Semiparametric Clustering Methods for Modeling Chronic Disease Dynamics

Abstract

The availability of clinical data repositories presents new opportunities for improving health care quality and reducing costs. However, the secondary function of these data sources as research tools presents challenges for longitudinal data analysis. Rather than use the sequences directly, we extend the semiparametric clustering framework to build probabilistic models, abstractions, of these sequences using continuous-time Markov models. This provides a principled way of transforming arbitrarily sampled, irregular length clinical observations to serve as input for a nonparametric Bayesian clustering method. Our results indicate over a 20% relative improvement on a benchmark and recognizable differences that can be visualized.

1. Motivation

The most significant issues facing the US health care system in the coming years includes: major aging, the massive growth of chronic diseases, and not enough caregivers. Increasing health care costs and quality issues already pose substantial issues and unless sustainable solutions can be developed our the US healthcare system will become increasingly stressed.

One proposition for transforming healthcare is to make it 'data-driven' and has resulted in numerous government and private initiatives aiming to make meaningful use of digital health data. Advocates hope that the richness of available patient data contained in these collections will enable a feedback loop of new knowledge discovery and translation to practice, supporting the engineering of a better and better health care system. There is an urgent need to demonstrate the cost-benefit of maintaining petabytes of patient data, and

------

an important role for probabilistic learning algorithms that can assist in the discovery new knowledge from these noisy, heterogenous, fragmented data collections.

Here we provide one way to approach data-driven care. To model chronic disease dynamics, which may evolve slowly, over years, we extend the semiparametric clustering framework of Jebara et al. (Jebara et al., 2007) for learning patient and population level disease characteristics from arbitrarily sampled longitudinal patient data. Specifically, we use parametric models based on continuous time Markov models paired with nonparmetric Bayesian clustering and we show results for two distinct data sets.

We describe the background for our work in Section 2. Our contribution and methods are detailed in Section 3. The rest of the paper describes: two applications in Section 5, experimental results in Section 6, and conclusions for our work in Section 7.

2. Preliminaries

Clustering is a pervasive and natural human activity that is used for a variety of tasks. Typically, we use it to group similar objects together so that we can assign characteristics that are useful for their definition. Computational clustering algorithms aim to divide data into groups that are meaningful or useful, and improving existing techniques has been the focus of considerable research in machine learning. At the minimal level, automated clustering can be viewed as a preprocessing method, or as an exploratory analysis technique that informs more targeted hypotheses.

2.1. Semiparametric Time Series Clustering

Temporal information provides critical context for diagnosis, prognosis and disease management, especially in the case of chronic conditions that can evolve at different rates among patients, and persist for years. Although clinically significant work applying exploratory techniques for patient and population level disease modeling has been demonstrated (Saria et al., 2010; Marlin et al., 2012), it is most appropriate for methods based on physiological signals collected in the crit-
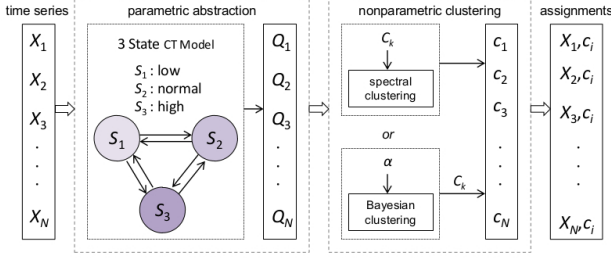
Figure 1. Semiparametric Clustering

ical care setting. Chronic disease progression can take months or years to manifest, and longer-term trends instead of significant features can have increased importance. Also, in the low frequency sampling setting, the sampling scheme may be unknown, data is more sparse, and collected over longer durations.

In the *semi-parametric clustering framework*, a parametric model of the underlying dynamic process provides useful assumptions for abstracting temporal measurement, and is paired with a nonparametric method used to cluster the abstractions. Figure 1 provides an overview of this approach.

The first step entails temporal abstraction using a parametric model to transform a set of raw time series, $X_1, X_2, ..., X_N$, into a more manageable form for traditional multivariate clustering algorithms. It entails learning each patients model-specific parameters, $Q_1, Q_2, ..., Q_N$, from the time series observations. The second step is nonparametric clustering. Spectral methods are typically used in the semiparametric framework, which requires determining the number of clusters, k, in advance. In this work, we extend the clustering component to the nonparametric Bayesian setting, allowing for the number of clusters to be expressed as a function of the sample size.

