Learning Temporal State of Diabetes Patients via Combining Behavioral and Demographic Data

Houping Xiao[*]
SUNY Buffalo
Buffalo, NY 14226
houpingx@buffalo.edu

Jing Gao
SUNY Buffalo
Buffalo, NY 14226
jing@buffalo.edu

Long Vu
IBM T.J. Watson Research Center
Yorktown Heights, NY 10598
lhvu@us.ibm.com

Deepak S. Turaga
IBM T.J. Watson Research Center
Yorktown Heights, NY 10598
turaga@us.ibm.com

ABSTRACT

Diabetes is a serious disease affecting a large number of people. Although there is no cure for diabetes, it can be managed. Especially, with advances in sensor technology, lots of data may lead to the improvement of diabetes management, if properly mined. However, there usually exists noise or errors in the observed behavioral data which poses challenges in extracting meaningful knowledge. To overcome this challenge, we learn the latent state which represents the patient’s condition. Such states should be inferred from the behavioral data but unknown a priori. In this paper, we propose a novel framework to capture the trajectory of latent states for patients from behavioral data while exploiting their demographic differences and similarities to other patients. We conduct a hypothesis test to illustrate the importance of the demographic data in diabetes management, and validate that each behavioral feature follows an exponential or a Gaussian distribution. Integrating these aspects, we use a Demographic feature restricted hidden Markov model (DfrHMM) to estimate the trajectory of latent states by integrating the demographic and behavioral data. In DfrHMM, the latent state is mainly determined by the previous state and the demographic features in a nonlinear way. Markov Chain Monte Carlo techniques are used for model parameter estimation. Experiments on synthetic and real datasets show that DfrHMM is effective in diabetes management.

CCS CONCEPTS

•Information systems → Data mining;

KEYWORDS

Diabetes Management; Hidden Markov Model; Hypothesis Testing

[*] This work was done while the author was an intern at IBM T.J. Watson Research Center.

KDD’17, August 13–17, 2017, Halifax, NS, Canada.
© 2017 ACM. 978-1-4503-4887-4/17/08…$15.00
DOI: http://dx.doi.org/10.1145/3097983.3098100

1 INTRODUCTION.

Diabetes is a progressive, chronic disease related to your body’s challenges with regulating blood sugar. Treating Diabetes has attracted lots of attentions recently as more and more people are being affected. Unfortunately, there is currently no cure. Hence, treating diabetes is not to cure you from the condition. Instead, the treatment of diabetes consists of an ongoing process of managing your condition and then urging you to take actions on time. Among existing treatments, insulin pump therapy becomes popular and has been widely used by people of all ages because of its improvement in glucose management. It does allow for more flexibility in lifestyle and the potential to even out the wide blood sugar fluctuations that are often experienced when injecting insulin. Especially, with advances in sensor technology, a variety of sensor-augmented insulin pumps have been developed for patients to have continuous, real-time glucose readings (i.e. referred to as dynamic behavioral features) that enhance the ability to monitor glucose, especially when making decisions that involve food, exercise and sick-day management. For instance, the insulin pump from Medtronic[1] streams live data about a patient’s glucose and insulin levels via the phone into the cloud. Coupled with other information from the phone such as activity, location, food etc., they provide patients live insights and recommendations to improve diabetes management.

One important task in diabetes management is to detect hyperglycemia and hypoglycemia, as they are dangerous conditions for diabetes patients, which should be detected in advance. This requires a precise modeling to predict the glucose and insulin levels. However, prediction is not easy, because there exists inevitably noise or errors in the behavioral data, due to many reasons, such as device faults, the carelessness of patients, etc. Thus, the behavioral data cannot be directly used for extracting meaningful knowledge about the patient’s state. To overcome this challenge, we assume that there exists K latent states on which the dynamic behavioral data does rely. The advantage of latent state learning [4, 9, 16, 24, 25] is that it reduces the dimensionality of data, i.e., transforming the high-dimensional and noisy dynamic behavioral data into low-dimensional and meaningful latent states. Moreover, the latent state represents the inherent conditions of patients which captures the characteristics of the behavioral data. Each latent state corresponds to a distribution on which the behavioral data is generated. We

[1] http://www.medtronic.com/us-en/index.html

propose to use hidden Markov model (HMM) to infer the trajectory of latent states for each patient. Besides the dynamic behavioral data, the demographic data of patients is also available, which plays an important roles in diabetes states prediction. Although some existing methods incorporate demographic data in their model, they do not provide explanations. Besides, they simply concatenate the demographic and behavioral features together, which ignores the difference between the demographic and behavioral features.

In this paper, we are thefirst to conduct a hypothesis testing to demonstrate that diabetes is indeed affected by the demographic data. Wefirst run K-means to cluster patients using their demographic data, and then conduct a hypothesis testing to see whether there exists distinction between the behavioral data from clusters. More specifically, the null hypothesis is that the distributions of the dynamic behavioral features for every pair of clusters, C_k and $C_{k'}$, are the same. Namely, $\mathbf{H}_0 : F_{C_k} = F_{C_{k'}}$. If two distributions are the same, they must share the samefirst and second moments.Thus, we reduce the origianl hypothesis testing to validate whether the clusters C_k and $C_{k'}$ share the same empirical variance and mean, i.e., $\mathbf{H}_0^{var} : \sigma^2_{C_k} = \sigma^2_{C_{k'}}$ and $\mathbf{H}_0^{mean} : \mu_{C_k} = \mu_{C_{k'}}$.The F-testing and T-testing are adopted for $\mathbf{H}_0^{var}$ and $\mathbf{H}_0^{mean}$, respectively.The testing results on the real DIAB-1000 data, collected by an enterprise from about 1,000 patients over two years, confirms that demographic data is one important factor of diabetes conditions which should be taken into consideration when predicting states of diabetes patients.

Another important contribution in this paper is that we propose to make prediction or clustering based on the trajectory of latent states that is learnt from the dynamic behavioral and static demographic data (i.e., raw data) rather than directly using the raw data.The latent states are learnt via the proposed Demographic feature restricted hidden Markov model (DfrHMM). In DfrHMM, the behavior data is mainly dependent on the latent state which depends on the previous latent state and the demographic features. Namely, different patients with different static demographic data have different transition patterns of latent states.The effect of latent state and demographic features is nonlinearly incorporated into the transition function via a logistic function. In fact, a similar idea was recently proposed by [25], where they proposed a higher-order hidden Markov model and applied it in educational setting to diagnose students' learning trajectory. Each dynamic behavioral feature follows an exponential or Gaussian distribution that is validated in the real data, and then the conditional distribution of each parameter can be derived. As the posterior distribution is complex, Markov Chain Monte Carlo (MCMC) simulation is used for model parameter estimation. We present the overall algorithm to learn the latent sate and predict the dynamic behavioral features in future timestamps.The proposed model is then tested on both synthetic dataset and real dataset, DIAB-1000, which is collected from 993 patients over two years. Experimental results show that the proposed method can outperform baselines in diabetes management.

2 METHODOLOGY.

Wefirst introduce two definitions that will be used across the paper.

Definition 2.1 (Dynamic Feature). A dynamic feature is an individual measurable property of a phenomenon being observed, where the value of this property is dynamically changed over time.

Definition 2.2 (Static Feature). A static feature is an individual measurable property of a phenomenon being observed, where the value of this property remain stable for a long time.

For example, the behavioral features capturing the patients' glucose and insulin levels are called *dynamic features*, because they are evolving over time.The demographic features, such as the patients age or gender, are called *static features*, because they does not change for a long time.

Next, we present the problem formulation in Section 2.1. We will introduce the proposed Demographic feature restricted hidden Markov model and Markov chain Monte Carlo techniques for model parameter estimation in Sections 2.2 and 2.3, respectively.

2.1 Temporal Latent Status Modeling.

Temporal data is commonly observed in many sensor applications. In the setting of diabetes management, the Medtronic company develop pumps for diabetes patients to monitor their behaviors. These behavioral features could be bolus deliver status, number of hypoglycemia events within a specific duration, number of low glucose events, etc. Besides, the demographic information for all patients is also available. To model the temporal latent status of patients, wefirst make the following assumptions.The validations of these assumptions are deferred to Section 3.1.

Assumption2.1. *If patients i and j share the same latent status, their behavioral features will not differ a lot.*

Assumption2.2. *If patients i and j share the same demographic data, their behavioral features will not differ a lot.*

Given Assumptions 2.1 and 2.2, the dynamic features of a patient are determined together by the effect of latent status as well as the demographic features. Simple concatenation between behavioral and demographic features cannot work. In the proposed model, we incorporate the demographic data in the following way:The current latent state is determined by the previous latent state and the patient's demographic data. Namely, if two patients have the same latent state at the previous timestamp and share similar demographic features, they will have the same latent status at the current timestamp. In return they share similar behavioral feature values in a high probability. In this sense, for each patient, we will learn the latent status trajectory from his temporal data. Specifically, suppose there are N diabetes patients which are regularly monitored.Their temporal data over T timestamps are denoted as $\mathbf{X}_i \in \mathbb{R}^{D\times T}$ $(i = 1, \cdots, N)$, where D is the dimension of dynamic features and T is the total timestamps. For each patient i, we denote by $\mathbf{X}_{it} \in \mathbb{R}^D$, his features collected at time t. As the status of the patient i is dynamic, $\mathbf{X}_{it}$ may evolve over time.Thus, $\mathbf{X}_i$ represents the dynamic features that we observe at every timestamp.The demographic information of patient i is also collected, referred to as the static feature, which is denoted as $\boldsymbol{Z}_i \in \mathbb{R}^M$ $(i = 1, \cdots N)$, where M is the dimension of static demographic features.

