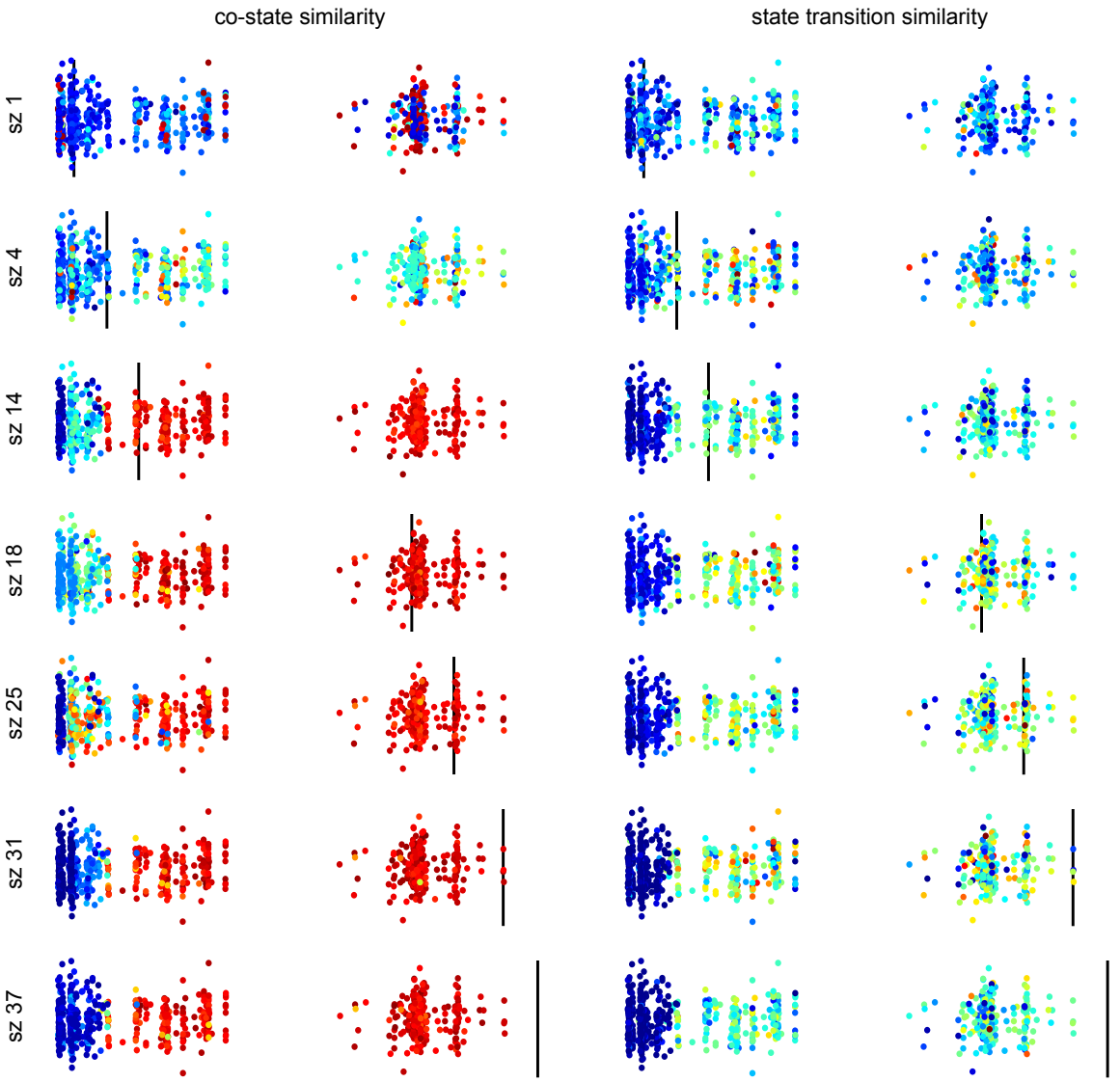

Figure 5.11. The similarities of all the bursts (small dots) with seven representative seizures for dog 002. The horizontal position of the dots and vertical black line denote the relative time of the bursts and seizure each occurred over the entire recording, respectively. Similarities are shown for two different calculation methods: average maximum event co-state similarity $P_{\mathbf{m}}^{(e,b)}$ and the event state transition matrix similarity $P_{\tilde{\mathbf{n}}}^{(e,b)}$. The color of the dots denotes the degree of similarity, with blue denoting less similarity and red more similarity. The scale of the colors is different for each seizure.

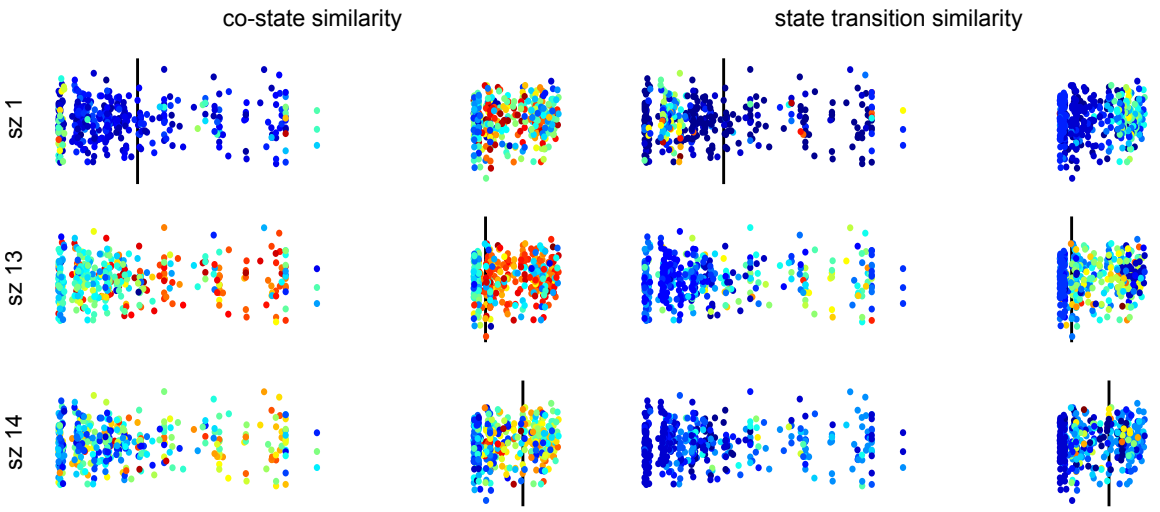

Figure 5.12. The similarities of all the bursts (small dots) with three representative seizures for dog 004 displayed in the same fashion as described in Figure 5.11.

■ 5.3 Understanding Physiologic Relationships from Bursts Alone

Some human epilepsy patients have very few seizures while their iEEG is recorded in an epilepsy monitoring unit. Since seizures provide the onset information epileptologists require to determine what area(s) of the brain to remove, a paucity of seizures means a paucity of information the physicians can use in their decision making process. Since our results indicated a similarity between sub-clinical bursts and seizure onsets, we wondered if we could get spatial information similar that derived from the seizure onsets by integrating over the bursts.

On some level, the bursts will never be able to fully describe the seizure onset because they are fundamentally a different class of event. In the previous section, we discussed the specific similarities present between the state transitions of the seizure onset and similar bursts. But these two classes of events were similar up to a point where the bursts ended and the seizure onset shifted into a new, different state transition progression. The event states in this new, seizure onset-only state progression were not well described in the bursts.

Despite this ultimate distinction, we were interested in exploring how similar state transition patterns of bursts match those of a seizure onset. We used the event state transitions of all the bursts that occurred before the representative seizure we have studied in previous sections. By summing the state transitions over these bursts, we produced a rough description of the transition patterns present in only the previous bursts.

Figure 5.14 shows the event state transition diagram for only the seizure onset and the diagram for the transitions of all the previous bursts. We have also included the event state covariances for a number of the prominent event states. We notice that much of the time points of the bursts are spent in states with low-correlation, low-magnitude covariances like

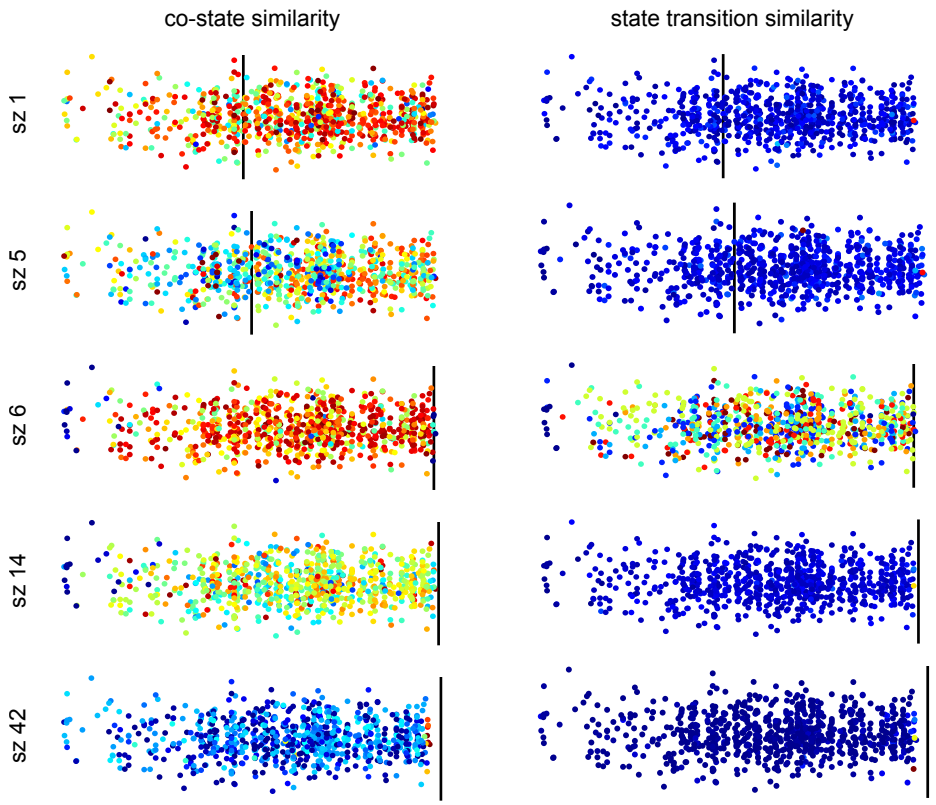

Figure 5.13. The similarities of all the bursts (small dots) with three representative seizures for dog 005 displayed in the same fashion as described in Figure 5.11.

29 and 13. During the bursts, these states often transition to some with more correlation and higher magnitude covariances like 12, 16, 27, and especially 20. But these correlations still usually do not progress outside a single 4-contact strip. On the other hand, in transitioning between states 27 and 8 and especially 27 and 23, the correlations move to the hemisphere-level along with an increase in magnitude.

This distinction is of course not surprising. Seizures are usually large-scale, high-magnitude events, whereas bursts are not. While the bursts do contain similarities with the seizure onset, their lack of transition to a larger, more global event fundamentally limits the extent to which they can convey all the information we desire in the seizure onset.

