A Spherical Hidden Markov Model for Semantics-Rich Human Mobility Modeling

Wanzheng Zhu*, Chao Zhang*, Shuochao Yao, Xiaobin Gao, Jiawei Han
University of Illinois at Urbana-Champaign, Urbana, IL, USA
{wz6, czh82, syao9, xgao16, hanj}@illinois.edu

Abstract
We study the problem of modeling human mobility from semantic trace data, wherein each GPS record in a trace is associated with a text message that describes the user’s activity. Existing methods fall short in unveiling human movement regularities for such data, because they either do not model the text data at all or suffer from text sparsity severely. We propose SHMM, a multi-modal spherical hidden Markov model for semantics-rich human mobility modeling. Under the hidden Markov assumption, SHMM models the generation process of a given trace by jointly considering the observed location, time, and text at each step of the trace. The distinguishing characteristic of SHMM is the text modeling part. We use fixed-size vector representations to encode the semantics of the text messages, and model the generation of the $l_2$-normalized text embeddings on a unit sphere with the von Mises-Fisher (vMF) distribution. Compared with other alternatives like multi-variate Gaussian, our choice of the vMF distribution not only incurs much fewer parameters, but also better leverages the discriminative power of text embeddings in a directional metric space. The parameter inference for the vMF distribution is non-trivial since it involves functional inversion of ratios of Bessel functions. We theoretically prove, for the first time, that: 1) the classical Expectation-Maximization algorithm is able to work with vMF distributions; and 2) while closed-form solutions are hard to be obtained for the M-step, Newton’s method is guaranteed to converge to the optimal solution with quadratic convergence rate. We have performed extensive experiments on both synthetic and real-life data. The results on synthetic data verify how possible it is to interpret human mobilities in a semantics-rich way. In addition to uncovering the spatiotemporal patterns of human movements, we could move one step further to understand what are people’s activities at different regions, and how and why people move from one region to another. Such semantics-rich knowledge not only enables us to understand human mobility in a more interpretable way, but also plays an important role for prediction and decision making in various downstream applications.

Unfortunately, learning semantics-rich human mobility models from semantic trace data is a challenging problem that remains largely unsolved by existing techniques. Traditional mobility modeling techniques (Giannotti et al. 2007; Li et al. 2010; Cho, Myers, and Leskovec 2011; Mathew, Raposo, and Martins 2012) mostly focus on mining pure spatiotemporal regularities and cannot handle the text information in semantic traces. Recently, there have been research efforts that attempt to integrate the text information into the mobility modeling process based on the bag-of-words model (Ying et al. 2011; Wu et al. 2015; Zhang et al. 2016a). Nevertheless, these methods are unable to make the best use of the text information. First, by considering each keyword as an independent dimension, they do not model the correlations between keywords (e.g., ‘car’, ‘taxi’ and ‘drive’) and may fail to correlate semantically similar messages. Further, since the vocabulary size is often large, their performance is limited by the high dimensionality and text sparsity, and meanwhile results in high computational overhead in the model learning process.

Introduction
Uncovering human mobility patterns is not only a fundamental task for human behavioral analysis, but also an important building block for urban planning, traffic forecasting, mobile health applications, and location-based recommender systems (Gonzalez, Hidalgo, and Barabasi 2008).

*Equal contribution.
To learn semantics-rich mobility models from semantic trace data, we propose SHMM, a method that uncovers human mobility regularities by statistically modeling the generation process of the given trace data. SHMM is a multi-modal spherical hidden Markov model (HMM). Under the hidden Markov assumption, it jointly models the generation of the observed location, time, and text at each step of an input trace. While the low-dimensional location and time can be modeled with Gaussian distributions, the key challenge is to capture the semantics of the high-dimensional text messages and model textual semantics. To address this challenge, we use fixed-size vector representations to encode the semantics of the text messages, which has been recently shown successful for a wide variety of NLP tasks (Mikolov et al. 2013a, Mikolov et al. 2013b). With the derived text embeddings, we further model the generation of the $l_2$-normalized text embeddings on a unit sphere with the von Mises-Fisher (vMF) distribution (Fisher 1953). Compared with other alternatives like multi-variate Gaussian, our choice of the vMF distribution not only incurs much fewer parameters, but also better unleashes the discriminative power of text embeddings in a directional metric space.

In the parameter inference process of our SHMM model, we use the classical Expectation-Maximization (EM) algorithm (Dempster, Laird, and Rubin 1977). However, since the vMF distribution has a complicated mathematical form, literature so far has not yet proved that EM algorithm is able to work with the vMF distribution. We, for the first time, theoretically prove the feasibility of using the EM algorithm on the vMF distribution. Furthermore, while obtaining closed-form solutions are hard to be obtained for the M-step, we prove that using Newton’s method is guaranteed to converge to the optimal solution with quadratic convergence rate.

To summarize, we make the following contributions:

1. We propose a spherical hidden Markov model for human mobility modeling with semantic trace data. Compared with existing mobility models, our method is novel in that it uses embeddings to encode the semantics of text messages and the von-Mises Fisher distribution to model the generation of text embeddings.

2. We provide rigorous theoretical proof to show that the EM algorithm is able to work with the vMF distribution. Also, while obtaining closed-form solutions for the M-step is intractable, we prove that Newton’s method is guaranteed to converge to the optimal solution with quadratic convergence rate. Some other properties of the vMF distribution and modified Bessel functions are also studied.

3. We perform extensive experiments on both synthetic and real-life data. The results on synthetic data verify our theoretical analysis; while the results on real-life data demonstrate that SHMM learns meaningful semantics-rich mobility models, outperforms state-of-the-art mobility models for next location prediction, and incurs lower training cost.

