THE DIGITAL PATIENT: MACHINE LEARNING TECHNIQUES FOR ANALYZING ELECTRONIC HEALTH RECORD DATA

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Suchi Saria

November 2011

Abstract

The current unprecedented rate of digitization of longitudinal health data — continuous device monitoring data, laboratory measurements, medication orders, treatment reports, reports of physician assessments — allows visibility into patient health at increasing levels of detail. A clearer lens into this data could help improve decision making both for individual physicians on the front lines of care, and for policy makers setting national direction.

However, this type of data is high-dimensional (an infant with no prior clinical history can have more than 1000 different measurements in the ICU), highly unstructured (the measurements occur irregularly, and different numbers and types of measurements are taken for different patients) and heterogeneous (from ultrasound assessments to lab tests to continuous monitor data). Furthermore, the data is often sparse, systematically not present, and the underlying system is non-stationary. Extracting the full value of the existing data requires novel approaches.

In this thesis, we develop novel methods to show how longitudinal health data contained in Electronic Health Records (EHRs) can be harnessed for making novel clinical discoveries. For this, one requires access to patient outcome data — which patient has which complications. We present a method for automated extraction of patient outcomes from EHR data; our method shows how natural languages cues from the physicians notes can be combined with clinical events that occur during a patient's length of stay in the hospital to extract significantly higher quality annotations than previous state-of-the-art systems.

We develop novel methods for exploratory analysis and structure discovery in bedside monitor data. This data forms the bulk of the data collected on any patient yet, it is not utilized in any substantive way post collection. We present methods to discover recurring *shape* and *dynamic* signatures in this data. While we primarily focus on clinical time series, our methods also generalize to other continuous-valued time series data.

Our analysis of the bedside monitor data led us to a novel use of this data for risk prediction in infants. Using features automatically extracted from physiologic signals collected in the first 3 hours of life, we develop Physiscore, a tool that predicts infants at risk for major complications downstream. Physiscore is both fully automated and significantly more accurate than the current standard of care. It can be used for resource optimization within a NICU, managing infant transport to a higher level of care and parental counseling. Overall, this thesis illustrates how the use of machine learning for analyzing these large scale digital patient data repositories can yield new clinical discoveries and potentially useful tools for improving patient care.

Preface

This document only presents the third chapter of the dissertation titled "The Digital Patient: Machine Learning techniques for analyzing Electronic Health Record data."

Contents

List of Tables

List of Figures

Chapter 1

Discovering Dynamic Signatures in Physiologic data

The task of discovering novel medical knowledge from complex, large scale and high-dimensional patient data, collected during care episodes, is central to innovation in medicine. This chapter, and the next addresses the task of discovery in the context of physiologic data from the bedside monitors.

In this chapter, we propose a method for exploratory data analysis and feature construction in continuous-valued physiologic time series. While our primary motivation comes from clinical data, our methods are applicable to other time series domain. Our method focuses on revealing shared patterns in corpora where individual time series differ greatly in their composition.

1.1 Introduction

Time series data is ubiquitous. The task of knowledge discovery from such data is important in many scientific disciplines including patient tracking, activity modeling, speech and ecology. For example, in our domain of seeking to understand disease pathogenesis from physiologic measurements (e.g., heart rate signal shown in figure 1.1), several interesting questions arise. Are there any repeating patterns or signatures in this data? How many such signatures exist and what their characteristics might be? Furthermore, are there collections of signatures that co-occur and are indicative of the underlying (disease) state? Such questions arise in other domains as well including surveillance and wild-life monitoring. In

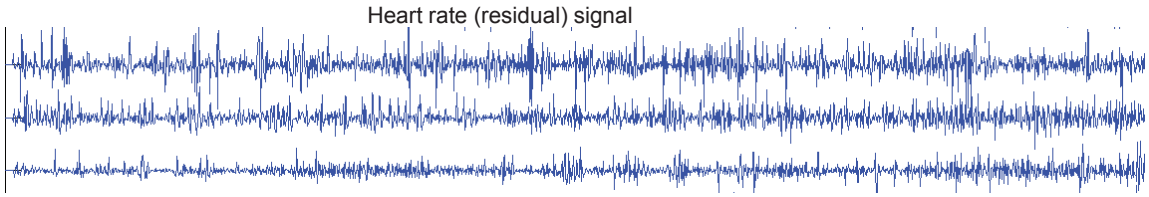

Figure 1.1: Heart signal (mean removed) from three infants in their first few hours of life.

such domains where the outcome of interest (e.g., health status) is difficult to measure directly and surrogate measurements are made instead (e.g., physiological variables), latent (hidden) variable models are a natural choice for knowledge discovery. Different diseases might be associated with multiple latent states that each generate data with distinct physiologic characteristics. Structure discovered with such a model can help reveal how diseases manifest, uncover novel disease associations, and highlight relationships between diseases.

In many temporal domains, individual series show significant variability and an a priori breakdown of data into distinct sets is unclear. In clinical data, for example, two patients are rarely alike; they may suffer from different sets of diseases and to varying extents. Traditional generative models for time series data, such as switching Kalman filters [Bar-Shalom and Fortmann, 1987] or mixtures of such models [Fine *et al.*, 1x998], assume the data to be generated from a discrete set of classes, each specifying the generation of a homogeneous population of i.i.d. time series. To see the shortcoming of such an approach, in our example, the patient state over time transitions over a large set of latent states (coughing, wheezing, sleeping and so on). Generation of all series from a single transition matrix over the set of latent states assumes that all series express these latent states in the same proportion (on expectation). But, in reality, different patients express these states in radically different proportions, depending on their combination of diseases and other physiological factors. While mixture models (inducing a distribution over different dynamic models) can generate additional variability, the set of possible combinations can grow combinatorially large. And, thus, a pre-imposed partition of the space of patients into a fixed number of classes limits our ability to model instance-specific variability.

Hierarchical Bayesian modeling [Kass and Steffey, 1989; Gelman *et al.*, 1995] has been proposed as a general framework for modeling variability between individual "units". As an example of this framework, in the domain of natural language processing, Latent Dirichlet

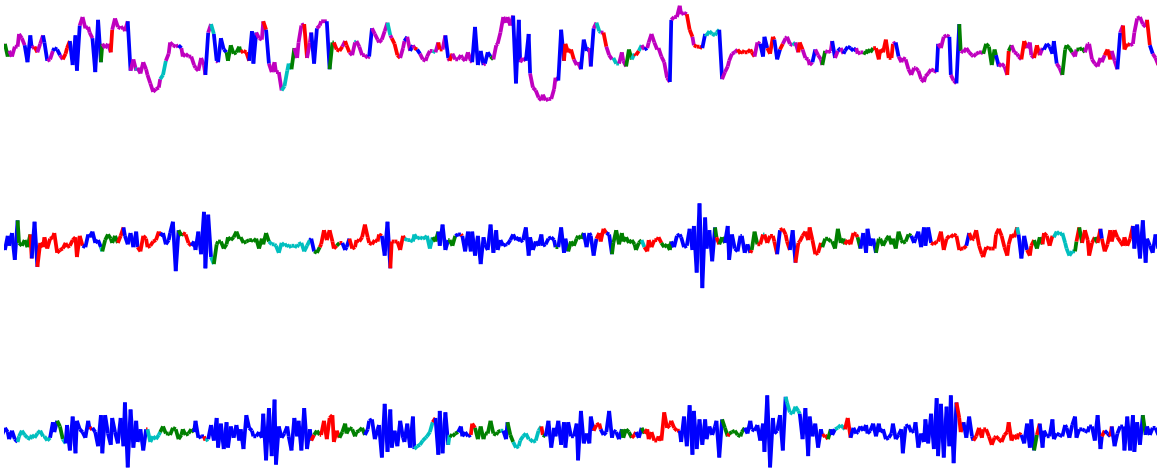

Figure 1.2: Example of three time series with shared signatures. Segments of each distinct color are generated by the same function, and thereby are instances of the same signature or *word*. Signatures here correspond to autoregressive functions. The choice of function used at any given time depends on the latent *topic* at that time. While the three series differ greatly in their composition, they contain shared structure to varying extents.

Allocation (LDA) [Blei *et al.*, 2003] has found success as a representation for uncovering the underlying structure of document corpora. Each document is associated with its own distribution over latent variables called topics, each of which is shared across the population and defines a distribution over words. Analogously, in our application[1], an individual patient maintains its own distribution over both latent (disease) topics and transitions between them. Each topic defines a distribution over temporal signatures (physiologic symptoms) observed in the time series and these behaviors play the role of words. However, unlike text data, in continuous-valued time series data, the notion of a word is non-obvious. A word could be specified as a segmented window of the data itself, but this allows for little compression, as most continuous-valued time series segments, unlike discrete text segments, do not repeat exactly. Our proposed model uses a more flexible representation of a word that specifies a parametric function to generate the temporal dynamics for the duration of that word. For example, in figure 1.2, autoregressive functions are used for generating the temporal dynamics. Each distinct color can be likened to a word and therefore, there are

[1]Our model is a more general instance of Hierarchical Bayes than LDA which models only discrete data. The analogy to LDA is made primarily to provide the readers a familiar overview of our model.

five "words" in this corpora. Moreover, the duration of the word also does not need to be fixed in advance, and as shown may vary from one occurrence to another. Hence, our model also postulates word boundaries.