We assume that there are K latent status for each patient, denoted as $\mathcal{S} = \{S_1, \cdots, S_K\}$. For the patient i, the trajectory of his latent status up to T timestamps is $\boldsymbol{\alpha}_i = (\alpha_{i1}, \cdots, \alpha_{iT})$ $(i = 1, \cdots, N)$, where $\alpha_{it} \in \mathcal{S}$ denotes the latent status at time t. We summary the notations in Table 1, where some notations will be introduced in next subsection. Next, we introduce our model, restricted hidden Markov model, to learn the latent status trajectories for patients.

Table 1: Notations.

Notation	Definition
$\mathbf{X}_i$	the dynamic behavioral data for patient i
$\mathbf{X}_i^e$	the data following an exponential dist. for patient i
$\mathbf{X}_i^n$	the data following an Gaussian dist. for patient i
$\mathcal{X}$	the collection of behavioral data
$\mathbf{Z}_i$	the static demographic data for patient i
$\mathcal{Z}$	the collection of demographic data
$\boldsymbol{\alpha}_i$	the trajectory of latent states for pateint i
$\mathcal{S}$	the space of latent state space, where $\lvert\mathcal{S}\rvert = K$
$\boldsymbol{\lambda}$	the coefficient vector of demographic features
$\boldsymbol{v}$	the coefficient vector of latent states
$\boldsymbol{c}$	the parameter vector of the exponential dist.
$\boldsymbol{\mu}$	the mean of the Gaussian dist.
$\boldsymbol{\sigma}^2$	the mean of the Gaussian dist.

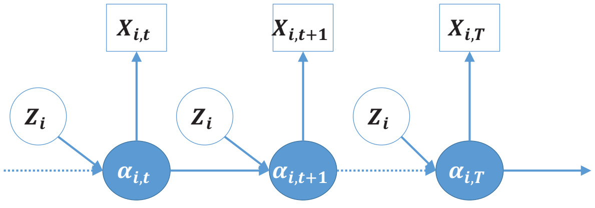


Figure 1: Flow of the proposed DfrHMM.

2.2 DfrHMM.

Demographic feature restricted Hidden Markov Model. Hidden Markov model is widely used to model the process where the behavior data is dependent on some invisible latent status. However, in our scenario, patients' latent state is also affected by their demographic features at each timestamp. Besides, the behavior data of patient i at time t, $\mathbf{X}_{it}$, are affected by both the demographic features $\mathbf{Z}_i$ and the corresponding latent state α_{it} at the specific time.The path diagram in Figure 1 illustrates such a process, which we refer to as Demographic feature estricted Hidden Markov Model (DfrHMM). DfrHMM **restricts** the observation not only to rely on the latent state at previous timestamp, like what the traditional HMM does, but only to depend on the demographic data. DfrHMM incorporates the demographic data into the modeling of transition probability.

Transition Probability Modeling. The transition probability matrix is defined as $\mathrm{P} \in \mathbb{R}^{K \times K}$. Different from traditional HMM approaches where the transition probability matrix is represented by a table with values residing in [0, 1], we adopt the idea from [25] that each transition probability is measured by a concrete function. More specifically, each element $\mathrm{P}_{kk'}$ of P represents the probability of $S_k \rightarrow S_{k'}$, i.e., the probability of being S_k' at the current timestamp given S_k at previous timestamp. In the proposed DfrHMM, we take the static feature $\mathbf{Z}_i$ into consideration. We believe that different patients with different static features will have different transition pattern. Namely,

$$\mathrm{P}_{kk'} = \mathrm{P}(\alpha_{it+1} = S_{k'} \mid \alpha_{it} = S_k, \mathbf{Z}_i) \doteq G(\mathbf{Z}_i, \alpha_{it}, \alpha_{it+1}), \tag{1}$$

where $G(\mathbf{Z}_i, \alpha_{it}, \alpha_{it+1})$is a link function. Similarly to [25], we choose a commonly used logistic function, that is,

$$logit[G(\mathbf{Z}_i, \alpha_{it})] = \lambda_0 + \boldsymbol{\lambda} \cdot \mathbf{Z}_i^\top + \boldsymbol{v} \cdot (\alpha_{it}, \alpha_{it+1})^\top, \tag{2}$$

where $\boldsymbol{v} = (v_1, v_2)$, and

- $\boldsymbol{\lambda}$ is the coefficient vector in which each entry represents the contribution of a static feature to the change of status;
- $v_1 > 0$ and $v_2 > 0$ are the change rates for previous and current underlying status, respectively.

In most conventional HMMs, P only depends on the latent states at both previous and current timestamps. Differing with them, in our scenario P is nonlinearly dependent on the latent states at both previous and current timestamps as well as the demographic data, as shown in Eq. (2). Next, we introduce an assumption about the distribution of dynamic behavioral features. We defer the validation of this assumption to Section 3.2.

Assumption 2.3 (DataDistribution). *The dynamic data* $\mathbf{X} \in \mathbb{R}^D$ *can be split into two independent sets:* $\mathbf{X}^e \in \mathbb{R}^{D_e}$ *and* $\mathbf{X}^g \in \mathbb{R}^{D_n}$*, such that 1)*$\mathbf{X} = (\mathbf{X}^e, \mathbf{X}^g)$ *with* $D = D_e + D_n$*, and 2)*$\mathbf{X}^e \dot\sim \mathrm{Exponential}(\boldsymbol{c})$*, and* $\mathbf{X}^g \dot\sim \mathrm{Gaussian}(\boldsymbol{\mu}, \boldsymbol{\sigma})$*, where* $\boldsymbol{c} \in \mathbb{R}^{D_e}$ *and* $\boldsymbol{\mu}, \boldsymbol{\sigma} \in \mathbb{R}^{D_n}$ *.*

Here, $\mathbf{X}^e \dot\sim \mathrm{Exponential}(\boldsymbol{c})$ means that for each element it follows $\mathbf{X}^e(d_e) \sim \mathrm{Exponential}(\boldsymbol{c}_{d_e}), \forall d_e \leq D_e$. Similarly, we have $\mathbf{X}^g(d_n) \sim \mathrm{Gaussian}(\boldsymbol{\mu}_{d_n}, \boldsymbol{\sigma}_{d_n}), \forall d_n \leq D_n$. So, for each patient i at time t, we have that $\mathbf{X}_{it} = (\mathbf{X}_{it}^e, \mathbf{X}_{it}^g)$ where $\mathbf{X}_{it}^e \dot\sim \mathrm{Exponential}(\boldsymbol{c})$, and $\mathbf{X}_{it}^g \dot\sim \mathrm{Gaussian}(\boldsymbol{\mu}, \boldsymbol{\sigma})$.These two types of features are commonly seen in the behavior data. For patients with different latent state k, they have different model parameters $\boldsymbol{c}_k$, $\boldsymbol{\mu}_k$ and $\boldsymbol{\sigma}_k$ to capture dynamic features.Therefore, the probability that the behavioral data of patient i at time t is $\mathbf{X}_{it}$ is as follows

$$\begin{aligned} \mathrm{P}_{\alpha_{it}}(\mathbf{X}_{it}) &\doteq \mathrm{P}(\mathbf{X}_{it} \mid \alpha_{it}, \boldsymbol{c}_{\alpha_{it}}, \boldsymbol{\mu}_{\alpha_{it}}, \boldsymbol{\sigma}_{\alpha_{it}}) \\ &= \mathrm{P}(\mathbf{X}_{it}^e \mid \alpha_{it}, \boldsymbol{c}_{\alpha_{it}}) \mathrm{P}(\mathbf{X}_{it}^g \mid \alpha_{it}, \boldsymbol{\mu}_{\alpha_{it}}, \boldsymbol{\sigma}_{\alpha_{it}}). \end{aligned} \tag{3}$$

The second equation holds because of Assumption 2.3. $\mathrm{P}_{\alpha_{it}}(X_{it})$ measures the probability that the value of the dynamic feature being X_{it} given the latent state of patient i being α_{it} at time t.The likelihood function of patient i $(i = 1, \cdots, N)$ can be written as:

$$\begin{aligned} &\mathrm{P}(\mathbf{X}_i, \alpha_i, \mathbf{Z}_i \mid \boldsymbol{c}_{\alpha_i}, \boldsymbol{\mu}_{\alpha_i}, \boldsymbol{\sigma}_{\alpha_i}) \\ &= \textstyle\prod_{t=2}^T \mathrm{P}(\mathbf{X}_{it} \mid \alpha_{it}, \boldsymbol{c}_{\alpha_{it}}, \boldsymbol{\mu}_{\alpha_{it}}, \boldsymbol{\sigma}_{\alpha_{it}}) \mathrm{P}(\alpha_{it} \mid \alpha_{it-1}, \mathbf{Z}_i) \\ &\qquad\qquad \times P(X_{i1} \mid \alpha_{i1}, \boldsymbol{c}_{\alpha_{i1}}, \mu_{\alpha_{i1}}, \sigma_{\alpha_{i1}}) \mathrm{P}(\alpha_{i1}) \\ &= \mathrm{P}(\alpha_{i1}) \textstyle\prod_{t=1}^T \mathrm{P}(\mathbf{X}_{it} \mid \alpha_{it}, \boldsymbol{c}_{\alpha_{it}}, \boldsymbol{\mu}_{\alpha_{it}}, \boldsymbol{\sigma}_{\alpha_{it}}) \prod_{t=2}^T \mathrm{P}(\alpha_{it} \mid \alpha_{it-1}, \mathbf{Z}_i) \end{aligned}$$

