Journal of Signal Processing Systems manuscript No.
(will be inserted by the editor)

Multimodal Event Detection in Twitter Hashtag Networks

Yasin Yılmaz · Alfred O. Hero

Received: date / Accepted: date

Abstract Event detection in a multimodal Twitter dataset is considered. We treat the hashtags in the dataset as instances with two modes: text and geolocation features. The text feature consists of a bag-of-words representation. The geolocation feature consists of geotags (i.e., geographical coordinates) of the tweets. Fusing the multimodal data we aim to detect, in terms of topic and geolocation, the interesting events and the associated hashtags. To this end, a generative latent variable model is assumed, and a generalized expectation-maximization (EM) algorithm is derived to learn the model parameters. The proposed method is computationally efficient, and lends itself to big datasets. Experimental results on a Twitter dataset from August 2014 show the efficacy of the proposed method.

Keywords Event detection · Twitter hashtag networks · Multimodal data fusion · Generative latent variable model · Variational EM algorithm

1 Introduction

Twitter is the most popular microblogging service and the second most popular social network with over 300 million active users generating more than 500 million tweets per day as of 2015 [1]. Its user-generated content from all over the world provides a valuable source of data for researchers from a variety fields such as machine learning, data mining, natural language processing, as well as social sciences.

This work was funded in part by the Consortium for Verification Technology under Department of Energy National Nuclear Security Administration award number DE-NA0002534, and the Army Research Office (ARO) under grants W911NF-11-1-0391 and W911NF-12-1-0443.

Y. Yılmaz
Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109, USA
Tel.: +1-734-763-5022
E-mail: yasiny@umich.edu

A. Hero
Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109, USA

Twitter data has been used for various tasks, e.g., event detection [2], sentiment analysis [3], breaking news analysis [4], rumor detection [5], community detection [6], election results prediction [7], and crime prediction [8].

Hashtags, which are keywords preceded by the hash sign #, are in general used to indicate the subject of the tweets. Hence, they provide useful information for clustering tweets or users. However, it is a noisy information source since hashtags are generated by users, and sometimes convey inaccurate or even counterfactual information. A small percentage of users (around 2%) also geotag their tweets. Given the 500 million tweets per day, geotags also constitute an important information source.

The detection of real-world events from conventional media sources has long been studied [9]. Event detection in Twitter is especially challenging because tweets use microtext, which is an informal language with a preponderance of abbreviated words, spelling and grammar errors. There are also many tweets of dubious value, consisting of nonsense, misrepresentations, and rumors. Much of the work on event detection in Twitter has considered a diversity of event types. For instance, [10] considers unsupervised breaking news detection; [11] considers supervised detection of controversial news events about celebrities; [12] addresses supervised musical event detection; and [13] deals with supervised natural disaster events monitoring. There are also a significant number of papers that consider unsupervised detection of events that do not require prespecification of the event type of interest, e.g., [14–18].

In this paper, we introduce a new unsupervised event detection approach to Twitter that exploits the multimodal nature of the medium. Data is pre-processed to form a network of hashtags. In this network, each unique hashtag is an instance with multimodal features, namely text and geolocation. For a hashtag, the text feature is given by the bag-of-words representation over the collection of words from tweets that use the hashtag. The geolocation feature of a hashtag consists of the geotags of the tweets that mention the hashtag. The proposed approach can detect events in terms of both topic and geolocation through multimodal data fusion. To fuse the multimodal data we use a probabilistic generative model, and derive an expectation-maximization (EM) algorithm to find the maximum likelihood (ML) estimates of the model parameters. The proposed model can be seen as a multimodal factor analysis model [19, 20]. However, it is more general than the model in [20] in terms of the considered probabilistic models, and also the temporal dimension that is inherent to our problem.

Fusing disparate data types, such as text and geolocation in our case, poses significant challenges. In [21], source separation is used to fuse multimodal data, whereas [22] follows an information-theoretic approach. Multimodal data fusion is studied for different applications such as multimedia data analysis [23] and brain imaging [24]. Multimodal feature learning via deep neural networks is considered in [25]. The literature on multi-view learning, e.g., [26–28], is also related to the problem of multimodal data fusion. Our contributions in this paper are twofold. Firstly, we propose a intuitive framework that naturally extends to the exponential family of distributions. Secondly, based on a simple generative model, the proposed algorithm is computationally efficient, and thus applicable to big datasets.

The paper is organized as follows. In Section 2, we formulate the multimodal event detection problem, and propose a generative latent variable model. Then, a generalized EM algorithm is derived in Section 3. Finally, experiment results on a

Twitter dataset are presented in Section 4, and the paper is concluded in Section 5. We represent the vectors and matrices with boldface lowercase and uppercase letters, respectively.

2 Problem Formulation

2.1 Observation Model

We consider P hashtags with text (i.e., collection of words used in tweets) and geotag (i.e., user geolocation) features, as shown in Table 1.

Table 1 Sample hashtags with text and geotag features.

Hashtag	**Text**	**Geotag (Latitude, Longitude)**
#Armstrong	#Oprah mag 'alles vragen' aan Lance #Armstrong. Uiteraard! Looking forward to the #Lance #Armstrong interview next week! ...	(52.4°N, 4.9°E) (43.5°N, 79.6°W) ...
#Arsenal	Sementara menunggu Team Power beraksi..#Arsenal First game of 2013, lets start it off with a our fifth win in a row! Come on you Gunners! #Arsenal	(8.6°S, 116.1°E) (23.7°N, 58.2°E) ...

We assume a model in which each word in a tweet that uses the i-th hashtag is independently generated from the multinomial distribution with a single trial (i.e., categorical distribution) $\mathcal{M}(1; p_{i1}, \ldots, p_{iD})$, where p_{id} is the probability of the d-th word for the i-th hashtag, and D is the dictionary size. In this model, the word counts $\boldsymbol{h}_i = [h_{i1}, \ldots, h_{iD}]^T$ for the i-th hashtag are modeled as

$$\boldsymbol{h}_i \sim \mathcal{M}(M_i; p_{i1}, \ldots, p_{iD}), \; i = 1, \ldots, P,$$

where M_i is the number of dictionary words used in the tweets for the i-th hashtag, i.e., $M_i = \sum_{d=1}^{D} h_{id}$. To this end, we use the bag of words representation for the hashtags (Fig. 1).

