Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Read the Silence: Well-Timed Recommendation via Admixture Marked Point Processes

Hideaki Kim, Tomoharu Iwata, Yasuhiro Fujiwara, Naonori Ueda
Ueda Research Laboratory, NTT Communication Science Laboratories
2–4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
{kin.hideaki, iwata.tomoharu, fujiwara.yasuhiro, ueda.naonori}@lab.ntt.co.jp

Abstract

Everything has its time, which is also true in the point-of-interest (POI) recommendation task. A truly intelligent recommender system, even if you don't visit any sites or remain silent, should draw hints of your next destination from the "silence", and revise its recommendations as needed. In this paper, we construct a well-timed POI recommender system that updates its recommendations in accordance with the silence, the temporal period in which no visits are made. To achieve this, we propose a novel probabilistic model to predict the joint probabilities of the user visiting POIs and their time-points, by using the admixture or mixed-membership structure to extend marked point processes. With the admixture structure, the proposed model obtains a low dimensional representation for each user, leading to robust recommendation against sparse observations. We also develop an efficient and easy-to-implement estimation algorithm for the proposed model based on collapsed Gibbs and slice sampling. We apply the proposed model to synthetic and real-world check-in data, and show that it performs well in the well-timed recommendation task.

1 Introduction

Everything has its time, a verity that also applies to point-of-interest (POI) recommendations. Places you want to visit today are not necessarily where you are likely to go tomorrow. Thus recommender systems should update recommendations daily or even hourly, by monitoring your behavior. Upon observing your visit to a site, a well-trained recommender system will update candidates for the next destination based on your visit history. But for a system to be truly intelligent, it should be able to draw hints of your next destination even if you are "silence", that is, don't visit any site or keep silent. We draw a parallel with your old friend who can read between the lines of your behavior, and kindly give well-timed suggestions despite your silence.

POI recommender systems usually predict the probability of a user visiting each of the candidates, rank them in the descending order, and recommend the higher ranked ones as the next destination. Traditional approaches calculate the ranking based on a user model, which essentially counts the POIs visited in the past (Koren et al. 2009; Kurashima et al. 2010; Ye et al. 2011; Cheng et al. 2012), and the temporal information of user behavior is entirely ignored. As a result, they never revise their recommendations until the user visits the next POI. To address this limitation, some pioneering methods take the circadian periodicity of user behavior or the time drift factor of user preferences into account (Iwata et al. 2009; Koren 2010; Gao et al. 2013; Yuan et al. 2013). Although these methods can change their suggestions hourly in a day, they ignore the periods in which the user doesn't visit any place. In other words, these methods do not consider the user's silent periods as being informative.

Recently in the field of e-commerce, some innovative recommendation approaches have been proposed that employ point process techniques (Wang, Sarwar, and Sundaresan 2011; Zhao et al. 2012; Wang and Zhang 2013). Because the point process provides the probability of each item being purchased at each point in time, item ranking for the user can be updated depending on not only the items purchased in the past, but also the temporal duration since the last (or some trigger) purchase time-point, that is, "silence". In these approaches, however, to address the sparsity in transaction data, all items are categorized into groups in advance, and a set of covariates related to the user and items should be prepared. Such requirements strongly limit their application to check-in data.

In this paper, by extending marked point processes (Jacobsen 2006) through the addition of the admixture or mixed-membership structure (Airoldi et al. 2014), we propose a novel probabilistic model to predict the joint probabilities of the user visiting POIs and their time-points. We use the proposed model to construct a well-timed recommender system that uses silent periods (no user visit activity) to update its recommendations. By virtue of its admixture structure, the proposed model obtains a low dimensional representation for each user, leading to robust recommendations against sparse observations without any recourse to category information or covariates. We call the proposed model, the *admixture Marked Point Process* (adMPP).

The rest of the paper is organized as follows. In Section 2, we outline related work. In Section 3, we propose the admixture marked point process, and construct a well-timed recommendation system by using adMPP. Also, we develop an efficient estimation algorithm based on collapsed Gibbs and slice samplings in Section 2. In Section 5, we apply adMPP to synthetic and real-world check-in data, and show its effectiveness in producing well-timed POI recommendations. Finally, Section 6 provides our conclusions.

2 Related Work

In the task of item recommendation (POIs, products, movies and so on), a recommender system calculates for the user a score for each item, ranks the items in the descending order, and recommends those in the higher ranks to the user. The simplest score is the frequency with which the user selected each of the items in the past, but usually the raw user-item score matrix has many missing values, resulting in low recommendation performance. To address the sparsity problem, matrix factorization methods (Koren et al. 2009) and topic models (Blei, Ng, and Jordan 2003; Griffiths and Steyvers 2004; Xu, Zhang, and Yi 2008), known as collaborative filtering techniques, have been employed intensively. They assume that several latent factors or states should be shared among all users, and reconstruct the user-item score matrix in a collaborative way. To further improve the recommendation accuracy, the utilization of additional information such as social and geographical influences has also been proposed (Ye et al. 2011; Cheng et al. 2012; Kurashima et al. 2010).

The above approaches focus on only what items the user selected, not when she/he did it. Thus their recommendations never change until the user selects the next item. Toward the development of timely recommender systems, some approaches have been proposed that take the effect of time into consideration. (Iwata et al. 2009; Koren 2010) modified conventional methods by modeling the factor of time drift in user preferences. To explain the daily periodicity of user's behavior, (Gao et al. 2013; Yuan et al. 2013) split a day into multiple equal time slots, and calculate a user's score for each item with respect to each time slot.

In contrast to these approaches, which focus on calendar times, (Wang, Sarwar, and Sundaresan 2011; Wang and Zhang 2013) recently proposed e-commerce recommender systems that utilize purchasing time intervals. Their systems update each user's recommendation score depending on the temporal duration since her/his last (or some triggering) purchase. Due to its explicit dependence on the user's previous action times, their recommendations are expected to be dynamically optimized in a more personalized manner. In these approaches, however, to address the sparsity of transaction data, the items must be categorized into some relevant groups in advance, and a set of covariates, which serves to explain the heterogeneity across users and item categories, should be prepared appropriately. Because check-in data usually consist of just a set of sequences of visited POIs and their time-points, applying these approaches to check-in data is difficult. In this paper, we propose a theoretically solid method that overcomes these limitations.