Moreover, based on the total probability formula, the joint probability of both static and dynamic features of patient i is

$$\mathrm{P}(\mathbf{X}_i, \mathbf{Z}_i \mid \boldsymbol{c}_{\alpha_i}, \boldsymbol{\mu}_{\alpha_i}, \boldsymbol{\sigma}_{\alpha_i}) = \textstyle\sum_{\boldsymbol{\alpha}_i} \mathrm{P}(\mathbf{X}_i, \boldsymbol{\alpha}_i, \mathbf{Z}_i \mid \boldsymbol{c}_{\alpha_i}, \boldsymbol{\mu}_{\alpha_i}, \boldsymbol{\sigma}_{\alpha_i}). \tag{4}$$

Denote collections of dynamic features and static features as $\mathcal{X} = \{\mathbf{X}_1, \cdots, \mathbf{X}_N\}$ and $\mathcal{Z} = \{\mathbf{Z}_1, \cdots, \mathbf{Z}_N\}$, respectively.Then, the full likelihood function of N patients over T timestamps is

$$\begin{aligned} &L(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Z}}, \boldsymbol{\alpha}, \boldsymbol{\lambda}, \boldsymbol{v}, \boldsymbol{c}, \boldsymbol{\mu}, \boldsymbol{\sigma}) \\ &= \mathrm{P}(\boldsymbol{\lambda})\mathrm{P}(\boldsymbol{v})\mathrm{P}(\boldsymbol{c})\mathrm{P}(\boldsymbol{\mu})\mathrm{P}(\boldsymbol{\sigma}) \textstyle\prod_{i=1}^N \sum_{k=1}^K \mathrm{P}(\alpha_{i1} = k) \\ &\quad \times \textstyle\prod_{t=1}^T \mathrm{P}(\mathbf{X}_{it} \mid \alpha_{it}, \boldsymbol{c}_{\alpha_{it}}, \boldsymbol{\mu}_{\alpha_{it}}, \boldsymbol{\sigma}_{\alpha_{it}}) \prod_{t=2}^T \mathrm{P}(\alpha_{it} \mid \alpha_{it-1}, \mathbf{Z}_i) \\ &= \mathrm{P}(\boldsymbol{\lambda})\mathrm{P}(\boldsymbol{v})\mathrm{P}(\boldsymbol{c})\mathrm{P}(\boldsymbol{\mu})\mathrm{P}(\boldsymbol{\sigma}) \textstyle\prod_{i=1}^N \sum_{k=1}^K \pi_k \\ &\quad \times \textstyle\prod_{t=1}^T \mathrm{P}(\mathbf{X}_{it} \mid \alpha_{it}, \boldsymbol{c}_{\alpha_{it}}, \boldsymbol{\mu}_{\alpha_{it}}, \boldsymbol{\sigma}_{\alpha_{it}}) \prod_{t=2}^T \mathrm{P}(\alpha_{it} \mid \alpha_{it-1}, \mathbf{Z}_i), \end{aligned} \tag{5}$$

where $\pi_k \doteq \mathrm{P}(\alpha_{i1} = k)$ measures the marginal probability of the latent state being k at the initialization step for patient i $(i = 1, \cdots, N)$.Then, the joint posterior distribution of $\boldsymbol{\alpha}$, $\boldsymbol{\lambda}$, $\boldsymbol{v}$, $\boldsymbol{c}$, $\boldsymbol{\mu}$, and $\boldsymbol{\sigma}$ given $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{Z}}$ is

$$\begin{aligned} &\mathrm{P}(\boldsymbol{\alpha}, \boldsymbol{\lambda}, \boldsymbol{v}, \boldsymbol{c}, \boldsymbol{\mu}, \boldsymbol{\sigma} \mid \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Z}}) \\ &\propto \mathrm{P}(\boldsymbol{\mathcal{X}} \mid \boldsymbol{\alpha}, \boldsymbol{c}, \boldsymbol{\mu}, \boldsymbol{\sigma}) P(\boldsymbol{\alpha} \mid \boldsymbol{\mathcal{Z}}, \boldsymbol{\lambda}, \boldsymbol{v}) \mathrm{P}(\boldsymbol{\lambda})\mathrm{P}(\boldsymbol{v})\mathrm{P}(\boldsymbol{c})\mathrm{P}(\boldsymbol{\mu})\mathrm{P}(\boldsymbol{\sigma}). \end{aligned} \tag{6}$$

Specifically, we have that

$$
\begin{aligned}
&\mathrm{P}(\boldsymbol{\alpha}, \boldsymbol{\lambda}, \boldsymbol{v}, \boldsymbol{c}, \boldsymbol{\mu}, \boldsymbol{\sigma} | \mathcal{X}, \mathcal{Z}) \\
&\propto \textstyle\prod_{i=1}^{N} \prod_{t=1}^{T} \mathrm{P}(\mathrm{X}_{it} | \alpha_{it}, \boldsymbol{c}, \boldsymbol{\mu}, \boldsymbol{\sigma}) \mathrm{P}(\alpha_{it} | \alpha_{it-1}, Z_i, \boldsymbol{\lambda}, \boldsymbol{v}) \\
&\qquad \times \mathrm{P}(\boldsymbol{\lambda}) \mathrm{P}(\boldsymbol{v}) \mathrm{P}(\boldsymbol{c}) \mathrm{P}(\boldsymbol{\mu}) \mathrm{P}(\boldsymbol{\sigma}) \\
&\propto \textstyle\prod_{i=1}^{N} \prod_{t=1}^{T} \mathrm{P}(\mathrm{X}_{it}^{e} | \alpha_{it}, \boldsymbol{c}_{\alpha_{it}}) \mathrm{P}(\mathrm{X}_{it}^{g} | \alpha_{it}, \mu_{\alpha_{it}}, \sigma_{\alpha_{it}}) \\
&\qquad \times \mathrm{P}(\alpha_{it} | \alpha_{it-1}, Z_i, \boldsymbol{\lambda}, \boldsymbol{v}) \mathrm{P}(\boldsymbol{\lambda}) \mathrm{P}(\boldsymbol{v}) \mathrm{P}(\boldsymbol{c}) \mathrm{P}(\boldsymbol{\mu}) \mathrm{P}(\boldsymbol{\sigma})
\end{aligned}
\tag{7}
$$

Denote the parameter set $\Omega = \{\boldsymbol{\lambda}, \boldsymbol{v}, \boldsymbol{c}, \boldsymbol{\mu}, \boldsymbol{\sigma}\}$ and $\forall \boldsymbol{s} \in \Omega$, $\Omega_{-\boldsymbol{s}} \triangleq \Omega \setminus \boldsymbol{s}$.Then, the full conditional distributions of each parameter in Ω given the data and the rest of the parameters are as follows:

$$\mathrm{P}(\boldsymbol{\lambda} | \mathcal{X}, \boldsymbol{\alpha}, \mathcal{Z}, \Omega_{-\boldsymbol{\lambda}}) \propto \textstyle\prod_{i=1}^{N} \prod_{t=1}^{T} \mathrm{P}(\alpha_{it} | \alpha_{it-1}, \mathrm{Z}_i, \boldsymbol{\lambda}, \boldsymbol{v}) \mathrm{P}(\boldsymbol{\lambda}); \tag{8}$$

$$\mathrm{P}(\boldsymbol{v} | \mathcal{X}, \boldsymbol{\alpha}, \mathcal{Z}, \Omega_{-\boldsymbol{v}}) \propto \textstyle\prod_{i=1}^{N} \prod_{t=1}^{T} \mathrm{P}(\alpha_{it} | \alpha_{it-1}, \mathrm{Z}_i, \boldsymbol{\lambda}, \boldsymbol{v}) \mathrm{P}(\boldsymbol{v}); \tag{9}$$

$$\mathrm{P}(\boldsymbol{c} | \mathcal{X}, \boldsymbol{\alpha}, \mathcal{Z}, \Omega_{-\boldsymbol{c}}) \propto \textstyle\prod_{i=1}^{N} \prod_{t=1}^{T} \mathrm{P}(\mathrm{X}_{it}^{e} | \alpha_{it}, \boldsymbol{c}) \mathrm{P}(\boldsymbol{c}); \tag{10}$$

$$\mathrm{P}(\boldsymbol{\mu} | \mathcal{X}, \boldsymbol{\alpha}, \mathcal{Z}, \Omega_{-\boldsymbol{\mu}}) \propto \textstyle\prod_{i=1}^{N} \prod_{t=1}^{T} \mathrm{P}(\mathrm{X}_{it}^{g} | \alpha_{it}, \boldsymbol{\mu}, \boldsymbol{\sigma}) \mathrm{P}(\boldsymbol{\mu}); \tag{11}$$

$$\mathrm{P}(\boldsymbol{\sigma} | \mathcal{X}, \boldsymbol{\alpha}, \mathcal{Z}, \Omega_{-\boldsymbol{\sigma}}) \propto \textstyle\prod_{i=1}^{N} \prod_{t=1}^{T} \mathrm{P}(\mathrm{X}_{it}^{g} | \alpha_{it}, \boldsymbol{\mu}, \boldsymbol{\sigma}) \mathrm{P}(\boldsymbol{\sigma}). \tag{12}$$