2.2. Continuous and Discrete-time Models

By default Markov models and dynamic Bayesian networks (DBNs), of which a hidden Markov model (HMM) is the simplest representation, discretize the time trajectory into uniform length contiguous segments, that are assumed to be approximately Markovian to enable tractable inference. The smallest temporal granularity among all sequences, Δ, denotes a fixed length interval that is repeated at each time slice, where a the template models is also repeated.

Although discrete-time models are suitable in many cases, there are two key limitations that have been noted in graphical modeling problems (Nodelman et al., 2003) and we describe their relevance to modeling EHR data. First, if the underlying health phenomena progresses in individuals at different rates, one granularity must be used to express time steps for the entire system. Second, when data is unavailable, intervening time slices must still be represented. When data is sparse, this forces the assumption of many unknown values that are propagated through the discrete time model's transition matrix at each Δt without support.

When there are no natural time slices, continuous-time DBNs (CT-DBNs) (Nodelman et al., 2003) can be used to more directly reflect sequential dependencies, and avoid discretizing the time intervals. Notably, the class of probabilistic graphical models known in epidemiology and biostatistics as Multi-state models (MSMs), some of which qualify as CT-DBNs, were developed independently of work in computer science and share their foundation in stochastic process theory.

MSMs have a set of rules that govern their design and interpretation and have be used to model a variety of diseases including cardiac disease, cancer, and HIV (Jackson, 2011). In terms of structure, nodes represent disease states that are ordered progressively to reflect stages in a disease trajectory. A patient with chronic diseases may traverse these nodes as their disease progresses, and typical MSM states correspond with 'healthy', 'diseased', and 'diseased with complications'. In the case of applications to survival analysis, a final absorbing state that has no outbound transitions is used to indicate death. Also, for modeling latent states, is common for an MSM emission matrix to reflect diagnostic misclassification error, and there are techniques to estimate initial values based on sensitivity and specificity.

3. Extending Semiparametric Clustering

Regardless of the temporal mining task, the first step of an algorithm is to transform, or abstract, the raw data into a more concise representation, preserving as much of the information contained in the original sequence as possible.

3.1. Continuous Time (CT) Model Abstraction

As mentioned above, Markov models and DBNs require that continuous time is represented as a series of uniform length units equal to the smallest time granularity in any observed sequence. When data is missing or otherwise incomplete, information is still prop-

agated throughout the model for every time step. It has been shown that when data is not missing at random, which often the case in clinical data, it can result in significantly more EM steps to learn the model parameters, and more importantly, lead to biased findings (Yeh et al., 2012).

CT Markov models avoid this limitation by enabling a more natural temporal representation for modelling dynamic changes that progress at different rates, and in non-liner time. A main distinguishing characteristic between DT and CT Markov models is that in the discrete-time case, the Markov process stays in a state i for a time distributed according to $F_i(t)$ and in the continuous-time case the holding time is *exponentially* distributed according to $F_i(t) = e^{q_i t}$ where q_i is the intensity of the transitions, or the tendency to change state.

3.1.1. MODEL DESCRIPTION

The specific class of CT Markov models we use for temporal abstraction are adapted from multi-state models (MSMs), which are described in more detail above. These models have been used in epidemiology and biostatistics for modeling population level dynamics for longitudinal data with missing values. We employ them for modeling patient level dynamics.

A 4 state MSM is shown in Figure 2, with states ordered by severity. The intensity matrix Q represents the instantaneous behavior of the process X. At a time t a patient is in the $S(t)$ state. For each pair of states, where $z(t)$ is a model variable, the set of transition intensities $q_{r,s}(t, z(t))$ is dependent on t and the instantaneous risk of transitioning from state r to state s

$$q_{r,s}(t, z(t)) = \lim_{\delta t \to +0} P(S(t + \delta t) = s | S(t) = r)/\delta t$$

In contrast to a discrete-time transition matrix, X with a domain of of $x_1, x_2, ..., x_n$, where n is the number of states, the intuition is that the intensity, q_i, no longer corresponds with the transition probability that is constant for the length of a time slice, but rather an *'instantaneous probability'* of leaving state x_i and the intensity of $q_{i,j}$ gives the instantaneous probability of transitioning from x_i to x_j.