In our approach, words are selected from an infinite dimensional latent space that corresponds to the possible real-valued instantiations to the parameters of the functions that generate the data. Naive sampling in this infinite-dimensional space given the data will result in no sharing of words across topics [Teh *et al.*, 2006]. For knowledge discovery tasks, sharing of words across topics is particularly desirable, as it allows us to uncover relationships between different latent states. For example, one can infer which diseases are physiologically similar based on the extent to which they share words. To enable sharing, we utilize *hierarchical Dirichlet processes* (HDPs) [Teh *et al.*, 2006], designed to allow sharing of mixture components within a multi-level hierarchy. Thus, our model discovers words and topics shared across the population while simultaneously modeling series-specific dynamics.

The chapter is structured as follows: we first give background on the existing building blocks of Dirichlet Processes and HDPs used in our model. We then describe the time series topic model (TSTM), a flexible hierarchical latent variable model for knowledge discovery in time series data, especially useful for domains when between series variability is significant. Next, we describe related work in models for processing continuous time-series data. Following this, we provide a block Gibbs sampler for TSTM. We present results on our target application of analyzing physiological time series data. We demonstrate usefulness of the model in constructing features within a supervised learning task. We also qualitatively evaluate the model output and derive new clinical insights that led to the development of a state-of-the-art personalized risk stratification score for morbidity in infants described in Chapter 5.

1.2 Background

Below, we briefly define Dirichlet Process and the Hierarchical Dirichlet Process. Though several texts have described these distributions in great detail before, for the sake of being comprehensive, we describe key properties here that give intuition about their use as priors on mixture models. The text for the following subsections is adapted from Fox et al. [2007].

1.2.1 Dirichlet Processes

The Dirichlet Process (DP) is commonly used as a prior on the parameters of a mixture model with a random number of components. A DP is a distribution on probability measures on a measurable space Θ. This stochastic process is uniquely defined by a base measure H on Θ and a concentration parameter γ. Consider a random probability measure $G_o \sim DP(\gamma, H)$. The DP is formally defined by the property that for any finite partition $\{A_1, \cdots, A_K\}$ of Θ,

$$(G_0(A_1), \cdots, G_0(A_K)) \mid \gamma, H \ \sim \ \mathrm{Dir}(\gamma H(A_1), \cdots, \gamma H(A_K))$$

That is, the measure of a random probability distribution $G_o \sim DP(\gamma, H)$ on every finite partition of Θ follows a finite-dimensional Dirichlet distribution [Ferguson, 1973]. A more constructive definition of the DP was given by Sethuraman [1994]. He shows that $G_0 \sim \mathrm{DP}(\gamma, H)$, a sample drawn from the DP prior, is a discrete distribution because, with probability one:

$$G_0 = \Sigma_{k=1}^{\infty} \beta_k \delta_{\theta_k} \tag{1.1}$$

where $\theta_k \sim H$. The sampling of β_k follows a *stick-breaking construction* defined as:

$$\beta_k = \beta'_k \prod_{l=1}^{k-1} (1 - \beta'_l) \qquad \beta'_k \sim \mathrm{Beta}(1, \gamma) \tag{1.2}$$

Essentially, we have divided a unit stick into lengths given by the weights β_k. β_k is a random proportion β'_k of the remaining stick after the previous $(k-1)$ weights have been defined. Generally, this construction is denoted by $\beta \sim \mathrm{GEM}(\gamma)$.

To give intuition for how the DP is used as a prior on the parameters of a mixture model with a random number of components, consider draws from H to be the description of candidate cluster centers. The weights β_k define the mixing proportions. γ controls the relative proportion of the mixing weights, and thus determines the model complexity in terms of the expected number of components with significant probability mass.

To see why the DP as a prior induces clustering, we visit another property of the DP, introduced by Blackwell and MacQueen [1973]. Consider a set of observations $\{y_i\}$ sampled from G_0. Consider z_i to be the variables that select for each data observation y_i the unique

value θ_k that the observation is sampled from i.e. $y_i \sim f(\theta_{z_i})$. Let K be the number of unique θ_k values that have data observations $y_1, \cdots, y_N$ associated with them.

$$p(z_{N+1} = z | z_1, \cdots, z_N, \gamma) = \frac{\gamma}{N+\gamma}\mathcal{I}(z = K+1) + \frac{1}{N+\gamma}\sum_{k=1}^{K}\sum_{i=1}^{N}\mathcal{I}(z_i = k)\mathcal{I}(z = k)$$

In other words, z_{N+1} samples its value based on how frequently these values have been used by previously sampled data observations and is more likely to sample a frequently sampled value. Thus, we see that the DP has a reinforcement property that leads to a clustering of the data. This property is essential in deriving finite and compact models.

Finally, we can also obtain the DP mixture model as the limit of a sequence of finite mixture models. It can be shown under mild conditions that if the data were generated by a finite mixture, then the DP posterior is guaranteed to converge (in distribution) to that finite set of mixture parameters [Ishwaran and Zarepour, 2002a]. Let us assume that there L components in a mixture model and we place a finite-dimensional Dirichlet prior on these mixture weights:

$$\beta | \gamma \sim \text{Dir}(\gamma/L, \cdots, \gamma/L) \tag{1.3}$$

Let $G_0^L = \sum_{k=1}^{L} \beta_k \delta_{\theta_k}$. Then, it can be shown [Ishwaran and Zarepour, 2000; 2002b] that for every measurable function f integrable with respect to the measure H, this finite distribution G_o^L converges weakly to a countably infinite distribution G_0 distribution according to a Dirichlet process. Similar to Fox et al., [2007], we use this truncation property in the development of our block Gibbs sampler.

1.2.2 Hierarchical Dirichlet Processes

There are many scenarios where groups of data are thought to be produced by related, yet distinct, generative processes. For example, in our target application of physiologic monitoring, different diseases and syndromes likely share physiologic traits, yet data for any single disease should be grouped and described by a similar but different model from that of another disease. Similarly, in document modeling, news articles may share topics in common. Yet, documents published in, say, the New York Times should be grouped and described by a similar but different model from that of the Wall Street Journal. Similarly, in surveillance, different activity trajectories may contain activities in common. Yet, data

collected at different times of the day should be modeled as different but similar groups.

The Hierarchical Dirichlet Process [Teh *et al.*, 2006] extends the DP to enable sharing in such scenarios by taking a hierarchical Bayesian approach: a global Dirichlet process prior $DP(\eta, G_0)$ is placed on Θ and group-specific distributions are drawn from a the global prior $G_j \sim DP(\eta, G_0)$, where the base measure G_o acts as an "average" distribution across all groups. When the base measure $G_0 \sim DP(\gamma, H)$ itself is distributed according to a Dirichlet process, the discrete atoms θ_k are shared both within and between groups. If the base measure G_0 were instead fixed and absolutely continuous with respect to Lebesgue measure, there would be zero probability of the group-specific distributions having overlapping support.

More formally, draws $G_d \sim \mathrm{DP}(\eta, G_o)$ from an HDP can be described as

$$G_0 = \Sigma_{k=1}^{\infty} \beta_k \delta_{\theta_k} \qquad \beta_k \sim \mathrm{GEM}(\gamma), \quad \theta_k \sim H \tag{1.4}$$

$$G_d = \Sigma_{k=1}^{\infty} \Sigma_{t=1}^{\infty} \hat{\beta}_{jt} \delta_{\theta_{jt}} I(\theta_{jt} = \theta_k) \qquad \hat{\beta}_{jt} \sim \mathrm{GEM}(\eta), \quad \theta_{jt} \sim G_0 \tag{1.5}$$

Essentially, since G_d samples its values θ_{jt} from a discrete distribution, any given atom θ_k from the base distribution can be sampled more than once. The corresponding weight for that atom θ_k in G_d is computed by aggregating the sampled weights $\hat{\beta}_{jt}$ for all atoms $\theta_{jt} = \theta_k$.

As with the DP, the HDP mixture model has an interpretation as the limit of a finite mixture model. Placing a finite Dirichlet prior on the global distribution induces a finite Dirichlet prior on the group-specific distribution and as $L \to \infty$, this model converges in distribution to the HDP mixture model [Teh *et al.*, 2006]:

$$\beta | \gamma \sim \mathrm{Dir}(\gamma/L, \cdots, \gamma/L) \tag{1.6}$$

$$\phi_j | \eta, \beta \sim \mathrm{Dir}(\eta\beta_1, \cdots, \eta\beta_L) \tag{1.7}$$

Our block Gibbs sampler for performing inference in the TSTM exploits this truncation property.

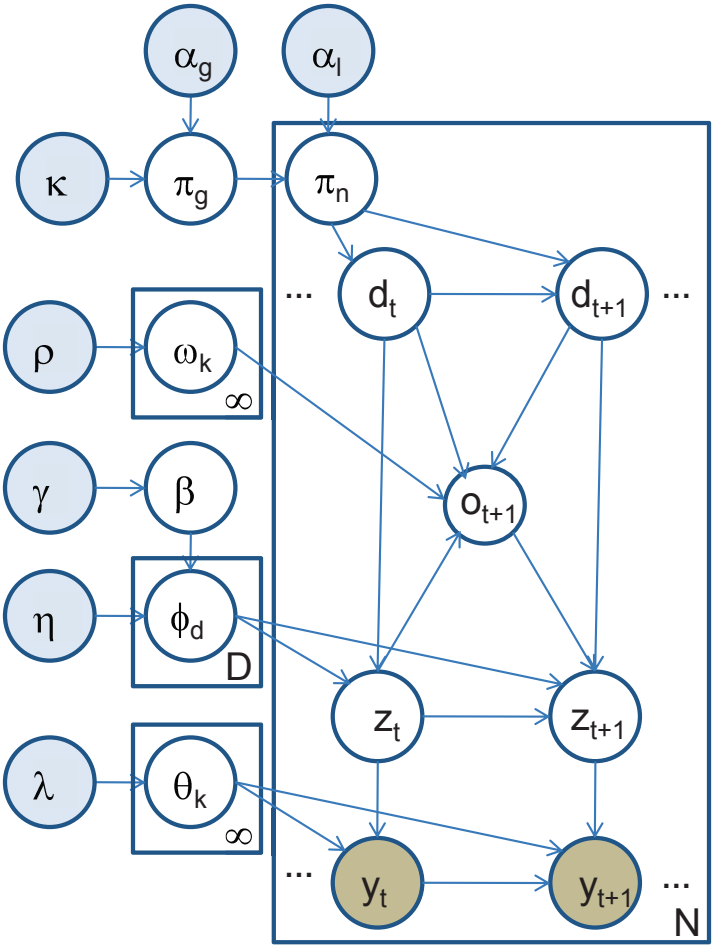

Figure 1.3: Graphical representation of the Time Series Topic Model (TSTM).