The full conditional distribution for α_{it} is

$$\mathrm{P}(\alpha_{it} | \mathrm{X}_i, Z_i, \boldsymbol{\pi}, \boldsymbol{\beta}) = \textstyle\prod_k \tilde{\pi}_{ik,t}^{I(\alpha_i = k)}, \tag{13}$$

where the estimation of the $\tilde{\pi}_{ik,t}$ ($\forall i = 1, \cdots, N, k = 1, \cdots, K, t = 1, \cdots, T$) takes the following forms in different scenarios:
(1) When $t = 1$,

$$\tilde{\pi}_{ik,t} = \frac{\pi_k \mathrm{P}(\mathrm{X}_{it} | \alpha_{it} = k, \boldsymbol{\beta})}{\sum_{k'} \pi_{k'} \mathrm{P}(\mathrm{X}_{it} | \alpha_{it} = k', \boldsymbol{\beta})}. \tag{14}$$

(2) For $1 < t \leq T$,

$$\tilde{\pi}_{ik,t} = \frac{\mathrm{P}(\mathrm{X}_{it} | \alpha_{it} = k, \boldsymbol{\beta}) \mathrm{P}(\alpha_{it} = k | \alpha_{it-1}, \mathrm{Z}_i)}{\sum_{k'} \mathrm{P}(\mathrm{X}_{it} | \alpha_{it} = k', \boldsymbol{\beta}) \mathrm{P}(\alpha_{it} = k' | \alpha_{it-1}, \mathrm{Z}_i)}. \tag{15}$$

As the posterior distribution is complex, one efficient approach to obtain the estimation of the parameters is Markov chain Monte Carlo (MCMC) simulation. Namely, we cannot obtain the closed form for the full conditional distribution of parameters in Ω, but samples for those type of parameters can be iteratively drawn from their distributions for estimation.

2.3 MCMC for Model Parameter Estimation.

To apply Markov chain Monte Carlo (MCMC) techniques [25] for estimating model parameters, we first introduce the following prior distributions for all parameters in Ω that are used in the model.

$$
\begin{aligned}
&\lambda_0 \sim \mathrm{Normal}(\mu_{\lambda_0}, \sigma_{\lambda_0}^2); \\
&\lambda_j \sim \mathrm{Lognormal}(\mu_{\lambda_j}, \sigma_{\lambda_j}^2), \forall, j = 1, \cdots D_M; \\
&v_j \sim \mathrm{Lognormal}(\mu_{v_j}, \sigma_{v_j}^2), j = 1, 2; \\
&\boldsymbol{c}_{kj} \sim \mathrm{Gamma}(\alpha_{\boldsymbol{c}_{kj}}, \beta_{\boldsymbol{c}_{kj}}), \forall j = 1, \cdots, D_e, k = 1, \cdots, K; \\
&\boldsymbol{\mu}_{kj} \sim \mathrm{Normal}(\mu_{\boldsymbol{\mu}_{kj}}, \sigma_{\boldsymbol{\mu}_{kj}}^2), \forall j = 1, \cdots, D_n, k = 1, \cdots, K; \\
&\boldsymbol{\sigma}_{kj}^2 \sim \mathrm{Inverse-Gamma}(\alpha_{\boldsymbol{\sigma}_{kj}^2}, \beta_{\boldsymbol{\sigma}_{kj}^2}), \forall j = 1, \cdots, D_n, k = 1, \cdots, K; \\
&\boldsymbol{\pi} \sim \mathrm{DiscreteUniform}(K);
\end{aligned}
$$

For simplicity, the hyper parameter set is denoted as

$$\boldsymbol{\Psi} = \{\mu_{\lambda_j}, \mu_{v_j}, \sigma_{v_j}, \alpha_{\boldsymbol{c}_{kj}}, \beta_{\boldsymbol{c}_{kj}}, \mu_{\boldsymbol{\mu}_{kj}}, \sigma_{\boldsymbol{\mu}_{kj}}, \alpha_{\boldsymbol{\sigma}_{kj}^2}, \beta_{\boldsymbol{\sigma}_{kj}^2}\}, \tag{16}$$

where subindexs are properly chosen. For any parameter vector $\boldsymbol{C} \in \mathbb{R}^d$, we denote that $\boldsymbol{C}_{-i}$ represents all elements in $\boldsymbol{C}$ except for $c_i, \forall i = 1, \cdots, d$.Then, the update at the r-th iteration of the MCMC simulation is shown as follows.

2.3.1 *Update $v_j \in \boldsymbol{v}, \forall j = 1, 2$.*

$\boldsymbol{v}$ is the coefficient parameter vector in (2) which captures the effect of latent status in both previous and current timestamps. To update v_j, we first draw v_j^* from $\mathrm{Uniform}(v_j^{r-1} - \delta_{v_j}, v_j^{r-1} + \delta_{v_j})$, and accept v_j^* with probability

$$\pi(v_j^*, v_j^{r-1}) = \frac{\prod_{i=1}^{N} \prod_{t=1}^{T} \mathrm{P}(\alpha_{it}^{r-1} | \alpha_{it-1}^{r-1}, \mathrm{Z}_i, \boldsymbol{\lambda}^{r-1}, \boldsymbol{v}_{-j}^{r-1}, v_j^*) \mathrm{P}(v_j^*)}{\prod_{i=1}^{N} \prod_{t=1}^{T} \mathrm{P}(\alpha_{it}^{r-1} | \alpha_{it-1}^{r-1}, \mathrm{Z}_i, \boldsymbol{\lambda}^{r-1}, \boldsymbol{v}^{r-1}) \mathrm{P}(v_j^{r-1})}. \tag{17}$$

Here, $\mathrm{P}(v_j^*)$ is the marginal p.d.f. of choosing v_j^*.

2.3.2 *Update $\lambda_j \in \boldsymbol{\lambda}, \forall j = 0, \cdots, D_M$.*

Note that λ_0 is the intersection and λ_j ($j = 1, \cdots, D_M$) is a coefficient which represents the contribution of the static demographic feature in the link function (2). To update λ_j, we first draw λ_j^* from a Uniform distribution, i.e., $\mathrm{Uniform}(\lambda_j^{r-1} - \delta_{\lambda_j}, \lambda_j^{r-1} + \delta_{\lambda_j})$, and accept λ_j^* with probability

$$\pi(\lambda_j^*, \lambda_j^{r-1}) = \frac{\prod_{i=1}^{N} \prod_{t=1}^{T} \mathrm{P}(\alpha_{it}^{r-1} | \alpha_{it-1}^{r-1}, \mathrm{Z}_i, \boldsymbol{\lambda}_{-j}^{r-1}, \boldsymbol{v}^{r-1}, \lambda_j^*) \mathrm{P}(\lambda_j^*)}{\prod_{i=1}^{N} \prod_{t=1}^{T} \mathrm{P}(\alpha_{it}^{r-1} | \alpha_{it-1}^{r-1}, \mathrm{Z}_i, \boldsymbol{\lambda}^{r-1}, \boldsymbol{v}^{r-1}) \mathrm{P}(\lambda_j^{r-1})}. \tag{18}$$

Here, $\mathrm{P}(\lambda_j^*)$ is the marginal p.d.f. of choosing λ_j^*.

2.3.3 *Update $\boldsymbol{c}_{kj} \in \boldsymbol{c}_k, \forall j = 1, \cdots, D_e; k = 1, \cdots, K$.*

Note that the prior of $\boldsymbol{c}_{kj}$ is Gamma distribution which is a conjugate distribution in terms of the exponential likelihood function.Thus, the posterior of $\boldsymbol{c}_{kj}$ is Lomax distribution with parameters $(\hat{\beta}, \hat{\alpha})$:

$$\widehat{\alpha} = \alpha_{\boldsymbol{c}_{kj}} + \sum_{i,t} I(\alpha_{it} = k); \quad \widehat{\beta} = \beta_{\boldsymbol{c}_{kj}} + \sum_{i,t} X_{it}^{e} I(\alpha_{it} = k). \tag{19}$$

Thus, $\boldsymbol{c}_{kj}$ is sampled following a two-step procedure:

$$
\begin{aligned}
&\text{(a)} \quad \text{Sample} \quad \lambda \sim \mathrm{Gamma}(\tfrac{1}{\hat{\beta}}, \hat{\alpha}); \\
&\text{(b)} \quad \text{Sample} \quad \boldsymbol{c}_{kj} \sim \mathrm{Exponential}(\lambda).
\end{aligned}
\tag{20}
$$

Note that the Lomax distribution arises as a mixture of exponential distributions where the mixing distribution of the rate is a gamma distribution. Namely, if $\lambda \sim Gamma(k, \theta)$ and $X \sim Exponential(\lambda)$, then the marginal distribution of X is $Lomax(\frac{1}{k}, \theta)$.