3 Model

We assume a set of U users, each of whom, $u \in \{1, \dots, U\}$, has a sequence of $(N^u + 1)$ observed pairs of visited POI and associated time-point, $D^u \equiv \{(y_n^u, t_n^u)\}_{n=0}^{N^u}$. Thus we have a set of visit event logs, denoted by $\boldsymbol{D} \equiv \{D^u\}_{u=1}^{U}$. Letting the number of POIs be Q, we denote the q-th POI by $q \in \{1, \dots, Q\}$. The notation is summarized in Table 1.

Table 1: Notation

Symbol	Definition
U, Q	# of users, and # of POIs
u, q	user and POI, $u \in \{1, \dots, U\}$, $q \in \{1, \dots, Q\}$
N^u	# of logs for each user
(y_n^u, t_n^u)	user u's n-th log of (visited POI, visit time)
T	definition range of inter-event interval
K	# of latent topics
k	latent topic, $k \in \{1, \dots, K\}$
z_n^u	latent topic assigned to u's n-th log,
θ_k^u	weights of topic k on user u, $\sum_{k=1}^{K} \theta_k^u = 1$
$\phi_{q'q}^k$	topic k's transition probability from q' to q, $\sum_{q=1}^{Q} \phi_{q'q}^k = 1$
$f_{q'q}^k(s)$	topic k's inter-event interval distribution of transition from q' to q
$L_{q'q}^k$	# of bins of $f_{q'q}^k(s)$
$\psi_{q'q\ell}^k$	ℓ-th probability mass of $f_{q'q}^k$, $\sum_{\ell=1}^{L_{q'q}^k} \psi_{q'q\ell}^k = 1$

User Model by Marked Point Process

We first introduce a theoretically feasible method for modeling visit events that occur in time. The theory of *marked point processes* (Jacobsen 2006), which has been applied in such diverse disciplines as seismology (Ogata 1988), computer vision (Ge and Collins 2009), and financial markets (Prigent 2001; Björk, Kabanov, and Runggaldier 1997), provides a powerful tool for modeling and analyzing the stochastic structure of point events with marks occurring in continuous time. Here, POIs can be treated as marks. In the marked point process, the visit behavior of a user is defined by the conditional intensity function (Jacobsen 2006), that is, the instantaneous probability of a user visiting a specific POI at each point in time. In this paper, we denote the conditional intensity function by the *visit rate*.

When we assume that the transition from one POI to another is Markovian, and the temporal interval between successive visit time-points is conditioned by both the previous and the next visit events, user u's visit rate, denoted by $\lambda^u(y, t)$, is given as follows,

$$\begin{aligned} \lambda^u(y, t) &\equiv \phi_{y_n^u y}\, \lambda^u(t - t_n^u | y_n^u, y), \\ \lambda^u(t - t_n^u | y_n^u, y) &\equiv -\partial_t \log\Big[1 - \int_0^{t - t_n^u} f_{y_n^u y}(s) ds\Big], \end{aligned} \tag{1}$$

where t represents each point in time between the previous (n-th) and the next ($(n+1)$-th) visit time point, ($t_n^u < t < t_{n+1}^u$), ∂_t represents the derivative with respect to t, y represents the next POI to be visited, $\phi_{q'q}$ represents the transition probability from q' to q, and $f_{q'q}(s)$ represents the inter-event interval distribution conditioned by the previous POI, q', and the next one, q. The transition matrix $\boldsymbol{\phi}$ satisfies a set of normalization conditions, $\sum_{q=1}^{Q} \phi_{q'q} = 1$, for

$q' = 1, 2, \ldots, Q$. The probability density of the observation D^u occurring is defined by using visit rate, $\lambda^u(y,t)$, as follows,

$$
\begin{aligned}
P(D^u|\lambda^u(y,t)) &= \prod_{n=1}^{N^u} \phi_{y_{n-1}^u y_n^u} \, \lambda^u(t_n^u - t_{n-1}^u | y_{n-1}^u, y_n^u) \\
&\quad \times \prod_{n=1}^{N^u} \exp\left[-\int_{t_{n-1}^u}^{t_n^u} \lambda^u(t - t_{n-1}^u | y_{n-1}^u, y_n^u) dt \right] \\
&= \prod_{n=1}^{N^u} \phi_{y_{n-1}^u y_n^u} \, f_{y_{n-1}^u y_n^u}(t_n^u - t_{n-1}^u), \qquad (2)
\end{aligned}
$$

where the exponential on the second line of Eq. (2), called the *survivor function* (Snyder 1975), represents the probability of no visit events occurring in interval (t_{n-1}^u, t_n^u), for $n = 1, \ldots, N^u$.

In the following, based on the marked point process (1), we derive the recommendation score that takes an observed silent duration of no events into consideration. Given user u and current time t, we can evaluate the probability that user u will visit the next POI, y, at future time-point t^* ($t^* > t$) for the first time since her/his last visit event, (y', t'), as

$$
\begin{aligned}
&\phi_{y'y} \, \lambda^u(t^* - t'|y', y) \exp\left[-\int_{t'}^{t^*} \lambda^u(s - t'|y', y) ds \right] \\
&= \phi_{y'y} \, f_{y'y}(t^* - t'). \qquad (3)
\end{aligned}
$$

From Eq. (3), we can evaluate the probability that user u will visit POI y within future interval $[t, t + \Delta)$, as

$$
\begin{aligned}
P_\Delta(y, \xi \equiv t - t'|y') &= \int_t^{t+\Delta} \phi_{y'y} \, f_{y'y}(t^* - t') dt^* \\
&= \phi_{y'y} \int_\xi^{\xi+\Delta} f_{y'y}(s) ds, \qquad (4)
\end{aligned}
$$

where $\xi \equiv (t - t')$ represents the silent duration of no visit events. Finally, from Eq. (4), we obtain the a posteriori probability of the next POI, y, conditioned by the observation of silent duration ξ, as

$$
P_\Delta(y|\xi, y') \propto P_\Delta(y, \xi|y') = \phi_{y'y} \int_\xi^{\xi+\Delta} f_{y'y}(s) ds. \quad (5)
$$

We adopt Eq. (5) as the score for well-timed recommendations. Equation (5) shows that the score is updated depending on the silent duration ξ, as well as the most recent POI y'. We generalize the score by developing the marked point process that is discussed in the last part of this section.

