Learning individual and population level traits from clinical temporal data

Suchi Saria[1], Daphne Koller[1] and Anna Penn[2]
[1]Department of Computer Science, [2]Department of Pediatrics
Stanford University
Stanford, CA 94305
ssaria@cs.stanford.edu

Abstract

This paper proposes a nonparametric Bayesian method for exploratory data analysis and feature construction in continuous time series such as longitudinal health data. Our method focuses on understanding shared characteristics in a set of time series that exhibit significant individual variability. Each series is characterized as switching between latent states ("topics"), where each topic is characterized as a distribution over generating functions ("words") that specify the series dynamics. Individual series maintain series-specific topic mixing proportions. The words are modeled as lying in an infinite dimensional space and the hierarchical Dirichlet Process prior allows selection of words that are shared across topics given data. Word and topic descriptions are shared across the entire population. We apply this model to the task of tracking the physiological signals of premature infants; our model obtains clinically significant insights as well as useful features for supervised learning tasks. Furthermore, based on these insights, we developed *Physiscore*, a personalized risk stratification score for preemies. Physiscore performs significantly better than APGAR, the current standard of care.

1 Introduction

To understand disease pathogenesis, longitudinal studies collect data by tracking the same set of patients over time. The ubiquity of small wireless health devices has heralded a new era in disease monitoring via the ability to continuously capture patient physiology. In addition, due to the recent push for the digitization of electronic health records (EHR), which aggregate such clinical data in one place, longitudinal studies are becoming increasingly feasible. Dynamic models that can succinctly capture generation of measured data as a function of underlying disease state, can serve as a lens into understanding how the set of diseases manifest and discover novel disease associations. This also provides an alternative method for gaining richer feature sets for clinical studies, majority of which focus only on feature selection and rely on clinical expertise for feature extraction (e.g., [19, 12]).

The set of clinical diagnoses for most diseases is not naturally discrete; patients suffer from diseases to varying extents, and even within disease labels, patients exhibit significant variability. Furthermore, often the diagnosis is the result of a clinical observation with significant uncertainty regarding both the onset and the nature of the disease. Traditional generative models for time series data (such as switching Kalman filters[1]) do not *explicitly* model generation of a continuum of heterogeneous exemplar series. Instead, the data is assumed to be generated from a discrete set of transition matrices, each specifying the generation of a homogeneous population of i.i.d. time series. Thus, for example, should a patient with benign symptoms of intraventricular hemorrhage (IVH, bleeding in the brain) be modeled as belonging to the healthy, IVH class (which can include patients with hemorrhages in multiple brain regions) or a class of its own? Increasing representation granularity by increasing the number of classes can help, but ad hoc discretization into a fixed set limits our ability to model instance-specific variability. Moreover, the combinatorics quickly get out of hand when one considers various combinations of diseases, a situation that is unfortunately common in practice.

Hierarchical Bayesian modeling [10, 6] 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 Allocation (LDA) [2, 7] 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. Similarly, in applications such as disease modeling, an individual patient maintains its own distribution over both latent (disease) topics and transitions between them. Each topic defines a distribution over dynamic behaviors (physiologic symptoms) observed in the time series and these behaviors play the role of words. However, unlike text data, in continuous 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. We propose a more flexible representation of a word which specifies instantiations to parametric functions that generate the temporal dynamics for the duration of the word. Moreover, the duration of the word also does not need to be fixed in advance, and may vary from one occurrence to another. Hence, our parameterization also postulates word boundaries.

In our approach, words lie in the infinite dimensional latent space specified by possible real-valued instantiations to the parametric functions that generate the data. Naive sampling in this infinite-dimensional space given the data will result in no sharing of words across topics [23]. For knowledge discovery tasks, sharing of words across topics is particularly desirable as it allows us to uncover relationships between different latent states. To enable this, we utilize *hierarchical Dirichlet processes* (HDPs), 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 evolution variability.

In the remainder of this paper, we define our generative *time series topic model* (TSTM), and provide an efficient block Gibbs sampling scheme for performing full Bayesian inference over the model. We then present results on our target application of analyzing physiological time series data, demonstrating its usefulness both for analyzing the behavior of different time series, and for constructing features that are subsequently useful in a supervised learning task. Finally, we briefly discuss a personalized risk stratification score[20] derived based on insights from this model.

2 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 [1], where one or more discrete state variables determine the momentary system dynamics, which can be either linear or in a richer parametric class. However, these methods typically utilize a single model 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, as required in our data.

