A Unifying Variational Inference Framework for Hierarchical Graph-Coupled HMM with an Application to Influenza Infection

Kai Fan
Duke University
Durham, NC 27708
kai.fan@stat.duke.edu

Chunyuan Li
Duke University
Durham, NC 27708
chunyuan.li@duke.edu

Katherine Heller
Duke University
Durham, NC 27708
kheller@stat.duke.edu

Abstract

The Hierarchical Graph-Coupled Hidden Markov Model (hGCHMM) is a useful tool for tracking and predicting the spread of contagious diseases, such as influenza, by leveraging social contact data collected from individual wearable devices. However, the existing inference algorithms depend on the assumption that the infection rates are small in probability, typically close to 0. The purpose of this paper is to build a unified learning framework for latent infection state estimation for the hGCHMM, regardless of the infection rate and transition function. We derive our algorithm based on a dynamic auto-encoding variational inference scheme, thus potentially generalizing the hGCHMM to models other than those that work on highly contagious diseases. We experimentally compare our approach with previous Gibbs EM algorithms and standard variational method mean-field inference, on both semi-synthetic data and app collected epidemiological and social records.

Introduction

The mainstay of computational epidemiology research for predicting patterns or discovering the risks of infectious disease has been at the population level. A prominent example is the Google flu trend website, which, leveraging the temporal searching key words associating with flu, forecasted the flu outbreak two weeks ahead of Center of Disease Control (CDC). However, the advent of personal health apps in mobile or wearable devices allows disease diffusion to be modeled in an individual level, e.g. data from cell phones is harnessed to identify the probability of contracting flu for every single person in a relatively close community (Dong, Pentland, and Heller 2012; Fan et al. 2015). They both incorporate the face-to-face contact information within a local area to adjust the transmission function and construct their models, thus constructing a hierarchical model with extra features with an enriched epidemiology dataset for more accurate prediction. The main idea in these works is to build a model in a fully Bayesian setting taking the advantage of efficient Gibbs sampling. However, the nice conjugacy property of such complicated models, to a great extent, relies on a small infection rate since a Taylor expansion trick is applied. This paper aims to break this assumption and develop a unifying inference method for a non-conjugate disease diffusion model.

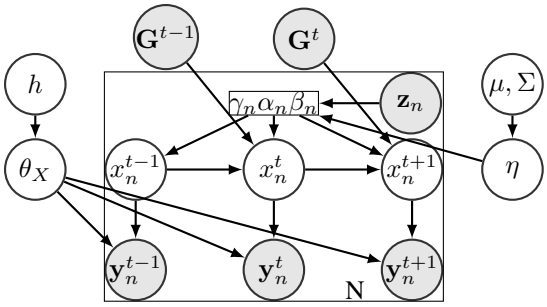


Figure 1: A template graphical representation of hGCHMM. See an unrolled but equivalent version in supplementary.

Along the lines of (Fan, Aiello, and Heller 2015), we adopt the hGCHMM model (Fig. 1) where the binary latent variables indicate whether the person is infected, and the observed variables are a binary vector which indicate multiple symptoms. The colored nodes are observables and the nodes $\mathbf{G}^t$ representing the dynamic social network, implicitly drawn (see explicit graph in(Fan et al. 2015) or supplementary[1]). The reason we emphasize this model is that the hGCHMM is a general and flexible framework to heterogeneously simulate the disease spread in a dynamic social network. It simulates a discretized Suspicious-Infections-Suspicious (SIS) model (Cooper and Lipsitch 2004), evolved from standard hidden Markov process, and is able to reduce to a homogeneous version, the Graph-coupled HMM or even latent Dirichlet allocation (LDA) (Gruber, Weiss, and Rosen-Zvi 2007). The generative process of Fig. 1 follows

$$\begin{aligned} \alpha_n, \beta_n, \gamma_n &\sim \text{Beta}\left(e^{\mathbf{z}_n^\top \boldsymbol{\eta}_{\cdot,1}}, e^{\mathbf{z}_n^\top \boldsymbol{\eta}_{\cdot,1}}\right) \\ x_n^t &\sim \text{Bernoulli}(\phi(\alpha_n, \beta_n, \gamma_n; \mathbf{x}^{t-1}, \mathbf{G}^{t-1})) \\ y_{n,s}^t &\sim \text{Bernoulli}(\theta_{x_n^t, s}) \end{aligned} \quad (1)$$

where the subscript $\cdot$ means either α, β or γ, and transition function ϕ takes the arguments $\alpha_n, \beta_n, \gamma_n$ and depends on