In a CT Markov model, the rows in the matrix Q sum to zero instead of one, with the sum of all transition intensities $q_{r,s}$ in the rth row, where $s \neq r$, equal to the absolute value of $q_{r,r}$

$$q_{r,r} = -\sum_{s \neq r} q_{r,s}$$
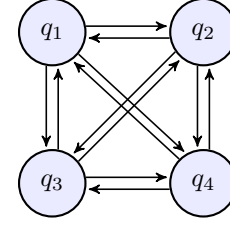
$$Q_X = \begin{bmatrix} -q_{1,1} & q_{1,2} & q_{1,3} & q_{1,4} \\ q_{2,1} & -q_{2,2} & q_{2,3} & q_{2,4} \\ q_{3,1} & q_{3,2} & -q_{3,3} & q_{3,4} \\ q_{4,1} & q_{4,2} & q_{4,3} & -q_{4,4} \end{bmatrix}$$

Figure 2. Multi-state Markov model with four discrete states that correspond with disease severity or risk, and the intensity matrix Q

and the probability of observing s immediately after state r is $q_{r,s}/q_{r,r}$.

3.1.2. MODEL ESTIMATION AND PARAMETER LEARNING

Ideally, expert knowledge is available to determine the number of disease states and the initial probabilities for the intensity matrix. For example, lab tests and biomarkers can have ranges that have been established in clinical guidelines corresponding with biological states that can be directly represented as discrete states in a probabilistic graphical model. However, in the case when there is no obvious translations of a measurement's value to discrete states, features can be learned. Non-parametric Bayesian methods are one approach that has been applied to discover features from physiological data (Saria et al., 2010) and we apply these methods to identify discrete states for disease modeling.

Specifically, for our hepatitis lab data, threshold values were provided by clinicians to indicate ranges for 'low', 'normal' and 'high' test values and directly translate to the model's state representation. In the case of glucose testing data set, which does not track test values but the incidence of physician ordered glucose tests over time, we learn a mapping from the sequence values to identify discrete model states.

To obtain initial values for the intensity matrix, a naive estimation was provided by counting the total number of transition pairs, and estimating their probability of occurrence. Although not all patient models can be learned by our method, this crude estimate consistently allowed more observations to converge using the our parameter estimation methods than the same naive initialization assumption at the population level.

For discrete models, the Expectation-Maximization algorithm (Dempster et al., 1977) is used to estimate the model parameters using a forward-backward method, and extensions exist for the continuous-time setting. To learn the model parameters in our work we use the BFGS quasi-Newton optimization algorithm (Avriel, 1976).

3.2. Nonparametric Bayesian Clustering

A main challenge posed by many traditional clustering algorithms is selecting the number of groups, k. Some approaches used to estimate k are based on the spectral gap, or predictive estimates. However, these are heuristics, and don't guarantee the choice of k, and more importantly, for many clustering problems k is unrealistic to assume that k is fixed.

By defining the clustering problem as identifying the components of infinite mixture, where k is random variable in the model, nonparametric Bayesian approaches allow for the definition of more flexible clustering models. Relative to other nonparametric clustering methods such as spectral clustering, they do not require that k is expressed a priori. Also, nonparametric Bayesian cluttering provides a generative model that can describe group structure at the population and subpopulations level, more easily lending itself to interpretation by a domain expert.

The density function of a finite mixture model is defined as:

$$p(x) = \sum_{k=1}^{K} \pi_k p(x|\theta_k)$$

where x is the data set, π is the mixing proportion, and θ_k are the model parameters for the cluster k.

In the nonparametric Bayesian application setting, we define the mixture model as that of one with *infinite components*. We can define the discrete case in the form of the integral $p(x) = \int p(x|\theta)G(\theta)d\theta$, where $G = \Sigma_{k=1}^{K} \pi_k \delta_{\theta_k}$ (**?**), which extends to the following in the case of infinite components for a Bayesian nonparametric models with a potentially infinite value of k.