■ 5.4 Future Work

Long-term continuous EEG records like those from dogs we examine in this chapter offer a unique perspectice on the relationship between sub-clinical bursts and clinical seizures. These recordings are not only about an order of magnitude longer than analagous recordings

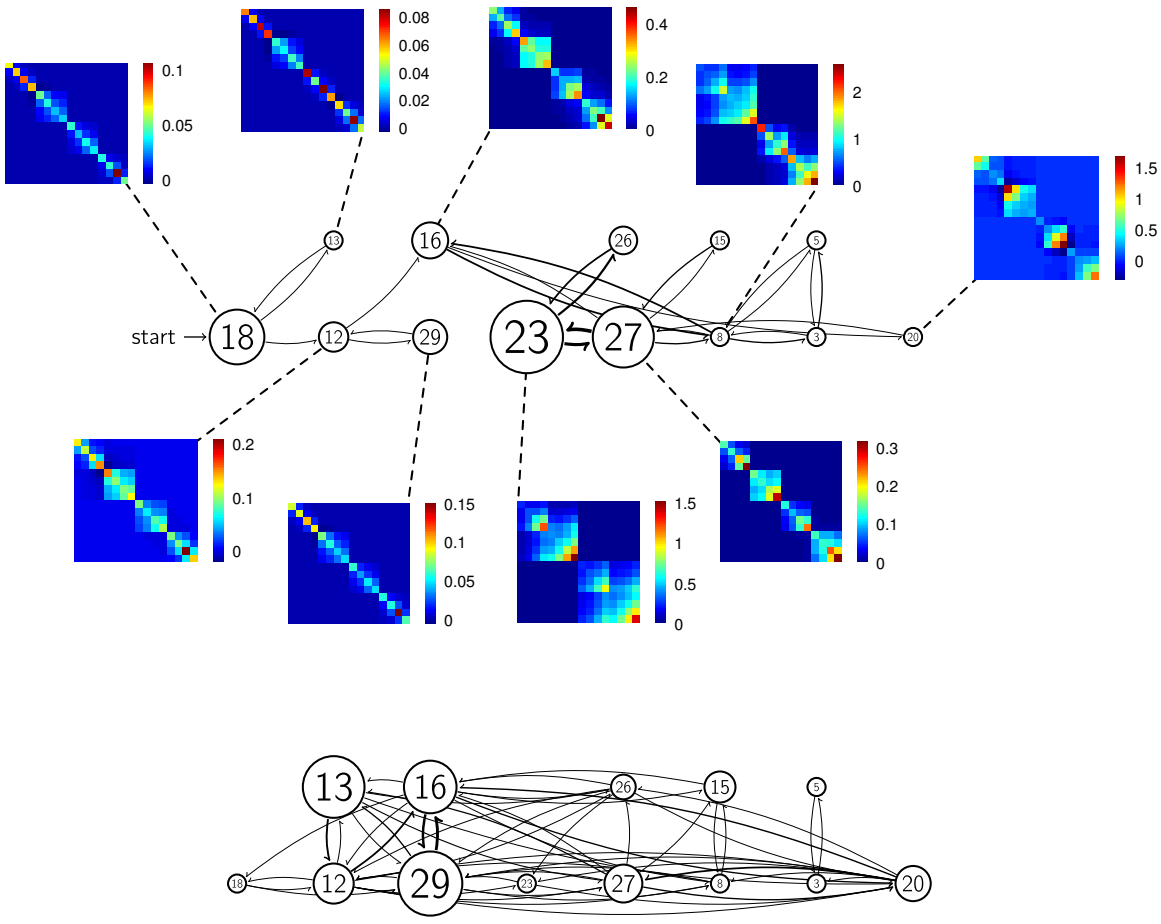

Figure 5.14. (**top**) The state transition diagram for the seizure onset shown in Figures 5.9 and 5.10, where the states shown describe the top 99% of all time points in the seizure onset. Select event state covariances are also shown with scalebars. (**bottom**) The aggregate state transition diagram over all the bursts that occurred before the seizure shown at the top of the figure. The states shown are the same (and have the same placement) as those determined by the seizure onset.

from the epilepsy monitoring unit, they are also are less influenced by the physiologic side effects of surgical implantation and and the anti-epileptic drug modulation commonly present during EMU recordings.

In this chapter, we use three of these recordings and our HIW-spatial AR-HMM model to quantitiatively explore the relationship between sub-clinical bursts and seizures. To our knowledge, this large-scale exploration is the first of its kind. In all three dogs examined, we show that subsets of the bursts for each dog display large similarities with the onsets of seizures. We show how our HIW-spatial AR-HMM model allows for a very fine-grained comparison between the states of the bursts and seizures. We explore the similarities and then the point of divergence between an example seizure and fifteen bursts. Despite these similarities, we find little temporal relationship between the bursts studied here and the seizures. Finally, we see that the bursts are unable to provide a complete picture of the spatial dyanmics present for our example seizure.

The results presented in Sections 5.2 and 5.3 should be qualified as a very preliminary analysis of a few subjects. The highly variable nature of epilepsy requires examining a number of patients before we can draw any broad conclusions about the relationship between epileptic sub-clinical bursts and clinical seizures.

In addition to simply increasing the number of subjects, however, we suggest avenues of future analysis that would improve the current analysis and also provide complementary perspectives on the burst-seizure relationship.

Improving quantitative and qualitative analysis Much of the quantitiative analysis presented in the previous sections relies on using a single representative MCMC sample, either for the model parameters (i.e., the channel state AR coefficients and event state innovation covariances), the event state sequences, or both. While computationally expedient, using a single MCMC sample in our quantitative analysis in many ways wastes the full posterior distribution available through the MCMC samples. Specifically, the analyses in Section 5.2.2 examining channel relationships common to both bursts and seizures would benefit from using the full posterior of the event states and parameters. The analysis in Section 5.2.3 for identifying the bursts most common to a seizure would similarly benefit.

Furthermore, qualitative displays that convey analysis summaries from multiple seizure would be preferable to those given above for representative seizures. The diversity of seizures even within a subject makes such summarization a challenge, lest important differences in various seizure types be lost in the averaging. One approach might involve integrating only over the seizures that tend to occur very close to each other, as they often do in these dogs.

Dynamic visualizations If the model and analyses described in this chapter are ever to be used on a regular bases in a clinical setting, the visualizations—even those suggested in the previous two paragraphs—must distill even further the essential channel relationships in the bursts and seizures. One approach which may be fruitful would involve displaying the event (and perhaps also the channel) states on channels shown on a 3-dimensional brain, where the state evolution is conveyed through the changing frames of a movie.

While seemingly trivial in principal, it is our experience that such visualizations can

ultimately be the difference between a non-technical audience understanding and using an analysis. A movie such as this is also quite non-trivial to generate, given that the 3-dimensional coordinates of the channel electrodes are not defined and would need to be estimated or inferred from existing imaging data.

Examining different classes of bursts The method we use to detect bursts of course reflects the type of bursts we wish to detect. Our detection method favored bursts with abnormal activity across most or all of the channels in an attempt to capture events most similar to seizures (which for these dogs had activity that spread to all the channels). Other classes of bursts may be of interest. For example, bursts with heightened activity over just a few of the channels may be of interest in identifying temporal states of heightened seizure susceptibility.

Investigating event relationships through the channel states In our analysis above, we have all but ignored the channel state sequences inferred for every channel of every event. Comparing the progression of the channel states in bursts and seizure onsets would augment our existing analysis of the event states alone. The channel states would allow for comparing the behavior of individual channels in bursts and seizures, providing a more fine-grained perspective not available from the event states alone.

Since we have 16 channel state sequences for every event state sequence, the computational aspects of quantitatively analyzing these channel state are certainly non-trivial. It is quite possible that the computational and storage space demands of the co-state analysis and certainly the Hamming distance analysis would require them to be modified significantly or substituted altogether for some other analysis. Nevertheless, we believe the channel and event state sequences together provide a far richer description of the events than either can alone.

Comparing seizure event state progression to clinical seizure onset zone Our analysis in this chapter has largely avoided examining the dynamics of individual channels, favoring the changing relationships of all channels together. Nevertheless, the early seizure onset activity epileptologists look for when determining which channels are involved with initiating seizures (and thus belong to the seizure onset zone) will almost certainly be manifested in the channel and events state sequence assignment and parameters. It is likely that one or more of the event states contains covariances that indicate heightened activity in individual channels before that seen in the others. In addition, those channels may transition earlier to channel states indicative of higher-magnitude activity.

Understanding how the model inferences correspond to the clinical seizure onset zone judgements would shed light on both and perhaps could lead to clinical decision support systems or at least augmented EEG visualizations for epileptologists determining what areas of a patient's brain should be surgically removed or perhaps stimulated, in the setting of an implanted device.

Exploring burst-burst similarity and seizure-seizure similarities Finally, we have focused in this work exclusively on burst-seizure similarities. But similarities between seizures and between

bursts are also quite clinically relevant. For example, identifying groups of bursts similar to each other, and how those groups change over time, may provided a more continuous measure of how a subject's underlying physgiologic state is changing (or not changing) between the times when seizures occur.

Similarity measures between seizures could also be used to augment how epileptologists visualize and organize a subject's seizures. In particular, organizing seizures by their similarities (or differences) could prove useful in reminding epileptologists of the diversity (or lack thereof) present in a subject's record that may be easy to lose track of when working with the EEG records alone.

Chapter 6

Discussion and Contributions

In this work, we develop models and produce analyses ultimately meant to aid clinicians in making better, more informed, more objective decisions when treating patients. We do not believe algorithms and models can or should replace the judgement, insight, and intuition that humans gain over years of training and experience. Rather, we believe the role of work like ours is to lend help where humans need it most and where algorithms excel: objective, repeatable analysis produced consistently over large amounts of data.

The physicians treating patients with epilepsy generally have at least a decade of post-college training. And yet, epileptologists currently spend much of their billable time reading EEGs, that is, manually paging through every second of a patient's continuous EEG record, noting relevant aspects in unstructured text notes with the patient's medical record. Surely there is a better way to use such highly-trained people. EEG reading and analysis will never be completely automated, nor should it be. Epilepsy is too nuanced and protean a disorder for machines to ever fully take over. But as other fields of medicine mechanize—from cardiology to genetics, anesthesiology to surgery—why does the study and treatment of epilepsy remain basically where it was thirty years ago? If anything, the vast amounts of data produced by EEG monitoring—both scalp and intracranial—should promote *early* adoption of automated analysis methods, not delayed adoption (if any at all), as seems to presently be the case.

Answers to this question are complex and multifaceted. But we believe one cause has been the relative dearth of analysis paradigms that provide both a flexible, extensible framework for understanding the data that also yields intuitive, straightforward insights to non-technical clinicians. Bayesian models, especially those of the nonparametric variety, provide this framework. It is surprising that they have received so little attention from the epilepsy community given their explicit acknowledgement and integration of the uncertainty present in all modeling problems, especially those of the epilepsy domain.

This work attempts to move the field of quantitative epilepsy analysis ever so slightly in the Bayesian direction. In Chapters 3 and 4, we develop new models that—while generally applicable to many application domains—were primarily motivated by essential problems in EEG analysis.

Our multi-level clustering hierarchical Dirichlet process (MLC-HDP) was inspired by the question, "How can we build a model that incorporates information from the seizures

of many other patients in its analysis of one particular seizure, as humans do?" When epileptologists examine an iEEG seizure, they do not forget about all the other seizures they have seen. Instead, they interpret the present seizure in the context of the others, examining the activity of the iEEG channels themselves, the types of channel activities present in the seizure, and the types of patients who manifest this type of seizure. Our MLC-HDP aims to replicate this form of information sharing.

We show that not only does the MLC-HDP yield superior models of a patient's seizures when information from other patients is included, but the seizure clusterings produced by the model are intuitive and agree reasonably well with the independent seizure clusterings of expert epileptologists. Thanks to our Bayesian framework, the natural uncertainty present in all of these estimates, as the human physicians readily acknowledge, is accessible and quantifiable.