1.3 Time Series Topic Model

Time Series Topic Model (TSTM) is a 4-level hierarchical Bayesian model. It makes the assumption that there is an underlying fixed set of topics that is common to the heterogeneous collection of time series in the corpus. A topic is a distribution over the vocabulary of all words present in the corpus. An individual time series is generated by first choosing a series-specific transition matrix over the topics. To sample each "word": sample a topic, and then sample a word from that topics' distribution over words.

As discussed above, unlike discrete sequence data (e.g., text), in time series data the

Symbol	Description
D	Number of topics.
$\phi_{1:D}$	Topics 1 through D.
ϕ_d	The dth topic.
π_g	Global topic transition matrix.
π_n	Series specific topic transition matrix.
$f_k(.;\theta_k)$	parametric functions that generate the data; k indexes these functions. Also, refered to as the kth word.
y_t	Data observed at time t.
$\mathcal{I}(E)$	Indicator function where E is the event.
$1:D$	Abbreviation for 1 through D.

Table 1.1: Notation.

features to be extracted are often not structurally obvious (see figure 1.1). Pre-segmenting the sequence data into "words" does not offer sufficient flexibility to learn from the data, especially in the realm of exploration for knowledge discovery. Thus, TSTM discovers words from an infinite-dimensional parametric function space while simultaneously learning topics and series-specific evolution parameters.

We describe below each of the TSTM components. We begin by giving a brief overview of the random variables in the model and then describe the generation process (in a bottom-up fashion). We define notation that is commonly used in this chapter in table 1.1.

1.3.1 Overview of the model variables

Random variable y_t denotes the observation at a given time t (see figure 1.3 for the graphical model). z_t is a latent variable that indexes the "word" used to generate the data observed at that time. $d_t \in \{1, \cdots, D\}$ tracks the latent topic at any given time. Binary variables o_t control the word length. The word at a given time t, z_t is generated from the topic distribution ϕ_{d_t}. Each series has a series specific topic transition matrix π_n from which d_t is sampled at each time t. The matrices π_n are sampled from a global topic transition matrix π_g.

1.3.2 Generative Process

Data generation model: Let $\mathcal{F} = \{f(. : \theta) : \theta \in \Theta\}$ be the set of all possible data generating functions and $f_k = f(. : \theta_k)$ be the fuction indexed by k. We assume that the continuous-valued data y_t at time t is generated using a function f_k These functions take as inputs $\vec{x}_t$, values dependent on current and previous time slices, and generate the output as $y_t = f(\vec{x}_t; \theta_k)$, denoted as f_k. f_k, an expressive characterization of the time series dynamics, can be thought of as the kth word in the time-series corpus vocabulary. The parameterization of f_k depends on the choice of the observation model. Below, we describe the Vector Autoregressive Process (VAR) observation model. VARs have been used extensively for temporal modeling in numerous domains, including medical time series of fMRI, EEG and physiologic data [Williams *et al.*, 2005]. We use this observation model for our target application and refer to functions discovered by applying TSTM to physiologic data as *dynamic signatures*.

Depending on the data, other observation models (such as the mixture model emissions utilized in [Fox *et al.*, 2007]) can be used instead within TSTM.

In an order p autoregressive process, given a function f_k with parameters $\{A^k, V^k\}$, observed data y_t is assumed to be generated as:

$$\vec{y}_t = A^k X_t^T + \vec{v}_t \qquad v_t \sim \mathcal{N}(0, V^k)$$

and $\vec{y}_t \in \mathcal{R}^m$ for an m-dimensional series. The inputs, $X_t = [\vec{y}_{t-1}, \ldots, \vec{y}_{t-p}]$. Parameters $A^k \in \mathcal{R}^{m \times p}$, and V^k is an $m \times m$ positive-semidefinite covariance matrix. The kth word then corresponds to a specific instantiation of the function parameters $\{A^k, V^k\}$. For TSTM, we want the words to persist for more than one time step. Thus, for each word, we have an additional parameter ω_k that specifies the mean length of the word as $1/\omega_k$. We describe how ω_k is used in data generation, below. For any $f_k \in \mathcal{F}$, we denote the function parameters more generally by $\vec{\theta}_k \in \Theta$.

Dynamics of words and topics: Given the words ($\mathcal{F}$), topics ($\phi_{1:D}$, D is the maximum number of topics) and series-specific transition matrices (π_n), the series generation is straightforward. For each time slice $t \in 1, \cdots, T$,

1. generate the current latent topic state given the topic at previous time-step, $d_t \sim \text{Mult}(\pi_n^{d_{t-1}})$,

2. generate the switching variables o_t, which determine whether a new word is selected. A new word is always generated ($o_t = 0$) if the latent state has changed from the previous time step; otherwise, o_t is selected from a Bernoulli distribution whose parameter determines the word length. Thus, $o_t \sim \mathcal{I}(d_t = d_{t-1})\text{Bernoulli}(\omega_{z_{t-1}})$, where I is the indicator function.

3. the identity of the word to be applied is generated; if $o_t = 1$, we have $z_t = z_{t-1}$, otherwise $z_t \sim \text{Mult}(\phi_{d_t})$.

4. the observation given the temporal function index z_t is generated as $y_t \sim f(x_t; \theta_{z_t})$.

The series specific topic transition distribution π_n is generated from the global topic transition distribution π_g. To generate π_n, each row i is generated from $\text{Dir}(\alpha_l \pi_g^i)$, where π_g^i is the ith row of the global topic transition distribution. Hyperparameter α_l controls the degree of sharing across series in our belief about the prevalence of latent topic states. A large α_l assigns a stronger prior and allows less variability across series. Given hyperparameters α_g and κ, $\pi_g^i \sim Dir(\alpha_g + \kappa\delta_i)$. κ controls the degree of self-transitions for the individual topics.

Word and Topic descriptions: To uncover the finite data generating parametric function set $\mathcal{F}$ where these functions are shared across latent topics ϕ_d, we use the hierarchical Dirichlet process (HDP) [Teh *et al.*, 2006]. Thus,

$$\phi_d \sim DP(\eta, \beta), \quad \beta \sim \text{GEM}(\gamma), \quad \theta_k \sim H \tag{1.8}$$

First, we define the base distribution H. Similar to [Fox *et al.*, 2009], we use a matrix-normal inverse-Wishart prior on the parameters $\{A^k, V^k\}$ of the autoregressive process and a symmetric Beta prior on ω_k as our base measures H. ϕ_d and β are easily generated using the truncation property, described in detail as part of the inference algorithm in section 1.5.

While we do not use the stick-breaking representation in our derivation of the inference algorithm, it is instructive to see how the HDP induces shared words between topics in the TSTM. Draws from H yield candidate words or data generating functions denoted by atoms δ_{θ_k} in Eq. 1.4. By associating each data sample y_t (time points in the series) through the latent variables z_t with a specific data generation function, the posterior distribution yields a probability distribution on different partitions of the data. The mixing proportion (the weights for each θ_k in Eq. 1.4) in the posterior distribution is obtained from aggregating

corresponding weights from the prior and the assigned data samples.

Since measures G_d in Eq. 1.5 are sampled from $\mathrm{DP}(\eta, G_0)$ with a discrete measure as its base measure, the resulting topic distributions ϕ_d have a non-zero probability of regenerating the same data generating functions θ_k, thereby sharing functions between related topics.

1.4 Related Work

An enormous body of work has been devoted to the task of modeling time series data. Probabilistic generative models, the category to which our work belongs, typically utilize a variant of a switching dynamical system [Bar-Shalom and Fortmann, 1987; Fine *et al.*, 1x998], where one or more discrete state variables determine the momentary system dynamics, which can be either linear or in a richer parametric class. The autoregressive hidden Markov model (AR-HMM) is one such example. Observations at any given time are generated via an autoregressive process. A hidden Markov model selects the choice of autoregressive process used at that time. However, these methods, as discussed above, typically utilize a single model (as is the case in an AR-HMM) for all the time series in the data set, or at most define a mixture over such models, using a limited set of classes. These methods are therefore unable to capture significant individual variations in the dynamics of the trajectories for different patients, required in our data.