2.3.4 *Update $\boldsymbol{\mu}_{kj}$ and $\boldsymbol{\sigma}_{kj}^2$, $\forall j = 1, \cdots, D_n; k = 1, \cdots, K$.*

The conjugate prior distribution of $\boldsymbol{\mu}_{kj}$ and $\boldsymbol{\sigma}_{kj}$ follows a Normal-inverse gamma distribution with parameters $(\mu_{\boldsymbol{\mu}_{kj}}, 1, \alpha_{\sigma_{kj}}, \beta_{\sigma_{kj}})$. As the likelihood is normal function, the posterior of $\boldsymbol{\mu}_{kj}$ and $\boldsymbol{\sigma}_{kj}$ also follows a Normal-inverse gamma distribution with posterior hyperparameters $(\hat{\mu}, \hat{\lambda}, \hat{\alpha}, \hat{\beta})$, where

$$
\begin{aligned}
&\hat{\mu} = \tfrac{\mu_{\boldsymbol{\mu}_{kj}} + \bar{X}_{it}}{1 + N_k}, \quad \hat{\lambda} = 1 + N_k, \quad \hat{\alpha} = \alpha_{\sigma_{kj}} + \tfrac{N_k}{2}, \\
&\hat{\beta} = \beta_{\sigma_{kj}} + \tfrac{1}{2} \textstyle\sum_{i=1}^{N_k} (X_{it} - \bar{X}_{it})^2 + \tfrac{N_k}{1 + N_k} \tfrac{(\bar{X}_{it} - \mu_{\boldsymbol{\mu}_{kj}})^2}{2}).
\end{aligned}
\tag{21}
$$

Here, $\bar{X}_{it}^{g} = \frac{1}{N_k} \sum_{it} X_{it}^{g} I(\alpha_{it} = k)$, $\widehat{\sigma}^2 = \frac{1}{N_k} \sum_{it} (X_{it}^{g} - \bar{X}_{it}^{g})^2$, and N_k is the total number of patients' status being k.Thus, $\boldsymbol{\mu}_{kj}$ and $\boldsymbol{\sigma}_{kj}$ are jointly generated as follows:

$$
\begin{aligned}
&\text{(a)} \quad \text{Sample} \quad \boldsymbol{\sigma}_{kj} \sim \mathrm{Inverse\text{-}Gamma}(\hat{\alpha}, \hat{\beta}); \\
&\text{(b)} \quad \text{Sample} \quad \boldsymbol{\mu}_{\boldsymbol{kj}} \sim \mathrm{Normal}(\hat{\mu}, \tfrac{\sigma_{kj}}{N_k + 1}).
\end{aligned}
\tag{22}
$$

2.3.5 *Update $\alpha_{it} \in \boldsymbol{\alpha}_i, \forall i = 1, \cdots, N, \forall t = 1, \cdots, T$.*

Note that α_{it} is a discrete distribution with values from $\{S_1, \cdots, S_K\}$.

(a) Draw α_{i1}^* from the discrete distribution with probability $\tilde{\pi}_{ik,1}$ based on Equation (14);
(b) Draw $\alpha_{it}^*, \forall t = 2, \cdots, T$ from the discrete distribution with probability $\tilde{\pi}_{ik,t}$ based on Equation (15).

The previous steps 1 ~ 5 iteratively proceed until the convergence condition is satisfied. The pseudo code of the restricted hidden Markov model is presented in Algorithm 1.

Algorithm 1 DfrHMM

```
Input: $\mathbf{X}^e$, $\mathbf{X}^n$, and hyperparameters $\mathbf{\Psi}$.
Output: $\boldsymbol{\alpha} \in \mathbb{R}^{N \times K}$.
1: while $\boldsymbol{\alpha}$ does not change do
2:     Sample $v_j^* \sim$ Uniform$(v_j^{r-1} - \delta_{v_j}, v_j^{r-1} + \delta_{v_j})$ and accept $v_j^*$ with probability (1(b)), $\forall j = 1, 2$;
3:     Sample $\lambda_j^* \sim$ Uniform$(\lambda_j^{r-1} - \delta_{\lambda_j}, \lambda_j^{r-1} + \delta_{\lambda_j})$ and accept $\lambda_j^*$ with probability (1(a)), $\forall j = 1, \cdots, D_M + 1$;
4:     Calculate the hyperparameters according to (1(e)), and sample $\boldsymbol{c}_{kj}$ according to (20);
5:     Calculate the hyperparameters according to (21), and jointly sample $\boldsymbol{\mu}_{kj}$ and $\boldsymbol{\sigma}_{kj}$ according to (22);
6:     Update $\alpha_{it}$ according to (14) and (15) properly.
7: end while
8: return $\boldsymbol{\alpha}$.
```

3 DIAB-1000 DATA.

The diabetes dataset, Diab-1000, is collected by an enterprise that monitors about 1000 patients over two years using sensors. Diab-1000 contains two types of data: dynamic behavioral data and static demographic data. The notations of the features in dynamic and static data are presented in Tables 2 and 3, respectively. In behavioral data, the details of all features are as follows:

- *Label* represents whether a patient has at least one hypoglycemia event after the current bolus. For example, if label2hr equals to 1 means that there is at least one hypoglycemia event after the patient takes the bolus.
- *Hypoglycemia Events*, i.e., Hypo Events, records how many times that hypoglycemia has occurred during days $t1 \sim t2$ for a patient.
- LGA is abbreviation of *Low Glucose Alerts* records the number of low glucose events during days $t1 \sim t2$.
- *Sensor Glucose*, abbreviated as SG, is measured with finger stick method while blood glucose is taken from blood. For SG, we record its total amount, the latest amount, and other statistics within a period, such as mean, deviation, and minimum slop. The length of the trajectory of the dynamic features for patient i is T_i. The range of T_i $(i = 1, \cdots, N)$ is 31 ~ 5786.

Table 2: Dynamic behavioral features.

Types	#	Details
Label	3	label2hr, label3hr, label4hr
Hypo Events	9	Hypo-t-Day, $t \in \{30, 14, 7, 3, 1\}$ Hypo-$t1$-to-$t2$-Day, $t1 \in \{1, 3, 7, 14\}$, $t2 \in \{3, 7, 14, 30\}$
Low Glucose Alert (LGA)	9	LGA-t-Day, $t \in \{1, 3, 7, 14, 30\}$ LGA-$t1$-to-$t2$-Day, $t1 \in \{1, 3, 7, 14\}$,$t2 \in \{3, 7, 14, 30\}$
Sensor Glucose (SG)	17	Total Number, SGlatest 30min: SGmean, SGsdev, SGmin, slope 2hr: SGmean, SGsdev, SGmin, slope 4hr: SGmean, SGsdev, SGmin, slope 2hr: SG2-4, SG30-2 SG0-30
Total number of features = 39		

Table 3: Static demographic features.

Types	Details
Age Gender Weight Height	Physiological features
Years_on_Insulin	number of years taking insulin
Diabetes_on_set_Age	how old when diagnosed as diabetes
Total number of features = 6	

3.1 Importance of Demographic Data

According to the "National Diabetes Statistics Report" [5], there exists great diversity in the demographic data of diabetes patients, such as age, gender. As stated in Assumption 2.2, patients with similar demographic features share the similar behavioral features of diabetes. We make hypothesis testings to validate the reasonability of Assumption 2.2 using the Diab-1000 data. In this dataset, there are 6 features related to the demography, including age, gender, height, weight, years on insulin, and diabetes on set age.

To do this, we first run K-means clustering on the data and then conduct hypothesis testing between pairs of clusters. The underlying principle is that if two patients are from two distinct clusters, their dynamic behavioral features are generated from the same type of distributions but with different parameters. So, to tell whether the static demographic features make a difference in the diabetes states, we need to confirm that the distribution from different clusters is distinguishable. Thus, for any pair of clusters $(C_k, C_{k'})$, we conduct the following hypothesis testing:

$$\text{null hypothesis:} \quad \mathbf{H}_0 : \mathrm{F}_{C_k} = \mathrm{F}_{C_{k'}}, \tag{23}$$

where F denotes the distribution of a specific cluster. To conduct the hypothesis testing (23), we can conduct the following two hypothesis testings step by step:

$$\text{null hypothesis:} \quad \mathbf{H}_0^{var} : \sigma_{C_k}^2 = \sigma_{C_{k'}}^2; \tag{24}$$

and

$$\text{null hypothesis:} \quad \mathbf{H}_0^{mean} : \mu_{C_k} = \mu_{C_{k'}}; \tag{25}$$

where σ^2 and μ denote the variance and mean of every cluster, respectively. The reason is that if two distributions have either different first moments (i.e., mean) or different second moments (i.e., variance), they cannot be the same distributions. Namely, $\mathbf{H}_0^{var}$ and $\mathbf{H}_0^{mean}$ are the necessary conditions for $\mathbf{H}_0$. Thus, $\mathbf{H}_0$ is rejected as long as either $\mathbf{H}_0^{var}$ or $\mathbf{H}_0^{mean}$ is rejected.