The MLC-HDP model of Chapter 3 in a way aims to mimic some of the high-level organization of iEEG activity an epileptologist may employ. In contrast, the HIW-spatial BP-AR-HMM of Chapter 4 approaches the analysis of iEEG from the opposite direction: how can we replicate the detailed, fine-grained parsing epileptologists do of epileptic events on the iEEG? Our model works with the same iEEG signals humans see when performing their analyses. It attempts to quantify the same state transitions present in individual channels and between the channels that epileptologists do in their manual reading. And again, instead of presenting a false confidence in a single "answer," it instead produces a distribution of answers, a distribution of parsings. While such uncertainty can sometimes be hard to visualize when using large, complex event datasets like ours, it is ultimately the honest answer to a question that any physician will acknowledge has great inherent ambiguity.

In Chapter 5, we use a variant of the HIW-spatial BP-AR-HMM to explore the relationship between epileptic sub-clinical bursts and clinical seizures in long-term iEEG records from dogs with naturally occurring epilepsy. These analyses, over hundreds of individual events in each dog would be impossible for even the most expert human EEG reader to perform. The number and diversity of both the bursts and seizures is simply too great. But it is exactly this same vast quantity of data that allows us the confidence to make more general claims about the relationship between these bursts and seizures. This data empowers us to begin superceding the anecdotal observations with more thorough, objective analyses over all the data available.

Our analysis of the bursts and seizures indicates that bursts are most similar with the onsets of seizures. The channel dynamics and relationships present in the initiating periods of seizures are often also found in a subset of the bursts. We show these similarities in detail in a few representative seizures and bursts. On one level, it is not surprising that these two classes of epileptic events share some similarities. They both are generated from the same disorder. It is equally unsurprising that these two classes of events diverge at the point where the bursts end and the seizure onsets escalate into full, clinical seizures, as our analyses have shown.

Our extensive analysis of these thousands of epileptic events across three dogs on some

level only confirms quantitatively what we knew qualitatively: bursts are similar to seizures in some ways but different in many others. Our findings are short on definitive conclusions and long on new questions, but in developing a framework for quantitative modeling and analysis of these events, our primary goal is to provide a starting point for future research in this valuable area of inquiry.

Future work should focus primarily on translating this work into real world, practical tools that physicians can use in the decision-making process of their everyday clinical practice. In one sense, we have tackled many of the hard problems associated with model construction, implementation, and analysis in this work, but another, equally-challenging set of tasks remains in producing a reliable, practical tool. Scaling and speeding up model inference to the large datasets of interest still remains a large challenge. Turn-key workflows of these analysis for non-technical users will also require a great deal of careful planning and execution. Without this important future work, the research we present here remains just that—research—and ultimately helps no one in the real world.

Appendix A

Posterior Derivations

■ A.1 Multivariate Normal Likelihood with a Conjugate Joint Prior

Below, we derive the posterior for a multivariate normal likelihood with a conjugate $\mathcal{N}$-IW prior.

Consider a set of N observations $\{\mathbf{x}_i\}_{i=1}^N$ in $\mathbb{R}^d$ modeled by a normal likelihood,

$$\mathbf{x}_i \mid \boldsymbol{\mu}, \Sigma \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma), \tag{A.1}$$

where for N i.i.d. observations yields a likelihood of

$$\begin{aligned} p(\mathbf{x}_1, \ldots, \mathbf{x}_N \mid \boldsymbol{\mu}, \Sigma) &\propto \prod_{i=1}^N |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^{\mathrm{T}} \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu})\right) \\ &\propto |\Sigma|^{-N/2} \exp\left(-\frac{1}{2} \sum_{i=1}^N (\mathbf{x}_i - \boldsymbol{\mu})^{\mathrm{T}} \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu})\right). \end{aligned} \tag{A.2}$$

The conjugate prior for unknown $\boldsymbol{\mu}$ and Σ is the normal inverse-Wishart [47, pg. 87],

$$\begin{aligned} \Sigma &\sim \mathrm{IW}_{\nu_0}\left(S_0^{-1}\right), \\ \boldsymbol{\mu} \mid \Sigma &\sim \mathcal{N}\left(\boldsymbol{\mu}_0, \Sigma/n_0\right), \end{aligned} \tag{A.3}$$

corresponding to the normal inverse-Wishart joint prior density $\mathcal{N}$-IW$(n_0, \boldsymbol{\mu}_0, \nu_0, S_0)$ with prior counts n_0, mean $\boldsymbol{\mu}_0$, degrees of freedom ν_0, and scale S_0,

$$\begin{aligned} p\left(\boldsymbol{\mu}, \Sigma\right) &\propto \left(|\Sigma|^{-(\nu_0+d+1)/2} \exp\left(-\frac{1}{2}\mathrm{tr}\left(S_0 \Sigma^{-1}\right)\right)\right) \cdot \\ &\quad \left(|\Sigma/n_0|^{-1/2} \exp\left(-\frac{1}{2}\left(\boldsymbol{\mu} - \boldsymbol{\mu}_0\right)^{\mathrm{T}} \left(\Sigma/n_0\right)^{-1} \left(\boldsymbol{\mu} - \boldsymbol{\mu}_0\right)\right)\right), \\ &\propto \left(|\Sigma|^{-(\nu_0+d+1)/2} \exp\left(-\frac{1}{2}\mathrm{tr}\left(S_0 \Sigma^{-1}\right)\right)\right) \cdot \\ &\quad \left(|\Sigma|^{-1/2} \exp\left(-\frac{n_0}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^{\mathrm{T}} \Sigma^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0)\right)\right), \\ &\propto |\Sigma|^{-(\nu_0+d)/2-1} \exp\left(-\frac{1}{2}\mathrm{tr}\left(S_0 \Sigma^{-1}\right) - \frac{n_0}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^{\mathrm{T}} \Sigma^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0)\right), \end{aligned} \tag{A.4}$$

We omit the prior parameters from the left hand side conditional for compactness. The product of the likelihood in Equation A.2 and joint prior in Equation A.4 yields the posterior

$$p(\boldsymbol{\mu}, \Sigma \mid \mathbf{x}_1, \ldots, \mathbf{x}_N) \propto \left[|\Sigma|^{-N/2} \exp\left(-\frac{1}{2} \sum_{i=1}^{N} (\mathbf{x}_i - \boldsymbol{\mu})^{\mathrm{T}} \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right) \right] \cdot \left[|\Sigma|^{-(\nu_0 + d)/2 - 1} \exp\left(-\frac{1}{2} \mathrm{tr}\left(S_0 \Sigma^{-1} \right) - \frac{n_0}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_0)^{\mathrm{T}} \Sigma^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0) \right) \right]$$

Since our prior is conjugate, we know the posterior will have the same form, i.e., a normal inverse-Wishart. We thus wish to combine the two exponential quadratic terms into a single exponential quadratic term and do so by completing the square. We start by expanding the likelihood's quadratic term,

$$\begin{aligned} -\frac{1}{2} \sum_{i=1}^{N} (\mathbf{x}_i - \boldsymbol{\mu})^{\mathrm{T}} \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) &= -\frac{1}{2} \sum_{i=1}^{N} \left(\mathbf{x}_i^{\mathrm{T}} \Sigma^{-1} \mathbf{x}_i + \boldsymbol{\mu}^{\mathrm{T}} \Sigma^{-1} \boldsymbol{\mu} - 2 \mathbf{x}_i^{\mathrm{T}} \Sigma^{-1} \boldsymbol{\mu} \right) \\ &= -\frac{1}{2} \left[\sum_{i=1}^{N} \mathbf{x}_i^{\mathrm{T}} \Sigma^{-1} \mathbf{x}_i + \sum_{i=1}^{N} \boldsymbol{\mu}^{\mathrm{T}} \Sigma^{-1} \boldsymbol{\mu} - 2 \sum_{i=1}^{N} \mathbf{x}_i^{\mathrm{T}} \Sigma^{-1} \boldsymbol{\mu} \right] \\ &= -\frac{1}{2} \sum_{i=1}^{N} \mathbf{x}_i^{\mathrm{T}} \Sigma^{-1} \mathbf{x}_i - \frac{N}{2} \boldsymbol{\mu}^{\mathrm{T}} \Sigma^{-1} \boldsymbol{\mu} + \sum_{i=1}^{N} \mathbf{x}_i^{\mathrm{T}} \Sigma^{-1} \boldsymbol{\mu}, \end{aligned} \tag{A.5}$$

and then expand the prior's quadratic term,

$$-\frac{n_0}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_0)^{\mathrm{T}} \Sigma^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0) = -\frac{n_0}{2} \boldsymbol{\mu}^{\mathrm{T}} \Sigma^{-1} \boldsymbol{\mu} - \frac{n_0}{2} \boldsymbol{\mu}_0^{\mathrm{T}} \Sigma^{-1} \boldsymbol{\mu}_0 + n_0 \boldsymbol{\mu}_0^{\mathrm{T}} \Sigma^{-1} \boldsymbol{\mu}. \tag{A.6}$$

From the last terms in Equations (A.5) and (A.6), we see that the posterior counts and mean are $n_N = n_0 + N$ and $\boldsymbol{\mu}_N = \frac{1}{n_N}(n_0 \boldsymbol{\mu}_0 + \sum_{i=1}^{N} \mathbf{x}_i)$, respectively. Our posterior quadratic term will thus be of the form,

$$\begin{aligned} -\frac{n_N}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_N)^{\mathrm{T}} \Sigma^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_N) = \\ -\frac{n_N}{2} \left(\boldsymbol{\mu} - \frac{1}{n_N} \left(n_0 \boldsymbol{\mu}_0 + \sum_{i=1}^{N} \mathbf{x}_i \right) \right)^{\mathrm{T}} \Sigma^{-1} \left(\boldsymbol{\mu} - \frac{1}{n_N} \left(n_0 \boldsymbol{\mu}_0 + \sum_{i=1}^{N} \mathbf{x}_i \right) \right) \end{aligned}$$

$$\begin{aligned} -\frac{n_N}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_N)^{\mathrm{T}} \Sigma^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_N) = \\ -\frac{n_N}{2} \boldsymbol{\mu} \Sigma^{-1} \boldsymbol{\mu} - \frac{1}{2 n_N} \left(n_0 \boldsymbol{\mu}_0 + \sum_{i=1}^{N} \mathbf{x}_i \right)^{\mathrm{T}} \Sigma^{-1} \left(n_0 \boldsymbol{\mu}_0 + \sum_{i=1}^{N} \mathbf{x}_i \right) + \\ \boldsymbol{\mu}^{\mathrm{T}} \Sigma^{-1} \left(n_0 \boldsymbol{\mu}_0 + \sum_{i=1}^{N} \mathbf{x}_i \right) \end{aligned} \tag{A.7}$$

Note that the first and last terms of Equation (A.7) are contained in the sum of the second and third terms of Equation (A.5) and the first and third terms of Equation (A.6). We can thus write Equation (A.7) as the sum of of the left hand side of Equation (A.5) minus its first right hand side term and the left hand side of Equation (A.5) minus its second right hand side term plus the additional quadratic $\boldsymbol{\mu}_N \Sigma^{-1} \boldsymbol{\mu}_N$ term on the right hand side of Equation (A.7),

$$
\begin{aligned}
-\frac{n_N}{2}\left(\boldsymbol{\mu}-\boldsymbol{\mu}_N\right)^{\mathrm{T}} \Sigma^{-1}\left(\boldsymbol{\mu}-\boldsymbol{\mu}_N\right) = \\
-\frac{1}{2}\sum_{i=1}^{N}(\mathbf{x}_i-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}_i-\boldsymbol{\mu}) - \frac{n_N}{2}\left(\boldsymbol{\mu}-\boldsymbol{\mu}_N\right)^{\mathrm{T}} \Sigma^{-1}\left(\boldsymbol{\mu}-\boldsymbol{\mu}_N\right) + \\
\frac{1}{2}\sum_{i=1}^{N}\mathbf{x}_i^{\mathrm{T}}\Sigma^{-1}\mathbf{x}_i + \frac{n_0}{2}\boldsymbol{\mu}_0^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}_0 - \frac{n_N}{2}\boldsymbol{\mu}_N^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}_N. \quad \text{(A.8)}
\end{aligned}
$$