The geolocation data of each tweet is a geographical location represented by a spherical coordinate (latitude and longitude). This coordinate is modeled using the 3-dimensional von Mises-Fisher (vMF) distribution, which is an extension of the Gaussian distribution to the unit sphere [29] (Fig. 2). We first convert the geographical coordinates (latitude, longitude) to the Cartesian coordinates (x, y, z), where $x^2 + y^2 + z^2 = 1$. Specifically, in our model, it is assumed that the geolocation of the n-th tweet that mentions the i-th hashtag is generated independently from the other tweets as follows

$$\boldsymbol{w}_{in} \sim \mathcal{V}(\boldsymbol{\alpha}_i, \kappa_i), \; i = 1, \ldots, P, \; n = 1, \ldots, N_i,$$

where $\boldsymbol{\alpha}_i \in \mathbb{R}^3$, $\boldsymbol{\alpha}_i^T \boldsymbol{\alpha}_i = 1$, is the mean direction, $\kappa_i \geq 0$ is the concentration parameter, and N_i is the number of geotagged tweets for the i-th hashtag. Larger

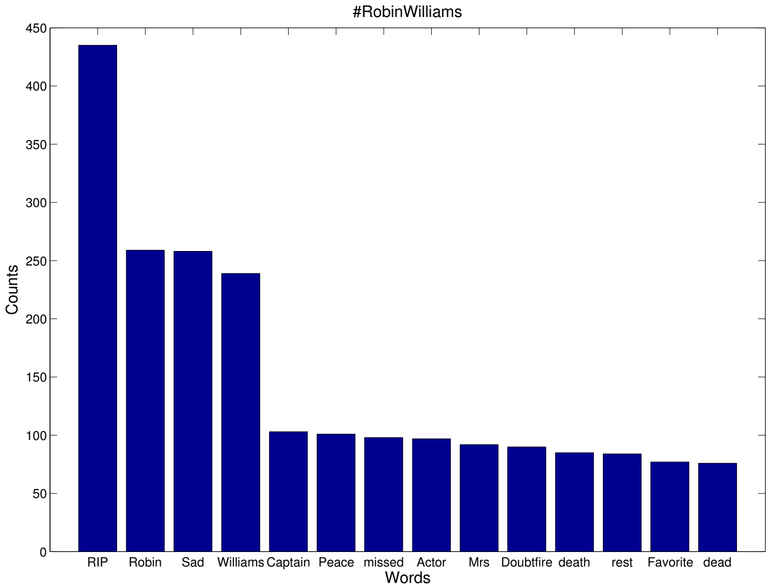

Fig. 1 A sample bag of words representation for the hashtag #RobinWilliams.

κ_i means more concentrated distribution around $\boldsymbol{\alpha}_i$. Therefore, a local hashtag, such as #GoBlue, which is used by supporters of the University of Michigan sports teams, requires a large κ, whereas a global hashtag, such as #HalaMadrid, which means "Go Madrid" and is used by the fans of the Real Madrid soccer team, requires a small κ (Fig. 3). This difference in κ is due to the fact that the Real Madrid supporters are well distributed around the globe, while the University of Michigan supporters are mostly confined to North America.

2.2 Generative Latent Variable Model

Some hashtags are created by users as a result of an underlying event in *time* and *space*, which we call a *generative event*. For instance, after Robin Williams' death, many hashtags such as #RobinWilliams, #RIPRobinWilliams, #RIPRobin, #mrsdoubtfire have been used to commemorate him. On the other hand, some hashtags are more spread out over time, such as #jobs, #love, #Healthcare, #photo. With a slight abuse of the terminology, we also consider such an underlying topic as a *generative event*. In addition to the topic/text feature, a generative event (time-dependent or -independent) possesses also a spatial feature due to the event's geolocation (e.g., Asia, America) or simply due to the language (e.g., English, Spanish).

We know that an event can generate multiple hashtags. Although there is usually a single event responsible for the generation of a hashtag, for generality, we let multiple events contribute to a single hashtag. In our generative model, K events linearly mix in the *natural parameters* of the multinomial and vMF distributions

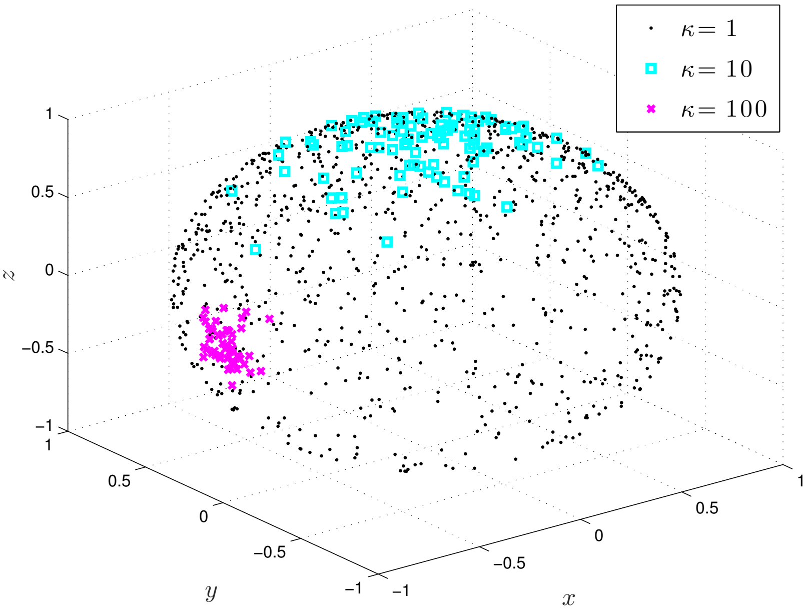

Fig. 2 Samples from the 3-dimensional von Mises-Fisher distribution with different concentration parameter values $\kappa = 1, 10, 100$ describing the spread of the distribution around random mean directions. The case $\kappa = 1$ produces the uniform distribution on the sphere.