[1] http://people.duke.edu/~kf96/docs/aaai2016supp.zip

all the latent states $\mathbf{x}^{t-1}$ (through turning $x_{1:N}^{t-1}$ into a column vector) coupled with social network $\mathbf{G}^{t-1}$ (a binary symmetric matrix) at the previous timestamp. In order to relate to epidemiological models, γ means the recovery probability if infectious at previous timestamp; α represents the probability of being infected from some one outside the network while β represents the probability of being infected from some one inside the network. Thus we can readily write ϕ as $p_n^t(1 \to 0) = \gamma_n$ and

$$p_n^t(0 \to 1) = 1 - (1 - \alpha_n)(1 - \beta_n)^{C_n^{t-1}} \quad (2)$$

where $C_n^{t-1} = \sum_{m:(m,n)\in G_{t-1}} \mathbb{I}_{\{x_m^{t-1}=1\}}$, in other words, C_n^{t-1} means the total number of infected people that person n had contacted with during the time interval $[t-1, t)$. This construction intuitively implies that if you have more friends with flu, you have a higher risk of becoming infected. This transition matrix makes biological sense but induces non-conjugacy. The previous studies circumvent this problem by using a Taylor expansion (Dong, Pentland, and Heller 2012) or polynomial decomposition (Fan et al. 2015) to approximate (2) and then apply Gibbs sampling, which gives acceptable performance if α and β are sufficiently small.

In our paper, we propose a novel learning approach in the variational inference (VI) framework. Differing from the traditional VI methods, such as mean-field (Beal 2003), we get rid of the complicated gradient computation for non-conjugate probabilistic models but build a tractable dynamic recognition model corresponding to the generative process. The recognition model is equivalent to the variational distribution in VI parlance. Therefore we can minimize the Kullback-Leibler (KL) divergence or maximize the corresponding lower bound for parameter estimation. Taking advantage of the binary variables in the hGCHMM, the recognition model can be completely constructed by a sigmoid belief network (SBN) (Sutskever and Hinton 2008), where it is straightforward to generalize to the categorical case by inducing softmax output, thus allowing our approach to reproduce the potential tasks introduced in (Fan et al. 2015). In addition, we overcome the major drawback – lacking the capability to simulate highly contagious disease, by proposing a dynamic auto-encoding variational inference method. In the experiments, we compare our algorithm with the Gibbs EM (Fan et al. 2015) and mean-field (Beal 2003) versions, and achieve competitive performance in the small infection rate case but outperform them in the large rate scenario.

Related Work

The idea of utilizing extra features like social contacts to enrich the epidemiology dataset has been extensively researched. (Christley et al. 2005) incorporated a fixed social network analysis on susceptible-infectious-recovered (SIR) models to identify high-risk individuals. (Salathé et al. 2010)'s work on close proximity interactions (CPIs) of dynamic social networks in a high school indicated immunization strategies are more credible if extra contact data is provided. Our studied model was first developed by (Fan et al. 2015) for heterogeneous personalized health data. The purpose of this work is to broaden the application range of this model to diversified epidemics even with larger infection rate. The novel auto-encoding variational inference method used here originated from the deep learning community. A variety of papers (Kingma and Welling 2014; Mnih and Gregor 2014; Rezende, Mohamed, and Wierstra 2014; Kingma et al. 2014) put forward generative neural networks and a recognition model by reversing the direction of the link between latent variables and visible variables, and further employed the stochastic variational inference to optimize the parameters of both models simultaneously. Unlike their work, which was mainly for a fixed network structure, we use a dynamic framework by introducing a Markov process with time-dependence between the temporal auto-encoders. Additionally, it is unnecessary to assume any conjugacy in the probabilistic model. We can directly use SBN to construct directed edges mapping to a binary variable. However, we prefer a gradient based parameter optimization method, such as Adagrad (Duchi, Hazan, and Singer 2011) or Rmsprop (Dauphin et al. 2015), rather than the slower wake-sleep algorithm (Hinton, Osindero, and Teh 2006).

Limitation of Gibbs Sampling

ϕ Approximation by Auxiliary Variable

Eq. (2) is an exception unsatisfying the Bernoulli-Beta conjugate in generative model (1). By inducing auxiliary variable R, which indicates the infection source from outside, inside or both of the surveyed community, an approximate Gibbs scheme can be developed. Particularly, (Dong, Pentland, and Heller 2012) used a simple Taylor expansion $\alpha_n + C_n^t\beta_n$ to represent (2), and (Fan et al. 2015) further applied a polynomial decomposition trick by rewriting it as the summation of three terms,

$$\alpha_n(1-\beta_n)^{C_n^{t-1}} + C_n^{t-1}(1-\alpha_n)\beta_n + C_n^{t-1}\alpha_n\beta_n \quad (3)$$

where $\left(1-(1-\beta_n)^{C_n^{t-1}}\right)$ is approximated by $C_n^{t-1}\beta_n$. Admittedly, (3) is a better estimation of (2) than $\alpha_n + C_n^t\beta_n$, and also favors the Bernoulli-Beta conjugate for potential Gibbs sampling. We have to notice that if the condition $\alpha, \beta \approx 0$ does not hold, the resulted error of (3) will be crucially non-negligible and violate the entire Bayesian scheme, thus leading a biased estimation. In Fig. 2, we plot the infection probability as the function of α, β with difference approximation.

The visualized comparison shows when α or β is sufficiently large, the approximation is not even a probability at all, which should be bounded by 1. One plausible solution is to rescale, but this hack trick lacks any theoretical or intuitive explanation (the rescaled figure is shown in supplementary). Therefore, the Taylor expansion restricts the generalization ability of this model, so we are intent to modify the inference method by using variational inference.