Rearranging some terms, we have

$$
\begin{aligned}
-\frac{n_N}{2}\left(\boldsymbol{\mu}-\boldsymbol{\mu}_N\right)^{\mathrm{T}} \Sigma^{-1}\left(\boldsymbol{\mu}-\boldsymbol{\mu}_N\right) - \frac{1}{2}\mathrm{tr}\left(\left(\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} + n_0\boldsymbol{\mu}_0\boldsymbol{\mu}_0^{\mathrm{T}} - n_N\boldsymbol{\mu}_N\boldsymbol{\mu}_N^{\mathrm{T}}\right)\Sigma^{-1}\right) = \\
-\frac{1}{2}\sum_{i=1}^{N}(\mathbf{x}_i-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}_i-\boldsymbol{\mu}) - \frac{n_N}{2}\left(\boldsymbol{\mu}-\boldsymbol{\mu}_N\right)^{\mathrm{T}} \Sigma^{-1}\left(\boldsymbol{\mu}-\boldsymbol{\mu}_N\right). \quad \text{(A.9)}
\end{aligned}
$$

Notice that the trace term fits perfectly with that of Equation (A.4). Our posterior thus has the form

$$
p(\boldsymbol{\mu}, \Sigma \mid \mathbf{x}_1, \ldots, \mathbf{x}_N) \propto |\Sigma|^{-(\nu_N+d)/2-1} \exp\left(-\frac{1}{2}\mathrm{tr}\left(S_N\Sigma^{-1}\right) - \frac{n_N}{2}\left(\boldsymbol{\mu}-\boldsymbol{\mu}_N\right)^{\mathrm{T}} \Sigma^{-1}\left(\boldsymbol{\mu}-\boldsymbol{\mu}_N\right)\right) \quad \text{(A.10)}
$$

where $\nu_N = \nu_0 + n$ and

$$
S_N = S_0 + n_0\boldsymbol{\mu}_0\boldsymbol{\mu}_0^{\mathrm{T}} + \sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} - n_N\boldsymbol{\mu}_N\boldsymbol{\mu}_N^{\mathrm{T}},
$$

which has the parameterization $\mathcal{N}$-IW$(n_N, \boldsymbol{\mu}_N, \nu_N, S_N)$. This posterior can be described via the parameters

$$
\begin{aligned}
n_N &= n_0 + N, \\
n_N\boldsymbol{\mu}_N &= n_0\boldsymbol{\mu}_0 + \sum_{i=1}^{N}\mathbf{x}_i, \\
\nu_N &= \nu_0 + N, \\
S_N &= S_0 + \sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} + n_0\boldsymbol{\mu}_0\boldsymbol{\mu}_0^{\mathrm{T}} - n_N\boldsymbol{\mu}_N\boldsymbol{\mu}_N^{\mathrm{T}}.
\end{aligned} \quad \text{(A.11)}
$$

Posterior predictive distribution The posterior predictive distribution for a new observation $\mathbf{x}^+$ is a multivariate t distribution,

$$\begin{aligned} p(\mathbf{x}^+ \mid \mathbf{x}_1, \ldots, \mathbf{x}_N) &\propto \int_\Theta p(\boldsymbol{\mu}, \Sigma | \mathbf{x}_1, \ldots, \mathbf{x}_N) p(\mathbf{x}^+ | \boldsymbol{\mu}, \Sigma) \, d\boldsymbol{\mu} d\Sigma \\ &\propto t_{(\nu_N - d + 1)} \left(\boldsymbol{\mu}_N, \frac{n_N + 1}{n_N (\nu_N - d + 1)}) S_N \right) \end{aligned}$$

Diagonal covariance When Σ has diagonal covariance, we can reparameterise our posterior scale matrix using vector prior scale $\mathbf{s}_0^2$,

$$\mathbf{s}_N^2 = \mathbf{s}_0 + n_0 \boldsymbol{\mu}_0 \circ \boldsymbol{\mu}_0 + \sum_{i=1}^N \mathbf{x}_i \circ \mathbf{x}_i - n_N \boldsymbol{\mu}_N \circ \boldsymbol{\mu}_N, \tag{A.12}$$

$$S_N = I s_N^2. \tag{A.13}$$

This formulation is equivalent to the product of d independent normal inverse-gamma distributions with paramters $\mathcal{N}$-IG$(n_N, \mu_{N,d}, \nu_N/2, s_{n,d}^2/2)$

■ A.2 (H)IW-spatial BP-AR-HMM Autoregressive State Coefficients

Recall that our prior on the autoregressive coefficients $\mathbf{a}_k$ is a multivariate normal with zero mean and covariance Σ_0,

$$\begin{aligned} p(\mathbf{a}_k \mid \Sigma_0) &\propto \mathcal{N}(\mathbf{0}, \Sigma_0) \\ \log p(\mathbf{a}_k \mid \Sigma_0) &\propto -\frac{1}{2} \mathbf{a}_k^{\mathrm{T}} \Sigma_0^{-1} \mathbf{a}_k. \end{aligned} \tag{A.14}$$

From Equation (4.24) the conditional event likelihood given the channel states $\mathbf{z}_{1:T}$ and the event states $Z_{1:T}$ is

$$\begin{aligned} p(\mathbf{y}_{1:T} \mid \mathbf{z}_{1:T}, Z_{1:T}, \{\mathbf{a}_k\}, \{\Delta_l\}) &\propto \prod_{t=1}^T \mathcal{N}(\mathbf{y}_t; \mathbf{A}_{\mathbf{z}_t} \widetilde{\mathbf{Y}}, \Delta_{Z_t}) \\ \log p(\mathbf{y}_{1:T} \mid \mathbf{z}_{1:T}, Z_{1:T}, \{\mathbf{a}_k\}, \{\Delta_l\}) &\propto -\frac{1}{2} \sum_{t=1}^T (\mathbf{y}_t - \mathbf{A}_{\mathbf{z}_t} \widetilde{\mathbf{Y}}_t)^{\mathrm{T}} \Delta_{Z_t}^{-1} (\mathbf{y}_t - \mathbf{A}_{\mathbf{z}_t} \widetilde{\mathbf{Y}}_t). \end{aligned} \tag{A.15}$$

The product of these prior and likelihood terms is the joint distribution over $\mathbf{a}_k$ and $\mathbf{y}_{1:T}$,

$$\begin{aligned} \log p(\mathbf{a}_k, \mathbf{y}_{1:T} \mid \mathbf{z}_{1:T}, Z_{1:T}, \{\mathbf{a}_{k'}\}_{k' \neq k}, \{\Delta_l\}) \propto \\ -\frac{1}{2} \mathbf{a}_k^{\mathrm{T}} \Sigma_0^{-1} \mathbf{a}_k - \frac{1}{2} \sum_{t=1}^T (\mathbf{y}_t - \mathbf{A}_{\mathbf{z}_t} \widetilde{\mathbf{Y}}_t)^{\mathrm{T}} \Delta_{Z_t}^{-1} (\mathbf{y}_t - \mathbf{A}_{\mathbf{z}_t} \widetilde{\mathbf{Y}}_t). \end{aligned} \tag{A.16}$$

We take a brief tangent to prove a useful identity,

Lemma A.2.1. *Let the column vector* $\mathbf{x} \in \mathbb{R}^m$ *and the symmetric matrix* $A \in \mathbb{S}^{m \times m}$ *be defined as*

$$\mathbf{x} = \begin{bmatrix} \mathbf{y} \\ \mathbf{z} \end{bmatrix} \quad \text{and} \quad A = \begin{bmatrix} B & C \\ C^{\mathrm{T}} & D \end{bmatrix},$$

where $\mathbf{y} \in \mathbb{R}^p$, $\mathbf{z} \in \mathbb{R}^q$, $B \in \mathbb{S}^{p \times p}$, $D \in \mathbb{S}^{q \times q}$, $C \in \mathbb{R}^{p \times q}$, *and* $m = p + q$. *Then*

$$\mathbf{x}^{\mathrm{T}} A \mathbf{x} = \mathbf{y}^{\mathrm{T}} B \mathbf{y} + \mathbf{z}^{\mathrm{T}} D \mathbf{z} + 2\mathbf{y}^{\mathrm{T}} C \mathbf{z}. \tag{A.17}$$

Proof.

$$\begin{aligned} \mathbf{x}^{\mathrm{T}} A \mathbf{x} &= \begin{bmatrix} \mathbf{y}^{\mathrm{T}} & \mathbf{z}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} B & C \\ C^{\mathrm{T}} & D \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{z} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{y}^{\mathrm{T}} & \mathbf{z}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} B\mathbf{y} + C\mathbf{z} \\ C^{\mathrm{T}}\mathbf{y} + D\mathbf{z} \end{bmatrix} \\ &= \mathbf{y}^{\mathrm{T}} B \mathbf{y} + \mathbf{y}^{\mathrm{T}} C \mathbf{z} + \mathbf{z}^{\mathrm{T}} C^{\mathrm{T}} \mathbf{y} + \mathbf{z}^{\mathrm{T}} D \mathbf{z} \\ &= \mathbf{y}^{\mathrm{T}} B \mathbf{y} + \mathbf{z}^{\mathrm{T}} D \mathbf{z} + 2\mathbf{y}^{\mathrm{T}} C \mathbf{z} \end{aligned}$$

■

Note that this identity also holds for any permutation p applied to the rows of $\mathbf{x}$ and the rows and columns of A. We now can manipulate the likelihood term of Equation (A.16) into a form that separates $\mathbf{a}_k$ from $\mathbf{a}_{k' \neq k}$. Suppose that $\mathbf{k}^+$ denotes the indices of the N channels where $z_t^{(i)} = k$ and $\mathbf{k}^- = \{1, \ldots, N\}/\mathbf{k}^+$ denotes those for whom $z_t^{(i)} \neq k$. Furthermore, we use the superscript indexing on these two sets of indices to select the corresponding portions of the $\mathbf{y}_t$ vector and the $\mathbf{A}_{\mathbf{z}_t}$, $\widetilde{\mathbf{Y}}_T$, and $\Delta_{Z_t}^{-1}$ matrices. We start by decomposing the likelihood term into three parts,