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$$

3.2.1. DIRICHLET PROCESS MIXTURE MODELS

One approach to nonparametric Bayesian clustering is *Dirichlet process Gaussian mixture modeling* (DPGMM). A Dirichlet process (DP) is the prior over the mixing distribution, G, and defines a measure on measures. It is characterized by two parameters: a base distribution G_0, from which samples are drawn, and is the parameter on which the nonparametric distribution is centered. The second is a positive scaling parameter α. More intuitively described as a 'splitting' criteria, α is a scaling factor that is associated with the probability of forming a new cluster. The base distribution is defined by G_0.

$$G \sim DP(G_0, \alpha)$$

For a sample, G, drawn from the base distribution G_0, if $G \sim DP(G_0, \alpha)$, then for any set of partitions $A_1 \cup A_2 \cup ... A_k$ of A:

$$(G(A_1), ..., G(A_k)) \sim Dir(\alpha G_0(A_1), ..., \alpha G_0(A_k))$$

In the the Dirichlet process mixture model (DPMM), the DP is used as nonparametric prior in a hierarchical Bayesian model, where G is portioned according to the prior. DPMMs were first applied for density estimation, but are now applied widely for clustering.

Dirichlet Process Gaussian Mixture Modeling (DPGMM) defines a DPMM by taking the limit of the number of mixture components, k as a hierarchical Gaussian mixture model approaches infinity. Two methods used to specify the priors are Markov chain Monte-Carlo (MCMC) and variational inference, the later of which is described in the literature (Blei, 2004) and implemented for our clustering step.

3.3. Software

We use two open source packages, and adapted functions to our specific needs. Python's sci-kit learn (Pedregosa et al., 2011) provides an implementation for Dirichlet process Gaussian mixture modeling. For temporal abstraction we modified the MSM package (Jackson, 2011) to learn continuous-time Markov models for individuals. Although we cannot make our data sets available to the public, we plan to release our implementation code and a example that calls the Python code from R, and uses R's ggplot2 (Wickham, 2009) plotting library to create some of the visualizations we show visualizing clustering results.

4. Data Sets

We apply our clustering approach to two fully de-identied clinical data sets. Each consists of arbitrarily sampled clinical data in the low frequency setting, spanning both short and multi-year patient observation durations, and subject to missing data. We describe the clustering application and the nature of the data in this section.

4.1. Liver Decline in Hepatitis Patients

Our first data set relates to liver disease. A liver biopsy is the gold standard for the prognosis and treatment of liver fibrosis and important for the health provider and the patient to guide management and treatment of hepatitis B and C. In addition to the invasive nature of liver biopsy, which involves extracting a tissue sample of at least 23 cm in length, obtained with a 16-gauge needle inserted between two of the patients ribs, it is costly, and associated with complications that can be potentially life-threatening. Also, it is subject to diagnostic error. For these reasons, alternatives for assessing the stage of liver fibrosis are in great demand and the lack of alternative assessment methods has been noted as a major limitation in both management and research in liver diseases.

The first data set consists of blood inspection and urinalysis laboratory data that was provided by the Chiba University Hospital in Japan, and was used for the ECML/PKDD-2003 and 2005 Discovery Challenges. One objective of this shared task was to evaluate if laboratory examinations can be used to estimate the stage of liver fibrosis. The data set consists of recorded data for 771 patients of type B and C, spanning the years 1982 through 2001. For many of the test types, values for low, normal and high values are indicated.

At the time of the challenge, medical research suggested some lab tests such as platelet count (PLT) were correlated at the time of biopsy; however, temporal analysis of the PLT data was rarely performed and limited by difficulties in time series comparison, irregular sampling intervals, and variable sequence lengths (Shoji & Shusaku, 2005). Using PLT lab tests, one system (Hirano & Tsumoto, 2007) demonstrated that medically relevant time series features associated with the progression of liver fibrosis could be learned from the patient records. Other lab tests reported by challenge participants as informative for predicting fibrosis stage included: ZTT, ALB, D-BIL and CHE.

4.2. Inpatient Glucose Testing

Our second data set relates to glucose tests. It contains patients admitted to [omitted] Hospital with at least one physician ordered glucose test indicated in their EHR. Similar to the hepatitis patient data, the glucose time series presents methodological challenges in that it is irregularly sampled, and variable in length. An additional complicating factor is an increased probability of record incompleteness.

National estimates report a 8.3% prevalence of diabetes in the Unites States, with over seven million undiagnosed (CDC, 2011). The disease can result in various health complications such as kidney failure and blindness. However, medical research shows that behavioral changes and other interventions can prevent or delay diabetes onset showing how importance of early diagnosis and treatment. Our glucose data set is derived from administrative data and our clustering application seeks to identify high-level patterns in physicians orders for glucose tests that corresponds with diabetes related symptoms.