For null hypothesis $\mathbf{H}_0^{var}$, F-testing is adopted and T-testing is used for $\mathbf{H}_0^{mean}$. For each hypothesis testing, we will obtain an $h \in \{0, 1\}$ score where the null hypothesis is accepted if $h = 1$ otherwise the null hypothesis is rejected. The importance of the demographic data is defined as follows:

$$\mathrm{IM} = \frac{\#(h_{\mathbf{H}_0^{var}} \odot h_{\mathbf{H}_0^{mean}} \neq 0)}{K^2 - K}, \tag{26}$$

where $\odot$ is the element-wise product. IM $\in [0, 1]$, that is, demographic data is the least important when IM $= 0$ and the most important when IM $= 1$. We conduct experiments on the demographic data in Diab-1000 dataset. We change the cluster number K from 5 to 11 and repeat K-means for 30 times. The result of IM

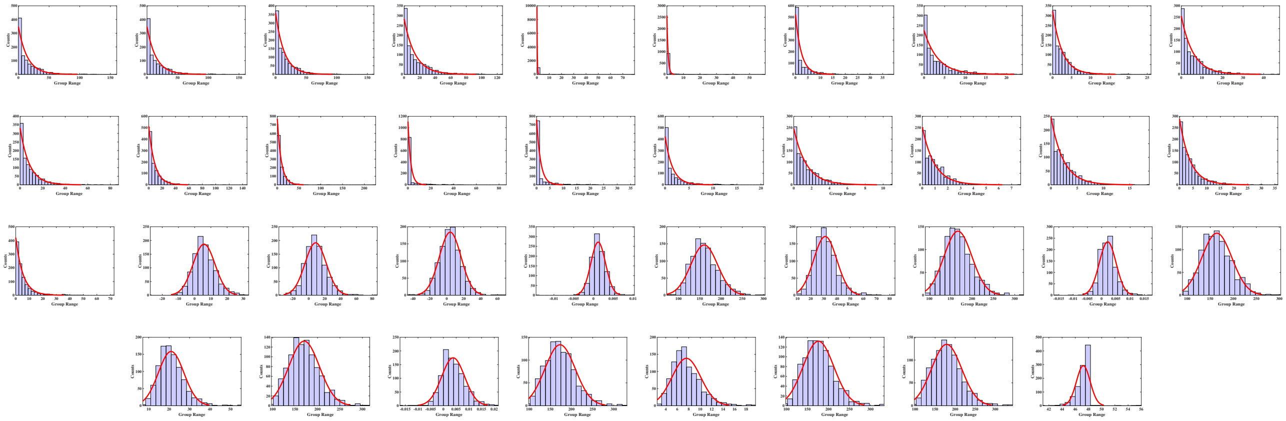


Figure 3: The empirical distributions of features, $f_1 \sim f_{38}$. The histogram is coded in blue and the corresponding distribution fitting is coded in red. For $f_1 \sim f_{21}$, each feature follows an exponential distribution while for the rest of features each one follows a Gaussian distribution.

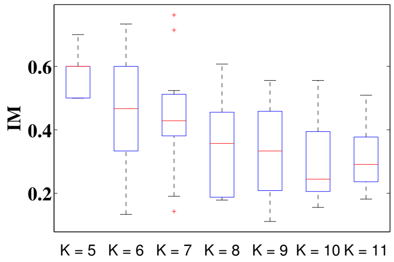


Figure 2: Importance of demographic data w.r.t. Cluster #, K.

measure is reported in Figure 2. From Figure 2, we can see that IM can achieve .7 when $K = 5$ or $K = 6$. In other cases, the average of IM remains .5. This phenomenon shows that there exists great distinction between some clusters. We show a case study when $K = 5$ in Table 4. We can see that C_1 is different from C_2, C_3, and C_5. C_3 is different from the rest of clusters. In this sense, we cannot treat C_1 or C_3 with other clusters as the same. Similar results can be obtained in other scenarios. Thus, the importance of using the demographic data in diabetes state prediction is obvious, which in return validates the proposed restricted hidden Markov model.

Table 4: A case study: Importance of demographic data when $K = 5$.

	C_1	C_2	C_3	C_4	C_5
C_1	0	1	1	0	1
C_2	1	0	1	0	0
C_3	1	1	0	1	1
C_4	0	0	1	0	0
C_5	1	0	1	0	0

3.2 Distribution Validation

In this part, we validate the distribution assumption (i.e. Assumption 2.3) introduced in Section 2.2, that is, any dynamic feature follows either an exponential distribution or a Gaussian distribution. Note that we have totally 42 dynamic behavioral features in Diab-1000 dataset. For each patient, we average his feature values over 31 timestamps. Then, for each feature, we plot its histogram and their corresponding distribution fitting in Figure 3. From Figure 3, we can see that each of the first 25 features follows an exponential distribution. They contains 3 types of dynamic behavioral features: Label, Hypo Events and Low Glucose Alert (as shown in Table 2), each of which has integer values. For the rest two types of features, RHD and SG, each of them follows a gaussian distribution whose value is continuous.

4 EXPERIMENTS.

In previous section, we have validated Assumptions 2.1 and 2.2, which are about the importance of demographic data, and Assumption 2.3 that each behavioral feature follows an exponential or a Gaussian distribution. Next, we will test the effectiveness of the proposed DfrHMM in both synthetic and real datasets in diabetes management when compared with baselines.

4.1 Experiment Setup.

In this part, we introduce the evaluation measures and the baselines. Evaluation measures we use are:

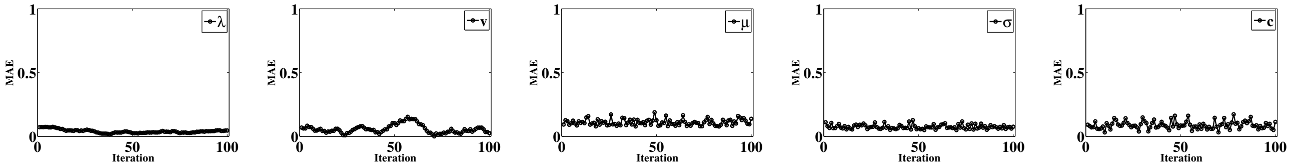
- *MAE.* Mean of Absolute Error (*MAE*) measures the L^1-norm between the real values and the predicted ones. *MAE* tends to penalize more on small errors.
- *RMSE.* Root of Mean Square Error (*RMSE*) measures the L^1-norm between the real values and the predicted values. *RMSE* penalizes more on the large difference and less on the small difference comparing with *MAE.*
- *Accuracy.* Accuracy is defined as the percentage of matched status between real latent status and the estimated one.

The accuracy of latent status is only used in synthetic data experiments, because no real state is available in real data. For Accuracy, the higher the better; while for MAE and RMSE, the lower the better.

For model comparison, we implement the following methods.

HMM. Hidden Markov model (HMM) is a widely used tool for estimating the sequence of states the model goes through to generate the observed data. HMM can estimate the emission distribution for each latent state. For each individual patient, HMM learns a specific trajectory of this latent status, as well as the emission distribution related to different latent status. Note that, HMM does not take the demographic data into consideration.

HMM+Clustering. The goal of HMM+Clustering is also to learn the trajectory of latent status for each patient. However, in



(a) $\boldsymbol{\lambda}$ (b) $\boldsymbol{v}$ (c) $\boldsymbol{\mu}$ (d) $\boldsymbol{\sigma}$ (e) $\boldsymbol{c}$

Figure 4: Markov Chain Monte Carlo mean of absolute error of models parameters: (a) λ, (b) $\boldsymbol{v}$, (c) μ, (d) σ, and (e) $\boldsymbol{c}$.

this scenario, we assume that patients belong to K' groups and patients within each group share the same sequence of status. The group information can be obtained from the demographic data. So, HMM+Clustering first clusters patients into several groups based on their static demographic features, and then estimate the trajectory of latent status for each patient groups using the dynamic behavior data. The emission distribution of different latent states is estimated. Therefore, the trajectory of each individual patient is the same as that of the group this patient belongs to.

DfrHMM. The restricted hidden Markov model (DfrHMM) is the proposed model. In DfrHMM, we incorporate the demographic data to model the transition probability between latent states. The emission distribution of different latent states is estimated.

When we want to predict the value at a future timestamp t, all methods will first learn the latent state at t and then sample the value from the corresponding emission distribution.

4.2 Experiments on Synthetic Data.

4.2.1 Data Generation. We first randomly generate the hyper parameters $\boldsymbol{\lambda} \in \mathbb{R}^{D_M+1}$, $\boldsymbol{v} \in \mathbb{R}^2$, $\boldsymbol{c} \in \mathbb{R}^{D_e \times K}$, $\boldsymbol{\mu}^d, \boldsymbol{\sigma}^d \in \mathbb{R}^{D_n \times K}$ and $\boldsymbol{\mu}^s, \boldsymbol{\sigma}^s \in \mathbb{R}^{D_M \times K'}$. Here, K is the number of latent status, K' is the number of clusters in static data, D_e is the dimensionality of dynamic features which follow the exponential distribution and D_n is the dimensionality of dynamic features following Gaussian distribution. Given the static data clustering label k', for each patient i $(i = 1, \cdots, N)$, its static data $\mathbf{Z}_i$ is randomly generated from a Gaussian distribution Gaussian$(\boldsymbol{\mu}^s_{k'}, \sigma^s_{k'})$. Given the initialized latent status, the trajectory of all patients status over T timestamps is determined by the link function (2). Based on the $\boldsymbol{\alpha}$, $\mathbf{X}^e_{it} \sim$ Exponential$(\boldsymbol{c}_{kj})$ and $\mathbf{X}^n_{it} \sim$ Gaussian$(\boldsymbol{\mu}^d_{kj}, \boldsymbol{\sigma}^d_{kj})$. We set $N = 200$ and $T = 100$, and K is changed from 2 to 4.