**Problem Description**

We study the problem of modeling human mobility from semantic trace data. Semantic traces are text-rich GPS traces where each GPS record is associated with a text message that describes the user’s activity. We provide the formal definition of a semantic trace as follows.

**Definition 1 (Semantic Trace).** The semantic trace of a user $u$ is a time-ordered sequence $S = [x_1, x_2, \ldots, x_R]$. Each record $x_i$ is a tuple $(t_i, l_i, m_i)$ where: (1) $t_i$ is the timestamp scalar; (2) $l_i$ is a two-dimensional vector representing the location of the user at time $t_i$; and (3) $m_i$ is a text message describing the activity of the user.

To capture the semantics of user activities, we use distributed representations for the text messages in our model. Specifically, we first use the CBOW model (Mikolov et al. 2013b) to obtain fixed-size vector representations (i.e., embeddings) for the keywords in the given corpus. The parameters used are: $\text{min-count}=10$, $\text{size}=30$, $\text{window}=5$, $\text{sample}=10^{-4}$, $\text{negative}=5$. As each text message usually consists of multiple keywords, we compute the TF-IDF weights of the keywords and use the weighted average of keyword embeddings to derive the embedding of the message $m_i$ (Le and Mikolov 2014, Arora, Liang, and Ma 2016).

Now we are ready to formulate our mobility modeling problem. Given the semantic traces for a group of users $D = \{S_1, S_2, \ldots, S_C\}$, our work aims to build semantics-rich mobility models for those users. The result mobility model is expected to address two aspects regarding human mobility: (1) **Discovering latent states.** The first aspect is to discover the latent states that govern people’s activities. A latent state is an abstraction of what the user is doing around where during when. Examples include shopping in the 5th Ave at 5pm, and watching a film at the AMC theater in the evening. (2) **Characterizing transition regularity.** The second aspect is to characterize how users move sequentially between the latent states. For example, after shopping in the 5th Ave, what activities will the users do next? We aim to characterize people’s transitions among the latent states in a concise and probabilistic way.

**The SHMM Model**

In this section, we describe SHMM in detail and describe the parameter inference procedure.

**Model Description**

Consider a sequence of chronologically ordered records $x_1, x_2, \ldots, x_R$ of a user $u$. In SHMM, we adopt the hidden Markov assumption, i.e., assuming each record $x_i$ is generated from a latent state $z_i$, and the sequence $z_1, z_2, \ldots, z_R$ follows a Markov process. The Markov process is parameterized by an initial probability matrix $\pi$ over the latent states, as well as a matrix $A$ that specifies the transition probabilities among the latent states. When generating $x_i$ from $z_i$, we assume the location $l_i$, the timestamp $t_i$, and the text embedding $m_i$ are generated independently. Therefore, the conditional probability $p(x_i | z_i)$ is given by $p(x_i | z_i) = p(t_i | z_i) \cdot p(l_i | z_i) \cdot p(m_i | z_i)$.

For each record $x_i$, we assume the following distributions for each component: 1) the timestamp $t_i$ is generated from a univariate Gaussian with mean $\mu_i$ and variance $\sigma_i$, i.e. $p(t_i | z_i) = N(t_i | \mu_i, \sigma_i)$, where $t_i$ indicates
the time in a day; 2) the location \( l_i \) is generated from a bi-
variate Gaussian with mean \( \mu_l \) and covariance matrix \( \Sigma_l \),
i.e. \( p(l_i|z_i) = N(l_i|\mu_l, \Sigma_l) \); 3) the message vector \( m_i \) is
generated from the von Mises-Fisher (vMF) distribution with mean direction \( \mu \) and concentration parameter \( \kappa \), i.e.
\( p(m_i|z_i) = \text{vMF}(m_i|\mu, \kappa) \).

While the Gaussian distribution is suitable for modeling
timestamps and locations, it is problematic for modeling text
embeddings. The reason is two-fold. First, using Gaussian
distributions to model text embeddings would lead to a large
covariance matrix with too many parameters. Second, pre-
vious research has demonstrated that the cosine distance better
reflects the semantic similarity between text embeddings
compared to the Euclidean distance, i.e., the discriminative
power of the text embeddings is stronger in a directional
metric space. Our choice of the vMF distribution addresses
the above two issues. A vMF distribution is defined on the
\( p \)-dimensional unit sphere, parameterized by two parame-
ters: a mean direction \( \mu \) and a concentration parameter \( \kappa \).
The mean \( \mu \) specifies the direction of the semantic focus of
the text embeddings, while \( \kappa \) controls how concentrated the
text embeddings are around the mean direction. The larger
\( \kappa \) is, the more concentrated the text embeddings are around
the mean direction. Formally, the probability density func-
tion of a vMF distribution for a \( p \)-dimensional unit vector \( m \)
is defined as:

\[
f_p(m; \mu, \kappa) = C_p(\kappa) \exp(\kappa \mu^T m),
\]

where \( ||\mu|| = 1 \), \( \kappa \geq 0 \), \( C_p(\kappa) = \frac{\kappa^{p/2-1}}{(2\pi)^{p/2} I_{p/2-1}(\kappa)} \), and \( p \)
is the dimension of the vector. \( I_v(\cdot) \) is the modified Bessel
function of the first kind at order \( v \) and is defined as
\( I_v(\kappa) = \sum_{q=0}^{\infty} \frac{1}{q!(q+v+1)} \left( \frac{\kappa}{2} \right)^{2q+v} \), and \( \Gamma(\cdot) \) is the gamma function.