Extension to Admixture Model

In the real world, the numbers of users and POIs are large, but the number of observations for each user is relatively small. In this situation, it is hard to estimate a user's model parameters, $\boldsymbol{\phi}$ and $\{f_{q'q}(s)\}_{q',q=1}^{Q}$ in (1), from her/his own data. To alleviate this sparsity problem, we propose a novel extension of the marked point process (1) into the admixture (or mixed-membership) model (Murphy 2012), where the number of parameters is drastically reduced by sharing parameters across different users, while the representative power of the model is maintained by assuming user-dependent factors.

We assume that a set of K latent states is shared among the users, where each latent state, denoted by $k \in \{1, \ldots, K\}$, is characterized by a set of transition matrices, $\boldsymbol{\phi}^k$, and inter-event interval distributions, $\{f_{q'q}^k(s)\}_{q',q=1}^{Q}$. We denote the latent state by *topic*, as with topic models (Blei, Ng, and Jordan 2003). Given the weights of the K topics for user u, $\boldsymbol{\theta}^u \equiv (\theta_1^u, \cdots, \theta_K^u)$, topic z is generated from $\boldsymbol{\theta}^u$, and the next POI, y, is generated from topic z's process characterized by $\boldsymbol{\phi}^z$ and $\{f_{q'q}^z(s)\}_{q',q=1}^{Q}$, with respect to each of N^u observations. The visit rate of each user is described as follows,

$$
\begin{aligned}
&\lambda^u(z, y, t) \equiv \theta_z^u \, \phi_{y_n^u y}^z \, \lambda^z(t - t_n^u | y_n^u, y), \\
&\lambda^z(t - t_n^u | y_n^u, y) \equiv -\partial_t \log\left[1 - \int_0^{t - t_n^u} f_{y_n^u y}^z(s) ds \right],
\end{aligned} \qquad (6)
$$

where t represents each point in time ($t_n^u < t < t_{n+1}^u$), z represents the topic from which the next ($(n+1)$-th) POI, y, is generated, and $\boldsymbol{\theta}^u$ satisfies the normalization condition, $\sum_{k=1}^{K} \theta_k^u = 1$. Note that topic z can be regarded as a mark in the marked point process (6), as well as POI y. By sharing a small number of topics among all the users, the admixture structure allows each user's behavior to be estimated in a collaborative way. We call the proposed model (6), the *admixture Marked Point Process* (adMPP).

Non-parametric Inter-event Interval Distribution

In adMPP (6), the inter-event interval distribution of each topic and each pair of POIs, $f_{q'q}^k(s)$, needs to be parameterized by an appropriate probability density function. The simplest candidate, the Gaussian distribution, is apparently unsuitable for inter-event intervals because the intervals should not have any negative value. Although other exponential families such as gamma and inverse Gaussian distributions could be adopted, such simple density functions strongly bias the forms of distribution considered, which would lead to inaccurate model estimation.

In this study, we employ the piecewise-constant density function, sometimes called *histogram density estimator* (Silverman 1986; Hall and Hannan 1988), as the inter-event interval distribution. The histogram density estimator is so flexible that it can model various types of density function characteristics like multi-modality and discontinuity. In the histogram, each inter-event interval distribution, $f_{q'q}^k(s)$, is represented by; (i) discretizing a half-open definition range of inter-event interval, $T = [T_0, T_1]$, into $L_{q'q}^k$ contiguous intervals (bins) of equal width, $\delta \equiv (T_1 - T_0)/L_{q'q}^k$; and (ii) assigning a uniform probability density, $\psi_{q'q\ell}^k \,/\, \delta$, to each bin region, $[x_\ell, x_{\ell+1}]$, for $\ell = 1, 2, \ldots, L_{q'q}^k$. Here the lower boundary of a bin, x_ℓ, is given by, $x_\ell = T_0 + (\ell - 1)\delta$, and the probability mass, $\boldsymbol{\psi}_{q'q}^k \equiv (\psi_{q'q1}^k, \ldots, \psi_{q'qL_{q'q}^k}^k)$, follows a normalization condition. Inter-event interval variable s follows the following piecewise-constant distribution,

$$
f_{q'q}^k(s) = \psi_{q'qw(s)}^k L_{q'q}^k / (T_1 - T_0), \quad \sum_{\ell=1}^{L_{q'q}^k} \psi_{q'q\ell}^k = 1, \quad (7)
$$

where $w(s)$ represents a discretization operator which transforms continuous variable s into the corresponding bin index, defined by

$$w(s) \equiv 1 + \lfloor L^k_{q'q}\,(s - T_0)/(T_1 - T_0)\rfloor. \tag{8}$$

Here $\lfloor x \rfloor$ represents the floor function. As a generative model, the histogram density estimator (7-8) can be represented by the following three processes: (i) Draw bin index w from a multinomial distribution with parameter, $\boldsymbol{\psi}^k_{q'q}$; (ii) Draw auxiliary variable η from a uniform distribution defined over a unit region, $[0, 1)$; (iii) A sample of inter-event interval is obtained as $s \equiv T_0 + (w + \eta - 1)\delta$.