to generate the text and geolocation features of each hashtag, respectively. Let $\boldsymbol{c}_i \in \mathbb{R}_+^K$ denote the mixture coefficients of K events for the i-th hashtag, where $\mathbb{R}_+$ is the set of nonnegative real numbers. Also let

$$\boldsymbol{U} = [\boldsymbol{u}_1 \dots \boldsymbol{u}_K] = \left[\boldsymbol{u}_{(1)}^T \dots \boldsymbol{u}_{(D)}^T\right]^T, \ \boldsymbol{u}_k \in \mathbb{R}^D, \ \boldsymbol{u}_{(d)} \in \mathbb{R}^{1\times K},$$

denote the event scores for the words in the dictionary; and

$$\boldsymbol{V} = [\boldsymbol{v}_1 \dots \boldsymbol{v}_K], \ \boldsymbol{v}_k \in \mathbb{R}^3,$$

denote the event geolocations in the Cartesian coordinates. Then, in our model, the mean of the vMF distribution is given by the normalized linear mixture

$$\boldsymbol{\alpha}_i = \frac{\boldsymbol{V}\boldsymbol{c}_i}{\|\boldsymbol{V}\boldsymbol{c}_i\|}, \quad i = 1, \dots, P,$$

where $\|\cdot\|$ is the l^2-norm normalization is required to ensure that $\boldsymbol{\alpha}_i$ is on the unit sphere; and the multinomial probabilities are given by the softmax function of the linear mixture $\boldsymbol{u}_{(d)}\boldsymbol{c}_i$, i.e.,

$$p_{id} = \frac{e^{\boldsymbol{u}_{(d)}\boldsymbol{c}_i}}{\sum_{j=1}^{D} e^{\boldsymbol{u}_{(j)}\boldsymbol{c}_i}}, \quad i = 1, \dots, P, \quad d = 1, \dots, D.$$

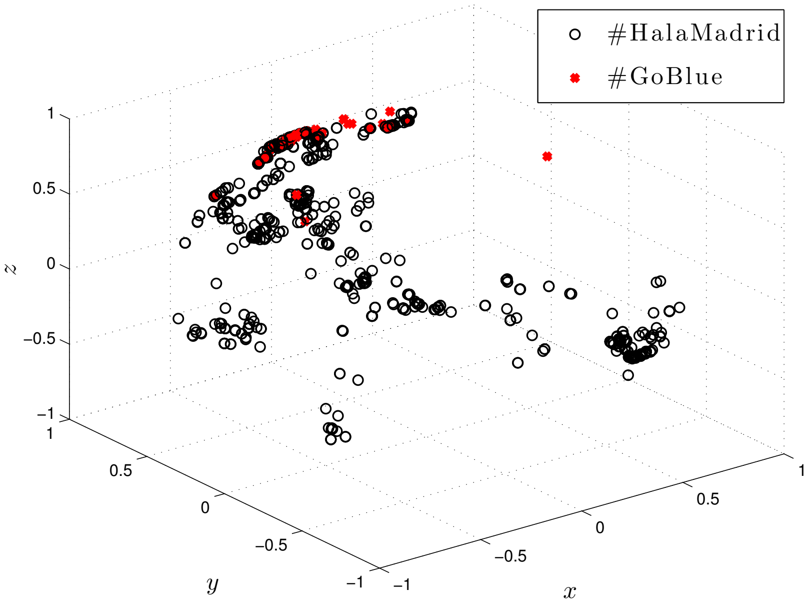

Fig. 3 Geolocations for the hashtags #HalaMadrid (used for the Real Madrid soccer team) and #GoBlue (used for the University of Michigan athletics) in terms of the Cartesian coordinates. The estimated concentration parameters for the von Mises-Fisher distribution are $\kappa_{\text{madrid}} = 1.3302$ and $\kappa_{\text{mich}} = 44.6167$, representing the wider global interest in Real Madrid soccer team as contrasted to the US-centric interest in University of Michigan sports teams.

That is,

$$\boldsymbol{h}_i \sim \mathcal{M}\left(M_i; \frac{e^{\boldsymbol{u}_{(1)}\boldsymbol{c}_i}}{\sum_{j=1}^{D} e^{\boldsymbol{u}_{(j)}\boldsymbol{c}_i}}, \ldots, \frac{e^{\boldsymbol{u}_{(d)}\boldsymbol{c}_i}}{\sum_{j=1}^{D} e^{\boldsymbol{u}_{(j)}\boldsymbol{c}_i}}\right), \quad i = 1, \ldots, P \tag{1}$$

$$\boldsymbol{w}_{in} \sim \mathcal{V}\left(\frac{\boldsymbol{V}\boldsymbol{c}_i}{\|\boldsymbol{V}\boldsymbol{c}_i\|}, \kappa_i\right), \quad i = 1, \ldots, P, \quad n = 1, \ldots, N_i, \tag{2}$$

where N_i is the number of geotagged tweets for the i-th hashtag. We assume a Gaussian prior for the latent variable vector $\boldsymbol{u}_k \in \mathbb{R}^D$

$$\boldsymbol{u}_k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \quad k = 1, \ldots, K, \tag{3}$$

and a vMF prior for $\boldsymbol{v}_k \in \mathbb{R}^3$

$$\boldsymbol{v}_k \sim \mathcal{V}(\boldsymbol{\beta}_k, s_k), \quad k = 1, \ldots, K, \tag{4}$$

since the conjugate prior to the vMF likelihood with unknown mean and known concentration is also vMF [30].