Parameter Inference

In our SHMM model, the parameters to be estimated are the
parameters \( \pi, A \) for the hidden states and the distribution
parameters \( \mu_t, \sigma_t, \mu_l, \Sigma_l, \mu, \kappa \). Since we are using standard
Gaussian distributions for modeling location and time, the
updating rules for all the parameters except \( \kappa \) can be easily
derived by the Baum-Welch algorithm (Baum et al. 1970) —
an Expectation-Maximization procedure for HMM. The challenge
of applying the Baum-Welch algorithm is how to estimate the parameter \( \kappa \).

Due to the complicated form of the vMF distribution, it is
intractable to derive closed-form solutions for \( \kappa \) in the
M-step of the Baum-Welch algorithm. However, we found
that one can use Newton’s method to find an approximate
solution of \( \kappa \) that asymptotically converges to the optimal
value. Below we first present our method for updating \( \kappa \)
based on Banerjee’s work (Banerjee et al. 2005) and New-
ton’s method, and then show our theoretical analysis that our
update rule achieves quadratic convergence rate. In the M-
step of the Baum-Welch algorithm, we estimate the \( \kappa \) value
with the following iterative procedure:

\[
\bar{\kappa} \leftarrow \frac{||\sum_{i=1}^N m_i||}{N}
\]

\[
\kappa \leftarrow \bar{\kappa} - \frac{\bar{\kappa} - \bar{\kappa}^2}{1 - A_p(\kappa) - \frac{p-1}{\kappa} A_p(\kappa)}
\]

repeat

\[
\kappa \leftarrow \kappa - \frac{A_p(\kappa) - \bar{\kappa}}{1 - A_p(\kappa) - \frac{p-1}{\kappa} A_p(\kappa)}
\]

until convergence

where \( m_i \) is a \( l \)-dimensional \( p \)-dimensional text embedding,
\( N \) is the total number of text embeddings belonging to the
current state, and \( A_p(\kappa) = \frac{I_{p/2}(\kappa)}{\Gamma_{p/2-1}(\kappa)} \).

Theoretical Analysis

Due to the complicated mathematical form of the vMF dis-
tribution, no existing literature has proved that the vMF dis-
tribution can work under the EM framework. In this section,
we theoretically prove that:

1. The EM algorithm is able to work with the vMF distri-
bution, because there exists a unique \( \kappa \) such that the Q-
function in the M-step can be maximized.

2. While closed-form solutions for \( \kappa \) are hard to be obtained,
one can use Newton’s method for obtaining an approxi-
mate solution, which is guaranteed to converge to the opti-
mal \( \kappa \) for M-step with quadratic convergence rate.

Theorem 1. There exists a unique \( \kappa \) that maximizes the Q-
function in the M-step of the EM algorithm.

Proof. To maximize the Q-function, it is equivalent to solve
\( A_p(\kappa) = \bar{\kappa} \) where \( \bar{\kappa} = \frac{||\sum_{i=1}^N m_i||}{N} \) (Banerjee et al. 2005).
Based on this result, we have the following claims:

1. Claim 1: \( 0 < \bar{\kappa} \leq 1 \).

   Proof. It is obvious that \( \bar{\kappa} > 0 \). Since \( m_1^2 + m_2^2 + \ldots + m_p^2 = 1 \), we have
   \( ||\sum_{i=1}^N m_i||^2 \leq (m_1 + \ldots + m_N)^2 + \ldots + (m_1 + \ldots + m_N)^2 \leq N (m_1^2 + \ldots + m_N^2) \).
   Hence, \( \bar{\kappa} = \frac{||\sum_{i=1}^N m_i||}{N} \leq 1 \).

2. Claim 2: \( \lim_{\kappa \to 0} A_p(\kappa) = 0 \), \( \lim_{\kappa \to \infty} A_p(\kappa) = 1 \).

   Proof. The first equation is given by Lemma 2.1 in (Seg-
gura 2011). With Corrollary 1 in (Balachandran, Viles,
and Kolaczyk 2013), we have \( \exp(-\frac{\bar{\kappa}}{\kappa^2}) \leq A_p(\kappa) \leq \exp(-\frac{\bar{\kappa}}{\kappa^2}) \),
where \( \alpha_0 = -\log(\sqrt{2} - 1) \), if \( p \leq 2 \kappa \).
Hence, we have \( \lim_{\kappa \to 0} \exp(-\frac{\bar{\kappa}}{\kappa^2}) \leq A_p(\kappa) \leq \lim_{\kappa \to \infty} \exp(-\frac{\bar{\kappa}}{\kappa^2}) \).
Therefore, \( \lim_{\kappa \to \infty} A_p(\kappa) = 1 \).

3. Claim 3: \( A_p(\kappa) \) is a continuous function if \( \kappa \) is real-valued
and positive.

   Proof. By the definition of the modified Bessel function and its recurrence relation (Equation 9.6.1 and 9.6.26 in
(Abramowitz and Stegun 1964)), we can get \( A_p(\kappa) = 1 - A_p(\kappa) - \frac{p-1}{\kappa} A_p(\kappa) \).
Since \( A_p(\kappa) \) is differentiable when \( \kappa \) is real-valued and positive, \( A_p(\kappa) \) is continuous at all real-valued positive \( \kappa \).
By the intermediate value theorem and the above claims, we have that there exists a solution for \( A_p(\kappa) = \bar{r} \). Since we have \( A_p'(\kappa) > 0 \) for all positive \( \kappa \) (Equation 15 in [Amos 1974]), there exists a unique \( \kappa \) such that \( A_p(\kappa) = \bar{r} \) and the Q-function is maximized. This completes the proof. \( \square \)

So far, we have proved that there exists one unique \( \kappa \) to converge to the solution. This implies the EM algorithm is able to work with vMF distributions. However, it is non-trivial to find the optimal \( \kappa \) since it involves calculating the ratios of modified Bessel functions. Next, we show that the solution can be found by Newton’s Method.