Recent work by Fox and colleagues [5, 3, 4] uses nonparametric Bayesian models for capturing generation of continuous-valued time series. Conceptually, the present work is most closely related to BP-AR-HMMs [5], which use Beta processes to share observation models, characterized by autoregressive processes (AR), across series. Thus, it captures variability between series by sampling subsets of low-level features that are specific to individual series. However, BP-AR-HMMs are aimed at capturing individual variation that manifests as words (features) that are unique to a particular time series. By comparison, we aim to capture higher-level concepts that occur broadly across a subpopulation (e.g., physiologic characteristics exhibited by *Lung disorders*) while modeling series-specific variability in terms of the extent to which these higher-level concepts are expressed (e.g., transient respiratory disorder versus long term respiratory distress). Modeling higher-level structure, can help identify structure such as the degree to which different topics share common low-level traits. It also gives the user finer control over the types of features extracted; for example, by using TSTM within a partially supervised setting, emphasis can be placed on discovering features that identify specific disease pairs. These aspects make TSTM more suitable as a model for knowledge discovery. Other works [3, 4] have utilized hierarchical Dirichlet process priors for inferring the number of features in hidden Markov models and switching linear dynamical systems, but these models do not try to represent variability across exemplar series. Temporal extensions of LDA [24, 25] 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 meaningful 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 [14]. However, this class of methods, as well as others [15], 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 aim to discover higher level structure where words are associated with different "topics" to different extents, which helps discover the underlying relationship between different disease topics.

Our specific domain of neonatal monitoring has been previously studied in a machine learning setting [26, 17], but focusing on the different task of monitoring for known events (e.g., sensor drops).

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 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, then sample from that topics' distribution over words.

Unlike text, in time series data often the features to be extracted are structurally not obvious (see figure 2). Pre-segmenting the sequence data into words does not offer sufficient flexibility to learn from the data, especially for knowledge discovery. Thus, we integrate feature discovery into our model. We describe below each of the TSTM components in a bottom-up fashion.

Data generation model: We assume that the continuous-valued data y_t at time t is generated using a function f_k from a finite set $\mathcal{F}$ of parametric functions. Note, this set is discovered from the data. These functions take as inputs x_t, values dependent on current and previous time slices, and generate the output as $y_t = f_k(x_t; \theta_k)$. f_k, an expressive characterization of the time series dynamics, can be thought of as th k*th* word in the time-series corpus vocabulary. From here on we call these *generating functions*, and use these interchangeably with words. The parameterization of f_k depends on the choice of the observation model.

We choose to use vector autoregressive processes, which are used for temporal modeling in numerous domains, including medical time series of fMRI and EEG. However, our framework is flexible and other observation models (such as an SLDS with mixture model emissions [3]) can also be used. In an order p autoregressive process, given a function f_k with parameters $\{A^k, V^k\}$, the observation is generated as: $\vec{y}_t = A^k X_t^T + \vec{v}_t$ where $v_t \sim \mathcal{N}(0, V^k)$ 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 k*th* word then corresponds to a specific instantiation of the generating 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 $1/\omega_k$. Our goal now is to uncover this set of functions $\mathcal{F}$ via instantiations of the generating function parameters, denoted more generally as $\vec{\theta}_k \in \Theta$.

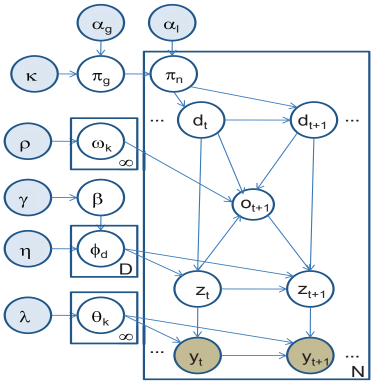

Figure 1: Graphical representation of the time series topic model

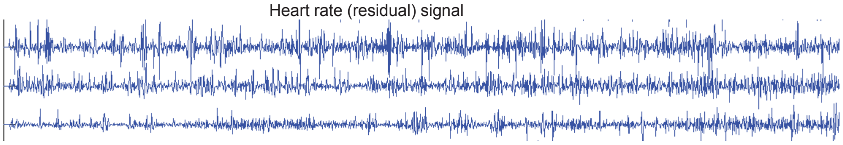

Figure 2: Heart signal (mean removed) from three infants in their first few hours of life