Recent work by Fox and colleagues [Fox *et al.*, 2009; 2007; 2008] uses nonparametric Bayesian models for capturing the generation of continuous-valued time series. Works [Fox *et al.*, 2007; 2008] have utilized hierarchical Dirichlet process priors for inferring the number of features in hidden Markov models and switching linear dynamical systems but, akin to our discussion of [Bar-Shalom and Fortmann, 1987] above, these models do not explicitly represent variability across exemplar series. Conceptually, the present work is most closely related to BP-AR-HMMs [Fox *et al.*, 2009], which captures variability between series by sampling subsets of words (AR processes) specific to individual series. However, unlike TSTM, BP-AR-HMM does not have a mechanism for modeling the high-level topics that hierarchically capture structure in the collection of words. We show example results from the BP-AR-HMM in the results section to further elucidate the benefits of the generation mechanism of TSTM over BP-AR-HMM. Temporal extensions of LDA [Wang *et al.*, 2008; Wang and McCallum, 2006] model evolution of topic compositions over time in text data but not continuous-valued temporal data.

A very different approach to analyzing time series data is to attempt to extract features from the trajectory without necessarily constructing a generative model. For example, one standard procedure is to re-encode the time series using a Fourier or wavelet basis, and then look for large coefficients in this representation. However, the resulting coefficients do not capture signals that are meaningful in the space of the original signal, and are therefore hard to interpret. Features can also be constructed using alternative methods that produce more interpretable output, such as the work on sparse bases [Lee *et al.*, 2009]. However, this class of methods, as well as others [Mueen *et al.*, 2009], require that we first select a window length for identifying common features, whereas no such natural granularity exists in many applications. Moreover, none of these methods aims to discover higher level structure where words are associated with different "topics" to different extents.

1.5 Approximate Inference

Several approximate inference algorithms have been developed for mixture modeling using the HDP; see [Teh *et al.*, 2006; Fox *et al.*, 2007; Kurihara *et al.*, 2007] for a discussion and comparison. We use a block-Gibbs sampler that relies on the *degree L weak limit* approximation presented in [Ishwaran and Zarepour, 2002b]. This sampler has the advantage of being simple, computationally efficient and shows faster mixing than most alternate sampling schemes [Fox *et al.*, 2007]. The block-Gibbs sampler for TSTM proceeds by alternating between sampling of the state variables $\{d_t, z_t\}$, the model parameters, and the series specific transition matrices.

We detail the update steps of our block-Gibbs inference algorithm below. To briefly describe new notation used below, n indexes individual series. We drop the index n when explicit that the variable refers to a single series. We drop sub-indices when all instances of a variable are used (e.g., $z_{1:N,1:T_n}$ is written as z for short).

Sampling latent topic descriptions β, ϕ_d: The DP can also be viewed as the infinite limit of the order L mixture model [Ishwaran and Zarepour, 2002b; Teh *et al.*, 2006]:

$$\beta|\gamma \sim \mathrm{Dir}(\gamma/L, \cdots, \gamma/L)$$

$$\phi_d \sim \mathrm{Dir}(\eta\beta) \qquad \theta_k \sim H$$

We can approximate the limit by choosing L to be larger than the expected number of words in the data set. The prior distribution over each topic-specific word distribution is then:

$$\phi_d|\beta,\eta \sim \text{Dir}(\eta\beta_1,\cdots,\eta\beta_L)$$

Within an iteration of the sampler, let $m_{d,l}$ be the counts for the number of times $z_{n,t}$ sampled the lth word[2] for the dth disease topic; that is, let

$$m_{d,l} = \sum_{n=1:N}\sum_{t=1:T_n} \mathcal{I}(z_{n,t}=l)\mathcal{I}(d_{n,t}=d)\mathcal{I}(o_{n,t}=0)$$

and

$$m_{\cdot,l} = \sum_{d=1:D} m_{d,l}$$

The posterior distribution for the global and individual topic parameters is:

$$\beta|z,d,\gamma \sim \text{Dir}(\gamma/L + m_{\cdot,1},\cdots,\gamma/L + m_{\cdot,L})$$

$$\phi_{d'}|z,d,\eta,\beta \sim \text{Dir}(\eta\beta_1 + m_{d',1},\cdots,\eta\beta_L + m_{d',L})$$

Sampling word parameters ω_l and θ_l: Loosely, the mean word length of the lth word is $1/\omega_l$. A symmetric Beta prior with hyperparameter ρ, conjugate to the Bernoulli distribution, is used as a prior over word lengths. The sufficient statistics needed for the posterior distribution of ω_l are the counts:

$$\bar{c}_{l,i} = \sum_{n=1:N}\sum_{t=1:T_n} \mathcal{I}(d_{n,t}=d_{n,t-1})\mathcal{I}(z_{n,t-1}=l)\mathcal{I}(o_{n,t}=i)$$

where $i \in \{0,1\}$, representing the number of time steps, across all sequences, in which the topic remained the same, the word was initially l, and the word either changed ($o_{n,t}=1$) or not ($o_{n,t}=0$). Thus, $\omega_l|\bar{c}_{l,\cdot},\rho \sim \text{Beta}(\rho/2+\bar{c}_{l,1},\rho/2+\bar{c}_{l,0})$.

[2]Within this approximation, words are ordered such that all words that are observed in the corpus are assigned indices less than L. Thus, l indexes the lth observed word, which can correspond to different parameter instantiations over different iterations.

For sampling the AR generating function parameters, note that, conditioned on the mode assignments z, the observations $y_{1:T,1:N}$ can be partitioned into sets corresponding to each unique $l \in L$. This gives rise to L independent linear regression problems of the form $Y^l = A^l X^l + E^l$ where Y^l is the target variable, with observations generated from mode l, stacked column-wise. X^l is a matrix with the corresponding r lagged observations and E^l is the corresponding noise matrix. The parameters A^l and V^l are sampled from the posterior given conjugate priors of the Matrix-Normal Inverse-Wishart, similar to [Fox *et al.*, 2009].

Sampling global and series-specific transition matrices, π_g and π_n: Since the number of topic states D is known, and we use conjugate priors of Dirichlet distribution for each row of the transition matrix, the posterior update simply involves summing up counts from the prior and the data. The relevant count vectors are computed as $c^i_{n,k} = \sum_{t=1}^{T_n} \mathcal{I}(d_{n,t-1} = i)\mathcal{I}(d_{n,t} = k)$ and $c^i_k = \sum_{n=1}^{N} c^i_{n,k}$ which aggregates over each series. $\vec{c^i} = \{c^i_1, \cdots, c^i_D\}$ and i indexes a row of the transition matrix:

$$\pi^i_g | d, \alpha_g, \kappa \sim \text{Dir}(\alpha_g + \kappa\delta_i + \vec{c^i})$$

$$\pi^i_n | \pi_g, d, \alpha_l \sim \text{Dir}(\alpha_l \pi^i_g + c^i_{n,1:D})$$

Sampling state variables: If all model parameters (topic and word descriptions) are specified, then one can exploit the structure of the dependency graph to compute the posterior over the state variables using a single forward-backward pass. This is the key motivation behind using block Gibbs. The joint posterior can be computed recursively. Forward sampling is used to sample the variables in each time slice given the samples from the previous time slice as $P(z_{1:T}, d_{1:T} | y_{1:T}, \vec{\pi}) = \prod_t P(z_t, d_t | z_{t-1}, d_{t-1}, y_{1:T}, \vec{\pi})$. Top-down sampling is used within a given time slice.

Let $\vec{\pi}$ represent the vector of all model parameter values $\{\pi_{1:N}, \omega_{1:L}, \theta_{l:L}, \phi_{1:D}\}$ instantiated in the previous Gibbs iteration. Since state variables for individual time series n can be sampled independently from the posterior, we drop this index and represent $\vec{s}_t = \{d_t, z_t, o_t\}$ as the state variables in time slice t for any given series. When obvious, we drop mention of the relevant model parameter to the right of the conditioning bar.

The joint posterior is:

$$P(z_{1:T}, d_{1:T}|y_{1:T}, \vec{\pi}) = \prod_t P(z_t, d_t|z_{t-1}, d_{t-1}, y_{1:T}, \vec{\pi})$$

To use forward sampling to sample variables in each time slice given samples for the previous time slice, we compute $P(\vec{s}_t|\vec{s}_{t-1}, y_{1:T}, \vec{\pi})$. This can derived recursively as:

$$P(z_t, d_t|z_{t-1}, d_{t-1}, y_{1:T}, \vec{\pi}) = \frac{P(z_t, d_t|z_{t-1}, d_{t-1}, \vec{\pi}) g(y_{t:T}|z_t, d_t, \vec{\pi})}{\sum_{z_t, d_t} P(z_t, d_t|z_{t-1}, d_{t-1}, \vec{\pi}) g(y_{t:T}|z_t, d_t, \vec{\pi})} \tag{1.9}$$

The first term in the numerator is:

$$\begin{aligned} P(z_t, d_t|z_{t-1}, d_{t-1}, \vec{\pi}) &= P(d_t|d_{t-1}, \pi_n^{d_{t-1}})(P(o_t = 1|\omega_{z_{t-1}}, z_{t-1})\mathcal{I}(z_{t-1} = z_t) + \\ &\quad P(o_t = 0|\omega_{z_{t-1}}, z_{t-1})P(z_t|d_t))^{\mathcal{I}(d_t = d_{t-1})} P(z_t|d_t)^{(1-\mathcal{I}(d_t = d_{t-1}))} \end{aligned}$$

The second term in Eq. 1.9 can be computed recursively using message passing starting at $t = T$ where $g(y_{t+1:T}|z_{t+1}, d_{t+1}, \vec{\pi}) = 1$ and moving backward

$$g(y_{t:T}|z_t, d_t, \vec{\pi}) = f(y_t|z_t) \sum_{z_{t+1}, d_{t+1}} P(z_{t+1}, d_{t+1}|z_t, d_t, \vec{\pi}) g(y_{t+1:T}|z_{t+1}, d_{t+1}\vec{\pi})$$

Once posteriors are computed, within a time step t, top-down sampling is used as:

$$d_t|d_{t-1}, z_{t-1}, y_{1:T}, \vec{\pi} \sim \textstyle\sum_{z_t} P(d_t, z_t|d_{t-1}, z_{t-1}, y_{1:T}, \vec{\pi})$$

Variable o_t is only sampled when $d_t = d_{t-1}$. Furthermore, z_t is sampled only when $o_t = 0$ or d_t is different from d_{t-1}, otherwise $z_t = z_{t-1}$.

$$o_t|z_{t-1}, y_{1:T}, \vec{\pi} \sim P(o_t|z_{t-1}, \omega_{z_{t-1}}) \sum_{z_t} P(z_t|o_t, \vec{\pi}) P(y_{t:T}|z_t, d_t, \vec{\pi}) \qquad d_t = d_{t-1}$$

$$z_t|d_t, o_t = 0, y_{1:T}, \vec{\pi} \sim P(z_t|d_t) P(y_{t:T}|z_t, d_t, \vec{\pi})$$

1.6 Experiments and Results

We demonstrate the utility of TSTM on physiologic heart rate (HR) and respiratory rate (RR) signals collected from 145 premature infants from our Stanford NICU dataset. Due to prematurity, these infants are extremely vulnerable, and complications during their stay in the NICU can adversely affect long term development. Our broader aim is to identify markers associated with and predictive of downstream health status.