Glucose tests are commonly ordered by physicians for hospitalized patients, and some hospitalises now suggest an initial test during admission should be part of standard care guidelines. A single order with out any follow up on succeeding days may corresponds with an one-day admission, or with a longer stay that does not require ongoing glucose monitoring. For this reason, a series of contiguous testing patterns for a patient suggests that their blood-glucose levels are being actively monitored by the attending physician.

For each patient that may or may not have diabetes, the sequence begins with the first physician ordered test on record and indicates the presence '1' or absence '0' of a physician's orders for each successive days, tracking multiple hospitalization and discharge periods. A patent's series ends when a censoring state is encountered (i.e. death).

For our experiments we select patients with a time series length in the range of 1000 to 1025 days, and consist of 1024 patient 0/1 measurement sequences. Our methods can be applied to larger subsets of the data set, but these constraints help visually assess results. Other subsets we have experimented with but are not reported in this work are random collections or patients and those with that fall within a specific visit range.

5. Experiments

We conducted two studies using patient data associated with chronic disease. The first experiment is designed to group hepatitis patients with similar stages of liver decline by modeling the dynamics of patient lab result. The second uses administrative indicating physicians orders for glucose tests during hospital stays and aims to group patients that are at high and low risk of diabetes associated hospitalizations.

5.1. Hepatitis Experiments

Using the data set described in Section 4, and given threshold for low, normal and high observation values for each test, we constructed the model structure.
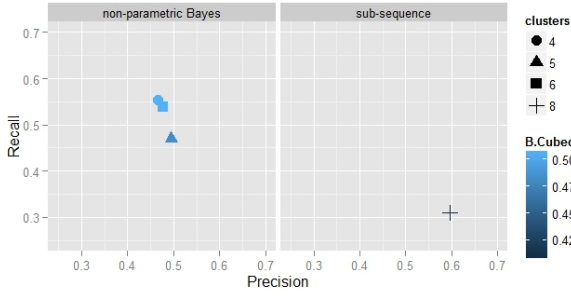
Figure 3. Comparison of semiparametric clustering with a reported benchmark

Figure 4. B-cubed value for different methods and k values for the hepatitis data set

Based on an initial run using six lab tests that were selected due to known associations with liver decline, three, the PLT, ALB, and ZZT tests, resulted in good cluster assignments, with PLT performing the best based on the b-cubed metric (Bagga & Baldwin, 1998), which is the harmonic mean of the average pointwise precision and recall for the data set. B-cubed is the only well-established metric that satisfies the full set of formal constraints proposed by researchers and validated by human assessors (Amigó et al., 2009).

5.1.1. EXTRINSIC VALIDATION

To validate the results of clustering platelet count values for the hepatitis data set, we used grading and staging data from liver biopsies a gold standard. Also, we compare our results of a previous method that used a feature extraction approach (Hirano & Tsumoto, 2007) and selected only those patients with hepatitis C and no indication of interferon therapy. The results of semiparametric clustering (SP-B) are reported for 94 of these patients based on clustering PLT (platelet) and ALB (albumin) lab tests, and for PLT only.

Figure 3 shows average pointwise precision on the x-axis, recall on the y-axis and the B-cubed value as a gradient value. Detailed scores appear in Table 1.

Table 1. Validation scores for spectral (SC) and nonparametric Bayes (Bayes) clustering on all hepatitis patients.

METHOD	K	P	R	B-CUBED
SUB-SEQUENCE	8	0.60	0.31	0.41
SP-B(PLT)	5	0.49	0.42	0.45
SP-B	**4**	**0.47**	**0.55**	**0.51**
SP-B	5	0.50	0.47	0.48
SP-B	**6**	**0.48**	**0.54**	**0.51**

Also, we apply our method to the entire data set, including those on interferon therapy and patients with hepatitis B, and compare it with an alternative non-parametric method, spectral clustering. The results appear in Figure 4 and are detailed in Table 2.

Table 2. B-cubed value for spectral (SP-SC) and nonparametric Bayes (SP-B) clustering, all hepatitis patients.