Prior of Model Parameters

The set of model parameters to be estimated consists of the topic weights $\boldsymbol{\theta}^u$, the transition matrix $\boldsymbol{\phi}^k$, the probability mass $\boldsymbol{\psi}^k_{q'q}$, and the number of bins $L^k_{q'q}$, for $u \in \{1, \ldots, U\}$, $k \in \{1, \ldots, K\}$, and $q', q \in \{1, \ldots, Q\}$. For each of them, adMPP assumes the following conjugate or feasible prior: $\boldsymbol{\theta}^u$ is generated from a symmetric Dirichlet distribution, Dirichlet(α), for each user u; $\boldsymbol{\phi}^k_{q':}$ is generated from Dirichlet(β) for each topic k and each POI q'; $\boldsymbol{\psi}^k_{q'q}$ is generated from Dirichlet($\gamma^k L^k_{q'q}$) for each POI pair (q', q); and discrete $L^k_{q'q}$ is generated from a uniform distribution defined over the range $[1, L_{\text{max}}]$, denoted by $\text{Uniform}_{\text{dis}}(1, L_{\text{max}})$. Here α, β and $\boldsymbol{\gamma} \equiv \{\gamma^k\}_{k=1}^K$ are Dirichlet parameters, and L_{max} is the maximum number of bins to be considered.

Admixture Marked Point Process

We summarize the generative process of adMPP for a set of visit logs, $\boldsymbol{D} \equiv \{\{(y^u_n, t^u_n)\}_{n=0}^{N^u}\}_{u=1}^U$, as follows:

1. For each topic $k = 1, \ldots, K$:
   (a) For each POI $q' = 1, \ldots, Q$:
      i. Draw transition matrix $\phi^k_{q':} \sim$ Dirichlet(β)
      ii. For each POI $q = 1, \ldots, Q$:
         · Draw number of bins $L^k_{q'q} \sim \text{Uniform}_{\text{dis}}(1, L_{\text{max}})$
         · Draw probability mass $\psi^k_{q'q} \sim$ Dirichlet($\gamma^k L^k_{q'q}$)
2. For each user $u = 1, \ldots, U$:
   (a) Draw topic weights $\boldsymbol{\theta}^u \sim$ Dirichlet(α)
   (b) For each visit event $n = 1, \ldots, N^u$:
      i. Draw topic assignment $z^u_n \sim$ Multi($\boldsymbol{\theta}^u$)
      ii. Draw POI $y^u_n \sim$ Multi($\phi^{z^u_n}_{y^u_{n-1}:}$)
      iii. Draw bin index $w^u_n \sim$ Multi($\boldsymbol{\psi}^{z^u_n}_{y^u_{n-1} y^u_n}$)
      iv. Draw auxiliary variable $\eta^u_n \sim \text{Uniform}_{\text{con}}(0, 1)$
      v. Give inter-event interval $s \equiv (t^u_n - t^u_{n-1})$ as
         $s = T_0 + (w^u_n + \eta^u_n - 1)(T_1 - T_0)/L^{z^u_n}_{y^u_{n-1} y^u_n}$

where $\text{Uniform}_{\text{con}}(x, y)$ is the continuous uniform distribution defined over a half-open interval $[x, y)$, and Multi(x) is the multinomial distribution with parameter x. Note that the initial visit event of each user u, (y^u_0, t^u_0), is assumed to be known.

Because the multinomial parameters, $\boldsymbol{\theta}^u$, $\boldsymbol{\phi}^k$ and $\psi^k_{q'q}$, are conjugate to Dirichlet priors, they can be marginalized out of the generative process, leading to the joint distribution of visit logs $\boldsymbol{D}$, latent topics $\boldsymbol{z} \equiv \{\{z^u_n\}_{n=1}^{N^u}\}_{u=1}^U$, and set of bin numbers $\boldsymbol{L} \equiv \{\{L^k_{q'q}\}_{q',q=1}^Q\}_{k=1}^K$, as follows:

$$\begin{aligned}
&p(\boldsymbol{D}, \boldsymbol{z}, \boldsymbol{L} | \alpha, \beta, \boldsymbol{\gamma}) \\
&= \prod_{u=1}^U \frac{\prod_{k=1}^K \Gamma(\alpha + N^u_k)}{\Gamma(K\alpha + N^u)} \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \\
&\times \prod_{k=1}^K \prod_{q'=1}^Q \frac{\prod_{q=1}^Q \Gamma(\beta + N_{kq'q})}{\Gamma(Q\beta + N_{kq'})} \frac{\Gamma(Q\beta)}{\Gamma(\beta)^Q} \\
&\times \prod_{k=1}^K \prod_{q',q=1}^Q \frac{\prod_{\ell=1}^{L^k_{q'q}} \Gamma(\gamma^k / L^k_{q'q} + N_{kq'q\ell})}{\Gamma(\gamma^k + N_{kq'q})} \\
&\times \prod_{k=1}^K \prod_{q',q=1}^Q \frac{\Gamma(\gamma^k)}{\Gamma(\gamma^k / L^k_{q'q})^{L^k_{q'q}}} \left(\frac{L^k_{q'q}}{T_1 - T_0}\right)^{N_{kq'q}},
\end{aligned} \tag{9}$$

where $\Gamma(x)$ is the gamma function, N^u_k is the number of visit events assigned to topic k in the history of user u, $N_{kq'q}$ is the number of transition events from POI q' to POI q assigned to topic k, $N_{kq'q\ell}$ is the number of transition events from POI q' to POI q addressed to the ℓ-th bin of topic k's histogram density estimator, and $N_{kq'} = \sum_q N_{kq'q}$.

Well-timed Recommendations

Based on adMPP, we construct a recommender system that can update its recommendations in response to the users' silent periods. With the same procedure as in the pure marked point process (5), we obtain user u's score for each POI, y, within future interval $[t, t + \Delta)$ as follows,

$$P_\Delta(y | \xi, y') \propto \sum_{k=1}^K \theta^u_k \, \phi^k_{y'y} \int_\xi^{\xi+\Delta} f^k_{y'y}(s) ds \tag{10}$$

where $\xi \equiv (t - t')$ represents the silent duration since the most recent visit event (y', t').