Clinicians and alert systems implemented on ICU monitors utilize coarse information such as signal means (e.g., is the HR > 160 beats per minute) and discard the remaining signal. We use TSTM to infer whether there is information contained in the signal dynamics. Transient events can manifest in the HR or the RR signal independently (e.g., bradycardia is observed in the HR signal while apnea is primarily observed in the RR signal). Thus, for our experiments below, we run TSTM on each signal independently; simultaneous processing of both signals is also easily possible with TSTM using the vector autoregressive process observation model as described in section 1.3.2.

Roadmap: We first evaluate the goodness of fit of TSTM on each physiologic signal. We also evaluate the utility of TSTM for feature construction on a supervised learning task of *grade assignment*. We then perform a qualitative analysis of the learned words, topics and inferred infant-specific distributions for clinical relevance. Finally, we show an experimental comparison between TSTM and BP-AR-HMM.

1.6.1 Experimental Setup

For all our experiments, we preprocess the physiologic signals to remove a 40 minute moving average window; this allows us to capture characteristics only related to the dynamics of the signal (resulting HR signal shown in figure 1.1). For TSTM, we fix the number of topics, $D = 4$. Although this choice is flexible, for our dataset, we chose this based on clinical bias. We identify four clinically meaningful topics: *Lung* for primarily lung related complications; *Head* for head related (neurological) complications; *Multi* as the catch-all class for severe complications that often affect multiple organ systems; and *Healthy*. We set the truncation level L to 15. We experimented with different settings of the hyperparameters for TSTM. Of particular interest is the choice of κ and ρ which control word and topic length and can,

as a result, force the words to be longer or shorter.[3]For the reported experiments, α_l, γ and η were each set to 10, $\kappa = 25$ and $\rho = 20$. Similar to [Fox *et al.*, 2009], we specify the priors on the observation model parameters A and Σ as a Matrix Normal Inverse-Wishart of $\mathcal{N}(0, 10 * I_p)$ and $IW(S_0, 3)$; S_0 is set to 0.75 times the empirical covariance of the first difference observations and p is the order of the autoregressive process. For all experiments with an AR observation model, we use this prior.

1.6.2 Quantitative Evaluation

Goodness-of-Fit

To evaluate the benefit of explicitly modeling heterogeneity between the time series, we compare the goodness-of-fit of TSTM with a switching-AR model on held out test data. Specifically, we compare with auto-regressive hidden Markov models (AR-HMM). AR-HMMs are HMMs where the observation model used are autoregressive processes [Poritz, 2009]. AR-HMMs, like other switching Markov models we discussed previously, assume that the data is generated i.i.d. from a single class model. Thus, an improved goodness-of-fit with TSTM will illustrate the benefit of modeling series-specific variability.

For the choice of the observation model, we experiment with both first (AR(1)-HMM) and second order (AR(2)-HMM) autoregressive processes. In figure 1.4, we illustrate the protocol for this experiment. As shown in figure 1.4a, to generate the test set on which we evaluate the fit, for each series, we hold out a sample of 4-hour blocks comprising 20% of the series. We keep the remaining as training data.

To evaluate the test log-likelihood, we average over three separate Gibbs chain for each model (see figure 1.4b). First, we run each chain for 2000 iterations on the training data. Chains for both models appear to mix by the $2000th$ iteration. Each chain is initialized by sampling model parameters (words, topics and topic-transition matrices for TSTM, and words and transition matrix for AR-HMM) from the prior. The training is done in an unsupervised way; the number of topics is initialized as $D = 4$ but no supervision is given regarding which infant contains which topics.

[3]We tested a few other settings of these hyperparameters. We qualitatively evaluated the word histograms (e.g., shown in figure 1.7a) derived from the series segmentations for whether the infants clustered based on their illness severity i.e. whether healthy infants have similar word profiles and profiles that are different from the unhealthy infants. This separation held consistently for our settings suggesting that with regard to discovery, the results are not sensitive to the specific choice of parameters. Since our primary goal is discovery, we ran all our experiments with only one setting of these hyperparameters.

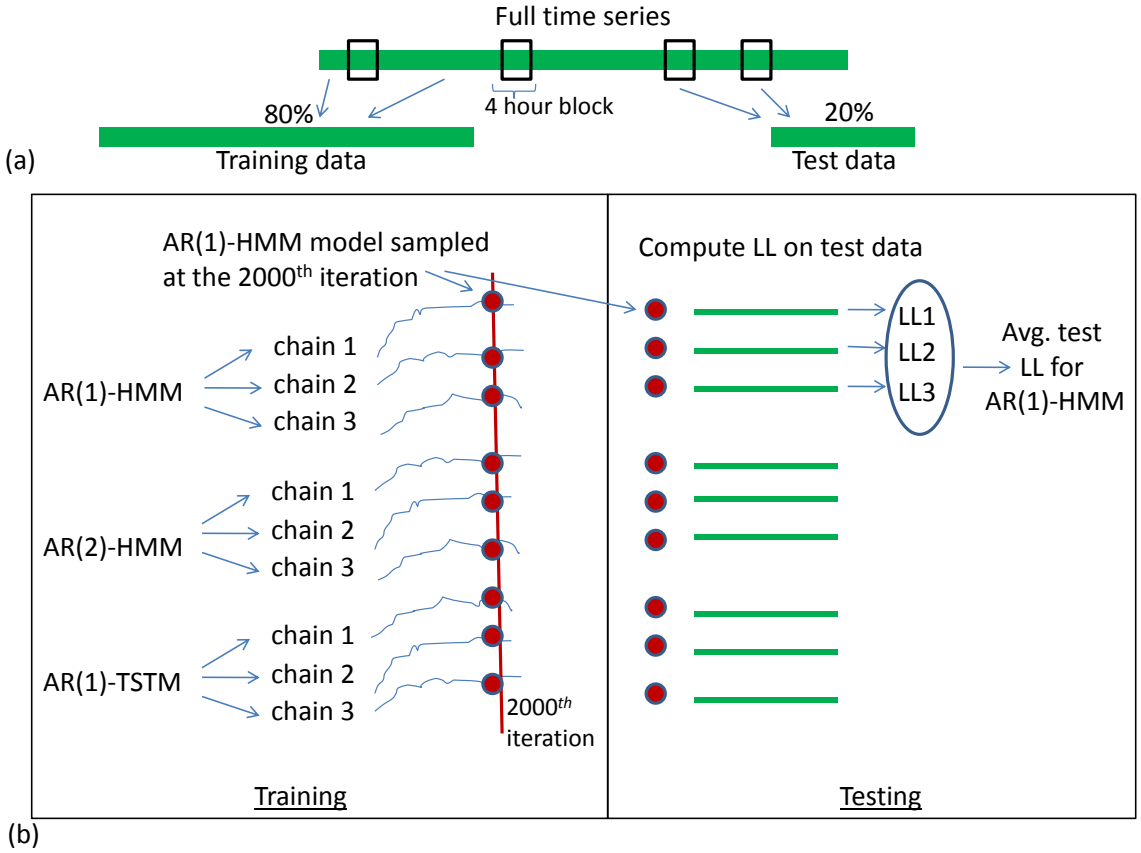

Figure 1.4: Experimental protocol for the evaluation of goodness-of-fit, a) the procedure for splitting each series into the train and test set, b) the pipeline for evaluating goodness of fit on the data.

We choose the model parameters sampled at the 2000th iteration as the model for which we evaluate the goodness-of-fit.[4] Each AR-HMM chain was initialized to have the same number of AR features as that inferred by the corresponding TSTM chain at the 2000th iteration. Test log-likelihood is computed with the forward-backward algorithm on the test sequences with model parameters fixed from the 2000th iteration of each Gibbs chain on the training data. Test log-likelihoods, averaged over 3 chains, for TSTM with an AR(1) observation model and an AR(1)-HMM are -1.425e+5 and -2.512e+5 respectively

[4]The 2000th iteration for each chain was chosen arbitrarily. Alternately, one may also choose more than one iteration and average over the iterations. This would require running inference on the test data with the model at each of these different iterations.