$$\begin{aligned} &(\mathbf{y}_t - \mathbf{A}_{\mathbf{z}_t} \widetilde{\mathbf{Y}}_t)^{\mathrm{T}} \Delta_{Z_t}^{-1} (\mathbf{y}_t - \mathbf{A}_{\mathbf{z}_t} \widetilde{\mathbf{Y}}_t) \propto \\ &\quad \left(\mathbf{y}_t^{(\mathbf{k}^+)} - \mathbf{A}_{\mathbf{z}_t}^{(\mathbf{k}^+, \mathbf{k}^+)} \widetilde{\mathbf{Y}}_t^{(\mathbf{k}^+, \mathbf{k}^+)} \right)^{\mathrm{T}} \Delta_{Z_t}^{-1(\mathbf{k}^+, \mathbf{k}^+)} \left(\mathbf{y}_t^{(\mathbf{k}^+)} - \mathbf{A}_{\mathbf{z}_t}^{(\mathbf{k}^+, \mathbf{k}^+)} \widetilde{\mathbf{Y}}_t^{(\mathbf{k}^+, \mathbf{k}^+)} \right) + \\ &\quad 2 \left(\mathbf{y}_t^{(\mathbf{k}^+)} - \mathbf{A}_{\mathbf{z}_t}^{(\mathbf{k}^+, \mathbf{k}^+)} \widetilde{\mathbf{Y}}_t^{(\mathbf{k}^+, \mathbf{k}^+)} \right)^{\mathrm{T}} \Delta_{Z_t}^{-1(\mathbf{k}^+, \mathbf{k}^-)} \left(\mathbf{y}_t^{(\mathbf{k}^-)} - \mathbf{A}_{\mathbf{z}_t}^{(\mathbf{k}^-, \mathbf{k}^-)} \widetilde{\mathbf{Y}}_t^{(\mathbf{k}^-, \mathbf{k}^-)} \right) + \\ &\quad \left(\mathbf{y}_t^{(\mathbf{k}^-)} - \mathbf{A}_{\mathbf{z}_t}^{(\mathbf{k}^-, \mathbf{k}^-)} \widetilde{\mathbf{Y}}_t^{(\mathbf{k}^-, \mathbf{k}^-)} \right)^{\mathrm{T}} \Delta_{Z_t}^{-1(\mathbf{k}^-, \mathbf{k}^-)} \left(\mathbf{y}_t^{(\mathbf{k}^-)} - \mathbf{A}_{\mathbf{z}_t}^{(\mathbf{k}^-, \mathbf{k}^-)} \widetilde{\mathbf{Y}}_t^{(\mathbf{k}^-, \mathbf{k}^-)} \right), \end{aligned} \tag{A.18}$$

which we then insert into our previous expression for the joint distribution of $\mathbf{a}_k$ and $\mathbf{y}_{1:T}$

(Equation (A.16)),

$$\begin{aligned}\log p(\mathbf{a}_k, \mathbf{y}_{1:T} \mid \mathbf{z}_{1:T}, Z_{1:T}, \{\mathbf{a}_{k'}\}_{k' \neq k}, \{\Delta_l\}) \propto -\tfrac{1}{2}\mathbf{a}_k^{\mathrm{T}}\Sigma_0^{-1}\mathbf{a}_k - \\ \tfrac{1}{2}\sum_{t=1}^{T}\Bigg\{\left(\mathbf{y}_t^{(\mathbf{k}^+)} - \mathbf{A}_{\mathbf{z}_t}^{(\mathbf{k}^+,\mathbf{k}^+)}\widetilde{\mathbf{Y}}_t^{(\mathbf{k}^+,\mathbf{k}^+)}\right)^{\mathrm{T}}\Delta_{Z_t}^{-1(\mathbf{k}^+,\mathbf{k}^+)}\left(\mathbf{y}_t^{(\mathbf{k}^+)} - \mathbf{A}_{\mathbf{z}_t}^{(\mathbf{k}^+,\mathbf{k}^+)}\widetilde{\mathbf{Y}}_t^{(\mathbf{k}^+,\mathbf{k}^+)}\right) + \\ 2\left(\mathbf{y}_t^{(\mathbf{k}^+)} - \mathbf{A}_{\mathbf{z}_t}^{(\mathbf{k}^+,\mathbf{k}^+)}\widetilde{\mathbf{Y}}_t^{(\mathbf{k}^+,\mathbf{k}^+)}\right)^{\mathrm{T}}\Delta_{Z_t}^{-1(\mathbf{k}^+,\mathbf{k}^-)}\left(\mathbf{y}_t^{(\mathbf{k}^-)} - \mathbf{A}_{\mathbf{z}_t}^{(\mathbf{k}^-,\mathbf{k}^-)}\widetilde{\mathbf{Y}}_t^{(\mathbf{k}^-,\mathbf{k}^-)}\right) + \\ \left(\mathbf{y}_t^{(\mathbf{k}^-)} - \mathbf{A}_{\mathbf{z}_t}^{(\mathbf{k}^-,\mathbf{k}^-)}\widetilde{\mathbf{Y}}_t^{(\mathbf{k}^-,\mathbf{k}^-)}\right)^{\mathrm{T}}\Delta_{Z_t}^{-1(\mathbf{k}^-,\mathbf{k}^-)}\left(\mathbf{y}_t^{(\mathbf{k}^-)} - \mathbf{A}_{\mathbf{z}_t}^{(\mathbf{k}^-,\mathbf{k}^-)}\widetilde{\mathbf{Y}}_t^{(\mathbf{k}^-,\mathbf{k}^-)}\right)\Bigg\}.\end{aligned} \tag{A.19}$$

Conditioning on $\mathbf{y}_{1:T}$ allows us to absorb the third term of the sum into the proportionality, and after replacing $\mathbf{A}_{\mathbf{z}_t}^{(\mathbf{k}^+,\mathbf{k}^+)}\widetilde{\mathbf{Y}}^{(\mathbf{k}^+,\mathbf{k}^+)}$ with a more explicit expression, we have

$$\begin{aligned}\log p(\mathbf{a}_k \mid \mathbf{y}_{1:T}, \mathbf{z}_{1:T}, Z_{1:T}, \{\mathbf{a}_{k'}\}_{k' \neq k}, \{\Delta_l\}) \propto -\tfrac{1}{2}\mathbf{a}_k^{\mathrm{T}}\Sigma_0^{-1}\mathbf{a}_k - \\ \tfrac{1}{2}\sum_{t=1}^{\mathrm{T}}\Bigg\{\left(\mathbf{y}_t^{(\mathbf{k}^+)} - \left[\widetilde{\mathbf{y}}_t^{(k_1^+)} \mid \cdots \mid \widetilde{\mathbf{y}}_t^{(k_{|\mathbf{k}^+|}^+)}\right]^{\mathrm{T}}\mathbf{a}_k\right)^{\mathrm{T}}\left(\Delta_{Z_t}^{-1(\mathbf{k}^+,\mathbf{k}^+)}\right) \cdot \\ \left(\mathbf{y}_t^{(\mathbf{k}^+)} - \left[\widetilde{\mathbf{y}}_t^{(k_1^+)} \mid \cdots \mid \widetilde{\mathbf{y}}_t^{(k_{|\mathbf{k}^+|}^+)}\right]^{\mathrm{T}}\mathbf{a}_k\right) + \\ 2\left(\mathbf{y}_t^{(\mathbf{k}^+)} - \left[\widetilde{\mathbf{y}}_t^{(k_1^+)} \mid \cdots \mid \widetilde{\mathbf{y}}_t^{(k_{|\mathbf{k}^+|}^+)}\right]^{\mathrm{T}}\mathbf{a}_k\right)^{\mathrm{T}}\left(\Delta_{Z_t}^{-1(\mathbf{k}^+,\mathbf{k}^-)}\right) \cdot \\ \left(\mathbf{y}_t^{(\mathbf{k}^-)} - \mathbf{A}_{\mathbf{z}_t}^{(\mathbf{k}^-,\mathbf{k}^-)}\widetilde{\mathbf{Y}}_t^{(\mathbf{k}^-,\mathbf{k}^-)}\right)\Bigg\},\end{aligned} \tag{A.20}$$

which we can further expand to yield

$$\begin{aligned}\log p(\mathbf{a}_k \mid \mathbf{y}_{1:T}, \mathbf{z}_{1:T}, Z_{1:T}, \{\mathbf{a}_{k'}\}_{k' \neq k}, \{\Delta_l\}) \propto -\tfrac{1}{2}\mathbf{a}_k^{\mathrm{T}}\Sigma_0^{-1}\mathbf{a}_k - \\ \tfrac{1}{2}\sum_{t=1}^{T}\Bigg\{\left(\mathbf{y}_t^{(\mathbf{k}^+)}\right)^{\mathrm{T}}\left(\Delta_{Z_t}^{-1(\mathbf{k}^+,\mathbf{k}^+)}\right)\left(\mathbf{y}_t^{(\mathbf{k}^+)}\right) + \\ \left(\mathbf{a}_k^{\mathrm{T}}\left[\widetilde{\mathbf{y}}_t^{(k_1^+)} \mid \cdots \mid \widetilde{\mathbf{y}}_t^{(k_{|\mathbf{k}^+|}^+)}\right]\right)\left(\Delta_{Z_t}^{-1(\mathbf{k}^+,\mathbf{k}^+)}\right)\left(\left[\widetilde{\mathbf{y}}_t^{(k_1^+)} \mid \cdots \mid \widetilde{\mathbf{y}}_t^{(k_{|\mathbf{k}^+|}^+)}\right]^{\mathrm{T}}\mathbf{a}_k\right) - \\ 2\left(\mathbf{y}_t^{(\mathbf{k}^+)}\right)^{\mathrm{T}}\left(\Delta_{Z_t}^{-1(\mathbf{k}^+,\mathbf{k}^+)}\right)\left(\left[\widetilde{\mathbf{y}}_t^{(k_1^+)} \mid \cdots \mid \widetilde{\mathbf{y}}_t^{(k_{|\mathbf{k}^+|}^+)}\right]^{\mathrm{T}}\mathbf{a}_k\right)\Bigg\} - \\ \sum_{t=1}^{T}\Bigg\{\mathbf{y}_t^{\mathrm{T}(\mathbf{k}^+)}\Delta_{Z_t}^{-1(\mathbf{k}^+,\mathbf{k}^-)}\left(\mathbf{y}_t^{(\mathbf{k}^-)} - \mathbf{A}_{\mathbf{z}_t}^{(\mathbf{k}^-,\mathbf{k}^-)}\widetilde{\mathbf{Y}}_t^{(\mathbf{k}^-,\mathbf{k}^-)}\right) - \\ \left(\mathbf{a}_k^{\mathrm{T}}\left[\widetilde{\mathbf{y}}_t^{(k_1^+)} \mid \cdots \mid \widetilde{\mathbf{y}}_t^{(k_{|\mathbf{k}^+|}^+)}\right]\right)\left(\Delta_{Z_t}^{-1(\mathbf{k}^+,\mathbf{k}^-)}\right)\left(\mathbf{y}_t^{(\mathbf{k}^-)} - \mathbf{A}_{\mathbf{z}_t}^{(\mathbf{k}^-,\mathbf{k}^-)}\widetilde{\mathbf{Y}}_t^{(\mathbf{k}^-,\mathbf{k}^-)}\right)\Bigg\}.\end{aligned} \tag{A.21}$$