Word and Topic descriptions: To uncover the finite generating function set $\mathcal{F}$, such that these functions are shared across latent topics, we use the hierarchical Dirichlet process[23]. A Dirichlet process (DP), denoted by DP(γ, H), yields a distribution on discrete measures. H is a base (continuous or discrete) probability measure on Θ and γ is the concentration parameter. Sethuraman [21] shows that $G_0 \sim$ DP(γ, H), a sample drawn from the DP, is a discrete distribution because, with probability one:

$$G_0 = \Sigma_{k=1}^{\infty} \beta_k \delta_{\theta_k} \qquad \theta_k \sim H \qquad \beta_k = \beta'_k \prod_{l=1}^{k-1} (1 - \beta'_l) \qquad \beta'_k \sim \text{Beta}(1, \gamma) \tag{1}$$

Draws from H yield the location of "sticks", δ_{θ_k}, in the discrete distribution. The generation of the stick weights β_k is denoted by $\vec{\beta} \sim$ GEM(γ). Thus, we use the DP to place priors on the mixture of generating functions by associating a generating function with each stick in G_0. In addition, by associating each data sample (time point in a series) with a specific generating function via indicator variables, the posterior distribution yields a probability distribution on different partitions of the data. The mixing proportion (the stick weights) in the posterior distribution are obtained from aggregating corresponding weights from the prior and the assigned data samples.

HDPs [23] extend the DP to enable sharing of generating functions between topics. If discrete measures G_j are sampled with the discrete measure G_0 as its base measure, the resulting distributions have non-zero probability of regenerating the same sticks, thereby sharing generating functions between related topics. Thus, a draw G_d from the HDP with G_0 as its base measure, $G_d \sim$ DP(η, G_o), can be described as:

$$G_d = \Sigma_{k=1}^{\infty} \phi_{dk} \delta_{\theta_k} \qquad \phi_d \sim DP(\eta, \beta), \quad \beta \sim GEM(\gamma), \quad \theta_k \sim H \tag{2}$$

where ϕ_d represents the topic specific mixing proportions over the generating functions and β represents the global mixing proportion. Similar to [5], we use a matrix-normal inverse-Wishart on the parameters $\{A^k, V^k\}$ and a symmetric Beta prior on ω_k as our base measures H.

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

1. Current latent topic state given topic at previous time-step, $d_t \sim$ Mult$(\pi_n^{d_{t-1}})$
2. 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, $o_t \sim I(d_t = d_{t-1})$Bernoulli$(\omega_{z_{t-1}})$, where I is the indicator function.
3. The identity of the generating function (word) to be applied; if $o_t = 1$, we have $z_t = z_{t-1}$, otherwise $z_t \sim$ Mult(ϕ_{d_t}).
4. Observation given the generating function index z_t as $y_t \sim f_{z_t}(x_{t-1}; \theta_{z_t})$.

The series specific topic transition distribution π_n is generated from the global topic transition distribution π_g. Hyperparameters 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. To generate π_n, each row i is generated from Dir$(\alpha_l \pi_g^i)$, where π_g^i is the *ith* row of the global topic transition distribution. 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.

4 Approximate Inference using block-Gibbs

Several approximate inference algorithms have been developed for mixture modeling using the HDP; see [23, 3, 13] for a discussion and comparison. We use a block-Gibbs sampler that relies on the *degree L weak limit* approximation presented in [8]. This sampler has the advantage of being simple, computationally efficient and shows faster mixing than most alternate sampling schemes [3].

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. To introduce notation briefly, we use $1:T$ to mean all indices 1 through T. n indexes individual 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). We drop the index n when explicit that the variable refers to a single series. We detail the update steps of our block-Gibbs inference algorithm below.

Sampling latent topic descriptions β, ϕ_d: The DP can also be viewed as the infinite limit of the order L mixture model [8, 23]:

$$\beta|\gamma \sim \text{Dirichlet}(\gamma/L, \cdots, \gamma/L) \qquad \phi_d \sim \text{Dirichlet}(\eta\beta) \qquad \theta_k \sim H \tag{3}$$

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:

$$\phi_d|\beta, \eta \sim \text{Dir}(\eta\beta_1, \cdots, \eta\beta_L) \tag{4}$$

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[1] for the dth disease topic; that is, let $m_{d,l} = \sum_{n=1:N} \sum_{t=1:T_n} I(z_{n,t} = l)I(d_{n,t} = d)I(o_{n,t} = 0)$ and $m_{.,l} = \sum_{d=1:D} m_{d,l}$. The posterior distribution for the global and individual topic parameters is:

$$\beta|z, d, \gamma \sim \text{Dir}(\gamma/L + m_{.,1}, \cdots, \gamma/L + m_{.,L}) \tag{5}$$

$$\phi_{d'}|z, d, \eta, \beta \sim \text{Dir}(\eta\beta_1 + m_{d',1}, \cdots, \eta\beta_L + m_{d',L}) \tag{6}$$