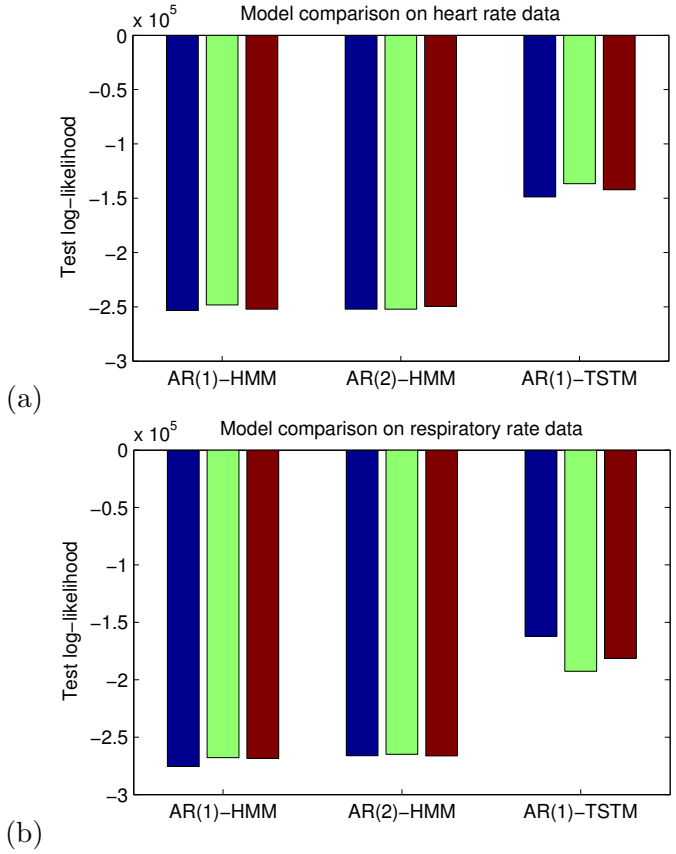

Figure 1.5: Test log-likelihood from three separate Gibbs chains for the AR(1)-HMM, AR(2)-HMM, and TSTM with an AR(1) observation model evaluated on a) the heart rate data (top), b) the respiratory rate data (bottom).

for the HR signal, and -1.787e+5 and -2.706e+5 for the RR signal. These results are also shown in figure 1.5a and figure 1.5b respectively. The significant gains in test log-likelihood using TSTM suggest that explicitly modeling heterogeneity between series is beneficial and that the topics and words generalize well to held-out data. In figure 1.5, we also see that the AR(2) observation model, albeit a more complex model than AR(1), does not benefit test log-likelihood. Thus, for all experiments that follow, we use an AR(1) observation model.

Feature Derivation

Deriving features is a common task one needs to tackle when using high-dimensional data (such as the physiologic signals) for supervised machine learning problems. We evaluate the usefulness of features obtained from the physiologic signals for the task of grade assignment. Grades $G_{1:N}$, representing an infant's health, are assigned to each infant based on his final outcome, as identified retrospectively by a clinician. Grade 0 is assigned to infants with no complications; grade 1 to isolated minor complications frequently associated with prematurity; grade 2 to multiple minor complications; and grades $3-5$ to major complications from low to severe grades. Features derived from any given model are used for predicting an infant's disease grade by combining these features with a rank support vector machine [Joachims, 2006]. The rank score for a ranking H is:

$$\text{rankscore}_H = \Sigma_{n=1}^{N}\Sigma_{m=1}^{N}\mathcal{I}(H(n) > H(m))(G_n - G_m)$$

We compute features from TSTM as follows. We run three Gibbs chains with TSTM (using unsupervised training as described in the goodness-of-fit experiment above) on the full data set. The features for each infant are derived as the frequency of each topic at the 2000th iteration of a Gibbs chain normalized by the length of the data sequence.

To report ranking test accuracy, for the set of 145 infants, we generate 20 random folds with $50-50$ train/test split and average performance over all folds. For each fold, the accuracy is computed as a percentage of the maximum achievable score for that test split. The SVM tradeoff parameter C for each model was set using cross-validation with features generated from the first Gibbs chain on 3 randomly sampled folds. Due to small sample size, we do not experiment with the choice of kernel and use the default choice of a linear kernel for all our experiments.

For comparison with other feature extraction methods from time series data, existing approaches can be divided into two broad classes: techniques in the frequency-domain and the time-domain [Shumway, 1988; Keogh *et al.*, 2000]. Frequency analysis using the discrete fourier transform is one of the most commonly used techniques for time series data analysis [Keogh *et al.*, 2000]. The frequencies of the resulting FFT coefficients span $1/v$ for $v \in \{1, \cdots, T\}$, which results in a large feature set. Traditionally, the large feature set size is not a concern in the presence of enough data. However, in our application, as is in most clinical applications, labeled data is often scarce. We experiment with using the raw features

within the rank SVM. Based on preliminary data analysis, we also compute transformed features by summing coefficients corresponding to time periods in increments of 4 minutes. This non-linear binning of features dramatically improves performance for HR data from near random to 63.5%. In the time domain, we compare with features derived from the AR(1)-HMM model; for this, we run three Gibbs chains on the full dataset (as described above). Features for each infant series are computed as the normalized proportion of words at the 2000th iteration of a Gibbs chain. Performance is reported as the average over all three chains. To get an assessment of the information contained in the dynamics compared to the signal mean (a simple measure usually used in standard care), we also grade infants based on deviation of their signal mean from the normal mean (normal specified as 125 beats/min for HR and 55 breaths/min for RR). We call this approach the clinical norm.

In table 1, we report results for all four methods. TSTM features yield higher performance for both the heart rate and respiratory rate signals compared to those derived from FFTs, AR-HMMs or the clinical norm (although statistical significance is not reached for difference between AR-HMM and TSTM performance). This suggests that the inferred topic proportions provide a useful feature representation scheme that can be employed within supervised objectives as an alternative or in addition to these existing methods.

Model	Heart Rate	Resp. Rate
TSTM	**74.45**%	**75.48**%
FFTs	63.5%	67.69%
AR-HMMs	71.29%	72.68%
Clinical norm	61.37%	68.93%

Table 1.2: Evaluating features from unsupervised training of TSTM.

Comparison to BP-AR-HMM

In figure 1.6a, we show an example run of the BP-AR-HMM model on the heart rate signal for a randomly selected set of 30 infants. For comparison, inference using TSTM on data from the same infants is shown in figure 1.7a. For the BP-AR-HMM run, we used a prior of Gamma$(1, 1)$ on α, the hyperparamter that controls the number of new features generated and Gamma$(100, 1)$ on κ, the self-transition parameter. The gamma proposals used $\sigma_\gamma^2 = 1$ and $\sigma_\kappa^2 = 200$. We refer the reader to [Fox *et al.*, 2009] for the definitions of these parameters.

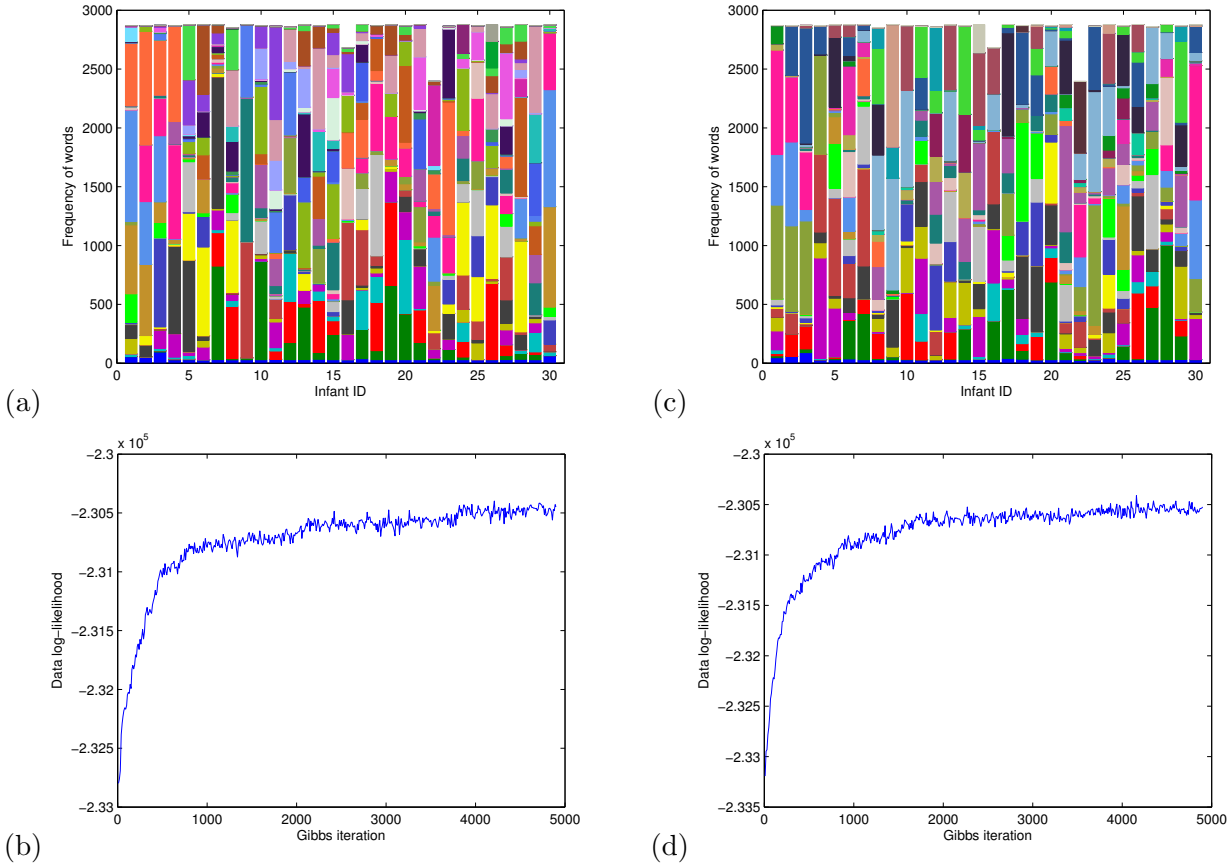

Figure 1.6: a) & c) Inferred word distributions from the heart rate signal for 30 infants during their first four days at the NICU with the BP-AR-HMM for two different initializations (initialization setting described in the text); distinct colors correspond to distinct words, b)& d) Corresponding data log-likelihood of the Gibbs chain for the first 5000 iterations.