Absorbing more terms unrelated to $\mathbf{a}_k$ into the proportionality, we have

$$\begin{aligned}\log p(\mathbf{a}_k \mid \mathbf{y}_{1:T}, \mathbf{z}_{1:T}, Z_{1:T}, \{\mathbf{a}_{k'}\}_{k' \neq k}, \{\Delta_l\}) \propto -\tfrac{1}{2}\mathbf{a}_k^{\mathrm{T}}\Sigma_0^{-1}\mathbf{a}_k - \\ \tfrac{1}{2}\sum_{t=1}^{T}\left\{\left(\mathbf{a}_k^{\mathrm{T}}\left[\widetilde{\mathbf{y}}_t^{(k_1^+)} \mid \cdots \mid \widetilde{\mathbf{y}}_t^{(k^+_{|\mathbf{k}^+|})}\right]\right)\left(\Delta_{Z_t}^{-1(\mathbf{k}^+,\mathbf{k}^+)}\right)\left(\left[\widetilde{\mathbf{y}}_t^{(k_1^+)} \mid \cdots \mid \widetilde{\mathbf{y}}_t^{(k^+_{|\mathbf{k}^+|})}\right]^{\mathrm{T}}\mathbf{a}_k\right) - \right. \\ \left. 2\left(\mathbf{y}_t^{(\mathbf{k}^+)}\right)^{\mathrm{T}}\left(\Delta_{Z_t}^{-1(\mathbf{k}^+,\mathbf{k}^+)}\right)\left(\left[\widetilde{\mathbf{y}}_t^{(k_1^+)} \mid \cdots \mid \widetilde{\mathbf{y}}_t^{(k^+_{|\mathbf{k}^+|})}\right]^{\mathrm{T}}\mathbf{a}_k\right)\right\} - \\ \sum_{t=1}^{T}\left\{-\left(\mathbf{a}_k^{\mathrm{T}}\left[\widetilde{\mathbf{y}}_t^{(k_1^+)} \mid \cdots \mid \widetilde{\mathbf{y}}_t^{(k^+_{|\mathbf{k}^+|})}\right]\right)\left(\Delta_{Z_t}^{-1(\mathbf{k}^+,\mathbf{k}^-)}\right)\left(\mathbf{y}_t^{(\mathbf{k}^-)} - \mathbf{A}_{\mathbf{z}_t}^{(\mathbf{k}^-,\mathbf{k}^-)}\widetilde{\mathbf{Y}}_t^{(\mathbf{k}^-,\mathbf{k}^-)}\right)\right\},\end{aligned} \tag{A.22}$$

which after some rearranging gives

$$\begin{aligned}\log p(\mathbf{a}_k \mid \mathbf{y}_{1:T}, \mathbf{z}_{1:T}, Z_{1:T}, \{\mathbf{a}_{k'}\}_{k' \neq k}, \{\Delta_l\}) \propto \\ -\tfrac{1}{2}\mathbf{a}_k^{\mathrm{T}}\left\{\Sigma_0^{-1} + \sum_{t=1}^{T}\left[\widetilde{\mathbf{y}}_t^{(k_1^+)} \mid \cdots \mid \widetilde{\mathbf{y}}_t^{(k^+_{|\mathbf{k}^+|})}\right]\left(\Delta_{Z_t}^{-1(\mathbf{k}^+,\mathbf{k}^+)}\right)\left[\widetilde{\mathbf{y}}_t^{(k_1^+)} \mid \cdots \mid \widetilde{\mathbf{y}}_t^{(k^+_{|\mathbf{k}^+|})}\right]^{\mathrm{T}}\right\}\mathbf{a}_k + \\ \mathbf{a}_k^{\mathrm{T}}\sum_{t=1}^{T}\left\{\left[\widetilde{\mathbf{y}}_t^{(k_1^+)} \mid \cdots \mid \widetilde{\mathbf{y}}_t^{(k^+_{|\mathbf{k}^+|})}\right]\left(\Delta_{Z_t}^{-1(\mathbf{k}^+,\mathbf{k}^+)}\right)\left(\mathbf{y}_t^{(\mathbf{k}^+)}\right) + \right. \\ \left.\left[\widetilde{\mathbf{y}}_t^{(k_1^+)} \mid \cdots \mid \widetilde{\mathbf{y}}_t^{(k^+_{|\mathbf{k}^+|})}\right]\left(\Delta_{Z_t}^{-1(\mathbf{k}^+,\mathbf{k}^-)}\right)\left(\mathbf{y}_t^{(\mathbf{k}^-)} - \mathbf{A}_{\mathbf{z}_t}^{(\mathbf{k}^-,\mathbf{k}^-)}\widetilde{\mathbf{Y}}_t^{(\mathbf{k}^-,\mathbf{k}^-)}\right)\right\}.\end{aligned} \tag{A.23}$$

Before completing the square, we will find it useful to introduce a bit more notation to simplify the expression,

$$\bar{\mathbf{Y}}_t^{(\mathbf{k}^+)} = \left[\widetilde{\mathbf{y}}_t^{(k_1^+)} \mid \cdots \mid \widetilde{\mathbf{y}}_t^{(k^+_{|\mathbf{k}^+|})}\right], \qquad \boldsymbol{\epsilon}_t^{(\mathbf{k}^-)} = \mathbf{y}_t^{(\mathbf{k}^-)} - \mathbf{A}_{\mathbf{z}_t}^{(\mathbf{k}^-,\mathbf{k}^-)}\widetilde{\mathbf{Y}}_t^{(\mathbf{k}^-,\mathbf{k}^-)},$$

yielding

$$\begin{aligned}\log p(\mathbf{a}_k \mid \mathbf{y}_{1:T}, \mathbf{z}_{1:T}, Z_{1:T}, \{\mathbf{a}_{k'}\}_{k' \neq k}, \{\Delta_l\}) \propto \\ -\tfrac{1}{2}\mathbf{a}_k^{\mathrm{T}}\left\{\Sigma_0^{-1} + \sum_{t=1}^{T}\bar{\mathbf{Y}}_t^{(\mathbf{k}^+)}\Delta_{Z_t}^{-1(\mathbf{k}^+,\mathbf{k}^+)}\bar{\mathbf{Y}}_t^{\mathrm{T}(\mathbf{k}^+)}\right\}\mathbf{a}_k + \\ \mathbf{a}_k^{\mathrm{T}}\left\{\sum_{t=1}^{T}\bar{\mathbf{Y}}_t^{(\mathbf{k}^+)}\left(\Delta_{Z_t}^{-1(\mathbf{k}^+,\mathbf{k}^+)}\mathbf{y}_t^{(\mathbf{k}^+)} + \Delta_{Z_t}^{-1(\mathbf{k}^+,\mathbf{k}^-)}\boldsymbol{\epsilon}_t^{(\mathbf{k}^-)}\right)\right\}.\end{aligned} \tag{A.24}$$

We desire an expression in the form $-\frac{1}{2}(\mathbf{a}_k - \boldsymbol{\mu}_k)^{\mathrm{T}}\Sigma_k^{-1}(\mathbf{a}_k - \boldsymbol{\mu}_k)$ for unknown $\boldsymbol{\mu}_k$ and Σ_k^{-1} so that it conforms to the multivariate normal density with mean $\boldsymbol{\mu}_k$ and precision Σ_k^{-1}. We already have our Σ_k^{-1} value from the quadratic term above,

$$\Sigma_k^{-1} = \Sigma_0^{-1} + \sum_{t=1}^{T}\bar{\mathbf{Y}}_t^{(\mathbf{k}^+)}\Delta_{Z_t}^{-1(\mathbf{k}^+,\mathbf{k}^+)}\bar{\mathbf{Y}}_t^{\mathrm{T}(\mathbf{k}^+)}, \tag{A.25}$$

which allows us to solve the cross-term for $\boldsymbol{\mu}_k$,

$$-\frac{1}{2}(-2\boldsymbol{\mu}_k^{\mathrm{T}}\Sigma_k^{-1}\mathbf{a}_k) = \mathbf{a}_k^{\mathrm{T}}\left(\sum_{t=1}^{T}\bar{\mathbf{Y}}_t^{(\mathbf{k}^+)}\left(\Delta_{Z_t}^{-1(\mathbf{k}^+,\mathbf{k}^+)}\mathbf{y}_t^{(\mathbf{k}^+)} + \Delta_{Z_t}^{-1(\mathbf{k}^+,\mathbf{k}^-)}\boldsymbol{\epsilon}_t^{(\mathbf{k}^-)}+\right)\right),$$

$$\Sigma_k^{-1}\boldsymbol{\mu}_k = \sum_{t=1}^{T}\bar{\mathbf{Y}}_t^{(\mathbf{k}^+)}\left(\Delta_{Z_t}^{-1(\mathbf{k}^+,\mathbf{k}^+)}\mathbf{y}_t^{(\mathbf{k}^+)} + \Delta_{Z_t}^{-1(\mathbf{k}^+,\mathbf{k}^-)}\boldsymbol{\epsilon}_t^{(\mathbf{k}^-)}+\right). \tag{A.26}$$

We can pull the final required $-\frac{1}{2}\boldsymbol{\mu}_k^{\mathrm{T}}\Sigma_k^{-1}\boldsymbol{\mu}_k$ term from the proportionality and complete the square. Thus, we have the form of the posterior for $\mathbf{a}_k$,

$$\begin{aligned} p(\mathbf{a}_k \mid \mathbf{y}_{1:T,}\mathbf{z}_{1:T}, Z_{1:T}, \{\mathbf{a}_{k'}\}_{k'\neq k}, \{\Delta_l\}) &\propto \exp\left(-\frac{1}{2}(\mathbf{a}_k-\boldsymbol{\mu}_k)^{\mathrm{T}}\Sigma_k^{-1}(\mathbf{a}_k-\boldsymbol{\mu}_k)\right) \\ &\propto \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k), \end{aligned} \tag{A.27}$$

where

$$\begin{aligned} \Sigma_k^{-1} &= \Sigma_0^{-1} + \sum_{t=1}^{T}\bar{\mathbf{Y}}_t^{(\mathbf{k}^+)}\Delta_{Z_t}^{-1(\mathbf{k}^+,\mathbf{k}^+)}\bar{\mathbf{Y}}_t^{\mathrm{T}(\mathbf{k}^+)} \\ \Sigma_k^{-1}\boldsymbol{\mu}_k &= \sum_{t=1}^{T}\bar{\mathbf{Y}}_t^{(\mathbf{k}^+)}\left(\Delta_{Z_t}^{-1(\mathbf{k}^+,\mathbf{k}^+)}\mathbf{y}_t^{(\mathbf{k}^+)} + \Delta_{Z_t}^{-1(\mathbf{k}^+,\mathbf{k}^-)}\boldsymbol{\epsilon}_t^{(\mathbf{k}^-)}\right). \end{aligned} \tag{A.28}$$

Appendix B

Schiff seizure features

The following six features were used by Schiff et al. [106] to quantify the spatiotemporal dynamics of a seizure. Let $x_t^{(i)}$ denote the ith channel's voltage at each sample t in a given time window. Let N denote the number of channels in the seizure and T the number of samples in a time window.