Sigmoid Variational Inference

Variational Inference Basics

Suppose we are interested in a latent variable model represented as distribution $P_\Phi(\mathbf{X}, \mathbf{Y})$ parametrized by Φ, where

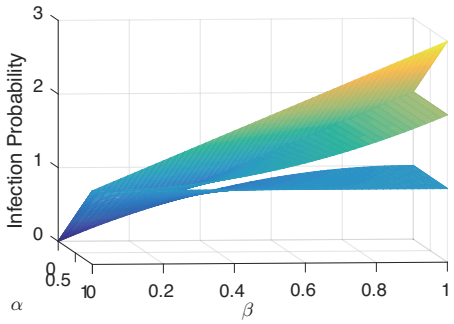


Figure 2: The bottom surface is the exact plot for $p_n^t(0 \to 1)$; the middle one is the approximation by (3); the top one is the approximation by $\alpha + C\beta$. In this plot, we set $C = 2$.

X is latent variable and Y is observed variable. The purpose of learning task is to estimate the posterior of latent variable $P_\Phi(\mathbf{X}|\mathbf{Y})$ and Φ. In most case, the exact inference is intractable and thus a variational lower bound on the marginal log-likelihood is often derived to be maximized, because optimization problem can be naively solved by gradient based algorithm. Particularly, we need induce a variational distribution $Q_\Psi(\mathbf{X}|\mathbf{Y})$ with parameters Ψ, which is selected with the intention of being similar to the true posterior $P_\Phi(\mathbf{X}|\mathbf{Y})$. A prevalent construction of Q will usually be assumed to factorize over some partition of the latent variables, i.e mean-field method.

With simple mathematical derivation, we have

$$\begin{aligned} &\log P(\mathbf{Y}) \qquad (4)\\ =&\mathbb{E}_Q\left[\log \frac{P_\Phi(\mathbf{X},\mathbf{Y})}{Q_\Psi(\mathbf{X}|\mathbf{Y})}\right] + KL(Q_\Psi(\mathbf{X}|\mathbf{Y})||P_\Phi(\mathbf{X}|\mathbf{Y}))\\ &\geq \mathbb{E}_Q\left[\log P_\Phi(\mathbf{X},\mathbf{Y}) - \log Q_\Psi(\mathbf{X}|\mathbf{Y})\right] \triangleq \mathcal{L}(\mathbf{Y},\Phi,\Psi)\end{aligned}$$

For the lower bound $\mathcal{L}$, maximizing it with respect to Φ, Ψ is equivalent to minimizing the KL divergence between proposed distribution and true posterior. The tightness of this bound holds when Q exactly recovers true posterior. By examining the bound, it does not rely on the form of Q. In our work, the variational distribution is restricted to belong to a family of distributions of simpler form than $P_\Phi(\mathbf{X}|\mathbf{Y})$, but preferably flexible enough to contain or close the true posterior as a solution. To be specific, we reverse the directed edge between latent and observed variable in generative process and enforce a sigmoid belief nets to simulate this link.

Dynamic Recognition Model

Before we describe the details of our recognition model, we first revisit the generative model and reformulate part of the process to be deterministic. Due to the fact that the expectation of distribution Beta $\left(e^{\mathbf{z}_n^\top \boldsymbol{\eta}_{\cdot,1}}, e^{\mathbf{z}_n^\top \boldsymbol{\eta}_{\cdot,1}}\right)$ is $\sigma\left((\boldsymbol{\eta}_{\cdot,1} - \boldsymbol{\eta}_{\cdot,2})^\top \mathbf{z}_n\right)$, where $\sigma(x) = \frac{1}{1+e^{-x}}$ is sigmoid function, we can simplify the parameterization by substituting a single $\boldsymbol{\eta}_\cdot$ for $\boldsymbol{\eta}_{\cdot,1} - \boldsymbol{\eta}_{\cdot,2}$ if we apply a sigmoid belief net with input $\mathbf{z}_n$. One advantage of this approximation is to reduce the number of parameters in generative model since $\alpha_n, \beta_n, \gamma_n$ can implicitly vanish from the arguments in ϕ, leading to a more direct form $\phi(\boldsymbol{\eta}_\cdot, \mathbf{z}_n)$. In the contrary, the disadvantage is apparently the deficiency of uncertainty for these intermediate variables. However, point estimation of associating parameters is also acceptable in the medical application, since the most concerned issue is to discover heterogenous infection probability on every single day and how the individual covariates $\mathbf{z}_n$ influence the personal physical constitution. The advantage of sigmoid formulation is to benefit our subsequent gradient based algorithm. Analogously, we can rewrite the generative process of $y_{n,t,s}$ by using sigmoid function, i.e. $P(y_{n,s}^t = 1) = \theta_{0,s}\mathbb{I}_{\{x_n^t=0\}} + \theta_{1,s}\mathbb{I}_{\{x_n^t=1\}} = \sigma(w_s x_n^t + b_s)$, where w_s and b_s become model parameters in Φ.