METHOD	K	P	R	B-CUBED
SP-SC	6	0.38	0.22	0.28
SP-SC	7	0.38	0.21	0.27
SP-B	**4**	**0.35**	**0.62**	**0.45**
SP-B	5	0.35	0.52	0.42
SP-B	6	0.36	0.46	0.40
SP-B	7	0.36	0.42	0.39

5.2. Diabetes Experiments

The glucose testing data was processed before state estimation. Creating a new vector, we set the observation value to the number of days contiguous tests were ordered. For example, and measurement sequence of $[1, 0, 0, 0, 1, 1, 1, 1]$ would consist of two observations, 1 and 4. We estimated n, the number of states for the models using a non-parametric Bayesian density estimator. Since the number of states and the continuous nature of the model should not be too large, the upper-bound on the number for density estimation was set to a maximum of five.

5.2.1. INTRINSIC VALIDATION

When a gold standard is unavailable to evaluate clustering performance, heuristics can be used to assess the intrinsic quality of clusters. For each point in the data set, the silhouette method is defined by the different of average dissimilarity of a point to members
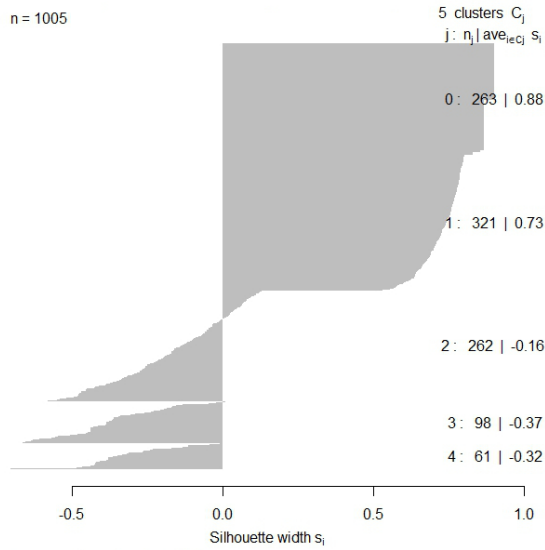
Figure 5. Silhouettes by cluster for 4-state model.

or its own cluster with that of the 'neighboring' cluster over the max of these two dissimilarity measures.

Based on the results of applying our clustering approach to the glucose testing data, Figure 5 shows the silhouette for the assignments and Table 3 shows cluster averages including: number of members, total glucose tests, number of hospital admissions, entropy of the measurement sequence, and fraction of days measured.

Table 3. Time series statistics aggregated by cluster

K	N	TESTS	STAYS	ENTROPY	FRACTION
0	263	8.6	7.81	0.05	0.01
1	321	39.64	16.98	0.14	0.04
2	262	22.00	13.53	0.10	0.02
3	98	24.77	11.38	0.10	0.02
4	61	16.74	7.69	0.08	0.02
T	1005	24.08	12.57	0.10	0.02

6. Discussion

Using biopsy results as a gold standard for clustering, our results indicate over a 20% relative improvement on a benchmark (.41-.51 b-cubed) for detecting liver fibrosis in a subset of hepatitis C patients from the liver disease data. Relative to the results based on the whole data set, better performance is achieved for the hepatitis C patients without interferon therapy. This

is not surprising. Interferon therapy is physiologically disruptive and the reason for the selection criteria in the benchmark set. However, despite the introduction of this noise, our method performs better on the group of all patients than the benchmark performs on only this subset of more predictable patients.

Of clinical relevance was an 'extreme' effect that could be viewed among different clustering runs. The lowest risk cluster, designated by the highest proportion of patients with no or minor fibrosis, reported 80-94% purity and was composed of 15-25 percent of the patients for a $4 < k < 6$. This cluster represents patients with very low risk of fibrosis, and may be good candidates for delaying biopsy.

Additionally, for an inpatient population, we can detect recognizable differences in the incidence of physicians' orders for glucose tests among discovered groups that can be visualized. We also assess the performance of pairing temporal abstraction with a non-parametric Bayesian clustering. It conveniently eliminates the need to estimate k, and performed better that spectral clustering on the hepatitis data set.

The glucose test experiment demonstrates two distinct groups with average silhouette value of .73 and .88 and accounting for almost 60% of the population sampled. These two groups show the most dramatic differences in average sequence entropy (0.05-0.14), the fraction of days measured (0.01-0.04) and the number of tests (8.60-39.64). When the original sequence is viewed as a heat map, a typical patient's sequence in the low risk group will consists of sparse signals, with only the initial visit consisting of more that one contiguous tests. However, patients in group have testing patterns that are longer in duration, showing streaks of contiguous testing, and suggesting they are more prone to diabetes related morbidity. Patients in the remaining clusters represent an intermediate between these two groups and mixtures of testing patterns.