**Lemma 1.** \( A_p(\kappa) \) is a strictly concave function, when \( \kappa > 0 \) and \( p \geq 2 \).

**Proof.** It is shown in Theorem 3 in [Simpson and Specter 1984] that \( \kappa \frac{\kappa}{A_p(\kappa)} \) is strictly convex for \( \kappa \geq 0, p \geq 2 \). Thus, \( \kappa \frac{A''(\kappa)}{A_p(\kappa)} - 2 \frac{A_p(\kappa)}{A_p(\kappa)} + \frac{A_p'(\kappa)}{A_p(\kappa)} - 1 > 0 \). Since \( A_p(\kappa) > 0 \), we get \( A_p''(\kappa) > 0 \). Therefore, \( \kappa \frac{A''(\kappa)}{A_p(\kappa)} - 2 \frac{A_p(\kappa)}{A_p(\kappa)} + \frac{A_p'(\kappa)}{A_p(\kappa)} - 1 > 0 \).

Next, we have: 1) \( A_p''(\kappa) > 0 \) and \( \kappa A_p''(\kappa) - A_p'(\kappa) < 0 \). (Equation 15 in [Amos 1974]); 2) \( A_p(\kappa) = \frac{\kappa}{A_p(\kappa)} > 0 \), since \( I_v(\kappa) \) is positive when \( v > 0 \) and \( \kappa > 0 \) (Equation 9.6.1 in [Abramowitz and Stegun 1964]). Therefore, \( A_p''(\kappa) < 0 \) and it is strictly concave. This completes the proof of Lemma 1. \( \square \)

Building on Lemma 1, we proceed to show that Newton’s Method is guaranteed to converge to the solution for \( A_p(\kappa) = \bar{r} \).

**Theorem 2.** Newton’s method is guaranteed to converge to the solution for \( A_p(\kappa) = \bar{r} \).

**Proof.** Assume \( \kappa = r \) is the solution, i.e. \( A_p(r) = \bar{r} = 0 \). Let’s start with a point \( \kappa_0 \) and \( 0 < \kappa_0 \leq r \). We define \( e_n = \kappa_n - r \). By Newton’s updating rule, we have

\[
e_{n+1} = e_n - \frac{A_p(\kappa_n) - \bar{r}}{A_p'(\kappa_n)}
\]  

(1)

By Taylor’s Theorem, we have \( A_p(r) = A_p(\kappa_n - e_n) = A_p(\kappa_n) - e_n A_p'(\kappa_n) + \frac{1}{2} e_n^2 A_p''(\xi_n) = \bar{r} \), where \( \kappa_n \leq \xi_n \leq r \). Then we have

\[
e_{n+1} = \frac{e_n A_p''(\kappa_n) - A_p'(\kappa_n) + \bar{r}}{A_p'\kappa_n)} \frac{1}{2} e_n^2
\]  

(2)

Therefore, \( e_n \leq 0 \) for all \( n \) since \( A_p''(\xi_n) < 0 \) and \( A_p''(\kappa_n) > 0 \). That implies \( \kappa_n \leq r \) and \( A_p(\kappa_n) \leq p - \bar{r} \) for all \( n \). Therefore, from Equation (1), we have \( e_{n+1} \geq e_n \). Thus, the sequence \( e_0, e_1, \ldots, e_n \) is an increasing sequence and bounded by 0, and therefore, it must have a limit \( e^* \). Accordingly, the sequence \( \kappa_0, \kappa_1, \ldots, \kappa_n \) must have a limit \( \kappa^* \). From Equation (1), we have \( \lim_{n \to \infty} e_{n+1} = \lim_{n \to \infty} (e_n - \frac{A_p(\kappa_n) - \bar{r}}{A_p'(\kappa_n)} \), and thus \( A_p(\kappa^*) = \bar{r} \) and \( \kappa^* = r \). Therefore, if we choose any positive starting point \( \kappa_0 \), such that \( \kappa_0 \leq r \), Newton’s method is guaranteed to converge to the solution. \( \square \)

We have shown that Newton’s method is guaranteed to converge to the solution for \( A_p(\kappa) = \bar{r} \). The update rule is simply given by \( \kappa_{n+1} = \kappa_n - \frac{A_p(\kappa_n) - \bar{r}}{A_p'(\kappa_n)} \), where \( A_p'(\kappa_n) = 1 - A_p''(\kappa_n) \). Next, we show the convergence rate of Newton’s method.

**Lemma 2.** The rate of convergence for calculating \( A_p(\kappa) = \bar{r} \) using Newton’s Method is quadratic and \( e_{n+1} \approx C e_n^2 \) with \( C \in (-1, 0) \).