The graphical model in Fig. 4 depicts the proposed generative latent variable model. The proposed model can be regarded as a *multimodal factor analysis* model [19] since it combines features from two disparate domains(geotag and text). In

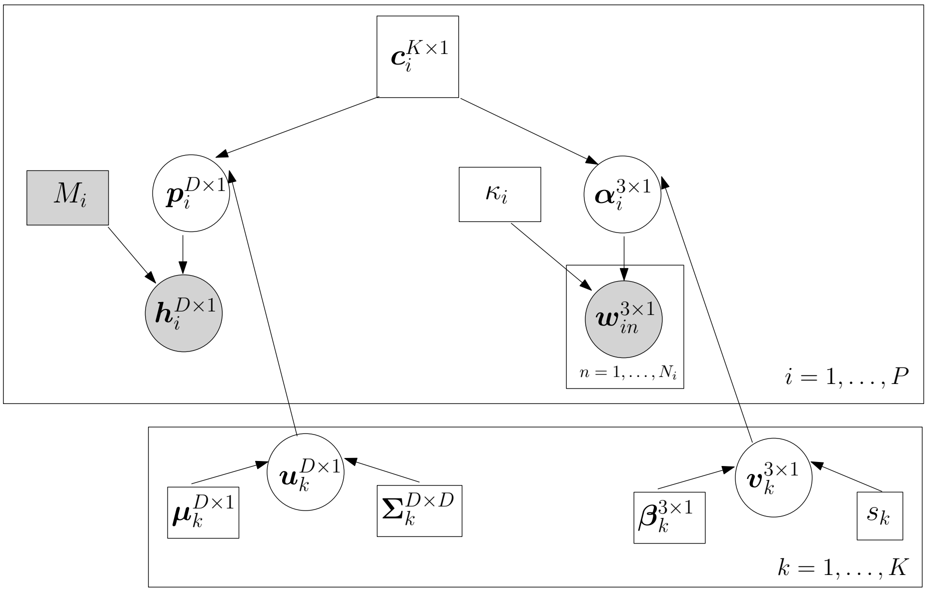

Fig. 4 Generative graphical model. Plate representation is used to show repeated structures. Circles and rectangles represent random and deterministic variables, respectively. Observed variables are shaded.

classical factor analysis [31], the mean of a Gaussian random variable is modeled with the linear combination $\boldsymbol{c}^T\boldsymbol{u}$ of factor scores in $\boldsymbol{u}$, where the coefficients in $\boldsymbol{c}$ are called factor loadings.The number of factors is typically much less than the number of variables modeled, as $K \ll P$ in our case. In the proposed model, the generative events correspond to the factors with the multimodal scores $\{\boldsymbol{u}_k\}$ and $\{\boldsymbol{v}_k\}$ for the multinomial and vMF observations, respectively. For both modalities, the natural parameters are modeled with the linear combination of the factor scores using the *same factor loading vector* $\boldsymbol{c}_i$ for the i-th hashtag. In the multinomial distribution, the softmax function maps the natural parameters to the class probabilities, whereas in the vMF distribution, the natural parameter coincides with the (scaled) mean. For each hashtag i, the factor loading vector $\boldsymbol{c}_i$ correlates the two observation modalities: text and geolocation.

Next we will present an EM algorithm to learn the parameters of the proposed model from the data.

3 EM Algorithm

We propose a generalized EM (GEM) algorithm that consists of two separate EM steps for the two modalities, and a coordinating M step for the mixture coefficients $\{\boldsymbol{c}_i\}$. Specifically, at each iteration of the GEM algorithm, the vMF EM steps are followed by the multinomial EM steps, which are followed by the M step for $\{\boldsymbol{c}_i\}$. The individual EM steps for vMF and multinomial are coupled only through $\{\boldsymbol{c}_i\}$, and independent otherwise. In the proposed GEM algorithm, the global likelihood function is monotonically increasing.

3.1 Von Mises-Fisher Parameters

We would like to find the ML estimates of the parameters $\boldsymbol{\beta}_k, s_k, \kappa_i$ under the graphical model depicted in the right branch of Fig. 4. We take a variational EM approach to deal with the latent variable vectors $\{\boldsymbol{v}_k\}$.

3.1.1 E-Step

Starting with the E-step we seek the posterior probability density function (pdf) $\mathsf{P}(\{\boldsymbol{v}_k\}|\{\boldsymbol{w}_{in}\},\theta)$, where $\theta = \{\boldsymbol{\beta}_k, s_k, \kappa_i, \boldsymbol{c}_i\}$. From (2) and (4), we know that the likelihood $\mathsf{P}(\{\boldsymbol{w}_{in}\}|\{\boldsymbol{v}_k\},\theta)$ and the prior $\mathsf{P}(\{\boldsymbol{v}_k\}|\theta)$ are both vMF, hence the joint distribution is given by

$$\begin{aligned}\mathsf{P}(\{\boldsymbol{v}_k\},\{\boldsymbol{w}_{in}\}|\theta) &= \mathsf{P}(\{\boldsymbol{w}_{in}\}|\{\boldsymbol{v}_k\},\theta)\ \mathsf{P}(\{\boldsymbol{v}_k\}|\theta)\\ &= \prod_{i=1}^{P} C(\kappa_i)^{N_i} \prod_{n=1}^{N_i} \exp\left(\kappa_i \boldsymbol{w}_{in}^T \frac{\boldsymbol{V}\boldsymbol{c}_i}{\|\boldsymbol{V}\boldsymbol{c}_i\|}\right) \prod_{k=1}^{K} C(s_k)\exp\left(s_k \boldsymbol{v}_k^T \boldsymbol{\beta}_k\right),\end{aligned}$$

where

$$C(x) = \frac{x^{1/2}}{(2\pi)^{3/2} I_{1/2}(x)} = \frac{x}{2\pi(e^x - e^{-x})} \tag{5}$$

is the normalization factor in the 3-dimensional vMF pdf, with $I_y(x)$ being the modified Bessel function of the first kind at order y. Reorganizing the terms we get

$$\begin{aligned}\mathsf{P}(\{\boldsymbol{v}_k\},\{\boldsymbol{w}_{in}\}|\theta) &= \prod_{i=1}^{P} C(\kappa_i)^{N_i} \prod_{k=1}^{K} C(s_k)\\ &\quad \exp\left(\sum_{i=1}^{P}\sum_{n=1}^{N_i} \kappa_i \boldsymbol{w}_{in}^T \sum_{k=1}^{K} \frac{c_{ik}}{\|\boldsymbol{V}\boldsymbol{c}_i\|}\boldsymbol{v}_k + \sum_{k=1}^{K} s_k \boldsymbol{v}_k^T \boldsymbol{\beta}_k\right)\\ &= \prod_{i=1}^{P} C(\kappa_i)^{N_i} \prod_{k=1}^{K} C(s_k)\\ &\quad \prod_{k=1}^{K} \exp\left(\boldsymbol{v}_k^T \left(\sum_{i=1}^{P}\sum_{n=1}^{N_i} \frac{c_{ik}}{\sqrt{\boldsymbol{c}_i^T \boldsymbol{V}^T \boldsymbol{V} \boldsymbol{c}_i}} \kappa_i \boldsymbol{w}_{in} + s_k \boldsymbol{\beta}_k\right)\right).\end{aligned} \tag{6}$$