Sampling word parameters ω_l and θ_l: Intuitively, 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 $\bar{c}_{l,i} = \sum_{n=1:N} \sum_{t=1:T_n} I(d_{n,t} = d_{n,t-1})I(z_{n,t-1} = l)I(o_{n,t} = i)$ 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,

$$\omega_l|\bar{c}_{l,.}, \rho \sim \text{Beta}(\rho/2 + \bar{c}_{l,1}, \rho/2 + \bar{c}_{l,0}) \tag{7}$$

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 [5].

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 $c^i_{n,k} = \sum_{t=1}^{T_n} I(d_{n,t-1} = i)I(d_{n,t} = k)$ 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:

$$\pi^i_g|d, \alpha_g, \kappa \sim \text{Dir}(\alpha_g + \kappa\delta_i + \vec{c^i}) \qquad \pi^i_n|\pi_g, d, \alpha_l \sim \text{Dir}(\alpha_l\pi^i_g + c^i_{n,1:D}) \tag{8}$$

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 samples for the previous time slice:

$$P(z_{1:T}, d_{1:T}|y_{1:T}, \vec{\pi}) = \prod_t P(z_t, d_t|z_{t-1}, d_{t-1}, y_{1:T}, \vec{\pi}) \tag{9}$$

Top-down sampling is used within a given time slice. (See extended report online for details).

[1] Within this approximation, words get re-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.

5 Experiments and Results

We demonstrate the utility of TSTM on physiologic heart rate (HR) signal collected from premature infants of gestational age $\leq$ 34 weeks and birth weight $\leq$ 2000 grams admitted to the Stanford neonatal ICU within the first few hours of life. Our inclusion criteria resulted in data from 145 infants (IRB approved).These infants are continuously monitored as part of routine care and data aggregated at the minute-level granularity was stored. These infants are extremely vulnerable, and complications during their stay in the NICU can adversely affect long term neurological development. Clinicians and alert systems implemented on ICU monitors utilize coarse information based on thresholding of the signal mean (e.g., HR > 160 bpm) and discard the dynamics as noise.

We use TSTM to discover whether there is information contained in the signal dynamics. Towards this end, we evaluate the usefulness of the features generated from TSTM for the supervised learning task of *grade assignment*. We also analyze the learned words, topics and inferred infant-specific word and topic distributions for clinical relevance and briefly describe a clinical application developed based on insights derived from this analysis.

For all our experiments, we preprocess the physiologic signals to remove the *basal* signal computed from taking a 40 minute moving average window; this allows us to capture characteristics only related to the dynamics of the signal (see raw HR signal in figure 2). We fix the number of topics, $D = 4$. Although this choice is flexible, for our dataset, we select this based on clinical bias. We identify four broad clinically meaningful topics: *Lung* for primarily lung related complications such as RDS; *Hemorrhage* for head related complications such as IVH; *Multi* as the catch-all class for severe complications that often affect multiple organ systems; and *Healthy*. All reported results use a first-order autoregressive process as the observation model. 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 length and can force the words to be longer or shorter. Overall, within a large reasonable range, the results were not sensitive to the choice of these hyperparameters. For the reported experiments, we set α_l, γ and η were each set to 10, $\kappa = 25$ and $\rho = 20$. We run 2000 iterations for each Gibbs chain.

Feature Discovery: We evaluate the utility of TSTM features derived from the residual 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.

Derived features are combined in a supervised regime using a support vector machine with a ranking objective [9] to predict the infant's disease grade. The rank score for a ranking H is:

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

We report results as a percentage of the maximum score achievable on our data. Reported results are averaged over 20 random folds with 50–50 train/test splits. The SVM tradeoff parameter C was set using cross-validation on 3 randomely sampled folds upfront.

For TSTM, we infer the infant specific topic distributions in an unsupervised setting[3]. The topic proportions are used as features for the grading task. 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 [22, 11].

Frequency analysis using the discrete fourier transform is one of the most commonly used techniques for time series data analysis [11]. 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 a grid of time periods in increments of 4 minutes. This non-linear

[3] Due to the size of our corpus, to speed up inference, the topic descriptions were inferred from only 20% of the data and fixed at the 2000*th* iteration. Thereafter, we infer series-specific topic proportions based on these fixed settings of the model parameters.