4.2.2 Result Analysis. Since the real latent status of each patient at every timestamp is known, we can calculate the accuracy of the predicted latent state returned by all methods. The accuracy comparison is presented in Figure 5. From Figure 5, we can draw the following conclusions. (1) HMM method only considers the dynamic data, and thus its accuracy is the lowest. It indicates that the static data is indeed helpful in recovering the latent status. (2) The proposed DfrHMM outperforms baselines in terms of accuracy.

Parameters Estimation. We also test the ability of the proposed DfrHMM in recovering the model parameters. As the real value of the model parameters are known as ground truths, we can report the error between the estimated value and the true value of model parameters. We have five sets of parameters, that is, $\boldsymbol{\lambda}$, $\boldsymbol{v}$, $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, and $\boldsymbol{c}$. Note that each set of model parameters is a vector. Instead of reporting the error for each element of each set, we use the mean of absolute errors to measure the overall deviation. The *MAE* of five parameter sets are presented in Figure 4. From Figure 4, we can see that after the burn-in step, the estimated values of all model parameters are consistently close to the true ones. It shows the effectiveness of the proposed DfrHMM.

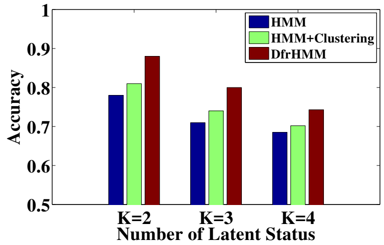


Figure 5: Comparison on Synthetic Data with different K.

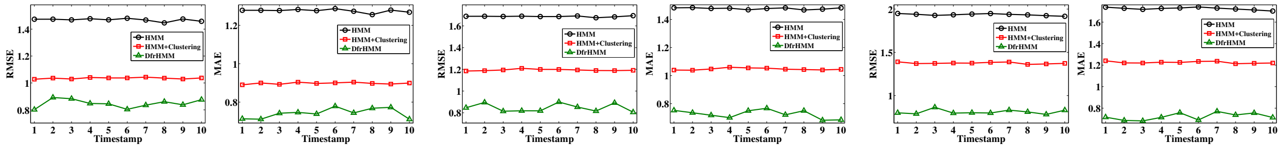
Table 5: RMSE and MAE comparison on synthetic data w.r.t K.

		DfrHMM	HMM	HMM+Clustering
$K = 2$	RMSE	.8200	1.325	.9768
	MAE	.7365	1.092	.8107
$K = 3$	RMSE	.8339	1.402	1.013
	MAE	.6929	1.137	.8156
$K = 4$	RMSE	.7899	1.354	.9670
	MAE	.6968	1.080	.7754

Prediction. We test the performance of all methods on the effectiveness of behavioral prediction. We first report the *RMSE* and *MAE* of all methods on prediction that is made for the next timestamp in Table 5. In this experiment, we use the data from $1 \sim T - 1$ timestamps as input and predict the behavioral value at time T. The results are reported in Table 5. From Table 5, we can see that the proposed DfrHMM framework outperforms HMM and HMM+Clustering methods in terms of both *RMSE* and *MAE*. Moreover, Table 5 also shows the performance of all methods in terms of the number of latent status. For HMM, it has the best performance when $K = 2$. For DfrHMM and HMM+Clustering, they achieves best results when $K = 4$. Further, we show the results on predicting behavioral data at more than one timestamps in Figure 6. In this experiment, when we predict the value at time $t+1$, we use the predicted value at t time as input. In Figure 6, we present both *RMSE* and *MAE* on $T = 10$ timestamps. From Figure 6, we can see that the proposed DfrHMM outperforms baselines in all scenarios in terms of both *RMSE* and *MAE*. Moreover, for all methods, the prediction error over T timestamps is stable. One possible reason is due to the procedure of generating the predicted value used by these methods: We first estimate the the latent status, along with their emission distributions. Then, based on the emission distribution, the behavioral data is sampled and treated as the predicted value. Once the latent status is correctly obtained, the sampled value is close to

Table 6: *RMSE* and *MAE* Comparison on Prediction on Diab-1000 data for 5 timestamps.

	RMSE					*MAE*				
	*t*1	*t*2	*t*3	*t*4	*t*5	*t*1	*t*2	*t*3	*t*4	*t*5
HMM	860.2	1203.3	-	-	-	583.2	982.4	-	-	-
HMM+Clustering	30.25	30.05	29.48	30.45	29.97	16.22	16.13	15.98	16.37	16.15
DfrHMM	14.26	8.27	9.33	9.61	11.04	10.36	5.890	6.256	5.65	7.96



(a) Latent Status $K = 2$

(b) Latent Status $K = 3$

(c) Latent Status $K = 4$

Figure 6: ***RMSE*** **and** ***MAE*** **comparison on all methods on synthetic data in prediction that are made for the next** 10 **timestamps when (a) the number of latent status** $K = 2$**, (b)** $K = 3$**, and (c)** $K = 4$**.**

the real value. Therefore, both Table 5 and Figure 6 demonstrate the effectiveness of the proposed DfrHMM in predicting the behavioral value for future timestamps.

4.3 Experiments on Diab-1000 Data.

In this section, we test the proposed DfrHMM framework on the real data, Diab-1000. The details of Diab-1000 can be found in Section 3.

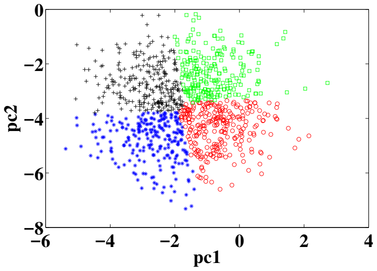


Figure 7: Visualization of the clustering based on the trajectory of latent state.

Prediction. Note that Diab-1000 collects data for all the patients over 31 timestamps. We use the data until the 26-th timestamp as input and predict the value at the remaining 5 timestamps. We report the average of *RMSE* and *MAE* over 20 repetitions in Table 6. From Table 6, we can have the following conclusions. (1) HMM has the worst performance. Especially, after prediction that made for the next two timestamps, the error becomes larger than 1000. (2) For both HMM+clustering and RHMM, their prediction error is stable and both HMM+clustering and DfrHMM outperform HMM. One possible reason is that the variance obtained by HMM in Diab-1000 data is very large. The average of variance over 42 features is around 850, while that of HMM+Clustering and DfrHMM is about one digit. The outperformance of HMM+Clustering and DfrHMM emphasizes the importance of demographic data. Moreover, DfrHMM achieves the best performance. Table 6 demonstrates the effectiveness of the proposed HMM in prediction.

Based on the trajectory of latent status, we can cluster the patients for better treatment. This is a by-product of the proposed method. We use *K*-means clustering method to group the similar patients. To better visualize the cluster results, we first apply principle component analysis on the trajectory of latent status, and then plot the clustering results in Figure 7. In Figure 7, the X-axis and Y-axis are the first and second principle components, respectively. We can see that all patients are clearly partitioned into 4 clusters. Next, similar to the procedure of testing the importance of demographic data in Section 3.1, we also conduct hypothesis to test the deviations between clusters. The goal of the hypothesis is to tell whether the data from different clusters are indeed different. We use the F-testing and T-testing to tell whether the variance and mean of different clusters are different, respectively. The final test results are obtained by following Equation (2) which combines the results of both F-test and T-test. The results are reported in Table 7. The first two rows of Table 7 show that (1) $C_1 \nsim C_3$ and $C_1 \nsim C_4$, (2) $C_2 \nsim C_3$ and $C_2 \nsim C_4$.

Table 7: Hypothesis testing results on the clustering results.

	C_1	C_2	C_3	C_4
C_1	0	0	1	1
C_2	0	0	1	1
C_3	1	1	0	0
C_4	1	0	1	0

5 RELATED WORK.

Latent variable model [4, 9, 15, 16] is a statistical model that relates a set of observed variables to a set of latent ones. One popular model is hidden Markov model (HMM), which was first introduced in the 1970s as a tool for speech recognition [7, 18]. Recently, the popularity of HMM has increased in bioinformatics domain [1, 6, 17, 19, 22, 26] primarily because of its strong mathematical basis and the ability to adapt to unknown data. In [6], a data clustering algorithm based on a single HMM has been proposed to identify the number of clusters in a dataset and also label the data item to its respective clusters. [17] use a HMM with four exercise levels as hidden states and the blood glucose levels (and insulin dosages) as observable data for activity prediction task. In [19], the authors use HMMs to characterize disability states and use mixed-effects ordinal logistic regression to estimate the probability of functional decline for type 2 diabetes patients. [26] proposes to model the linear block dependency in the SNP data using HMMs and further develop a compound decision-theoretic framework for testing HMM-dependent hypothesis. All these work only deals with behavioral data and does not consider demographic data. In our proposed DfrHMM, however, we propose to incorporate the demographic data into the link function between transition status.

Another relevant topic is diabetes prediction, which has attracted lots of attention in past decades [8, 11, 13, 14, 20, 21, 23]. In [23], the authors identify that a 62-locus genotype risk score ($\mathrm{GRS_t}$) can improve type 2 diabetes. [21] propose to use multiple metabolic and genetic markers to improve the prediction of type 2 diabetes. In contrast, in this paper we propose a novel framework DfrHMM to predict the behavioral data for diabetes patients, which is not discussed in these works.