In the alternative expression for the joint pdf

$$\mathsf{P}(\{\boldsymbol{v}_k\},\{\boldsymbol{w}_{in}\}|\theta) = \mathsf{P}(\{\boldsymbol{v}_k\}|\{\boldsymbol{w}_{in}\},\theta)\ \mathsf{P}(\{\boldsymbol{w}_{in}\}|\theta),$$

the dependency on $\{\boldsymbol{v}_k\}$ appears only in the posterior pdf, hence $\mathsf{P}(\{\boldsymbol{v}_k\}|\{\boldsymbol{w}_{in}\},\theta)$ lies in the exponential term in (6), which resembles the vMF pdf except the dependence of the normalization factor on $\{\boldsymbol{v}_k\}$. The diagonal entries of $\boldsymbol{V}^T\boldsymbol{V}$ are $\boldsymbol{v}_k^T\boldsymbol{v}_k = 1$; and the off-diagonal entries are $\boldsymbol{v}_j^T\boldsymbol{v}_k \leq 1,\ j \neq k$. Since $c_{ik} \geq 0,\ k = 1,\dots,K$, the inequality $\boldsymbol{c}_i^T\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{c}_i \leq \boldsymbol{c}_i^T\boldsymbol{1}_K\boldsymbol{1}_K^T\boldsymbol{c}_i$ holds, where $\boldsymbol{1}_K$ is the vector of

K ones. To make (6) tractable we replace $\boldsymbol{c}_i^T \boldsymbol{V}^T \boldsymbol{V} \boldsymbol{c}_i$ with $\boldsymbol{c}_i^T \mathbf{1}_K \mathbf{1}_K^T \boldsymbol{c}_i$ and obtain the lower bound

$$\begin{aligned} \mathsf{P}(\{\boldsymbol{v}_k\}, \{\boldsymbol{w}_{in}\}|\theta) &\geq Q_v(\{\boldsymbol{v}_k\}, \theta) \\ &= \prod_{i=1}^{P} C(\kappa_i)^{N_i} \prod_{k=1}^{K} C(s_k) \\ &\quad \prod_{k=1}^{K} \exp\left(\boldsymbol{v}_k^T \left(\sum_{i=1}^{P} \sum_{n=1}^{N_i} \frac{c_{ik}}{\sqrt{\boldsymbol{c}_i^T \mathbf{1}_K \mathbf{1}_K^T \boldsymbol{c}_i}} \kappa_i \boldsymbol{w}_{in} + s_k \boldsymbol{\beta}_k \right) \right). \end{aligned} \tag{7}$$

To put (7) into the standard form of the vMF pdf we normalize the term in the inner parentheses and obtain

$$Q_v(\{\boldsymbol{v}_k\}, \theta) = \prod_{i=1}^{P} \prod_{n=1}^{N_i} \prod_{k=1}^{K} \frac{C(\kappa_i)^{N_i}\, C(s_k)}{C(r_k)} \underbrace{C(r_k) \exp\left(r_k \boldsymbol{v}_k^T \boldsymbol{b}_k \right)}_{q_v(\boldsymbol{v}_k)}, \tag{8}$$

$$\boldsymbol{b}_k = \frac{\sum_{i=1}^{P} \frac{c_{ik}}{\sum_{k=1}^{K} c_{ik}} \kappa_i \sum_{n=1}^{N_i} \boldsymbol{w}_{in} + s_k \boldsymbol{\beta}_k}{\left\| \sum_{i=1}^{P} \frac{c_{ik}}{\sum_{k=1}^{K} c_{ik}} \kappa_i \sum_{n=1}^{N_i} \boldsymbol{w}_{in} + s_k \boldsymbol{\beta}_k \right\|}, \tag{9}$$

$$r_k = \left\| \sum_{i=1}^{P} \frac{c_{ik}}{\sum_{k=1}^{K} c_{ik}} \kappa_i \sum_{n=1}^{N_i} \boldsymbol{w}_{in} + s_k \boldsymbol{\beta}_k \right\|, \tag{10}$$

where $\boldsymbol{b}_k$ is the mean direction and r_k is the concentration parameter. We approximate the posterior $\mathsf{P}(\boldsymbol{v}_k|\{\boldsymbol{w}_{in}\}, \theta)$ with the vMF distribution $q_v(\boldsymbol{v}_k)$ for $k = 1, \ldots, K$.

3.1.2 M-Step

In the M-step, we find the parameters $\boldsymbol{\beta}_k, s_k, \kappa_i$ that maximize the expected value of the lower bound for the complete-data log-likelihood, which from (7) is given by

$$\begin{aligned} \mathsf{E}_{q_v(\boldsymbol{v}_k)} \left[\log Q_v(\{\boldsymbol{v}_k\}, \theta) \right] = \sum_{k=1}^{K} &\left(\sum_{i=1}^{P} \frac{c_{ik}}{\sum_{k=1}^{K} c_{ik}} \kappa_i \sum_{n=1}^{N_i} \boldsymbol{w}_{in} + s_k \boldsymbol{\beta}_k \right)^T \boldsymbol{b}_k \\ &+ \sum_{k=1}^{K} \log C(s_k) + \sum_{i=1}^{P} N_i \log C(\kappa_i), \end{aligned} \tag{11}$$

where the expectation is taken over $q_v(\boldsymbol{v}_k)$, which approximates the posterior pdf $\mathsf{P}(\boldsymbol{v}_k|\{\boldsymbol{w}_{in}\}, \theta)$ (see (8)).