For the observation model parameter priors, we use the same specification as that for the TSTM and AR-HMM experiments. At initialization, each series was segmented into five contiguous blocks, with feature labels unique to that sequence. BP-AR-HMM is sampling in the space of binary matrices; as a result, the birth-death sampler in [Fox *et al.*, 2009] takes much longer to mix compared to our block-Gibbs sampler. Due to the computational requirements of BP-AR-HMM, each chain is run on only the first four days of data; one such run takes approximately 34 hours. In the figure 1.6a, we show results from the 5000th iteration of a Gibbs chain. Each unique color corresponds to a distinct AR-feature (word) and the bar graph displays the distribution of words for each series.

In comparing the inferred word distributions from the BP-AR-HMM (figure 1.6a) with that from TSTM (figure 1.7a), we see that the inferred word distributions from the BP-AR-HMM are not as clearly amenable to clinical interpretation. Specifically, most series have features that are not shared with other series. To investigate whether the fragmentation was due to lack of mixing, in figure 1.6b, we plot the data log-likelihood as $\log \prod_i \sum_{Z^i} P(Y^i, Z^i | \pi^i, \vec{\theta})$; the chain appears to have mixed by the 5000th iteration. To encourage more sharing across the time series with the BP-AR-HMM, we ran several different Gibbs chains with different variants of the hyperparameters but the results were visually similar or even more fragmented. We show another example run in figure 1.6c where we use the prior for κ as Gamma$(200, 1)$ and at initialization, segment each series into only two blocks instead of five.[5]

This behavior is not entirely surprising, since the notion of individual series variation in BP-AR-HMM is quite different from that in TSTM: TSTM encourages sharing by having all series utilize the same set of topics, but to different extents; by contrast, BP-AR-HMMs uses the Beta prior for generation of series which posits that each series samples some features that are shared and others that are explicitly series-specific. The abundance of unique features makes comparison between series based on these features difficult.

In order to perform a quantitative comparison with the BP-AR-HMM features, since we only have a small set of 30 samples and therefore, grading does not make sense, we train an SVM classifier using leave-one-out cross-validation, to distinguish *Healthy* vs *Not healthy*, using the labels shown at the bottom of figure 1.7a. For the BP-AR-HMM, we compute frequencies of words extracted from the final iteration of the Gibbs chain shown in figure 1.6a as features. For TSTM, we run a Gibbs chain on the data for 30 infants without any supervision. The topic proportions at the 2000th iteration are used as features within the SVM. The inferred 4-topic proportions from TSTM yields an accuracy of 80% for HR and 60% for RR data. In contrast, BP-AR-HMM word proportions used as features yields a lower performance of 70% and 53%. Thus, the fragmentation of the data across multiple individual features hurts BP-AR-HMM's performance.

[5]The colors in these figures are generated randomly for each run so the colors are not comparable between figure 1.6a and figure 1.6c.

1.6.3 Qualitative Evaluation

We now analyze the words and topics inferred from TSTM in more detail. We focus on the model inferred from the heart rate signal.

Partially-supervised training: We experiment with the partially-supervised training regime of labeled LDA [Ramage *et al.*, 2009], which has the advantage of biasing the topics into categories that are coherent and more easily interpreted. During training, we constrain infant-specific transition matrices to *not* have topics corresponding to complications that they did not show symptoms for. This type of negative evidence imposes minimal bias, particularly relevant in clinical tasks, because of the uncertainty associated with the diagnosis of the onset and severity of the complication. For each infant in a randomly chosen subset of 30 infants, we assign a vector λ_n of length D, where we have a 0 at index i when this infant is known not to have complications related to the ith category. All infants are marked to allow having the healthy topic, representing the assumption that there may be some fraction of their time in the NICU during which they have recovered and are healthy. Each row of the infant-specific transition matrix is generated as:

$$\pi_n^i \sim \text{Dir}\left(\alpha_l \frac{\pi_g^i \otimes \lambda_n}{< \pi_g^i, \lambda_n >}\right) \qquad \lambda_n(i) = 1 \tag{1.10}$$

where $\otimes$ denotes the element-wise vector product. Under this regime, we run a Gibbs chain (G1) for the 30 infants. Next, we fix the words and topic distributions $\phi_{1:D}$ to that of the $2000th$ Gibbs iteration (as discussed in previous experiments) and run inference on our entire set of 145 infants. Here, no supervision is given; that is, both π_g and π_n are initialized from the prior and are left unconstrained during the inference process (using block Gibbs). We run a Gibbs chain to 400 iterations. Given the words and topics, the block-Gibbs sampler appears to mix within 200 iterations.

Qualitative analysis: In figure 1.7a, we show 30 randomly selected infants from this test set at the $400th$ iteration from chain 1. These infants are not the same as the infants used in the training set. In panel 3(a), we plot the word distribution for days 1,2 (top) and days 7,8 (bottom). Infants with no complications are shown as red squares at the bottom of this panel. In panel 3(b), we plot the degree to which a word is associated with each of the four topics.

First, we examine the inferred topic posteriors to track the clinical evolution of three

sample infants 2, 16 and 23 chosen to be illustrative of different trajectories of the word distributions over time. In figure 1.7c, the bold line shows the smoothed posterior over the infant being healthy over time. To compute this posterior, for each of the three infants we run 30 test Gibbs chains with words and topics fixed from the final iteration of G1 (described above). For each infant, at time t within its sequence, we compute h_t as the proportion of times latent state d_t = Healthy from the final iteration of all 30 chains. We smooth h_t using an 8 hour window around t. Thus, a posterior value of 1 at t implies the infant only expressed words associated with the Healthy topic within an 8 hour window of t.

Infant 2 (I2) was born with a heart defect (small VSD). I2's heart condition was treated on day 4. On day 7, her state started to resolve significantly, and on day 8 her ventilator settings were minimal and she was taken off the drug. Her empirical evolution closely tracks her medical history; in particular, her state continually improves after day 4. Infant 16 was a healthier preemie with few complications of prematurity and was discharged on day 4. Infant 23, on the other hand, got progressively sicker and eventually died on day 4. The figure shows that their inferred posterior prediction closely tracks their medical history as well.

Next, we analyze the words and word histograms. Loosely interpreting, words with AR parameter $a > 1$ represent heart rate accelerations (e.g., word 8 shown in gray), words where a is positive and close to 0 represent periods with significantly lower dynamic range (e.g., word 2 shown in purple) and words with large V represent higher dynamic range or high entropy. The word frequencies vary greatly across infants. Respiratory distress (RDS), a common complication of prematurity, usually resolves within the first few days as the infant stabilizes and is transitioned to room air. This is reflected by the decrease in relative proportion of word 2, only associated with the Lung topic (as seen in figure 1.7b). Exceptions to this are infants 3 and 30, both of whom have chronic lung problems. Overall, the inferred word histograms highlights separability between healthy and other infants based on the word mixing proportions, suggesting different dynamics profiles for these two populations. Words 3, 9 and 10, associated primarily with the healthy topic, occur more frequently in infants with no complications. These three words also have the highest V^k values suggesting entropy as a signature for health in neonates. Thus, we developed a new risk stratification score [Saria *et al.*, 2010], that predicts based on data from the first three hours of life, infants at risk for major complications. We describe this score in detail in

Chapter 5.

1.7 Discussion and Future work

The primary contribution of this chapter is a new class of models for time-series data that emphasizes the modeling of instance specific variability while discovering population level characteristics. Unlike BP-AR-HMM, its closest counterpart, TSTM has a mechanism for modeling high-level topics that hierarchically capture structure in collections of words. For a knowledge discovery task, modeling higher-level structure has several advantages. First, it can help discover novel semantic structure in the data such as the degree to which two topics (e.g., diseases) share common words (e.g., physiologic traits). It also gives the used finer control over the types of features extracted from the data; for example, by using TSTM within a partially supervised setting, emphasis can be placed on discovering features that identify specific disease pairs.

We demonstrate the use of TSTM in a novel and useful application of modeling heterogeneous patient populations over time. We believe that TSTM provides a significant departure from current practices and a flexible tool for exploratory time series data analysis in novel domains. Furthermore, learned topic or word distributions can serve as features within supervised tasks. We demonstrated the utility of TSTMs on medical time series, but the framework is broadly applicable to other time-series applications such as financial or human-activity data [Liao *et al.*, 2007].

Several extensions of TSTM could yield additional domain insight. In particular, modeling individual topic distributions as evolving over time (analogous to [Wang *et al.*, 2008]) should highlight how the characteristics of the expressed temporal signatures vary as diseases evolve over longer periods of time. Modeling the data as the composition of repeating signatures expressed at varying temporal granularity (seconds versus minutes versus hours) would highlight the granularity at which diseases alter measured physiology. We leave these next steps for future work. In the next chapter, we extend the notion of words from dynamics related signtures to richer shape related signatures.