In practice, at current time $t = \tau$, our recommender system ranks the POIs according to the score (10), and recommends those in the higher ranks during interval $[\tau, \tau + \Delta]$. When t reaches $(\tau + \Delta)$, the system re-calculates the scores with $\xi \to \xi + \Delta$ and $[\tau, \tau + \Delta] \to [\tau + \Delta, \tau + 2\Delta]$, and updates its recommendation; this operation is repeated until the user visits the next POI. The updating window, Δ, can have arbitrary positive values, but should be chosen appropriately: When Δ is small, the system updates its recommendation frequently, and its recommended POIs are likely to be what the user is about to visit. However, the risk that she/he has little time to arrange or prepare for the next visit would increase; When Δ is large, the user is more likely to get useful information well in advance, but the recommendation accuracy might not be good.

4 Estimation Procedure

By using the simple and easy-to-implement collapsed Gibbs sampling (Griffiths and Steyvers 2004) and slice sampling (Neal 2003), we obtain posterior samples of $\boldsymbol{z}$ and $\boldsymbol{L}$ following $p(\boldsymbol{z}, \boldsymbol{L} | \boldsymbol{D}, \alpha, \beta, \boldsymbol{\gamma})$, from which the topic weights $\boldsymbol{\theta}^u$, the transition $\boldsymbol{\phi}^k$, and the probability mass $\boldsymbol{\psi}^k_{q'q}$, as well as the hyperparameter $(\alpha, \beta, \boldsymbol{\gamma})$, can be estimated efficiently.

Collapsed Gibbs Sampling of Topic

Given a set of bin numbers $\boldsymbol{L}$ and the current state of all but one latent variable z_j, where $j=(u,n)$, the assignment of a latent topic to the n-th visit event of user u is sampled from the following multinomial distribution:

$$p(z_j=k|\boldsymbol{D},\boldsymbol{z}^{\setminus j},\boldsymbol{L},\alpha,\beta,\boldsymbol{\gamma}) \propto (\alpha+N_{ku}^{\setminus j})\,\frac{\beta+N_{ky^u_{n-1}y^u_n}^{\setminus j}}{Q\beta+N_{ky^u_{n-1}}^{\setminus j}}\,\frac{\gamma^k/L^k_{y^u_{n-1}y^u_n}+N^{\setminus j}_{ky^u_{n-1}y^u_n w^u_{kn}}}{\gamma^k+N^{\setminus j}_{ky^u_{n-1}y^u_n}} \quad k=1,\ldots,K, \quad (11)$$

where $N^{\setminus j}$ represents the count that does not include the current assignment of z_j, and w^u_{kn} represents the bin index of the histogram specified by latent topic, k, and pair of POIs, (y^u_{n-1},y^u_n), within which $(t^u_n-t^u_{n-1})$ falls, calculated as

$$w^u_{kn}\equiv 1+\lfloor L^k_{y^u_{n-1}y^u_n}(t^u_n-t^u_{n-1}-T_0)/(T_1-T_0)\rfloor. \quad (12)$$

Equation (11) and the next equation (13) are derived from the joint distribution (9).

Slice Sampling of Bin Number

Given $\boldsymbol{z}$, each of the bin numbers, $\{L^k_{q'q}\}_{k,p',p}$, is sampled from the following discrete probability distribution:

$$p(L^k_{q'q}=V|\boldsymbol{D},\boldsymbol{z},\alpha,\beta,\boldsymbol{\gamma}) \propto \frac{\prod_{\ell=1}^V\Gamma(\frac{\gamma^k}{V}+N_{kq'q\ell})}{\Gamma(\gamma^k+N_{kq'q})}\frac{\Gamma(\gamma^k)}{\Gamma(\frac{\gamma^k}{V})^V}V^{N_{kq'q}}, \quad V=1,..,L_{\max}. \quad (13)$$

Because a direct sampling from Eq. (13) is time-consuming, we adopt the slice sampling with stepping-out procedure (Neal 2003). If the right-hand side of Eq. (13) is denoted by $g(V)$, the sampling procedure is represented as follows:

0. Let the current value of $L^k_{q'q}$ be L, and initialize slice $[S_0,S_1]\leftarrow[L,L]$. Set a width parameter dW.
1. Draw $a\sim \text{Uniform}_{\text{con}}(0,g(L))$
2. Expand slice as $[S_0,S_1]$ $\leftarrow[\max(1,S_0-dL),\min(L_{\max},S_1+dL)]$
3. If $g(S_i)<a$ for $i=0$ and 1, go to step 4, otherwise return to step 2.
4. Draw candidate $L^*\sim\text{Uniform}_{\text{dis}}(S_0,S_1)$
5. If $g(L^*)>a$, accept L^* as the next sample of $L^k_{q'q}$. Otherwise, update slice as $[S_0,S_1]\leftarrow[L^*,S_1]$ when $L^*<L$, or $[S_0,S_1]\leftarrow[S_0,L^*]$ when $L^*>L$, and return to step 4.

We set the width parameter, dL, at 100 in this study.

Large bin number $L^k_{q'q}$ gives adMPP better fitting to training data, while too large $L^k_{q'q}$ would deteriorate the predictive performance due to over-fitting. It should be noted here that Eq. (13) finds an appropriate bin number, $L^k_{q'q}$, by the technique of Bayesian model selection (Bishop 2006): The posterior distribution of $L^k_{q'q}$ (13) is obtained by the marginalization over $(L^k_{q'q})$-dimensional parameter $\psi^k_{q'q\ell}$, in which excessive values of $L^k_{q'q}$ are penalized.