**Proof.** Equation (2) shows the convergence rate is quadratic. Then, we have \( e_{n+1} = \frac{1}{2} A_p''(\kappa_n) e_n^2 \approx \frac{1}{2} A_p''(\kappa_n) e_n^2 \) and

\[
A_p''(\kappa_n) = \left( -\frac{2A_p(\kappa_n)}{A_p(\kappa_n)} - \kappa \right) A_p'(\kappa_n) + \frac{p - 1}{\kappa^2} A_p(\kappa_n)
\]

where the last inequality uses Equation 15 in [Amos 1974]. Therefore, \( A_p''(\kappa_n) \in (-2A_p(\kappa_n), 0) \). We have shown that \( \lim_{\kappa \to 0} A_p(\kappa) = 0 \), \( \lim_{\kappa \to \infty} A_p(\kappa) = 1 \) and \( A_p'(\kappa) > 0 \) in Theorem 1, and hence \( A_p(\kappa) \in (0, 1) \). Therefore, \( e_{n+1} \approx C e_n^2 \), where \( C \in (-1, 0) \). \( \square \)

**Experiments**

In this section, we empirically evaluate the performance of SHMM. We implemented SHMM and the baseline methods in JAVA and conducted all the experiments on a computer with 2.9 GHz Intel Core i7 CPU and 16GB memory.

**Experimental Setup**

**Data** We use both synthetic and real-life datasets to evaluate SHMM. The real-life datasets are semantic traces from Twitter users collected by Zhang et al. [Zhang et al. 2016a]. The first dataset (LA) consists of million-scale geo-tagged tweets created by Los Angeles users from 2014.08.01 to 2014.11.30. Following the preprocessing steps in [Zhang et al. 2016a], we first group the tweets by user ID to obtain the location history for each user. Since two consecutive records in a raw location history can be large, we further segment the location history into dense semantic traces with a time threshold \( \Delta t = 6 \) h, such that the time gap between any two consecutive records is no larger than 6 hours. After preprocessing, we obtain approximately 30 thousand semantic traces. The second dataset (NY) consists of the geo-tagged tweets in New York City during 2014.08.01 and 2014.11.30. We preprocess the NY data in a similar manner and obtain approximately 42 thousand trajectories in total.

We also generate multiple synthetic data to verify the theoretical analysis of our SHMM model. For generating points from the vMF distribution, we use the code from Chen et al. [Chen et al. 2013]. Given the synthetic points, we apply our proposed Newton’s method for estimating the parameters of the SHMM model, and evaluate the approximation errors and convergence speed.
Baseline Methods  We compare our SHMM model with the following baseline methods:

1. Law (Brockmann, Hufnagel, and Geisel 2006) is a widely used mobility model based on the Lévy flight law with long-tailed distributions.

2. HMM (Mathew, Raposo, and Martins 2012) uses HMM to model the spatial locations in the trace data for mobility modeling.

3. ST-HMM is an extension of HMM that models both spatial and temporal information in the trace data.

4. GMove (Zhang et al. 2016a) is the state-of-the-art mobility model for semantic traces. It differs from SHMM in the text modeling part. It uses the bag-of-words model to represent text messages and uses multinomial distributions to generate the observed messages.

5. GHMM is an adaption of SHMM. It uses independent Gaussians to model text vectors instead of the vMF distribution.

Evaluation Protocol  We use next location prediction as a downstream task for evaluating the quality of the SHMM model. Given a semantic trace dataset, we randomly select 70% traces for model training and use the rest 30% for testing. For a test trajectory $[x_1, x_2, \ldots, x_R]$, we assume the first $R-1$ locations $[x_1, x_2, \ldots, x_{R-1}]$ are observed and attempt to recover the last record $x_R$. Specifically, we first form a candidate pool by mixing $x_R$ with other records whose creating time and distances are close to $x_R$. After the candidate pool is formed, we use the SHMM model to select the top-$K$ most likely visited records and see whether the ground-truth appears in the top-$K$ list. We use the prediction accuracy $@K$ to measure the performance of different models, i.e., the percentage of test traces for which the ground-truth record is recovered by the top-$K$ list.

Results on Synthetic Data

Figure 1 shows the performance of our used Newton’s method for estimating the parameters of a vMF distribution on synthetic data. As shown in Figure 1(a), our estimation method converges extremely fast for estimating $\kappa$, achieving approximation errors smaller than $10^{-13}$ after three iterations. Figure 1(b) shows the $\mu$ and $\kappa$ estimation performance on synthetic datasets with different sizes. We can see the approximation error tends to become smaller on synthetic data sets with more samples. This is expected, as a small number of samples may lead to biased estimations of the true parameter values. The results in Figure 2 shows the estimation performance for different $\kappa$ and dimension $p$ (the number of samples is 100,000). Generally, we do not observe obvious patterns showing how the approximation errors change with different $\kappa$ and $p$, but the relative approximation errors are quite small under different $\kappa$ and $p$ values.

Results on Real-Life Data

Visualization of the Mobility Models  In this set of experiments, we set the number of states to 50 on LA and NY to obtain mobility models. After parameter inference, each state is characterized by: (1) a two-dimensional Gaussian distribution for the spatial location; (2) a one-dimensional Gaussian distribution for the time; and (3) a 30-dimensional vMF distribution for the semantics.

Figure 3 visualizes a number of representative states and some frequent transitions among them. We plot the mean location of some states as well as the top-10 keywords from the vocabulary whose embeddings are the closest to the vMF mean directions. Most of the top-10 keywords for the same state carry consistent and clear semantics. For example, for the BASEBALL state on the LA dataset: *homeruns* is a specific baseball term; *dodgers* is the baseball team in LA; *giants* is the baseball team in San Francisco; *162* indicates that there are 162 games for each team in the Major League Baseball (MLB) season; and the rest six keywords are all related to baseball too. We have examined the center locations of the states, and found that the geographical locations well match the semantic meanings of different states.

Another interesting finding is that in LA dataset, the mean direction of the General Sports state lies in-between the Basketball state and the Baseball state in the embedding space. Also, the concentration parameter $\kappa$ for General Sports state is lower than Basketball and Baseball state. Such phenomenon intuitively makes sense since General Sports is a broader topic and the semantics of the tweets are more scattered.