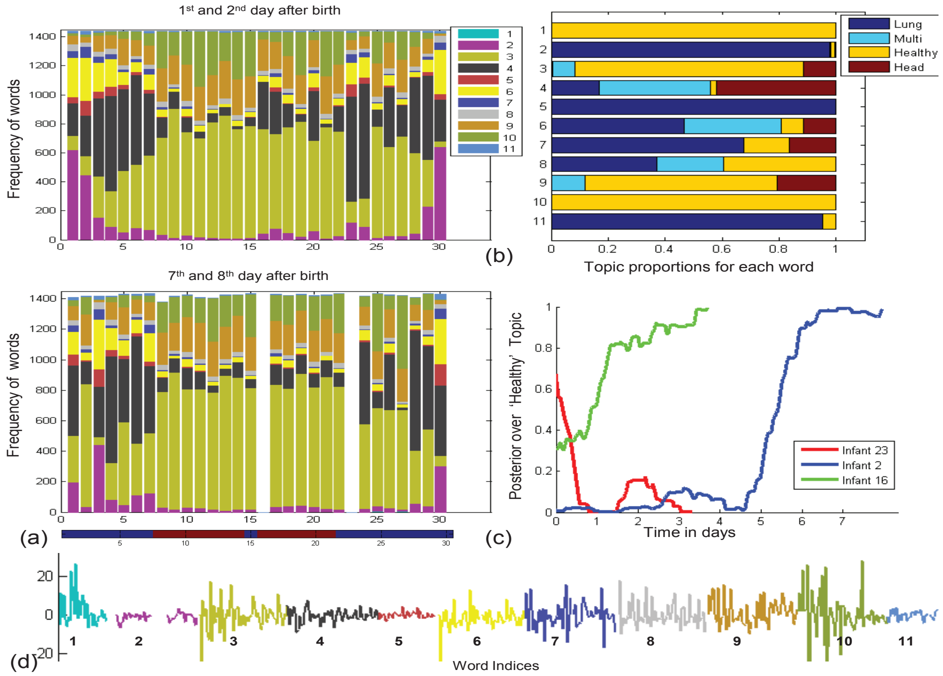

Figure 3: (a) Inferred word distributions for 30 infants during their stay at the NICU. At the bottom of the word panel, infants marked with red squares have no complications, (b) distribution over disease topic given words for the population, (c) posterior over latent state, *Healthy*, (d) examples of inferred features extracted from the data.

binning of features dramatically improves performance for HR data from near random to 63.5%. In the time domain, a range of techniques from local heuristic based piecewise approximations [11] to more global likelihood based compression (e.g., Hidden Markov models), have been developed. We compare with feature distributions from training a hidden Markov model with an Autoregressive Process(1) observation model. We fix the number of features to be the number inferred by TSTM.

Ranking using the TSTM features yields higher performance of 74.43% compared with both FFTs and AR-HMMs which yield 63.5% and 71.38% respectively.

Clinical Evaluation: We analyze the learned words and topics in greater detail. For this, we experiment with the partially-supervised training regime of labeled LDA [18], 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 i*th* category. All infants are marked to have the healthy topic, representing the assumption that there is some fraction of their time in the NICU at which they exhibit signatures of a healthy baby. Each row of the infant-specific transition matrix is generated as:

$$\pi_n^i \sim \text{Dir}\left(\alpha_l \frac{\pi_g^i \otimes \lambda_n}{< \pi_g^i, \lambda_n >}\right) \qquad \lambda_n(i) = 1 \tag{10}$$

where $\otimes$ denotes the element-wise vector product. Next, we fix the topic distributions $\phi_{1:D}$ to that of the 2000*th* Gibbs iteration 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 ran three separate Gibbs chains to 400 iterations. Given the topics, the block-Gibbs sampler mixes within 200 iterations.

In figure 3, we analyze 30 randomly selected infants from this test set at the 400*th* iteration from chain 1. 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. We compute the posterior for the Healthy topic being expressed from averaging over 30 Gibbs chains. In figure 3c, the bold line shows the smoothed posterior over time. Infant 2 (I2) was born with a heart defect called VSD (can be acutely life-threatening) and a moderate size patent ductus arteriosus (PDA), both of which cause left to right shunting and disrupt blood flow to the body. I2 was ligated on day 4, a procedure performed to resolve PDA. She was also on dopamine starting day 2 due to hypotension, had a renal failure on day 3 post indomethacine (medication for PDA) and was on a ventilator during this entire time. On day 7, her state started to resolve significantly, and on day 8 her ventilator settings were minimal and she was taken off dopamine. 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. Of note, the ventilator controls only oxygen supply and does not directly control the heart rate, so the inferred topics are not just uncovering the different ventilator settings.