With the reformulated generative process, we can readily construct the a simple dynamic recognition model $Q_\Psi(\mathbf{X}|\mathbf{Y},\mathbf{Z})$ by applying sigmoid belief nets. Notice that for clarity we write the observed variables $\mathbf{Y}, \mathbf{Z}$ separately, which is a slightly different from the general discussion.

$$x_{n,t} \sim \text{Bernoulli}\left(\sigma(\boldsymbol{\omega}^\top \tilde{\mathbf{x}}_n^{t-1} + \boldsymbol{\nu}^\top \mathbf{y}_n^t + \boldsymbol{\kappa}^\top \mathbf{z}_n + b)\right) \quad (5)$$

where the vector $\tilde{\mathbf{x}}_n^{t-1} = \mathbf{g}_n^{t-1} \odot \mathbb{I}_{\mathbf{x}^{t-1}=1}$, and $\mathbf{g}_n^{t-1}$ is the nth column of $\mathbf{G}^{t-1}$ and $\odot$ is element-wise multiplication operator. (5) indicates the graphical representation of recognition model has only two modifications of Fig.1: getting rid of redundant intermediate variable by inducing link between $\mathbf{z}_n$ and x_n^t; reversing the direction from x_n^t to $\mathbf{y}_n^t$. The powerful approximation ability of sigmoid belief nets allows it to obtain perfect estimation of any transition function. This is key reason why we did not make the same assumption as (Fan et al. 2015). Additionally, we can further plug in one more hidden layer $\mathbf{h}_n^t$ following the convention of (5), and then let $x_{n,t} \sim \text{Bernoulli}\left(\sigma(\boldsymbol{\omega}_h^\top \mathbf{h}_n^t + b_h)\right)$. The deep architecture of networks can enlarge the representative ability, and succeeds with great improvement in many machine learning areas, such as collaborative filtering (Salakhutdinov, Mnih, and Hinton 2007), or document modeling (Srivastava, Salakhutdinov, and Hinton 2013). In our paper, deep structure is not the main issue we discuss, since we found it had no significant improvement and brought more variational parameters to estimate.

Parametrization and Optimization

In previous section, we did not specify the index or subscript of variational parameters $\boldsymbol{\omega}, \boldsymbol{\nu}, \boldsymbol{\kappa}, b$ in recognition model. By observing the factorized structure of generative model (6), we analyze different parameterization possibilities for (5).

$$P_\Phi(\mathbf{X},\mathbf{Y}) = \prod_{n=1}^{N}\prod_{t=1}^{T} p(\mathbf{y}_n^t|x_n^t)p(x_n^t|\tilde{\mathbf{x}}_n^{t-1}) \qquad (6)$$

where we hide the dependence on $\mathbf{z}_n$ for simplicity but without ambiguousness, and x_n^0 can either be non-exsited node or the initial node without emission. Therefore, according to previous discussion, the corresponding recognition model for approximating $P_\Phi(\mathbf{X}|\mathbf{Y})$ has the following form.

$$Q_\Psi(\mathbf{X}|\mathbf{Y}) = \prod_{n=1}^{N}\prod_{t=1}^{T} q(x_n^t|\tilde{\mathbf{x}}_n^{t-1}, \mathbf{y}_n^t) \qquad (7)$$

Algorithm 1 Temporal Sigmoid Variational Inference

Initialization: By normal distribution with small variance;
1: **while** (Φ, Ψ, Ξ) Not Converge **do**
2: **for** $t = 1, \ldots, T$ **do**
3: Sample $x^t_{1:N} \sim q(x^t_{1:N}|x^{t-1}_{1:N}, \mathbf{y}^t_{1:N})$;
4: Compute temporal learning signal l^t_Ψ;
5: Subtract baseline $l^t_\Psi \leftarrow l^t_\Psi - B_{\Xi_t}(\mathbf{y}^t_{1:N})$;
6: $\Psi_t \leftarrow \Psi_t + \epsilon \cdot l^t_\Psi \nabla_{\Psi_t} \log q(x^t_{1:N}|x^{t-1}_{1:N}, \mathbf{y}^t_{1:N})$;
7: $\Xi_t \leftarrow \Xi_t + \epsilon \cdot l^t_\Psi \nabla_{\Xi_t} \log B_{\Xi_t}(\mathbf{y}^t_{1:N})$;
8: **end for**
9: $\Phi \leftarrow \Phi + \epsilon \cdot \sum_{t=1}^T \nabla_\Phi \log p(x^t_{1:N}, \mathbf{y}^t_{1:N})$;
10: **end while**

In order to take the advantage of variational inference principle established in (4), an ideal factor by factor optimization between the logarithm form of P and Q inspires three main parameterization methods.