Total power The total power is the sum of the squared values at each time point

$$P = \sum_{i=1}^{N} \sum_{t=1}^{T} x_t^{2(i)} \tag{B.1}$$

Total correlation at zero time lag The total correlation at zero time lag find the zero-lag correlation across channels after first-order detrending each channel's time series. Let $a_i t + b_i$ be the first order trend for chanel i and $y_t^{(i)} = x_t^{(i)} - a_i t - b_i$, so the normalized cross-correlation between channels i and j is given by

$$c_\tau^{(i,j)} = \frac{\sum_{t=-T/2}^{T/2} y_t^{(i)} y_{t+\tau}^{(j)}}{\left(\sum_{t=-T/2}^{T/2} y_t^{2(i)}\right)^{1/2} \left(\sum_{t=-T/2}^{T/2} y_t^{2(j)}\right)^{1/2}} \tag{B.2}$$

The total zero-lag cross-correlation is thus

$$S_0 = \sum_{\{i \neq j\}} |c_0^{(i,j)}| \tag{B.3}$$

Total correlation at arbitrary time lag The total correlation at arbitrary time lag sums the correlation values larger than twice the estimator of the standard deviation, which is given as

$$\sigma_{i,j} = \frac{1}{T+1-\tau} \left| \sum_{\tau=-T/2}^{T/2} c_\tau^{(i,i)} c_\tau^{(j,j)} \right|^{1/2}, \tag{B.4}$$

and thus the total correlation at arbitrary lag is

$$S_\tau = \sum_{\{i \neq j\}} \sum_{\tau=-T/2}^{T/2} c_\tau^{i,j} \mathbf{1}(c_\tau^{(i,j)} > 2\sigma_{i,j}). \tag{B.5}$$

Phase amplitude coherence The phase amplitude coherence reflects the average phase amplitude across all the channels and time points. We calculate the phase $\varphi_t^{(i)}$ of channel i at time t using the Hilbert transform. Let

$$r_t = \frac{1}{N}\sqrt{\left(\sum_{i=1}^{N}\cos(\varphi_t^{(i)})\right)^2 + \left(\sum_{i=1}^{N}\sin(\varphi_t^{(i)})\right)^2} \tag{B.6}$$

denote the average phase amplitude at time t over all the channels. The phase amplitude coherance is simply the average over all time points,

$$\mathbb{E}[r] = \frac{1}{T}\sum_{t=1}^{T} r_t. \tag{B.7}$$

Phase angle dispersion is a metric on the average phase angle across all the channels and time points. Let

$$\theta_t = \tan^{-1}\left(\left(\sum_{i=1}^{N}\cos(\varphi_t^{(i)})\right) + \left(\sum_{i=1}^{N}\sin(\varphi_t^{(i)})\right)\right) \tag{B.8}$$

denote the average phase angle at time t, with θ_t' representing the non-discontinuous, unwrapped version of the phase. The dispersion of the phase angle is given by the variance of the first-order differences,

$$\theta_{\text{disp}} = \mathbb{V}\left[(\theta_2 - \theta_1, \ldots, \theta_T - \theta_{T-1})\right] \tag{B.9}$$

Phase amplitude dispersion is a similar metric for the phase amplitude,

$$r_{\text{disp}} = \mathbb{V}\left[(r_2 - r_1, \ldots, r_T - r_{T-1})\right] \tag{B.10}$$

Bibliography

[1] Global Comparative Assessment in the Health Sector; Disease Burden, Expenditures, and Intervention Packages. Technical report, World Health Organization, Geneva, 1994.

[2] E. Acar, C. Aykut-Bingol, H. Bingol, R. Bro, and B. Yener. Multiway analysis of epilepsy tensors. *Bioinformatics*, 23:10–18, 2007.

[3] H. Adeli, Z. Zhou, and N. Dadmehr. Analysis of EEG records in an epileptic patient using wavelet transform. *Journal of Neuroscience Methods*, 123:69–87, 2003.

[4] G. Alarcon. Electrophysiological aspects of interictal and ictal activity in human partial epilepsy. *Seizure*, 5:7–33, 1996.

[5] G. Alarcon, C. D. Binnie, R. D. Elwes, and C. E. Polkey. Power spectrum and intracranial EEG patterns at seizure onset in partial epilepsy. *Electroencephalography and Clinical Neurophysiology*, 94:326–337, 1995.

[6] K. Alper, M. Raghavan, R. Isenhart, B. Howard, W. Doyle, R. John, and L. Prichep. Localizing epileptogenic regions in partial epilepsy using three-dimensional statistical parametric maps of background EEG source spectra. *NeuroImage*, 39:1257 – 1265, 2008.

[7] C. E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.

[8] F. Bartolomei, D. Cosandier-Rimele, A. McGonigal, S. Aubert, J. Régis, M. Gavaret, F. Wendling, and P. Chauvel. From mesial temporal lobe to temporoperisylvian seizures: a quantified study of temporal lobe seizure networks. *Epilepsia*, 51(10): 2147–2158, 2010.

[9] Matthew J Beal. Gene Expression Time Course Clustering with Countably Infinite Hidden Markov Models. In *Proceedings of the 22nd Conference Conference on Uncertainty in Artificial Intelligence*, 2006.

[10] M. Berendt and L. Gram. Epilepsy and seizure classification in 63 dogs: a reappraisal of veterinary epilepsy terminology. *Journal of Veterinary Internal Medicine*, 13:14–20, 1999.

[11] M. Berendt, H. Høgenhaven, A. Flagstad, and M. Dam. Electroencephalography in dogs with epilepsy: similarities between human and canine findings. *Acta Neurologica Scandinavica*, 99:276–283, 1999.

[12] Christopher M Bishop. *Pattern Recognition and Machine Learning*. Number 4 in Information Science and Statistics. Springer-Verlag, New York, New York, 2006.

[13] D. Blackwell and J. B. MacQueen. Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, 1(2):353–355, 1973.

[14] J. A. Blanco, M. Stead, A. Krieger, W. Stacey, D. Maus, Eric Marsh, J. Viventi, K. H. Lee, R. Marsh, B. Litt, and G. A. Worrell. Data mining neocortical high-frequency oscillations in epilepsy and controls. *Brain*, 134:2948–2959, 2011.

[15] D. M. Blei, T. L. Griffiths, and M. I. Jordan. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57 (2):1–30, 2010.

[16] P. Boon, M. D'Havé, C. Adam, K. Vonck, M. Baulac, T. Vandekerckhove, and J. De Reuck. Dipole modeling in epilepsy surgery candidates. *Epilepsia*, 38(2):208–218, 1996.

[17] A. Bragin, J. Engel, C. L. Wilson, I. Fried, and G. Buzsáki. High-frequency oscillations in human brain. *Hippocampus*, 9:137–142, 1999.

[18] A. Bragin, C. L. Wilson, J. Almajano, I. Mody, and J. Engel. High-frequency oscillations after status epilepticus: epileptogenesis and seizure genesis. *Epilepsia*, 45(9): 1017–1023, 2004.

[19] M. Breakspear, J. A. Roberts, J. R. Terry, S. Rodrigues, N. Mahant, and P. A. Robinson. A unifying explanation of primary generalized seizures through nonlinear brain modeling and bifurcation analysis. *Cerebral Cortex*, 16:1296–1313, 2006.

[20] S. P. Brooks and A. Gelman. General methods for monitoring convergence of iterative simulations general methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455, 1998.

[21] S. P. Brooks, A. Gelman, G. L. Jones, and X. Meng, editors. *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/CRC, Boca Raton, Florida, 2011.

[22] C. M. Carvalho, H. Massam, and M. West. Simulation of hyper-inverse Wishart distributions in graphical models. *Biometrika*, 94(3):647–659, 2007.

[23] G. Casella and C. P. Robert. Rao-Blackwellisation of sampling schemes. *Biometrika*, 83(1):81–94, 1996.

[24] A. M. Chan, F. T. Sun, E. H. Boto, and B. M. Wingeier. Automated seizure onset detection for accurate onset time determination in intracranial EEG. *Clinical Neurophysiology*, 119:2687–2696, 2008.

[25] K. Chandler. Canine epilepsy: what can we learn from human seizure disorders? *Veterinary Journal*, 172:207–217, 2006.

[26] W. A. Chaovalitwongse. Novel quadratic programming approach for time series clustering with biomedical application. *Journal of Combinatorial Optimization*, 15:225–241, 2008.

[27] A. W. L. Chiu, M. Derchansky, M. Cotic, P. L. Carlen, S. O. Turner, and B. L. Bardakjian. Wavelet-based Gaussian-mixture hidden Markov model for the detection of multistage seizure dynamics: a proof-of-concept study. *BioMedical Engineering OnLine*, 10(29), 2011.

[28] D. Cosandier-Rimele, J. Badier, P. Chauvel, and F. Wendling. A Physiologically Plausible Spatio-Temporal Model for EEG Signals Recorded With Intracerebral Electrodes in Human Partial Epilepsy. *IEEE Transactions on Biomedical Engineering*, 54 (3):380–388, 2007.

[29] K. A. Davis, B. K. Sturges, C. H. Vite, V. Ruedebusch, G. Worrell, A. B. Gardner, K. Leyde, W. D. Sheffield, and B. Litt. A novel implanted device to wirelessly record and analyze continuous intracranial canine EEG. *Epilepsy Research*, 96:116–122, 2011.

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

[31] P. Dayan and L. F. Abbott. *Theoretical neuroscience: computational and mathematical modeling of neural systems*. The MIT Press, Cambridge, Massachusetts, 2005.

[32] L. Ding, G. A. Worrell, T. D. Lagerlund, and B. He. Ictal source analysis: localization and imaging of causal interactions in humans. *NeuroImage*, 34:575–586, 2007.

[33] J. Engel, Jr. and T. A. Pedley, editors. *Epilepsy: a comprehensive textbook*. Lippincot Williams & Wilkins, Philadelphia, Pennsylvania, second edition, 2008.

[34] M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1995.

[35] R. Esteller, J. Echauz, T. Tcheng, B. Litt, and B. Pless. Line length: an efficient feature for seizure onset detection. In *Proceedings of the 23rd Engineering in Medicine and Biology Society Conference*, 2001.

[36] S. Faul, G. Gregorcic, G. Boylan, W. Marnane, G. Lightbody, and S. Connolly. Gaussian process modeling of EEG for the detection of neonatal seizures. *IEEE Transactions on Biomedical Engineering*, 54(12):2151–2162, 2007.

[37] T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.

[38] E. B. Fox. *Bayesian Nonparametric Learning of Complex Dynamical Phenomena*. PhD thesis, Massachusetts Institute of Technology, 2009.

[39] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky. An HDP-HMM for systems with state persistence. In *Proceedings of the 25th International Conference on Machine Learning*, 2008.

[40] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky. Nonparametric Bayesian learning of switching linear dynamical systems. In *Proceedings of Neural Information Processing Systems*, 2008.

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

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

[43] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky. Bayesian nonparametric inference of switching dynamic linear models. *IEEE Transactions on Signal Processing*, 59:1569–1585, 2011.

[44] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky. Joint Modeling of Multiple Related Time Series via the Beta Process. *Arxiv preprint arXiv:1111.4226v1*, 2011.

[45] J. A. French. Refractory epilepsy: clinical overview. *Epilepsia*, 48(Suppl 1):3–7, 2007.

[46] A. E. Gelfand and A. F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990.

[47] A. Gelman, J. B. Carlin, H. S. Stern, and D. S. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton, Florida, second edition, 2004.

[48] Z. Ghahramani. Nonparametric Bayesian learning. In *Uncertainty in Artificial Intelligence*, 2005.

[49] Z. Ghahramani and M. I. Jordan. Factorial hidden Markov models. *Machine learning*, 29:245–273, 1997.

[50] S. Ghosh-Dastidar, H. Adeli, and N. Dadmehr. Principal component analysis-enhanced cosine radial basis function neural network for robust epilepsy and seizure detection. *IEEE Transactions on Bio-Medical Engineering*, 55(2):512–518, 2008.

[51] P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

[52] F. Grenier, I. Timofeev, and M. Steriade. Neocortical very fast oscillations (ripples, 80-200 Hz) during seizures: intracellular correlates. *Journal of Neurophysiology*, 89: 841–52, 2003.

[53] J. E Griffin and M. F. J. Steel. Order-Based Dependent Dirichlet Processes. *Journal of the American Statistical Association*, 101(473):179–194, 2006.

[54] T. L. Griffiths and Z. Ghahramani. Infinite latent feature models and the Indian buffet process. *Gatsby Computational Neuroscience Unit, Technical Report #2005-001*, 2005.

[55] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer-Verlag, New York, New York, 2001.

[56] A. Hegde, D. Erdogmus, D. S. Shiau, J. C. Principe, and C. J. Sackellares. Clustering approach to quantify long-term spatio-temporal interactions in epileptic intracranial electroencephalography. *Computational Intelligence and Neuroscience*, 2007, 2007.

[57] A. L. Hodgkin and A. F. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology*, 117: 500–544, 1952.