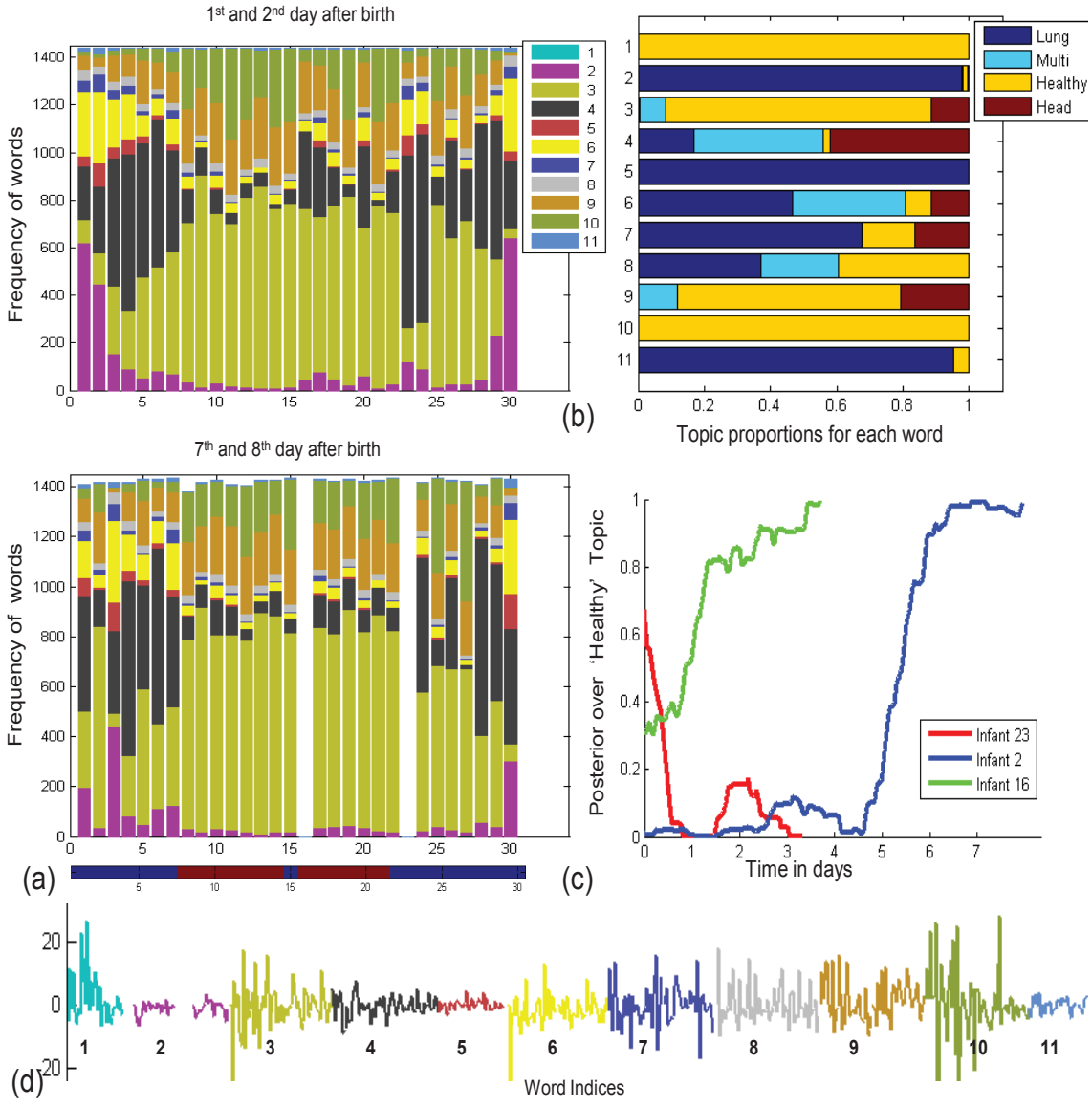

Figure 1.7: (a) Inferred word distributions for the heart rate data for 30 infants during their stay at the NICU. At the bottom of the word panel, infants marked with red squares have no complications, (b) distribution over disease topic given words for the population, (c) posterior over latent state, *Healthy*, (d) examples of inferred features extracted from the data.

Bibliography

[Bar-Shalom and Fortmann, 1987] Y. Bar-Shalom and T.E. Fortmann. *Tracking and Data Association*. Academic Press Professional, Inc., 1987.

[Blackwell and MacQueen, 1973] D. Blackwell and J.B. MacQueen. Ferguson distributions via polya urn schemes. In *Annals of Statistics, 1(2)*, 1973.

[Blei *et al.*, 2003] David Blei, Andrew Ng, and Michael Jordan. Latent Dirichlet allocation. In *Journal of Machine Learning Research*, 2003.

[Ferguson, 1973] T.S. Ferguson. A bayesian analysis of some nonparametric problems. The Annals of Statistics, 1(2), 1973.

[Fine *et al.*, 1x998] S. Fine, Y. Singer, and N. Tishby. The Hierarchical Hidden Markov Model: Analysis and applications. In *Journal of Machine Learning (JMLR)*, 1x998.

[Fox *et al.*, 2007] Emily Fox, Erik Sudderth, Michael Jordan, and Alan Willsky. The sticky HDP-HMM: Bayesian nonparametric hidden Markov models with persistent states. Technical Report P-2777, MIT LIDS, 2007.

[Fox *et al.*, 2008] Emily Fox, Erik Sudderth, Michael Jordan, and Alan Willsky. Nonparametric Bayesian learning of switching linear dynamical systems. In *Neural Information Processing Systems (NIPS)*, 2008.

[Fox *et al.*, 2009] E. Fox, E. Sudderth, M. Jordan, and A. Willsky. Sharing features among dynamical systems with Beta Processes. In *Neural Information Processing Systems (NIPS)*, 2009.

[Gelman *et al.*, 1995] A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian data analysis*. Chapman and Hall, 1995.

[Ishwaran and Zarepour, 2000] H. Ishwaran and M. Zarepour. Markov chain monte carlo in approximate dirichlet and beta twoparameter process hierarchical models. In *Biometrika, 87*, 2000.

[Ishwaran and Zarepour, 2002a] H. Ishwaran and M. Zarepour. Dirichlet prior sieves in nite normal mixtures. In *Statistica Sinica 12*, 2002.

[Ishwaran and Zarepour, 2002b] H. Ishwaran and M. Zarepour. Exact and approximate sum-representation for the Dirichlet process. In *Canadian Journal of Statistics*, 2002.

[Joachims, 2006] T. Joachims. Training linear SVMs in linear time. In *Knowledge Discovery and Datamining (KDD)*, 2006.

[Kass and Steffey, 1989] R. Kass and D. Steffey. Approximate bayesian inference in conditionally independent hierarchical models (parametric empirical bayes models). In *Journal of American Statistical Association*, 1989.

[Keogh *et al.*, 2000] E. Keogh, K. Chakrabarti, M. Pazzani, and S. Mehrotra. Dimensionality reduction for fast similarity search in large time series databases. In *Journal of Knowledge and Information Systems*, 2000.

[Kurihara *et al.*, 2007] K. Kurihara, M. Welling, and Y.W. Teh. Collapsed variational Dirichlet process mixture models. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2007.

[Lee *et al.*, 2009] Honglak Lee, Yan Largman, Peter Pham, and Andrew Ng. Unsupervised feature learning for audio classification using convolutional deep belief networks. In *Neural Information Processing Systems (NIPS)*, 2009.

[Liao *et al.*, 2007] L. Liao, D.J. Paterson, D. Fox, and H.A. Kautz. Learning and inferring transportation routines. International Joint Conference on Artificial Intelligence (IJCAI), 2007.

[Mueen *et al.*, 2009] A. Mueen, E. Keogh, Q. Zhu, S. Cash, and B. Westover. Exact discovery of time series motifs. In *Siam Conference on Data Mining*, 2009.

[Poritz, 2009] A.B. Poritz. Linear predictive hidden markov models and the speech signal. In *Symposium on Applications of Hidden Markov Models to Test and Speech*, 2009.

[Ramage *et al.*, 2009] D. Ramage, D. Hall, Nallapati R, and C. Manning. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2009.

[Saria *et al.*, 2010] S. Saria, A. Rajani, J. Gould, D. Koller, and A. Penn. Integration of early physiological responses predicts later illness severity in preterm infants. In *Science Trans. Med.*, 2010.

[Sethuraman, 1994] J. Sethuraman. A constructive definition of Dirichlet priors. In *Statistics Sinica*, 1994.

[Shumway, 1988] R. Shumway. *Applied statistical time series analysis.* Prentice Hall, 1988.

[Teh *et al.*, 2006] Y.W. Teh, M.I. Jordan, M. Beal, and D. Blei. Hierarchical Dirichlet processes. In *Journal of American Statistical Association*, 2006.

[Wang and McCallum, 2006] X. Wang and A. McCallum. Topics over Time: A non-Markov continuous time model of topical trends. In *Knowledge Discovery and Datamining (KDD)*, 2006.

[Wang *et al.*, 2008] C. Wang, D. Blei, and D. Heckerman. Continuous time dynamic topic models. In *Uncertainty in Artificial Intelligence (UAI)*. 2008.

[Williams *et al.*, 2005] C. Williams, J. Quinn, and N. McIntosh. Factorial switching Kalman filters for condition monitoring in neonatal intensive care. In *Neural Information Processing Systems (NIPS)*, 2005.