Temporal Parametrization We emphasize the description on the most natural temporal parametrization and demonstrate our learning signal (Sutton and Barto 1998) based optimization algorithm. In this setting, each factor q can be formulated as

$$q(x^t_n = 1) = \sigma(\boldsymbol{\omega}_t^\top \tilde{\mathbf{x}}^{t-1}_n + \boldsymbol{\nu}_t^\top \mathbf{y}^t_n + \boldsymbol{\kappa}_t^\top \mathbf{z}_n + b_t) \quad (8)$$

This allows the dynamic parameters shared by different chains, and is equivalent to train the model with a unique datapoint $\mathbf{Y}_{N\times T\times S}$. The reason is the temporal parameters $\Psi_t = \{\boldsymbol{\omega}_t, \boldsymbol{\nu}_t, \boldsymbol{\kappa}_t, b_t\}$ can be updated locally. Each set of temporal parameters is merely associated with the correspondent observed data $\mathbf{y}^t_{1:N}$. To make this argument concrete, it can be shown by integration by parts for the derivative of lower bound (4). For notation simplicity, we denote the learning signal $l_\Psi = \log P_\Phi - \log Q_\Psi$, and then have the following temporal decomposition.

$$\begin{aligned}\nabla_{\Psi_t}\mathcal{L} =& \mathbb{E}_Q[l_\Psi \nabla_{\Psi_t} \log q(x^t_n = 1|\tilde{\mathbf{x}}^{t-1}_n, \mathbf{y}^t_n)] \\ =& \mathbb{E}_{q(x^{1:t-1}_{1:N}|\mathbf{y}^{1:t-1}_{1:N})}\Big[\mathbb{E}_{q(x^{t:T}_{1:N}|x^{t-1}_{1:N}, \mathbf{y}^{t:T}_{1:N})} \\ & \big[l_\Psi \cdot \nabla_{\Psi_t} \log q(x^t_n = 1|\tilde{\mathbf{x}}^{t-1}_n, \mathbf{y}^t_n) \mid x^{t-1}_{1:N}\big]\Big] \quad (9)\end{aligned}$$

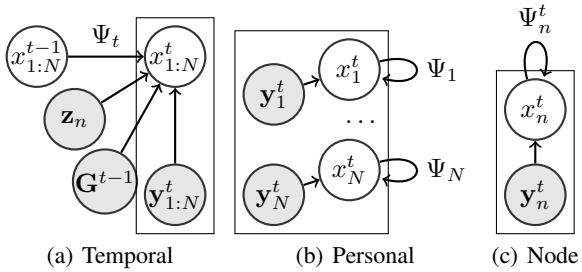
Additionally, (9) can also enable the possibility of temporal learning signal, $l^t_\Psi = \log P\left(x^{t:T}_{1:N}, \mathbf{y}^{t:T}_{1:N}|x^{t-1}_{1:N}\right) - \log Q\left(x^{t:T}_{1:N}|\mathbf{y}^{t:T}_{1:N}, x^{t-1}_{1:N}\right)$. Thus, Ψ_t can be locally updated by l^t_Ψ. To fully adopt the reinforcement learning trick, we also induce an observation dependent but latent variable independent signal baseline $B_{\Xi_t}(\mathbf{y}^t_{1:N})$ temporally parametrized by Ξ_t, which is also implemented by SBN. Since the identity property

$$\mathbb{E}_Q[(l_\Psi - B)\nabla_{\Psi_t} \log q_{\Psi_t}] = \mathbb{E}_Q[l_\Psi \nabla_{\Psi_t} \log q_{\Psi_t}] \quad (10)$$

it works practically in gradient variance reduction (Mnih and Gregor 2014; Mnih et al. 2015). Therefore, we summarize the Algorithm 1. It is noticed that we only provide the basic gradient ascent algorithm, whereas many existed advanced trick can be explored as well, such as the adaptive learning rate, RMSprop, or Adagrad. Compared with mean-field

Algorithm 2 Personal Sigmoid Variational Inference

1: **while** (Φ, Ψ, Ξ) Not Converge **do**
2: **for** $t = 1, \ldots, T$ **do**
3: Sample $x^t_{1:N} \sim q(x^t_{1:N}|x^{t-1}_{1:N}, \mathbf{y}^t_{1:N})$;
4: Compute temporal learning signal l^t_Ψ;
5: Subtract baseline $l^t_\Psi \leftarrow l^t_\Psi - B_\Xi(\mathbf{y}^t_{1:N})$;
6: **end for**
7: Normalize $\{l^t_\Psi\}^T_{t=1}$;
8: $\nabla_{\Psi_n}\mathcal{L} = \sum_{t=1}^T l^t_\Psi \nabla_{\Psi_n} \log q(x^t_n|\tilde{\mathbf{x}}^{t-1}_n, \mathbf{y}^t_n)$;
9: $\nabla_\Xi \mathcal{L} = \sum_{t=1}^T l^t_\Psi \nabla_\Xi \log B_\Xi(\mathbf{y}^t_{1:N})$;
10: Update Φ (as Alg. 1), Ξ and Ψ_n by gradient ascent;
11: **end while**



(a) Temporal (b) Personal (c) Node

Figure 3: Recognition models, where (b) and (c) omit some side information nodes.

algorithm, the burden of derivative computation in our algorithm is relatively simple. Due to the limited space, the detailed derivation shows in the supplementary materials. Even if we deploy a deep SBN, the convenience of deriving sigmoid function makes our algorithm efficient. As we discussed before, if a hidden layer plugs into the SBN, we can actually sample this layer similarly as Line 3 in Algorithm 1. When it comes to compute the gradient associated with hidden layer, it is not necessary to use the back-propagation, which is a standard method in feed forward neural networks. We merely need to compute the gradient of sigmoid function twice, since given the sampled hidden layer the input and the output are independent (see supplementary).