Since diabetes is undiagnosed in millions of Americans, and preventative treatment can help avoid serious effects and avoidable costs to providers, using administrative data, such as glucose testing pattern, may prove useful for disease surveillance. For example, an insurance provider does not have access to lab results, but they will have a signal for per patient tests, information on demographic risk factors, and other billing diagnoses for patients. The ability to leverage high level signals with demographic and other claims data to flag prediabetes or undiagnosed diabetes could be useful for developing cost-savings strategies that improve health outcomes in parallel.

7. Conclusions and Future Work

We describe a new method to model patient disease dynamics with several key features. First, we apply continuous-time Markovian models for modeling disease dynamics, which avoids some of the limitations of discrete-time approaches when a dynamic process evolves at different time granulations, and when observations are irregularly sampled and missing not at random. Second, non-parametric Bayesian clustering methods avoid the problem of identifying the number of clusters a priori, inferring the appropriate number of mixture component as a function of the sample size.

The limitations of this work are mainly attributed to the temporal modeling steps. Continuous-time models bring us closer to a natural representation, but they are still inconsistent with the real-world. For example, the in the model the instantaneous probability for a state transition is the same for the entire duration of occupation. Another issue is that not all patient models converged during the abstraction step. Although this impacted only small fraction of the total patients, it is a key limitation to the method.

Immediate next steps are to extend temporal abstraction to continuous-time HMMs. Also, we feel that external validation metrics for cluster assessment are fundamentally weak for many problems where clusters are not categorical, and rather a graded interval. In terms of intrinsic evaluation, heuristics such as the silhouettes are also limited. Instead of developing one more metric, we propose a visualization tool to enable the browsing of temporal clustering results and feel this would be more useful for system development and is another direction for our future work.

References

Amigó, E., Gonzalo, J., Artiles, J., and Verdejo, F. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 12(4):461–486, 2009.

Avriel, M. *Nonlinear Programming: Analysis and Methods.* Prentice-Hall, Englewood Cliffs, N.J., 1976.

Bagga, A. and Baldwin, B. Algorithms for scoring coreference chains. In *International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pp. 563–566, 1998.

Blei, D. Variational methods for the dirichlet process. In *International Conference on Machine Learning*, 2004.

CDC. *National diabetes fact sheet.* U.S. Department of Health and Human Services, 2011.

Dempster, A., Laird, N., and Rubin, D. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistics Society, Series B*, 39(1):1–38, 1977.

Hirano, S. and Tsumoto, S. Cluster analysis of trajectory data on hospital laboratory examinations. *AMIA Annu Symp Proc*, pp. 324–328, 2007.

Jackson, C. Multi-state models for panel data: The msm package for r. *Journal of Statistical Software*, 38(8):1–28, 2011.

Jebara, T., Song, Y., and Thadani, K. Spectral clustering and embedding with hidden markov models. In *Procedings of the European Conference on Machine Learning*, 2007.

Marlin, B., Kale, D., Khemani, R., and Wetzel, R. Unsupervised pattern discovery in electronic health care data using probabilistic clustering models. In *Proceedings of the ACM SIGHIT International Health Informatics Symposium*, pp. 389–398, 2012.

Nodelman, U., Shelton, C., and Koller, D. Learning continuous time Bayesian networks. In *Proceedings of the conference on Uncertainty in Artificial Intelligence*, pp. 451–458, 2003.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Saria, S., Koller, K., and Penn, A. Learning individual and population level traits from clinical temporal data. In *In NIPS, Predictive Models in Personalized Medicine workshop*, 2010.

Shoji, H. and Shusaku, T. Temporal analysis of platelet data in chronic viral hepatitis dataset. In *Procedings of the European Conference on Machine Learning*, 2005.

Wickham, H. *ggplot2: elegant graphics for data analysis.* Springer New York, 2009.

Yeh, H., Chan, W., and Symanski, E. Intermittent missing observations in discrete-time hidden markov models. *Communications in Statistics - Simulation and Computation*, 41(2):167–181, 2012.