We start with the estimator $\hat{\kappa}_i$ which is given by

$$\hat{\kappa}_i = \arg\max_{\kappa_i} \; \kappa_i \left(\sum_{n=1}^{N_i} \boldsymbol{w}_{in} \right)^T \sum_{k=1}^{K} \frac{c_{ik}}{\sum_{k=1}^{K} c_{ik}} \boldsymbol{b}_k + N_i \log C(\kappa_i).$$

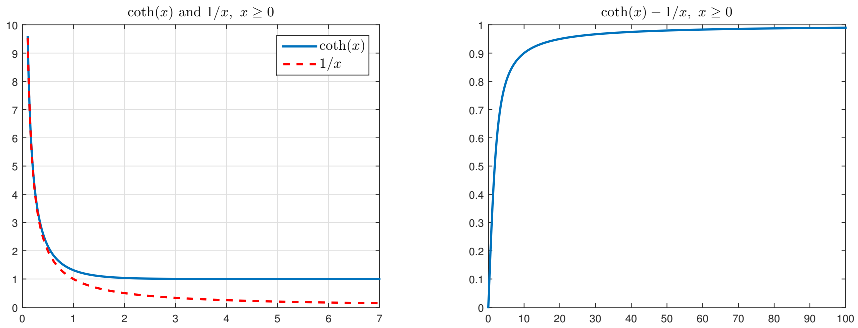

Fig. 5 For nonnegative numbers, the hyperbolic cotangent and the multiplicative inverse (left figure); and their difference (right figure).

We next show that $\hat{\kappa}_i$ always makes the derivative with respect to κ_i zero, i.e.,

$$-\frac{C'(\hat{\kappa}_i)}{C(\hat{\kappa}_i)} = \left(\frac{1}{N_i} \sum_{n=1}^{N_i} \boldsymbol{w}_{in} \right)^T \sum_{k=1}^{K} \frac{c_{ik}}{\sum_{k=1}^{K} c_{ik}} \boldsymbol{b}_k \triangleq \tau_i. \tag{12}$$

From (26), we write the derivative $C'(\hat{\kappa}_i)$ as

$$C'(\hat{\kappa}_i) = \frac{\hat{\kappa}_i^{1/2}}{(2\pi)^{3/2} I_{1/2}(\hat{\kappa}_i)} \left(\frac{1}{2\hat{\kappa}_i} - \frac{I'_{1/2}(\hat{\kappa}_i)}{I_{1/2}(\hat{\kappa}_i)} \right) = C(\hat{\kappa}_i) \left(\frac{1}{2\hat{\kappa}_i} - \frac{I'_{1/2}(\hat{\kappa}_i)}{I_{1/2}(\hat{\kappa}_i)} \right).$$

Hence,

$$-\frac{C'(\hat{\kappa}_i)}{C(\hat{\kappa}_i)} = \frac{\hat{\kappa}_i I'_{1/2}(\hat{\kappa}_i) - I_{1/2}(\hat{\kappa}_i)/2}{\hat{\kappa}_i I_{1/2}(\hat{\kappa}_i)}.$$

Using (12) and the following recurrence relation [32, Section 9.6.26]

$$x I_{3/2}(x) = x I'_{1/2}(x) - I_{1/2}(x)/2$$

we get

$$\frac{I_{3/2}(\hat{\kappa}_i)}{I_{1/2}(\hat{\kappa}_i)} = \tau_i, \tag{13}$$

which can be rewritten as [32, Section 10.2.13]

$$\coth(\hat{\kappa}_i) - \frac{1}{\hat{\kappa}_i} = \tau_i. \tag{14}$$

We can also obtain (14) using (26) and (12). Fig. 5 shows that the left side of (14) is a continuous function taking values in $[0, 1]$. Since τ_i, defined in (12), is also in $[0, 1]$, there always exists a unique solution to (14). However, there is no analytical solution to (13) or (14); hence we resort to approximating $\hat{\kappa}_i$. A numerical solution easily follows using a root-finding method such as the bisection method.

Alternatively, using the following continuing fraction representation

$$\frac{I_{3/2}(\hat{\kappa}_i)}{I_{1/2}(\hat{\kappa}_i)} = \frac{1}{\frac{3}{\hat{\kappa}_i} + \frac{1}{\frac{5}{\hat{\kappa}_i} + \cdots}} = \tau_i$$

we can approximate $\hat{\kappa}_i$ as [33]

$$\begin{aligned}\frac{1}{\tau_i} &\approx \frac{3}{\hat{\kappa}_i} + \tau_i \\ \hat{\kappa}_i &\approx \frac{3\tau_i}{1-\tau_i^2}.\end{aligned}$$

Furthermore, an empirical correction is also provided in [33]:

$$\hat{\kappa}_i \approx \frac{3\tau_i - \tau_i^3}{1-\tau_i^2}, \tag{15}$$

which is constrained to be nonnegative for feasibility. We introduce a Lagrange multiplier $\lambda > 0$, replacing τ_i with $\tilde{\tau}_i = \tau_i + \lambda$ to enforce this non-negativity constraint. Due to complementary slackness, this leads to the following estimator

$$\hat{\kappa}_i \approx \max\left\{0, \frac{3\tau_i - \tau_i^3}{1-\tau_i^2}\right\}. \tag{16}$$

Similar to κ_i (12)–(16), from (11), we estimate s_k with

$$\begin{aligned}\hat{s}_k &= \arg\max_{s_k} \; s_k \boldsymbol{\beta}_k^T \boldsymbol{b}_k + \log C(s_k). \\ &\approx \max\left\{0, \frac{3\boldsymbol{\beta}_k^T \boldsymbol{b}_k - (\boldsymbol{\beta}_k^T \boldsymbol{b}_k)^3}{1-(\boldsymbol{\beta}_k^T \boldsymbol{b}_k)^2}\right\}. \end{aligned}\tag{17}$$