Personal Parametrization Analogously, the subscript of variational parameters can be heterogeneously indexed by n.

$$q(x^t_n = 1) = \sigma(\boldsymbol{\omega}_n^\top \tilde{\mathbf{x}}^{t-1}_n + \boldsymbol{\nu}_n^\top \mathbf{y}^t_n + \boldsymbol{\kappa}_n^\top \mathbf{z}_n + b_n) \quad (11)$$

The setting means the variational parameters $\Psi = \{\boldsymbol{\omega}_n, \boldsymbol{\nu}_n, \boldsymbol{\kappa}_n, b_n\}^N_{n=1}$ would not change dynamically but differ chain by chain, thus being equivalent to train the model with dataset $\{\mathbf{y}_{1:N}\}^T_{t=1}$. It also indicates Ψ can not be trained temporally as previous discussion but allows centralization and normalization of learning signal for variance reduction. For Algorithm 2, the baseline becomes global, and l^t_Ψ has the different meaning from previous one.

Node-wise Parametrization More specifically, we can even construct the node-wise parameterization $\Psi^t_n = \{\boldsymbol{\omega}^t_n, \boldsymbol{\nu}^t_n, \boldsymbol{\kappa}^t_n, b^t_n\}$. The formulation and the detailed algorithm framework are similar to a combination of previous

two settings, so we include the **Algorithm** in the supplementary. All three parameterized recognition models do not assume any quantitative magnitude on the parameters, and the only approximation induced in this framework is due to the inequality (4). This gap is hopefully filled by the gradient based optimization algorithm. Furthermore, the recognition model of hGCHMM is particularly counter-intuitive, since its parameterization shows a trade-off between the complexity of graphical model and the number of variational parameters. To approximate the same true posterior, our approach is to propose either a simple model with more parameters or complex one with less parameters (See Fig. 3 for a graphcial illustration). In our experiments, the $\mathcal{L}_2$ norm penalty on parameters is also implemented.

Experiments

Data Description

We apply our inference method to two flu diffusion datasets. The first experiment is based on MIT social evolution data. The dynamic social contacts can be summarized from the daily bluetooth data, thus resulting in $\mathbf{G}_t$, $t = 1, \ldots, 107$. In addition, the personal health habits $\mathbf{z}_n$ for each person contain 9 features, weight, height, salads per week, veggies fruits per day, healthy diet level, aerobics per week, sports per week, smoking indicator, and default feature 1. Since the ground truth of infectious state is latent variable or unknown, we need to simulate consistent $\mathbf{X}$ and $\mathbf{Y}$ for evaluation by generative model, though self-reported symptom $\mathbf{Y}$ is in fact provided in the dataset.

Another dataset is exFlu survey. This study is conducted in a college dormitory during a chain referral recruitment process carried out from September 2012 to January 2013. 103 enrolled students participated flu survey by the smartphone installed with health apps. Besides the covariates described in MIT data, the temporally unchanged features of exFlu also include gender, age, average times of hand washing by sanitizer, and indicator for vaccination or flu shot. The data type for dynamic social networks $\mathbf{G}$ and daily symptoms $\mathbf{Y}$ is exactly the same as above dataset. However, this survey has a special treatment on participants with severe symptoms to diagnose whether the specific person is infected and record the flu duration since onsite. Thus, it allows us to evaluate the performance of our approach comparing with expertise.

Partially Synthetic Datasets

In this experiment, we mainly study the generalization ability of our approach for various diseases by synthesizing 3 different infection rates. The data we used in this section is partially simulated MIT data with true $\mathbf{G}_t$ and $\mathbf{z}_n$. Since the infection rate is crucially heterogenous in our model, the magnitude mentioned later or the severity of contagiosity is virtually the mean value of these person-specific rates. Basically, the three contagious diseases are usual flu with recovery rate 0.3 and low infection rate 0.01 (outside) and 0.02 (inside), severe flu (such as H1N1) with lower recovery rate 0.2 and high infection rate 0.34 and 0.24, and a completely artificial flu with high recovery rate 0.6 but relatively high

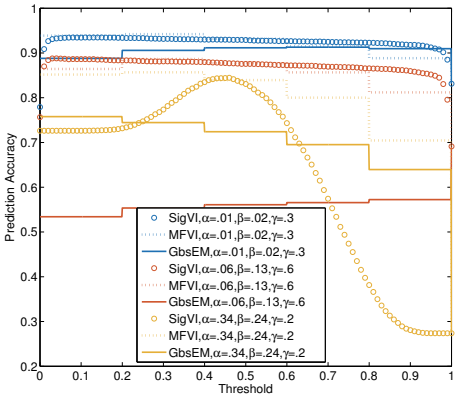


Figure 4: Accuracy v.s threshold.