Determination of Hyperparameter

Based on the empirical Bayes method, the set of hyperparameters, $(\alpha,\beta,\boldsymbol{\gamma})$, is determined by maximizing the marginal likelihood,

$$\hat\alpha,\hat\beta,\hat{\boldsymbol{\gamma}}=\arg\max_{\alpha,\beta,\boldsymbol{\gamma}}\sum\nolimits_{\boldsymbol{z},\boldsymbol{L}}\big[p(\boldsymbol{D},\boldsymbol{z},\boldsymbol{L}|\alpha,\beta,\boldsymbol{\gamma})\big], \quad (14)$$

which can be performed by the stochastic EM algorithm. In each posterior sampling of $\boldsymbol{z}$ and $\boldsymbol{L}$ (11-13), the set of hyperparameters, $(\alpha,\beta,\boldsymbol{\gamma})$, is updated by using the fixed-point iteration method (Minka 2000) as follows:

$$\alpha\leftarrow\alpha\,\frac{\sum_u\sum_k\Psi(\alpha+N^u_k)-UK\Psi(\alpha)}{K\sum_u\Psi(K\alpha+N^u)-UK\Psi(K\alpha)},$$
$$\beta\leftarrow\beta\,\frac{\sum_{k,q',q}\Psi(\beta+N_{kq'q})-KQ^2\Psi(\beta)}{Q\sum_{k,q'}\Psi(Q\beta+N_{kq'})-KQ^2\Psi(Q\beta)},$$
$$\gamma^k\leftarrow\gamma^k\,\frac{\sum_{q',q,\ell}\Psi\Big(\frac{\gamma^k}{L^k_{q'q}}+N_{kq'q\ell}\Big)-\sum_{q'q}L^k_{q'q}\Psi\Big(\frac{\gamma^k}{L^k_{q'q}}\Big)}{\sum_{q'q}L^k_{q'q}\big[\Psi(\gamma^k+N_{kq'q})-\Psi(\gamma^k)\big]}, \quad (15)$$

where $\Psi(x)$ is the digamma function defined by the derivative of $\log\Gamma(x)$.

Estimation of Model Parameters

By repeating (11-15) until convergence is achieved, we estimate the set of bin numbers, $\hat{\boldsymbol{L}}$, and the set of hyperparameters, $(\hat\alpha,\hat\beta,\hat{\boldsymbol{\gamma}})$, as the last updated values. Next, we further draw M samples of latent topics, $[\boldsymbol{z}^{(1)},\boldsymbol{z}^{(2)},\ldots,\boldsymbol{z}^{(M)}]$, according to Eq. (11), and obtain the posterior mean estimate of the other model parameters as follows:

$$\hat\theta^u_k\simeq\frac{1}{M}\sum_{m=1}^M\frac{\hat\alpha+N^{u(m)}_k}{K\hat\alpha+N_u}, \quad (16)$$

$$\hat\phi^k_{q'q}\simeq\frac{1}{M}\sum_{m=1}^M\frac{\hat\beta+N^{(m)}_{kq'q}}{Q\hat\beta+N^{(m)}_{kq'}}, \quad (17)$$

$$\hat\psi^k_{q'q\ell}\simeq\frac{1}{M}\sum_{m=1}^M\frac{\hat\gamma^k/\hat L^k_{q'q}+N^{(m)}_{kq'q\ell}}{\hat\gamma^k+N^{(m)}_{kq'q}}, \quad (18)$$

where $N^{u(m)}_k$, $N^{(m)}_{kq'q}$ and $N^{(m)}_{kq'q\ell}$ represent the sufficient statistics in the m-th sample, $\boldsymbol{z}^{(m)}\equiv\{\{z^{u(m)}_n\}^{N^u}_{n=1}\}^U_{u=1}$. In the following experiment, M was set at 100. We can select the number of topics, K, by maximizing the cross validated likelihood.

adMPP has only two parameters to be set, the definition range, $T\equiv[T_0,T_1]$, and the maximum of bin number to be considered, $L_{\max}$. Unless the observed intervals lie out of T, adMPP is expected to provide similar estimations regardless of T, because the broad T's effect on estimation of interval distribution can be offset by large bin number. In practice, we set $[T_0,T_1]$ = [zero, the maximum of observed intervals] to ensure that observations lie inside T. Also, $L_{\max}$ does not affect the estimation of bin number unless it's smaller than the optimal bin number. Thus it should be set at a value large enough ($L_{\max}$ = 200 in this paper).

Table 2: Data Statistics

Data set	SYN_{10}	SYN_{50}	SF	NY
# of users	1,000	1,000	557	624
# of POIs	4	4	984	975
# of logs	10,000	50,000	27,687	44,193
mean IEI	0.474	0.474	0.557	0.357

IEI: inter-event interval (week)

Scalability

The most demanding part of the estimation is the Gibbs sampling (11) and the slice sampling (13). Gibbs sampling and slice sampling have computational complexities of $O(KN)$ and $O(KQ^2RN)$, respectively, where N is the total number of logs, and $R \leq L_{\text{max}}^2$. Thus adMPP is applicable to large-scale ($N \gg 1$) data, while being somewhat time-consuming when the number of POIs, Q, is extremely large.

5 Experiments

We examine the effectiveness of the adMPP-based recommendation system by applying it to synthetic and real-world check-in data.

About Data Used

Synthetic Data We make synthetic data for the scenario that a user generates a sequence of visit logs from a mixture of three marked point processes, each of which has a set of distinct inter-event interval distributions $\{f_{q'q}^k(s)\}_{k=1}^3$: gamma distributions in the first (k = 1), beta distributions in the second (k = 2), and two-component gamma mixtures in the last (k = 3). The parameters of the inter-event interval distributions, as well as the set of transition matrices $\{\phi^k\}_{k=1}^3$, are chosen randomly. Each user is characterized by the mixing proportions, which are sampled from a flat Dirichlet distribution. Assuming that the number of users, U, is 1000 and each user has C visit logs, we make two datasets of C = 10 and 50, denoted by SYN_{10} and SYN_{50}, respectively.

Check-in Data From the real-world check-in datasets collected by Brightkite[1] (Cho, Myers, and Leskovec 2011), we make two subsets comprising the visit logs in San Francisco (SF) and New York City (NY), denoted by SF and NY, respectively. In each dataset, we rank the POIs according to the number of unique users who visited them, and focus on the top-1000 popular POIs in the experiment. Also, we omit users who had less than 10 visit logs.