We have also observed some interesting state transitions. As shown in Figure 3, the following transitions receive high...
Average Accuracy

0.4
0.6
0.8

Accuracy

0.2
0.4
0.6
0.8

top-1 top-2 top-3 top-4 top-5

State 1: Basketball
lakers grizzles celtics nba bryant home runs giants dodgers 2-May

State 2: Baseball

State 3: General Sports

State 4: Restaurants/ Food

State 5: Beach

State 6: Universal Studios

State 7: Airport

(ny)

Figure 3: The visualization of the mobility models. The digits (1-7) indicate different state locations and the letters (A-D) indicate frequent human movements.

Figure 4: The accuracies of top-$K$ location prediction.

Figure 5: Prediction accuracy v.s. the number of states.

(b) NY dataset state keywords and transitions

(b) NY dataset state keywords and transitions

(b) NY dataset state keywords and transitions

(b) NY dataset state keywords and transitions

(b) NY dataset state keywords and transitions

(b) NY dataset state keywords and transitions

(b) NY dataset state keywords and transitions

Figure 3: The visualization of the mobility models. The digits (1-7) indicate different state locations and the letters (A-D) indicate frequent human movements.

Figure 4: The accuracies of top-$K$ location prediction.

Figure 5: Prediction accuracy v.s. the number of states.

Effects of Parameters In Figure 5, we study the performance of SHMM and GMove when the number of states varies. We find that the performance of both models generally increases with the number of states. One major reason is that the semantics of people’s activities are separated at finer granularities when the number of states is large. For example, we can see from the LA data set that General Sports, Basketball and Baseball are three separated topics. If the number of states is not large enough, these states may be clustered as one single topic. On the other hand, one caveat for choosing the number of states is that a large number of states could incur high computational overhead, and also harm the interpretability of the result model because the same semantics may be split into duplicate ones.

Efficiency Comparison Finally, we compare the modeling training time between SHMM and GMove when the
number of states varies. Generally speaking, the training time of both models increases quadratically with the number of states. The training time of SHMM is obviously smaller than that of GMove, e.g., when the number of states is 100, training SHMM is 25.7% faster on LA and 40.9% faster on NY, and the speedups are even larger when the number of states or the data size increases. This is because SHMM models low-dimensional text embeddings instead of high-dimensional bag-of-words, and thus involves much fewer parameters. In addition, the estimation of the parameters for the vMF distribution is cheap and converges fast.

![Figure 6: Running time v.s. the number of states.](a) LA
(b) NY

**Related Work**

**Human mobility modeling.** Classic human mobility modeling methods focus on mining the spatiotemporal regularities underlying human movements. Generally, existing mobility modeling methods can be divided into two categories: pattern-based methods and model-based methods. Pattern-based methods aim at discovering specific mobility patterns that occur regularly. Different mobility patterns have been introduced to capture people’s movement regularities, such as frequent sequential patterns (Giannotti et al. 2007), periodic patterns (Li et al. 2010), and co-location patterns (Kalnis, Mamoulis, and Bakiras 2005). Model-based methods use statistical models to characterize the human mobility, and learn the parameters of the designed model from the observed trace data. Mathew et al. (Mathew, Raposo, and Martins 2012) use the hidden Markov model to capture the sequential transition regularities of human mobility; Brockmann et al. (Brockmann, Hufnagel, and Geisel 2006) proposed that human mobility can be modeled by a continuous-time random-walk model with long-tail distribution; Cho et al. (Cho, Myers, and Leskovec 2011) introduce periodic mobility models to discover the periodicity underlying human movements.

While the above mobility modeling methods focus on spatiotemporal regularities without considering text data, recent years are witnessing growing interest in modeling human mobility from semantic trace data (Ying et al. 2011; Wu et al. 2015; Zhang et al. 2016a; Zhang et al. 2014). Among these works, the state-of-the-art GMove model (Zhang et al. 2016a) is the most relevant to our model. Both GMove and SHMM use hidden Markov models to model the generation process of the observed semantic trace data. However, SHMM is different from GMove in that it encodes the semantics of user activities with text embeddings, and uses the vMF distribution to model the text embeddings in the HMM model. As such, the SHMM involves much fewer parameters and well unleashes the discriminative power of text embeddings in a directional metric space.

It is worth mentioning that, there are quite a number of works that uses human trace data for the location prediction problem (Wang et al. 2015; Liu and et al. 2016). Typically, they extract features that are important for predicting which place the user tends to visit next based on discriminative models such as recurrent neural networks. While we use location prediction as an evaluation task in our experiments, our work is quite different from these works. Instead of optimizing the performance of location prediction, our focus is to learn interpretable models that reveals the regularities underlying human movements. Besides location prediction, our learned mobility models can be used for many other downstream tasks as well.

**vMF-based learning.** There are some existing works that utilize vMF distribution for different learning tasks. Dhillon et al. (Dhillon and Sra 2003) and Banerjee et al. (Banerjee et al. 2005) are two pioneering works that use the vMF distributions to handle directional data, which demonstrate inspiring results for text categorization and gene expression analysis. Besides, Gopal and Yang (Gopal and Yang 2014) recently applied vMF distributions for clustering analysis, and proposed variational inference procedures for estimating the parameters of the vMF clustering model. Batmanghelich et al. (Batmanghelich et al. 2016) proposed a spherical topic model based on the vMF distribution, which accounts for word semantic regularities in language and has been demonstrated to be superior than multi-variate Gaussian distributions. However, there are no previous works that integrate the vMF distribution with HMMs for semantic trace data. To the best of our knowledge, we are the first to demonstrate that the vMF distribution can work well with HMM for directional data with theoretical guarantees.