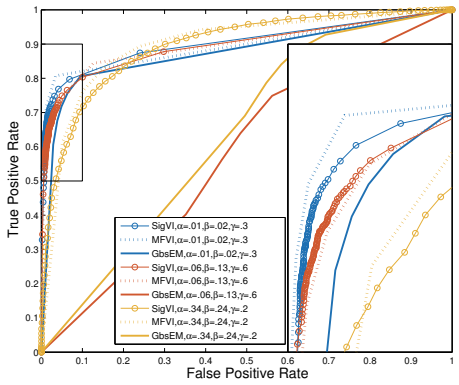
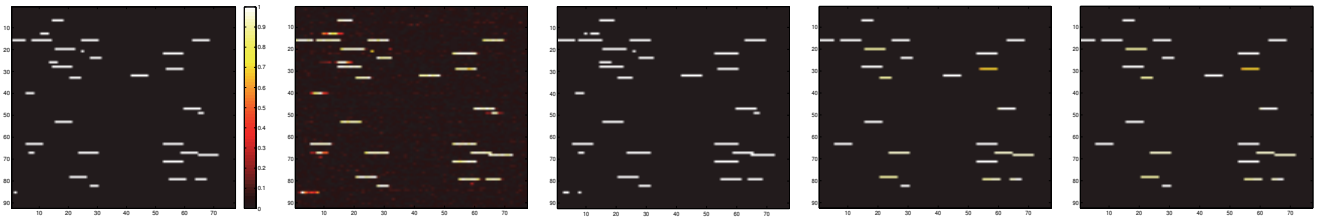


Figure 5: ROC curve comparison

infection rate 0.06 and 0.13. The demo of dynamic prediction is recorded in the `.avi` files of supplementary.

In this setting, we evaluate the predicting performance on simulated infected states $\mathbf{X}_{N\times T}$, with N participants in T days. The algorithms we tested are node-wise sigmoid variational infection, mean-field variational bayesian inference and Gibbs sampling with EM (Fan et al. 2015). Mean-field is the most standard variational inference approach; however, its derivation is extremely complicated for the non-conjugate model. In our case, some variables like γ, $\mathbf{Y}$, belonging to conjugate exponential (CE) family, have a nice variational EM updating formula, which is actually similar to the full conditional in Gibbs sampling. However, for other variables, we need to derive the gradient based optimization method. Unlike much simpler gradient computation with respect to sigmoid function, the gradient of some logarithm term may become quite annoying (we provide the mean field VB derivation in the supplementary material).

Fig.4 and 5 display the measurement of accuracy and ROC curve. For usual flu setting when α, β are extremely small (≈ 0), the performance on three algorithms has no significant difference, no matter on which criterion. However, as mentioned previously, the flaw of Taylor expansion will be exaggeratedly amplified if α, β increases. On this circumstance, the advantage of variational inference is exemplified



(a) Expertise Diagnosis (b) SigVI (c) SigVI(> 0.5) (d) Mean-field (e) Gibbs+EM

Figure 6: Posterior $P(\mathbf{X}|\mathbf{Y})$ estimation: x-axis represents day and y-axis means id number of participants. (a) A binary matrix with 1 indicating infected; (b) Posterior mean estimated by SigVI. (c) Resulted binary matrix by applying a 0.5 threshold on (v); (d) Posterior mean estimated by mean-filed; (e) Posterior mean estimated by Gibbs EM.

on the other two flu settings. Both the accuracy and ROC illustrate that Gibbs EM almost deteriorates as a random classifier, while mean field and sigmoid variational infection can still obtain reasonable result. Additionally, the behavior of SigVI is smoother than mean-field. This phenomenon is more obvious in yellow circle of Fig.4 in the setting with low recovery and high infection rate, thus empirically leading to a reasonable hard threshold for decision. Since mean-field VBEM is similar to full conditional of Gibbs sampling, its curves behave analogously as Gibbs EM. The 1000 thresholds we used to plot are uniformly distributed on interval $[0, 1]$, but the corresponding points of ROC except SigVI are almost located in the leftmost region of coordinates.

Short Pattern Capture on Real Flu Case

In the second experiment, we validate our approach in the real flu diffusion dataset with professional diagnosis. In previous work, the Gibbs sampling based algorithm tends to find the long duration flu while the short pattern is usually omitted by averaging when computing posterior mean. However, we found the node-wise sigmoid variational inference can eliminate this deficiency due to the structure of parameterization. It favors a node-wise estimation rather than temporal dependency. Fig.6 illustrates the comparison between the expertise diagnosis and different inference results. We notice that the short duration after onsite of flu does exist in accordance to Fig.6(a). The posterior mean of SigVI in Fig.6(b) shows that even the short duration is less than 5 days, such patterns can be captured by assigning high risk on these days, accompanying with less risk of infection appearing before or after these days. The finding actually reflects the common sense. For example, Participant 13 was diagnosed as flu during day 13 to 15. Comparing with other algorithms, only SigVI detected this patient with prediction during the whole week. The first and last 2 days are predicted as low risk period, while the middle 3 days are high risk period which is corresponding to expertise.