[58] M. C. Hughes, E. B. Fox, and E. B. Sudderth. Effective split-merge Monte Carlo methods for nonparametric models of sequential data. In *Advances in Neural Information Processing Systems*, 2012.

[59] Intel. Intel hyper-threading technology. Technical report, 2003.

[60] H. Ishwaran and L. F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

[61] H. Ishwaran and M. Zarepour. Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika*, 87(2):371–390, 2000.

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

[63] J. Jacobs, M. Zijlmans, R. Zelmann, C.-E. Chatillon, J. Hall, A. Olivier, F. Dubeau, and J. Gotman. High-frequency electroencephalographic oscillations correlate with outcome of epilepsy surgery. *Annals of Neurology*, 67(2):209–20, 2010.

[64] B. Jones, C. Carvalho, A. Dobra, C. Hans, C. Carter, and M. West. Experiments in Stochastic Computation for High-Dimensional Graphical Models. *Statistical Science*, 20(4):388–400, 2005.

[65] M. I. Jordan, Z. Ghahramani, T. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

[66] M. Jordon. Dirichlet processes, Chinese restaurant processes, and all that. In *Neural Information Processing Systems*, 2005.

[67] Z. P. Kilpatrick and P. C. Bressloff. Spatially structured oscillations in a two-dimensional excitatory neuronal network with synaptic depression. *Journal of Computational Neuroscience*, 28:193–209, 2010.

[68] John Kingman. Completely random measures. *Pacific Journal of Mathematics*, 21 (1):59–78, 1967.

[69] A. Klatchko, G. Raviv, W. R. Webber, and R. P. Lesser. Enhancing the detection of seizures with a clustering algorithm. *Electroencephalography and Clinical Neurophysiology*, 106:52–63, 1998.

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

[71] M. Le Van Quyen, J. Soss, V. Navarro, R. Robertson, M. Chavez, M. Baulac, and J. Martinerie. Preictal state identification by synchronization changes in long-term intracranial EEG recordings. *Clinical Neurophysiology*, 116:559–68, 2005.

[72] P. Liang and D. Klein. Structured Bayesian nonparametric models with variational inference. In *Association for Computational Linguistics*, 2007.

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

[74] J. S. Liu, W. H. Wong, and A. Kong. Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, 81(1):27–40, 1994.

[75] X. Liu, D. B. McCreery, R. R. Carter, L. A. Bullara, T. G. Yuen, and W F Agnew. Stability of the interface between neural tissue and chronically implanted intracortical microelectrodes. *IEEE Transactions on Rehabilitation Engineering*, 7(3):315–26, 1999.

[76] B. A. Lopour and A. J. Szeri. A model of feedback control for the charge-balanced suppression of epileptic seizures. *Journal of Computational Neuroscience*, 28:375–387, June 2010.

[77] D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Cambridge, UK, 2003.

[78] I. Merlet and J. Gotman. Reliability of dipole models of epileptic spikes. *Clinical Neurophysiology*, 110:1013–1028, 1999.

[79] P. Müller and F. A. Quintana. Nonparametric Bayesian Data Analysis. *Statistical Science*, 19(1):95–110, 2004.

[80] J. Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1):32–38, 1957.

[81] K. P. Murphy. Conjugate Bayesian analysis of the Gaussian distribution. Technical report, 2007.

[82] T. I. Netoff, R. Clewley, S. Arno, T. Keck, and J. A. White. Epilepsy in small-world networks. *The Journal of Neuroscience*, 24(37):8075–8083, September 2004.

[83] J. J. Niederhauser, R. Esteller, J. Echauz, G. Vachtsevanos, S. Member, and B. Litt. Detection of seizure precursors from depth-EEG using a sign periodogram transform. *IEEE Transactions on Biomedical Engineering*, 51(4):449–458, 2003.

[84] P. Orbanz and Y. W. Teh. Modern Bayesian nonparametrics. In *Neural Information Processing Systems*, 2011.

[85] A. Ossadtchi, R. E. Greenblatt, V. L. Towle, M. H. Kohrman, and K. Kamada. Inferring spatiotemporal network patterns from intracranial EEG data. *Clinical Neurophysiology*, 121(6):823–35, 2010.

[86] P. Paramanathan and R. Uthayakumar. Application of fractal theory in analysis of human electroencephalographic signals. *Computers in Biology and Medicine*, 38: 372–8, 2008.

[87] E. E. Patterson, J. J. Howbert, M. Stead, V. Vasoli, D. Crepeau, C. Vite, B. Sturges, V. Ruedebusch, L. D. Coles, J. C. Cloyd, J. Mavoori, W. D. Sheffield, B. Litt, and G. Worrell. Forecasting seizures in dogs with naturally occurring epilepsy. 2013.

[88] J. Pitman. Poisson-Dirichlet and GEM invariant distributions for split-and-merge transformations of an interval partition. *Combinatorics, Probability & Computing*, 11 (5):501–514, 2002.

[89] J. Pitman and M. Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2):855–900, 1997.

[90] S. R. Platt, V. Adams, L. S. Garosi, C. J. Abramson, J. Penderis, A De Stefani, and L Matiasek. Treatment with gabapentin of 11 dogs with refractory idiopathic epilepsy. *Veterinary Record*, 159:881–884, 2006.

[91] V. S. Polikov, P. A. Tresco, and W. M. Reichert. Response of brain tissue to chronically implanted neural electrodes. *Journal of Neuroscience Methods*, 148:1–18, 2005.

[92] R. Prado and M. West. *Time series modeling, computation, and inference*. Chapman & Hall/CRC, Boca Raton, FL, 2010.

[93] R. Prado, M. West, and A. D. Krystal. Multichannel electroencephalographic analyses via dynamic regression models with time-varying lag/lead structure. *Journal of the Royal Statistical Society*, 50(1):95–109, 2001.

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

[95] L. A. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[96] M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846– 850, 1971.

[97] J. C. Reijneveld, S. C. Ponten, H. W. Berendse, and C. J. Stam. The application of graph theoretical analysis to complex networks in the brain. *Clinical Neurophysiology*, 118:2317–2331, 2007.

[98] S. Richardson and P. J. Green. On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Series B*, 59(4): 731–792, 1997.

[99] C. P. Robert. *The Bayesian Choice*. Springer-Verlag, New York, New York, second edition, 2007.

[100] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer-Verlag, New York, New York, second edition, 2004.

[101] P. A. Robinson, C. J. Rennie, D. L. Rowe, S. S. O'Connor, J. J. Wright, E. Gordon, and R. W. Whitehouse. Neurophysical modeling of brain dynamics. *Neuropsychopharmacology*, 28:74–79, July 2003.

[102] A. Rodriguez and K. Ghosh. Nested Partition Models. 2011.

[103] A. Rodríguez, D. B. Dunson, and A. E. Gelfand. The nested Dirichlet process. *Journal of the American Statistical Association*, 103(483):1131–1154, 2008.

[104] S. Santaniello, S. P. Burns, A. J. Golby, J. M. Singer, W. S. Anderson, and S.V. Sarma. Quickest detection of drug-resistant seizures: an optimal control approach. *Epilepsy & Behavior*, 22:49–60, 2011.

[105] S. Schiff, X. Huang, and J. Wu. Dynamical evolution of spatiotemporal patterns in mammalian middle cortex. *Physical Review Letters*, 98:1–4, 2007.

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

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

[108] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4: 639–650, 1994.

[109] C. E. Shannon. A mathematical theory of communication. *The Bell Systems Technical Journal*, 27:379–423, 1948.

[110] D. Spencer and J. Inserni. *Epilepsy Surgery.* Raven Press, New York, New York, 1991.

[111] V. Srinivasan, C. Eswaran, and N. Sriraam. Approximate entropy-based epileptic EEG detection using artificial neural networks. *IEEE Transactions on Information Technology in Biomedicine*, 11(3):288–95, 2007.

[112] W. C. Stacey and D. M. Durand. Noise and coupling affect signal detection and bursting in a simulated physiological neural network. *Journal of Neurophysiology*, 88: 2598–2611, 2002.

[113] W. C. Stacey, A. Krieger, and B. Litt. Network recruitment to coherent oscillations in a hippocampal computer model. *Journal of Neurophysiology*, 105:1464–1481, 2011.

[114] C. J. Stam. Nonlinear dynamical analysis of EEG and MEG: review of an emerging field. *Clinical Neurophysiology*, 116(10):2266–301, October 2005.

[115] T. Stepleton. Understanding the Antoniak equation. Technical report, 2008.

[116] E. B. Sudderth. *Graphical models for visual object recognition and tracking.* PhD thesis, Massachusetts Institute of Technology, 2006.

[117] F. T. Sun, M. J. Morrell, and R. E. Wharen. Responsive cortical stimulation for the treatment of epilepsy. *Neurotherapeutics*, 5(1):68–74, 2008.

[118] Y. W. Teh. Dirichlet processes: tutorial and practical course. In *Machine Learning Summer School*, 2007.

[119] Y. W. Teh and M. I. Jordan. Hierarchical Bayesian nonparametric models with applications. In N. Hijort, P. Holmes, P. Muller, and S. Walker, editors, *Bayesian Nonparametrics*, pages 158–207. Cambridge University Press, Cambridge, UK, 2010.

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

[121] O. Temkin. *The Falling Sickness*. Johns Hopkins University Press, Baltimore, Maryland, 1971.

[122] R. Thibaux and M. I. Jordan. Hierarchical beta processes and the Indian buffet process. In *Proceedings of the 10th Conference on Artificial Intelligence and Statistics*, 2007.

[123] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B*, 63: 411–423, 2001.

[124] R. D. Traub, M. A. Whittington, E. H. Buhl, F. E. LeBeau, A. Bibbig, S. Boyd, H. Cross, and T. Baldeweg. A possible role for gap junctions in generation of very fast EEG oscillations preceding the onset of, and perhaps initiating, seizures. *Epilepsia*, 42(2):153–170, 2001.

[125] P. Van Hese, B. Vanrumste, H. Hallez, G. J. Carroll, K. Vonck, R. D. Jones, P. J. Bones, and I. Lemahieu. Detection of focal epileptiform events in the EEG by spatio-temporal dipole clustering. *Clinical Neurophysiology*, 119:1756–1770, 2008.

[126] H. A. Volk, L. A. Matiasek, A. Luján Feliu-Pascual, S. R. Platt, and K. E. Chandler. The efficacy and tolerability of levetiracetam in pharmacoresistant epileptic dogs. *The Veterinary Journal*, 176:310–319, 2008.

[127] T. von Klopmann, B. Rambeck, and A. Tipold. Prospective study of zonisamide therapy for refractory idiopathic epilepsy in dogs. *The Journal of Small Animal Practice*, 48:134–138, 2007.

[128] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

[129] S. G. Walker, P. Damien, P. W. Laud, and A. F. M. Smith. Bayesian Nonparametric Inference for Random Distributions and Related Functions. *Journal of the Royal Statistical Society, Series B*, 61:485–527, 1999.

[130] F. Wendling, F. Bartolomei, J. J. Bellanger, and P. Chauvel. Interpretation of interdependencies in epileptic signals using a macroscopic physiological model of the EEG. *Clinical Neurophysiology*, 112:1201–1218, 2001.

[131] M. West, R. Prado, and A. D. Krystal. Evaluation and comparison of EEG traces: latent structure in nonstationary time series. *Journal of the American Statistical Association*, 94(446):1083–1095, 1999.

[132] J. V. Wilson and E. H. Reynolds. Translation and analysis of a cuneiform text forming part of a Babylonian treatise on epilepsy. *Medical History*, 34:185–98, 1990.