**Conclusion**

We proposed a spherical hidden Markov model for learning interpretable human mobility model from semantic trace data. Our model uses text embeddings to capture the semantics of text messages and integrate the vMF distribution into the hidden Markov model for generating such text embeddings. We have theoretically proved that the Expectation-Maximization algorithm is able to work with vMF distribution, and that the Newton’s method can be applied for efficiently solving the M-step with quadratic convergence rate. Our experiments on synthetic data simulations verify our theoretical analysis. Furthermore, by applying our model to real-life semantic trace datasets, we are able to obtain highly interpretable mobility models, which intuitively make sense and outperform baseline models for downstream tasks like location prediction.

**Acknowledgments**

This work was sponsored in part by the U.S. Army Research Lab. under Cooperative Agreement No. W911NF-09-2-
References

[Brockmann, Hufnagel, and Geisel 2006] Brockmann, D.; Hufnagel, L.; and Geisel, T. 2006. Handbuch of mathematical functions: with formulas, graphs, and mathematical tables; volume 55. Courier Corporation.

[Amos 1974] Amos, D. E. 1974. Computation of modified bessel functions and their ratios. Mathematics of Computation 28(125):239–251.

[Amos 1974] Amos, D. E. 1974. Computation of modified bessel functions and their ratios. Mathematics of Computation 28(125):239–251.

[Arora, Liang, and Ma 2016] Arora, S.; Liang, Y.; and Ma, T. 2016. A simple but tough-to-beat baseline for sentence embeddings.

[Balachandran, Viles, and Kolaczyk 2013] Balachandran, P.; Viles, W.; and Kolaczyk, E. D. 2013. Exponential-type inequalities involving ratios of the modified bessel function of the first kind and their applications. arXiv preprint arXiv:1311.1450.

[Banerjee et al. 2005] Banerjee, A.; Dhillon, I. S.; Ghosh, J.; and Sra, S. 2005. Clustering on the unit hypersphere using von mises-fisher distributions. Journal of Machine Learning Research 6(Sep):1345–1382.

[Batmanghelich et al. 2016] Batmanghelich, K.; Saeedi, A.; Narasimhan, K.; and Gershman, S. 2016. Nonparametric spherical topic modeling with word embeddings. arXiv preprint arXiv:1604.00126.

[Baum et al. 1970] Baum, L. E.; Petrie, T.; Soules, G.; and Weiss, N. 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. The annals of mathematical statistics 41(1):164–171.

[Brockmann, Hufnagel, and Geisel 2006] Brockmann, D.; Hufnagel, L.; and Geisel, T. 2006. The scaling laws of human travel. arXiv preprint cond-mat/0605511.

[Chen et al. 2015] Chen, Y.-H.; Wei, D.; Newstadt, G.; De Graef, M.; Simmons, J.; and Hero, A. 2015. Parameter estimation in spherical symmetry groups. IEEE Signal Processing Letters 22(8):1152–1155.

[Cheng et al. 2011] Cheng, Z.; Caverlee, J.; Lee, K.; and Sui, D. Z. 2011. Exploring millions of footprints in location sharing services. In ICWSM.

[Cho, Myers, and Leskovec 2011] Cho, E.; Myers, S. A.; and Leskovec, J. 2011. Friendship and mobility: User movement in location-based social networks. In KDD, 1082–1090.

[Dempster, Laird, and Rubin 1977] Dempster, A. P.; Laird, N. M.; and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the em algorithm. Journal of the royal statistical society. Series B (methodological) 1–38.

[Dhillon and Sra 2003] Dhillon, I. S., and Sra, S. 2003. Modeling data using directional distributions. Technical report, Technical Report TR-03-06, Department of Computer Sciences, The University of Texas at Austin.

[Dhillon et al. 2003] Dhillon, I. S., and Sra, S. 2003. Modeling data using directional distributions. Technical report, Technical Report TR-03-06, Department of Computer Sciences, The University of Texas at Austin.

[Dhillon, Sra, and Ma 2003] Dhillon, I. S., and Sra, S. 2003. Modelling data using directional distributions. Technical report, Technical Report TR-03-06, Department of Computer Sciences, The University of Texas at Austin.

[Dhillon et al. 2003] Dhillon, I. S., and Sra, S. 2003. Modelling data using directional distributions. Technical report, Technical Report TR-03-06, Department of Computer Sciences, The University of Texas at Austin.

[Dhillon, Sra, and Ma 2003] Dhillon, I. S., and Sra, S. 2003. Modelling data using directional distributions. Technical report, Technical Report TR-03-06, Department of Computer Sciences, The University of Texas at Austin.

[Dhillon, Sra, and Ma 2003] Dhillon, I. S., and Sra, S. 2003. Modelling data using directional distributions. Technical report, Technical Report TR-03-06, Department of Computer Sciences, The University of Texas at Austin.

[Dhillon, Sra, and Ma 2003] Dhillon, I. S., and Sra, S. 2003. Modelling data using directional distributions. Technical report, Technical Report TR-03-06, Department of Computer Sciences, The University of Texas at Austin.

[Dhillon, Sra, and Ma 2003] Dhillon, I. S., and Sra, S. 2003. Modelling data using directional distributions. Technical report, Technical Report TR-03-06, Department of Computer Sciences, The University of Texas at Austin.