For the overall predication accuracy, all algorithms achieve a more than 99% accuracy. This is mainly due to two reasons. The underlying infection rate should be close to 0, since this dataset is collected during a normal flu season, not from SARS or HxNx outbreaking period. Another obvious reason is that the negative cases (0) dominate the binary matrix, thus meaning that a more than 95% accuracy can be obtained even if we predict all zeros. Therefore, we also prefer to examine recall (equivalently, sensitivity or true positive rate). Table 1 shows our results compared with other papers' report. Mean-field variational inference has the highest accuracy while it is not significantly better than the others. However, SigVI achieved at least 5% improvement on recall than any other algorithm. This is, in turn, consistent with the short pattern capture property, since more positive cases are detected by SigVI.

Table 1: Recall and Accuracy of exFlu Epidemics

Model	Recall	Accuracy
SigVI	**0.9548**	0.9946
MfVI	0.9032	**0.9979**
GibEM (Fan et al. 2015)	0.8974	0.9978
GCHMMs+LogReg (Dong, Pentland, and Heller 2012)	0.7436	0.9912

Conclusion and Future Work

In this paper, we propose a unified variational inference framework to learn a general flu diffusion model – hGCHMMs. Our VI algorithm is based on minimizing the KL divergence between true posterior of generative model and the proposed recognition model. Differing from standard variations EM, our approach can learn the parameters of both models simultaneously even in a dynamic and heterogenous set-up. In particular, the experimental results imply that our inference method is possible to generalize the application of hGCHMMs to more broader diseases, such as high contagious avian influenza, which proves difficult to model previously. Like deep neural networks, our developed variational inference may suffer the problem of blow-up parameters even regularization is imposed during training. The recent success of deep learning lies on the sufficient training data, which is usually impossible to obtain in the research of health informatics. An important avenue of future research might explore the MCMC method which can likewise overcome the problem of non-conjugacy, especially the efficient Hamiltonian Monte Carlo (Neal 2011). In addition, (Salimans, Kingma, and Welling 2015) bridges the gap between variational inference and MCMC, thus making it more possible to develop robust inference method.

References

Beal, M. J. 2003. *Variational algorithms for approximate Bayesian inference*. University of London.

Christley, R. M.; Pinchbeck, G.; Bowers, R.; Clancy, D.; French, N.; Bennett, R.; and Turner, J. 2005. Infection in social networks: using network analysis to identify high-risk individuals. *American journal of epidemiology* 162(10):1024–1031.

Cooper, B., and Lipsitch, M. 2004. The analysis of hospital infection data using hidden markov models. *Biostatistics* 5(2):223–237.

Dauphin, Y. N.; de Vries, H.; Chung, J.; and Bengio, Y. 2015. Rmsprop and equilibrated adaptive learning rates for non-convex optimization. *arXiv preprint arXiv:1502.04390*.

Dong, W.; Pentland, A.; and Heller, K. A. 2012. Graph-coupled hmms for modeling the spread of infection. *Association for Uncertainty in Artificial Intelligence*.

Duchi, J.; Hazan, E.; and Singer, Y. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research* 12:2121–2159.

Fan, K.; Aiello, A. E.; and Heller, K. A. 2015. Bayesian models for heterogeneous personalized health data. *arXiv preprint arXiv:1509.00110*.

Fan, K.; Eisenberg, M.; Walsh, A.; Aiello, A.; and Heller, K. 2015. Hierarchical graph-coupled hmms for heterogeneous personalized health data. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 239–248. ACM.

Gruber, A.; Weiss, Y.; and Rosen-Zvi, M. 2007. Hidden topic markov models. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 163–170.

Hinton, G. E.; Osindero, S.; and Teh, Y.-W. 2006. A fast learning algorithm for deep belief nets. *Neural computation* 18(7):1527–1554.

Kingma, D. P., and Welling, M. 2014. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*.

Kingma, D. P.; Mohamed, S.; Rezende, D. J.; and Welling, M. 2014. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems (NIPS)*, 3581–3589.

Mnih, A., and Gregor, K. 2014. Neural variational inference and learning in belief networks. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, 1791–1799.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518(7540):529–533.

Neal, R. M. 2011. Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo* 2.

Rezende, D. J.; Mohamed, S.; and Wierstra, D. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of The 31st International Conference on Machine Learning (ICML)*, 1278–1286.

Salakhutdinov, R.; Mnih, A.; and Hinton, G. 2007. Restricted boltzmann machines for collaborative filtering. In *Proceedings of the 24th International Conference on Machine Learning (ICML)*, 791–798. ACM.

Salathé, M.; Kazandjieva, M.; Lee, J. W.; Levis, P.; Feldman, M. W.; and Jones, J. H. 2010. A high-resolution human contact network for infectious disease transmission. *Proceedings of the National Academy of Sciences* 107(51):22020–22025.

Salimans, T.; Kingma, D. P.; and Welling, M. 2015. Markov chain monte carlo and variational inference: Bridging the gap. In *Proceedings of The 32st International Conference on Machine Learning (ICML)*, 1278–1286.

Srivastava, N.; Salakhutdinov, R. R.; and Hinton, G. E. 2013. Modeling documents with deep boltzmann machines. *arXiv preprint arXiv:1309.6865*.

Sutskever, I., and Hinton, G. E. 2008. Deep, narrow sigmoid belief networks are universal approximators. *Neural Computation* 20(11):2629–2636.

Sutton, R. S., and Barto, A. G. 1998. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.