Since $\boldsymbol{\beta}_k$ is a mean direction on the unit sphere, it has to satisfy $\boldsymbol{\beta}_k^T \boldsymbol{\beta}_k = 1$.Therefore, from (11), our estimator is given by

$$\hat{\boldsymbol{\beta}}_k = \arg\max_{\boldsymbol{\beta}} \boldsymbol{\beta}_k^T s_k \boldsymbol{b}_k \quad \text{subject to} \quad \boldsymbol{\beta}_k^T \boldsymbol{\beta}_k = 1.$$

Maximum of $\boldsymbol{\beta}_k^T s_k \boldsymbol{b}_k$ is attained when the angle between $\boldsymbol{\beta}_k$ and $s_k \boldsymbol{b}_k$ is zero, i.e., $\hat{\boldsymbol{\beta}}_k = c \; s_k \boldsymbol{b}_k$. Since the feasible set is the unit sphere, $\hat{\boldsymbol{\beta}}_k = \frac{s_k \boldsymbol{b}_k}{\|s_k \boldsymbol{b}_k\|} = \frac{\boldsymbol{b}_k}{\|\boldsymbol{b}_k\|}$. The posterior mean direction $\boldsymbol{b}_k$, given by (9), is already on the unit sphere, hence

$$\hat{\boldsymbol{\beta}}_k = \boldsymbol{b}_k. \tag{18}$$

3.2 Multinomial Parameters

Note that there are $D-1$ degrees of freedom in the multinomial class probabilities due to the constraint $\sum_{d=1}^{D} p_{id} = 1$. For identifiability, we set the D-th word as pivot, and deal with the latent event scores

$$\tilde{\boldsymbol{u}}_{(d)} = \boldsymbol{u}_{(d)} - \boldsymbol{u}_{(D)}, \quad d = 1, \ldots, D-1,$$

and accordingly $\tilde{\boldsymbol{U}} = [\tilde{\boldsymbol{u}}_1 \ldots \tilde{\boldsymbol{u}}_K]$, where from (3)

$$\tilde{\boldsymbol{u}}_k^{D-1\times 1} \sim \mathcal{N}(\tilde{\boldsymbol{\mu}}_k, \tilde{\boldsymbol{\Sigma}}_k). \tag{19}$$

3.2.1 E-Step

We seek the posterior pdf $\mathsf{P}\left(\{\tilde{\boldsymbol{u}}_k\}|\{\boldsymbol{h}_i\},\theta\right)$ where $\theta = \{\tilde{\boldsymbol{\mu}}_k, \tilde{\boldsymbol{\Sigma}}_k, \boldsymbol{c}_i\}$. From (1) and (19),

$$\begin{aligned}
\mathsf{P}\left(\{\tilde{\boldsymbol{u}}_k\},\{\boldsymbol{h}_i\}|\theta\right) &= \mathsf{P}\left(\{\boldsymbol{h}_i\}|\{\tilde{\boldsymbol{u}}_k\},\theta\right)\mathsf{P}\left(\{\tilde{\boldsymbol{u}}_k\}|\theta\right)\\
&= \prod_{i=1}^{P} \frac{M_i!}{h_{i1}!\cdots h_{iD}!} \prod_{d=1}^{D} \exp\left(h_{id}\left[\eta_{id} - \mathrm{lse}(\boldsymbol{\eta}_i)\right]\right)\\
&\quad \prod_{k=1}^{K} \frac{\exp\left(-\frac{1}{2}(\tilde{\boldsymbol{u}}_k - \tilde{\boldsymbol{\mu}}_k)^T \tilde{\boldsymbol{\Sigma}}_k^{-1}(\tilde{\boldsymbol{u}}_k - \tilde{\boldsymbol{\mu}}_k)\right)}{(2\pi)^{(D-1)/2}\,|\tilde{\boldsymbol{\Sigma}}_k|^{1/2}},
\end{aligned} \tag{20}$$

where $\eta_{id} = \tilde{\boldsymbol{u}}_{(d)}\boldsymbol{c}_i,\ d = 1,\dots,D-1,\ \eta_{iD} = 0,\ \boldsymbol{\eta}_i = [\eta_{i1}\dots\eta_{iD-1}]^T = \tilde{\boldsymbol{U}}\boldsymbol{c}_i$, and the log-sum-exp function

$$\mathrm{lse}(\boldsymbol{\eta}_i) = \log\left(1 + \sum_{d=1}^{D-1}\exp\left(\eta_{id}\right)\right). \tag{21}$$

As in the vMF case (6), the normalization factor in (20), which is the lse function, prevents a tractable form. Following [20] we use a quadratic upper bound of the lse function based on the Taylor series expansion to obtain a lower bound for the complete-data likelihood, given in (20). The second order Taylor series expansion around a fixed point $\boldsymbol{\psi}_i$ is given by

$$\mathrm{lse}(\boldsymbol{\eta}_i) = \mathrm{lse}(\boldsymbol{\psi}_i) + (\boldsymbol{\eta}_i - \boldsymbol{\psi}_i)^T \nabla\mathrm{lse}(\boldsymbol{\psi}_i) + \frac{1}{2}(\boldsymbol{\eta}_i - \boldsymbol{\psi}_i)^T \nabla^2\mathrm{lse}(\tilde{\boldsymbol{\psi}}_i)(\boldsymbol{\eta}_i - \boldsymbol{\psi}_i),$$

where $\tilde{\boldsymbol{\psi}}_i \in (\boldsymbol{\eta}_i, \boldsymbol{\psi}_i)$. From (21),