[Dhillon, Sra, and Ma 2003] Dhillon, I. S., and Sra, S. 2003. Modelling data using directional distributions. Technical report, Technical Report TR-03-06, Department of Computer Sciences, The University of Texas at Austin.

[Dhillon, Sra, and Ma 2003] Dhillon, I. S., and Sra, S. 2003. Modelling data using directional distributions. Technical report, Technical Report TR-03-06, Department of Computer Sciences, The University of Texas at Austin.

[Dhillon, Sra, and Ma 2003] Dhillon, I. S., and Sra, S. 2003. Modelling data using directional distributions. Technical report, Technical Report TR-03-06, Department of Computer Sciences, The University of Texas at Austin.

[Dhillon, Sra, and Ma 2003] Dhillon, I. S., and Sra, S. 2003. Modelling data using directional distributions. Technical report, Technical Report TR-03-06, Department of Computer Sciences, The University of Texas at Austin.

[Dhillon, Sra, and Ma 2003] Dhillon, I. S., and Sra, S. 2003. Modelling data using directional distributions. Technical report, Technical Report TR-03-06, Department of Computer Sciences, The University of Texas at Austin.

[Dhillon, Sra, and Ma 2003] Dhillon, I. S., and Sra, S. 2003. Modelling data using directional distributions. Technical report, Technical Report TR-03-06, Department of Computer Sciences, The University of Texas at Austin.

[Dhillon, Sra, and Ma 2003] Dhillon, I. S., and Sra, S. 2003. Modelling data using directional distributions. Technical report, Technical Report TR-03-06, Department of Computer Sciences, The University of Texas at Austin.

[Dhillon, Sra, and Ma 2003] Dhillon, I. S., and Sra, S. 2003. Modelling data using directional distributions. Technical report, Technical Report TR-03-06, Department of Computer Sciences, The University of Texas at Austin.

[Dhillon, Sra, and Ma 2003] Dhillon, I. S., and Sra, S. 2003. Modelling data using directional distributions. Technical report, Technical Report TR-03-06, Department of Computer Sciences, The University of Texas at Austin.

[Dhillon, Sra, and Ma 2003] Dhillon, I. S., and Sra, S. 2003. Modelling data using directional distributions. Technical report, Technical Report TR-03-06, Department of Computer Sciences, The University of Texas at Austin.

[Dhillon, Sra, and Ma 2003] Dhillon, I. S., and Sra, S. 2003. Modelling data using directional distributions. Technical report, Technical Report TR-03-06, Department of Computer Sciences, The University of Texas at Austin.

[Dhillon, Sra, and Ma 2003] Dhillon, I. S., and Sra, S. 2003. Modelling data using directional distributions. Technical report, Technical Report TR-03-06, Department of Computer Sciences, The University of Texas at Austin.

[Dhillon, Sra, and Ma 2003] Dhillon, I. S., and Sra, S. 2003. Modelling data using directional distributions. Technical report, Technical Report TR-03-06, Department of Computer Sciences, The University of Texas at Austin.

[Dhillon, Sra, and Ma 2003] Dhillon, I. S., and Sra, S. 2003. Modelling data using directional distributions. Technical report, Technical Report TR-03-06, Department of Computer Sciences, The University of Texas at Austin.

[Dhillon, Sra, and Ma 2003] Dhillon, I. S., and Sra, S. 2003. Modelling data using directional distributions. Technical report, Technical Report TR-03-06, Department of Computer Sciences, The University of Texas at Austin.

[Dhillon, Sra, and Ma 2003] Dhillon, I. S., and Sra, S. 2003. Modelling data using directional distributions. Technical report, Technical Report TR-03-06, Department of Computer Sciences, The University of Texas at Austin.

[Dhillon, Sra, and Ma 2003] Dhillon, I. S., and Sra, S. 2003. Modelling data using directional distributions. Technical report, Technical Report TR-03-06, Department of Computer Sciences, The University of Texas at Austin.

[Dhillon, Sra, and Ma 2003] Dhillon, I. S., and Sra, S. 2003. Modelling data using directional distributions. Technical report, Technical Report TR-03-06, Department of Computer Sciences, The University of Texas at Austin.

[Dhillon, Sra, and Ma 2003] Dhillon, I. S., and Sra, S. 2003. Modelling data using directional distributions. Technical report, Technical Report TR-03-06, Department of Computer Sciences, The University of Texas at Austin.
[Ying et al. 2011] Ying, J. J.; Lee, W.; Weng, T.; and Tseng, V. S. 2011. Semantic trajectory mining for location prediction. In GIS, 34–43.

[Zhang et al. 2014] Zhang, C.; Han, J.; Shou, L.; Lu, J.; and Porta, T. F. L. 2014. Splitter: Mining fine-grained sequential patterns in semantic trajectories. PVLDB 7(9):769–780.

[Zhang et al. 2016a] Zhang, C.; Zhang, K.; Yuan, Q.; Zheng, L.; Hanratty, T.; and Han, J. 2016a. Gmove: Group-level mobility modeling using geo-tagged social media. In KDD, 1305–1314.

[Zhang et al. 2016b] Zhang, C.; Zhou, G.; Yuan, Q.; Zhuang, H.; Zheng, Y.; Kaplan, L. M.; Wang, S.; and Han, J. 2016b. Geoburst: Real-time local event detection in geo-tagged tweet streams. In SIGIR, 513–522.

[Zhang et al. 2017] Zhang, C.; Zhang, K.; Yuan, Q.; Peng, H.; Zheng, Y.; Hanratty, T.; Wang, S.; and Han, J. 2017. Regions, periods, activities: Uncovering urban dynamics via cross-modal representation learning. In WWW, 361–370.