For all four datasets, the first 80% visit logs for each user, denoted by $\boldsymbol{D}^{\text{fit}}$, are used for model fitting, and the remaining 20%, denoted by $\boldsymbol{D}^{\text{test}}$, for model evaluation. Note that this represents per-user splitting. The data statistics are summarized in Table 2.

Evaluation

Based on the above datasets, we compare adMPP's recommendation performance against the results achieved by the following three baseline methods, denoted by Multi, Markov and LDA: (i) Multi is a non-collaborative method based on a multinomial distribution, in which user's scores for POIs are represented by the choice probabilities estimated from the user's visit logs; (ii) Markov is also a non-collaborative method based on a Markov model, in which user's scores are calculated by a transition matrix estimated from the user's visit history; (iii) LDA is the latent Dirichlet allocation (Blei, Ng, and Jordan 2003; Griffiths and Steyvers 2004), a widely used collaborative method in a Bayesian framework, where user's scores are defined by a weighted sum of latent choice probabilities. In Multi and LDA, once the model parameters are estimated based on $\boldsymbol{D}^{\text{fit}}$, the score doesn't change in $\boldsymbol{D}^{\text{test}}$. In contrast, Markov and adMPP update their scores in $\boldsymbol{D}^{\text{test}}$ according to the last POI that a user visited. adMPP also updates the score depending on the duration since the last visit, where the updating window, Δ, can be chosen arbitrarily (see Section 3).

We evaluate the performance of each method based on recall for the top-R recommendation task,

$$\text{rec@}R \equiv \#\text{hits} \;/\; |\boldsymbol{D}^{\text{test}}|, \tag{19}$$

where $\#$hits represents the number of test data logs that were listed in the top-R recommendation list, and $|\boldsymbol{D}^{\text{test}}|$ represents the total number of logs in the test data. R is specified as one and two in the experiment. Figure 1(A) displays the obtained performances with respect to updating window Δ, where Δ is varied from 0.1 to 10 (unit is the mean inter-event interval). The results show that adMPP performs better than the other models for all data sets, which indicates that the silent duration since the last visit is helpful in precisely predicting users' future destinations.

In adMPP and LDA, we select the number of topics, K, by maximizing the two-fold cross validated likelihood. Figure 1(B) shows that the true value, K = 3, is obtained in synthetic data SYN_{50} (we also obtained K = 3 in SYN_{10}),

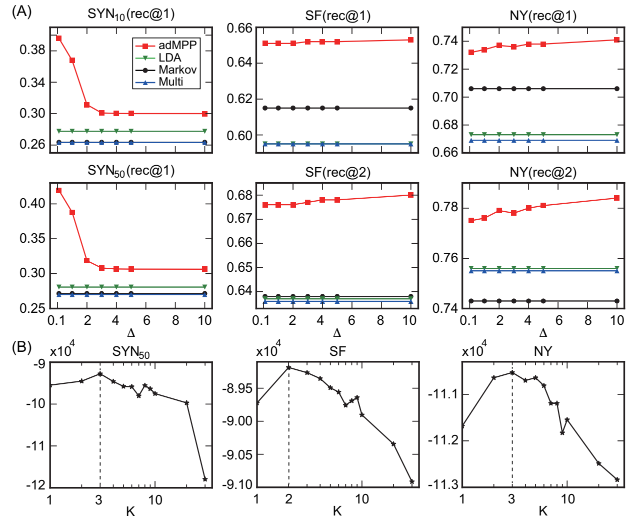

Figure 1: (A) Predictive performance. (B) Cross validated likelihood for adMPP.

[1] https://snap.stanford.edu/data/loc-brightkite.html

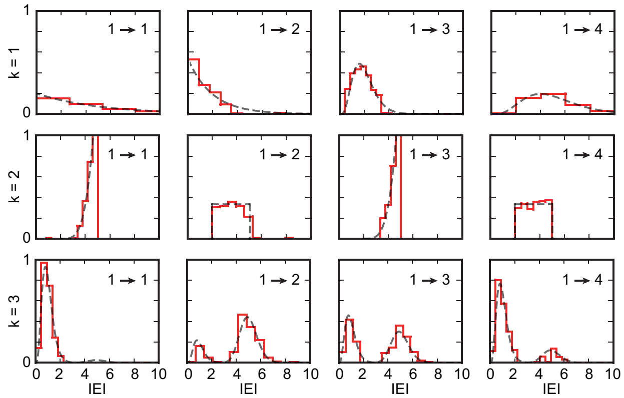

Figure 2: Inter-event interval (IEI) distributions estimated in SYN_{50}. Only the transition from POI q = 1 is displayed. The dashed and solid lines represent the true distribution and its estimation by adMPP, respectively.

where the various kinds of distributions are estimated precisely in a non-parametric way (see Figure 2). Also, Figure 1(B) shows that the number of topics is selected as K = 2 and K = 3 in real-world datasets SF and NY, respectively. The latent transition probabilities, $\{\phi^k\}$, and inter-event interval distributions, $\{f^k_{q'q}(s)\}$, estimated by adMPP are partially displayed in Figure 3, which shows that each of the latent topics represents a distinctive temporal pattern of user's visit behavior.

6 Conclusion

In this paper, we constructed a well-timed recommender system that can update its recommendations in response to silence, the temporal duration in which user does not visit any place or select any item. To achieve this, we proposed a novel marked point process with admixture structure, in which a small number of latent marked point processes are assumed to be shared by all users, leading to robust recommendation against sparse observations. We also derived an efficient estimation algorithm based on collapsed Gibbs and slice sampling. We confirmed experimentally that our proposed model achieved high predictive performance when challenged with synthetic and real-world data.

adMPP is not limited to visit event data, but is widely applicable to any sequences of event points. Thus our model helps provide timely recommendations or navigations in such diverse fields as e-commerce and information retrieval. A possible extension of our model is to incorporate the self-exciting property that events in the past induce events occurring in the future, which was observed in purchasing and querying behaviors (Kim, Takaya, and Sawada 2014; Li et al. 2014). It's a fact of great interest that "silence" should be a negative signal of the next event in self-exciting behaviors.