Next, we analyze the word histograms; several interesting observations arise. The infants follow a continuum of word distribution profiles. 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. 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. 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 entropy. 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 called *Physiscore* [20], that predicts based on data from the first three hours of life, infants at risk for major complications. Our score combines entropy from the base and residual physiological signals of heart-rate and respiratory rate with birth weight, gestational age and measures computed from the oxygen saturation signals (e.g., the amount of hypoxia). When validated on 138 infants with the leave-one-out method to prospectively identify infants at risk of short- and long-term morbidity, our score provided higher accuracy prediction of overall morbidity (area under the receiver operating curve (AUC) of 0.91) than other neonatal scoring systems, including the Apgar score (AUC 0.69) which is the current standard of care and SNAP (AUC 0.82) [19], a machine learning based score that requires multiple invasive tests.

6 Discussion and Future work

In summary, the primary contribution of this paper is a new class of models for time-series data that emphasizes the modeling of instance specific variability while discovering population level characteristics, especially useful for clinical data. We demonstrate its use 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. Early insights gathered from TSTM has already led to the development of a useful clinical tool.

There are several avenues for future work. Extensions of TSTM that can model disease evolution within a single patient should provide interesting insight. Furthermore, in our current effort we limited the TSTM observation model to AR(1) processes. Bayesian model selection for the order of the AR process could highlight how complexity of word varies with different disease topics. Finally, the use of non-markovian observation models can allow discovery of longer duration motifs such as apnea and bradycardia[16], known to be clinically relevant.

References

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

[2] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. In *J. Mach. Learn. Res.* 2003.

[3] E. Fox, E. Sudderth, M. Jordan, and A. Willsky. The sticky HDP-HMM: Bayesian nonparametric hidden Markov models with persistent states. Technical Report P-2777, MIT LIDS, 2007.

[4] E. Fox, E. Sudderth, M. Jordan, and A. Willsky. Nonparametric Bayesian learning of switching linear dynamical systems. In *NIPS*. 2008.

[5] E. Fox, E. Sudderth, M. Jordan, and A. Willsky. Sharing features among dynamical systems with Beta Processes. In *NIPS*. 2009.

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

[7] T. Griffiths and M. Steyvers. Finding scientific topics. In *PNAS*. 2004.

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

[9] T. Joachims. Training linear SVMs in linear time. In *KDD*. 2006.

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

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

[12] A. Khosla, Y. Cao, C. Lin, H. Chiu, J. Hu, and H. Lee. An integrated machine learning approach to stroke prediction. In *KDD*, 2010.

[13] K. Kurihara, M. Welling, and Y. W. Teh. Collapsed variational Dirichlet process mixture models. In *IJCAI*, 2007.

[14] H. Lee, Y. Largman, P. Pham, and A. Ng. Unsupervised feature learning for audio classification using convolutional deep belief networks. In *NIPS*, 2009.

[15] A. Mueen, E. Keogh, Q. Zhu, S. Cash, and B. Westover. Exact discovery of time series motifs. In *SDM*, 2009.

[16] A. A. of Pediatrics Committee on Fetus and Newborn. Apnea, sudden infant death syndrome, and home monitoring. In *Pediatrics*. Apr 2003.

[17] J. Quinn, C. Williams, and N. McIntosh. Factorial switching linear dynamical systems applied to physiological condition monitoring. In *IEEE Trans. Pattern Analysis Machine Intelligence*, 2009.

[18] D. Ramage, D. Hall, R. Nallapati, and C. Manning. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *EMNLP*, 2009.

[19] D. Richardson, J. Gray, M. McCormick, K. Workman, and D. Goldmann. Score for neonatal acute physiology: a physiologic severity index for neonatal intensive care. In *Pediatrics*. 1993.

[20] 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.* Sept 2010.

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

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

[23] Y. W. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical Dirichlet processes. In *JASA*. 2006.

[24] C. Wang, D. Blei, and D. Heckerman. Continuous time dynamic topic models. In *UAI*. 2008.

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

[26] C. Williams, J. Quinn, and N. McIntosh. Factorial switching Kalman filters for condition monitoring in neonatal intensive care. In *NIPS*, 2005.