This paper is also related to multi-source or multi-view task learning [2, 3, 10, 12, 27, 29]. In [2, 27, 29], the authors propose to simultaneously decompose dynamic data from multiple sources with consensus constraints for anomaly detection task.These work clusters the timestamps into a pre-fixed number of clusters, which ignores the importance of the order of timestamps. In contrast, in the proposed DfrHMM, we use the link function to model the transition of status between every different pair of timestamps. The importance of the timestamp order is naturally captured via DfrHMM. In [3, 10, 12, 28], different models for multi-view clustering are proposed. However, these work treat the data from different views or different sources equally important.

6 CONCLUSIONS.

In this paper, we propose a Demographic feature restricted hidden markov model (DfrHMM) based framework for the task of diabetes management by combining both behavioral and demographic data. Using the proposed DfrHMM model, we can learn the trajectory of latent status for each patient. Different from conventional HMM, we take the demographic data into consideration when modeling the transition between different status. To estimate the model parameters, we propose to use Markov Chain Monte Carlo techniques. Moreover, we conduct hypothesis testing on the real data, Diab-1000, to test the importance of demographic data on the patients' behavioral data. Experiments on synthetic and Dial-1000 data demonstrates the effectiveness of the proposed DfrHMM.

7 ACKNOWLEDGEMENTS.

We would like to thank the anonymous reviewers for their valuable comments, which help us tremendously in improving the quality of the paper's quality. We also thank Dr. Shiyu Wang from University of Georgia for her discussions and suggestions for the modeling design and solution development. Any opinions,findings, and conclusions or recommendations expressed in this paper are those of the authors.

REFERENCES

[1] Nicola Bartolomeo, Paolo Trerotoli, and Gabriella Serio. 2011. Progression of liver cirrhosis to HCC: an application of hidden Markov model. *BMC Med. Res. Methodol.* 11, 1 (2011), 38.

[2] Alain E Biem, Jing Gao, Deepak S Turaga, Long H Vu, and Houping Xiao. 2015. Unsupervised multisource temporal anomaly detection. (2015). US Patent App.

[3] Kamalika Chaudhuri, Sham M Kakade, Karen Livescu, and Karthik Sridharan. 2009. Multi-view clustering via canonical correlation analysis. In *Proc. of ICML.*

[4] Bruce E Compas, Margaret C Boyer, Catherine Stanger, Richard B Colletti, Alexandra HThomsen, Lynette M Dufton, and David A Cole. 2006. Latent variable analysis of coping, anxiety/depression, and somatic symptoms in adolescents with chronic pain. *J. Consult. Clin. Psychol.* 74, 6 (2006).

[5] Centers for Disease Control, Prevention, and others. 2014. National diabetes statistics report: estimates of diabetes and its burden in the United States. *Atlanta, GA: US Department of Health and Human Services* 2014 (2014).

[6] Md Rafiul Hassan, Baikunth Nath, and Michael Kirley. 2006. A data clustering algorithm based on single hidden markov model. In *Proc. of ISSN,* Vol. 1896. 7094.

[7] Xuedong D Huang, Yasuo Ariki, and Mervyn A Jack. 1990. *Hidden Markov models for speech recognition.* Vol. 2004.

[8] T Kimpimaki, Antti Kupila, A-M Hamalainen, Marika Kukko, Petri Kulmala, Kaisa Savola, Tuula Simell, P Keskinen, Jorma Ilonen, Olli Simell, and others. 2001.The first signs of β-cell autoimmunity appear in infancy in genetically susceptible children from the general population: the Finnish Type 1 Diabetes Prediction and Prevention Study. *J. Clin. Endocrinol. Metab.* 86, 10 (2001), 4782–4788.

[9] Martin Knott and David J Bartholomew. 1999. *Latent variable models and factor analysis.* Number 7.

[10] Abhishek Kumar, Piyush Rai, and Hal Daume. 2011. Co-regularized multi-view spectral clustering. In *Proc. of NIPS.* 1413–1421.

[11] A Kupila, P Muona, T Simell, P Arvilommi, H Savolainen, A-M Hämäläinen, S Korhonen, T Kimpimäki, M Sjöroos, J Ilonen, and others. 2001. Feasibility of genetic and immunological prediction of type I diabetes in a population-based birth cohort. *Diabetologia* 44, 3 (2001), 290–297.

[12] Jialu Liu, Chi Wang, Jing Gao, and Jiawei Han. 2013. Multi-view clustering via joint nonnegative matrix factorization. In *Proc. of SDM.* 252–260.

[13] Maria Lönnrot, Karita Korpela, Mikael Knip, Jorma Ilonen, Olli Simell, Sari Korhonen, Kaisa Savola, Päivi Muona, Tuula Simell, Pentti Koskela, and others. 2000. Enterovirus infection as a risk factor for beta-cell autoimmunity in a prospectively observed birth cohort: the Finnish Diabetes Prediction and Prevention Study. *Diabetes* 49, 8 (2000), 1314–1318.

[14] Fenglong Ma, Radha Chitta, Jing Zhou,Quanzeng You, Tong Sun, and Jing Gao. 2017. Dipole: Diagnosis Prediction in Healthcare via Attention-based Bidirectional Recurrent Neural Networks. In *Proc. of KDD.*

[15] Fenglong Ma, Chuishi Meng, Houping Xiao, Qi Li, Jing Gao, Lu Su, and Aidong Zhang. 2017. Unsupervised Discovery of Drug Side-Effects From Heterogeneous Data Sources. In *Proc. of KDD.*

[16] Bengt Muthén. 2004. Latent variable analysis. *The Sage Handbook ofQuantitative Methodology for the Social Sciences* (2004), 345–68.

[17] Davy Preuveneers and Yolande Berbers. 2008. Mobile phones assisting with health self-care: a diabetes case study. In *Proc. of MobileHCI.* 177–186.

[18] Lawrence Rabiner and B Juang. 1986. An introduction to hidden Markov models. *IEEE ASSP Magazine* 3, 1 (1986), 4–16.

[19] W Jack Rejeski, Edward H Ip, Alain G Bertoni, George A Bray, Gina Evans, Edward W Gregg, and Qiang Zhang. 2012. Lifestyle change and mobility in obese adults with type 2 diabetes. *N. Engl. J. Med.* 366, 13 (2012).

[20] Eugene P Rhee, Susan Cheng, Martin G Larson, Geoffrey A Walford, Gregory D Lewis, Elizabeth McCabe, Elaine Yang, Laurie Farrell, Caroline S Fox, Christopher J O'Donnell, and others. 2011. Lipid profiling identifies a triacylglycerol signature of insulin resistance and improves diabetes prediction in humans. *J. Clin. Invest.* 121, 4 (2011), 1402–1411.

[21] Matthias B Schulze, Cornelia Weikert, Tobias Pischon, Manuela M Bergmann, Hadi Al-Hasani, Erwin Schleicher, Andreas Fritsche, Hans-Ulrich Häring, Heiner Boeing, and Hans-Georg Joost. 2009. Use of multiple metabolic and genetic markers to improve the prediction of type 2 diabetes: the EPIC-Potsdam Study. *Diabetes care* 32, 11 (2009), 2116–2119.

[22] Adam Siepel and David Haussler. 2004. Combining phylogenetic and hidden Markov models in biosequence analysis. *J. Comp. Biol.* 11, 2-3 (2004), 413–428.

[23] Jason L Vassy, Marie-France Hivert, Bianca Porneala, Marco Dauriz, Jose C Florez, Josée Dupuis, David S Siscovick, Myriam Fornage, Laura J Rasmussen-Torvik, Claude Bouchard, and others. 2014. Polygenic type 2 diabetes prediction at the limit of common variant detection. *Diabetes* (2014), DB_131663.

[24] Shiyu Wang and Jeff Douglas. 2015. Consistency of nonparametric classification in cognitive diagnosis. *Psychometrika* 80, 1 (2015), 85–100.

[25] Shiyu Wang, Yan Yang, Steven Culpepper, and Jeff Douglas. 2016. Tracking skill acquisition with cognitive diagnosis models: A higher-order hidden markov model with covariates. *Submitted Manuscript* (2016).

[26] Zhi Wei, Wenguang Sun, Kai Wang, and Hakon Hakonarson. 2009. Multiple testing in genome-wide association studies via hidden Markov models. *Bioinformatics* 25, 21 (2009), 2802–2808.

[27] Houping Xiao, Jing Gao, Deepak S Turaga, Long H Vu, and Alain Biem. 2015. Temporal multi-view inconsistency detection for network traffic analysis. In *Proc. of WWW.* 455–465.

[28] Houping Xiao, Jing Gao, Long Vu, and Deepak S Turaga. 2017. Detecting Malicious Behavior in Computer Networks via Cost-Sensitive and Connectivity Constrained Classification. In *Proc. of SDM.*

[29] Houping Xiao, Yaliang Li, Jing Gao, Fei Wang, Liang Ge, Wei Fan, Long H Vu, and Deepak S Turaga. 2015. Believe it today or tomorrow? detecting untrustworthy information from dynamic multi-source data. In *Proc. of SDM.* 397–405.