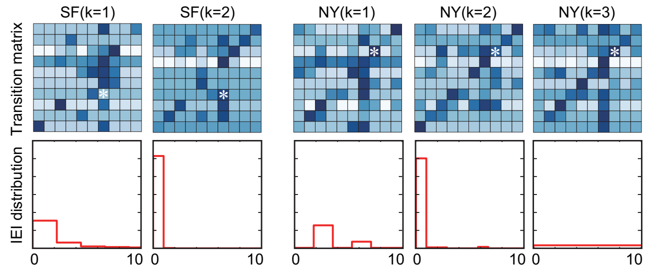

Figure 3: (Upper figures) Heatmap visualization of transition matrix for each latent topic in SF and NY. The values of probabilities are plotted on a log scale, where larger values are represented by darker squares. Only a subset of the whole POIs are displayed. (Lower figures) Inter-event interval (IEI) distribution of a transition specified by white asterisk in the corresponding upper figure.

References

Airoldi, E. M.; Blei, D.; Erosheva, E. A.; and Fienberg, S. E. 2014. *Handbook of Mixed Membership Models and Their Applications*. CRC Press.

Bishop, C. 2006. *Pattern Recognition and Machine Learning*. Springer New York.

Björk, T.; Kabanov, Y.; and Runggaldier, W. 1997. Bond Market Structure in the Presence of Marked Point Processes. *Mathematical Finance* 7(2):211–239.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3(Jan):993–1022.

Cheng, C.; Yang, H.; King, I.; and Lyu, M. R. 2012. Fused Matrix Factorization with Geographical and Social Influence in Location-based Social Networks. In *AAAI*, volume 12, 17–23.

Cho, E.; Myers, S. A.; and Leskovec, J. 2011. Friendship and Mobility: User Movement in Location-based Social Networks. In *KDD*, 1082–1090. ACM.

Gao, H.; Tang, J.; Hu, X.; and Liu, H. 2013. Exploring Temporal Effects for Location Recommendation on Location-based Social Networks. In *RecSys*, 93–100. ACM.

Ge, W., and Collins, R. T. 2009. Marked Point Processes for Crowd Counting. In *CVPR*, 2913–2920. IEEE.

Griffiths, T. L., and Steyvers, M. 2004. Finding Scientific Topics. *Proceedings of the National Academy of Sciences* 101(suppl 1):5228–5235.

Hall, P., and Hannan, E. 1988. On Stochastic Complexity and Nonparametric Density Estimation. *Biometrika* 75(4):705–714.

Iwata, T.; Watanabe, S.; Yamada, T.; and Ueda, N. 2009. Topic Tracking Model for Analyzing Consumer Purchase Behavior. In *IJCAI*, volume 9, 1427–1432.

Jacobsen, M. 2006. *Point Process Theory and Applications: Marked Point and Piecewise Deterministic Processes*. Springer Science & Business Media.

Kim, H.; Takaya, N.; and Sawada, H. 2014. Tracking

Temporal Dynamics of Purchase Decisions via Hierarchical Time-Rescaling Model. In *CIKM*, 1389–1398. ACM.

Koren, Y.; Bell, R.; Volinsky, C.; et al. 2009. Matrix Factorization Techniques for Recommender Systems. *Computer* 42(8):30–37.

Koren, Y. 2010. Collaborative Filtering with Temporal Dynamics. *Communications of the ACM* 53(4):89–97.

Kurashima, T.; Iwata, T.; Irie, G.; and Fujimura, K. 2010. Travel Route Recommendation Using Geotags in Photo Sharing Sites. In *CIKM*, 579–588. ACM.

Li, L.; Deng, H.; Dong, A.; Chang, Y.; and Zha, H. 2014. Identifying and Labeling Search Tasks via Query-based Hawkes Processes. In *KDD*, 731–740. ACM.

Minka, T. 2000. Estimating a Dirichlet Distribution.

Murphy, K. P. 2012. *Machine Learning: a Probabilistic Perspective*. MIT press.

Neal, R. M. 2003. Slice Sampling. *Annals of Statistics* 705–741.

Ogata, Y. 1988. Statistical Models for Earthquake Occurrences and Residual Analysis for Point Processes. *Journal of the American Statistical Association* 83(401):9–27.

Prigent, J.-L. 2001. Option Pricing with a General Marked Point Process. *Mathematics of Operations Research* 26(1):50–66.

Silverman, B. W. 1986. *Density Estimation for Statistics and Data Analysis*, volume 26. CRC press.

Snyder, D. L. 1975. *Random Point Processes*. Wiley New York.

Wang, J., and Zhang, Y. 2013. Opportunity Model for E-commerce Recommendation: Right Product; Right Time. In *SIGIR*, 303–312. ACM.

Wang, J.; Sarwar, B.; and Sundaresan, N. 2011. Utilizing Related Products for Post-purchase Recommendation in E-commerce. In *RecSys*, 329–332. ACM.

Xu, G.; Zhang, Y.; and Yi, X. 2008. Modelling User Behaviour for Web Recommendation Using LDA Model. In *Web Intelligence and Intelligent Agent Technology*, volume 3, 529–532. IEEE.

Ye, M.; Yin, P.; Lee, W.-C.; and Lee, D.-L. 2011. Exploiting Geographical Influence for Collaborative Point-of-interest Recommendation. In *SIGIR*, 325–334. ACM.

Yuan, Q.; Cong, G.; Ma, Z.; Sun, A.; and Thalmann, N. M. 2013. Time-aware Point-of-interest Recommendation. In *SIGIR*, 363–372. ACM.

Zhao, G.; Lee, M. L.; Hsu, W.; and Chen, W. 2012. Increasing Temporal Diversity with Purchase Intervals. In *SIGIR*, 165–174. ACM.