$$\begin{aligned}
\nabla\mathrm{lse}(\boldsymbol{\psi}_i) &= \left[\frac{\exp\left(\psi_{i1}\right)}{1+\sum_{d=1}^{D-1}\exp\left(\psi_{id}\right)}\cdots\frac{\exp\left(\psi_{iD-1}\right)}{1+\sum_{d=1}^{D-1}\exp\left(\psi_{id}\right)}\right] = \boldsymbol{p}_{\boldsymbol{\psi}_i}\\
\nabla\mathrm{lse}^2(\tilde{\boldsymbol{\psi}}_i) &= \boldsymbol{\Lambda}_{\tilde{\boldsymbol{\psi}}_i} - \boldsymbol{p}_{\tilde{\boldsymbol{\psi}}_i}\boldsymbol{p}_{\tilde{\boldsymbol{\psi}}_i}^T,\quad \boldsymbol{\Lambda}_{\tilde{\boldsymbol{\psi}}_i} = \mathrm{diag}(\boldsymbol{p}_{\tilde{\boldsymbol{\psi}}_i}),
\end{aligned}$$

where $\boldsymbol{\Lambda}_{\tilde{\boldsymbol{\psi}}_i}$ is the diagonal matrix form of $\boldsymbol{p}_{\tilde{\boldsymbol{\psi}}_i}$. In [34], it is shown that the matrix

$$\boldsymbol{A} = \frac{1}{2}\left(\boldsymbol{I}_{D-1} - \frac{\mathbf{1}_{D-1}\mathbf{1}_{D-1}^T}{D}\right) \succcurlyeq \nabla\mathrm{lse}^2(\tilde{\boldsymbol{\psi}}_i),\quad \forall\tilde{\boldsymbol{\psi}}_i, \tag{22}$$

in the positive semi-definite sense, where $\boldsymbol{I}_d$ is the d-dimensional identity matrix. That is,

$$\begin{aligned}
\mathrm{lse}(\boldsymbol{\eta}_i) &\le \frac{1}{2}\boldsymbol{\eta}_i^T\boldsymbol{A}\boldsymbol{\eta}_i + \boldsymbol{g}_{\tilde{\boldsymbol{\psi}}_i}^T\boldsymbol{\eta}_i + c_{\tilde{\boldsymbol{\psi}}_i},\\
\boldsymbol{g}_{\tilde{\boldsymbol{\psi}}_i} &= \boldsymbol{p}_{\boldsymbol{\psi}_i} - \boldsymbol{A}\boldsymbol{\psi}_i,\\
c_{\tilde{\boldsymbol{\psi}}_i} &= \mathrm{lse}(\boldsymbol{\psi}_i) + \frac{1}{2}\boldsymbol{\psi}_i^T\boldsymbol{A}\boldsymbol{\psi}_i - \boldsymbol{\psi}_i^T\boldsymbol{p}_{\boldsymbol{\psi}_i}.
\end{aligned} \tag{23}$$

In (20), replacing $\mathrm{lse}(\boldsymbol{\eta}_i)$ with the quadratic upper bound in (23) we get the following lower bound for the likelihood $\mathsf{P}\left(\{\boldsymbol{h}_i\}|\{\tilde{\boldsymbol{u}}_k\},\theta\right)$

$$\begin{aligned}\mathsf{P}\left(\{\boldsymbol{h}_i\}|\{\tilde{\boldsymbol{u}}_k\},\theta\right) \geq \prod_{i=1}^{P} & \frac{M_i!}{h_{i1}!\cdots h_{iD}!} \\ & \exp\left(\sum_{d=1}^{D} h_{id}\eta_{id} - \left(\frac{1}{2}\boldsymbol{\eta}_i^T\boldsymbol{A}\boldsymbol{\eta}_i + \boldsymbol{g}_{\tilde{\boldsymbol{\psi}}_i}^T\boldsymbol{\eta}_i + c_{\tilde{\boldsymbol{\psi}}_i}\right)\sum_{d=1}^{D} h_{id}\right) \\ = \prod_{i=1}^{P} & \frac{M_i!}{h_{i1}!\cdots h_{iD}!} \\ & \exp\left(-\frac{1}{2}\left(\boldsymbol{\eta}_i^T M_i\boldsymbol{A}\boldsymbol{\eta}_i - 2M_i\left(\frac{\boldsymbol{h}_{i\setminus D}}{M_i} - \boldsymbol{g}_{\tilde{\boldsymbol{\psi}}_i}\right)^T\boldsymbol{\eta}_i + 2M_i c_{\tilde{\boldsymbol{\psi}}_i}\right)\right),\end{aligned}$$

where $\boldsymbol{h}_{i\setminus D} = [h_{i1}\dots h_{iD-1}]^T$ is the count vector of the i-th hashtag for the first $D-1$ words. Defining a new observation vector

$$\tilde{\boldsymbol{h}}_i = \boldsymbol{A}^{-1}\left(\frac{\boldsymbol{h}_{i\setminus D}}{M_i} - \boldsymbol{g}_{\tilde{\boldsymbol{\psi}}_i}\right) = \boldsymbol{A}^{-1}\left(\frac{\boldsymbol{h}_{i\setminus D}}{M_i} - \boldsymbol{p}_{\boldsymbol{\psi}_i}\right) + \boldsymbol{\psi}_i$$

we write

$$\begin{aligned}\mathsf{P}\left(\{\tilde{\boldsymbol{h}}_i\}|\{\tilde{\boldsymbol{u}}_k\},\theta\right) &\geq \prod_{i=1}^{P} f_{\tilde{\boldsymbol{\psi}}_i}\exp\left(-\frac{1}{2}\left(\boldsymbol{\eta}_i - \tilde{\boldsymbol{h}}_i\right)^T M_i\boldsymbol{A}\left(\boldsymbol{\eta}_i - \tilde{\boldsymbol{h}}_i\right)\right), \qquad (24)\\ f_{\tilde{\boldsymbol{\psi}}_i} &= \frac{M_i!}{h_{i1}!\cdots h_{iD}!}\exp\left(\frac{\tilde{\boldsymbol{h}}_i^T M_i\boldsymbol{A}\tilde{\boldsymbol{h}}_i}{2} - M_i c_{\tilde{\boldsymbol{\psi}}_i}\right).\end{aligned}$$

Recall that $\boldsymbol{\eta}_i = \tilde{\boldsymbol{U}}\boldsymbol{c}_i = \sum_{k=1}^{K} c_{ik}\tilde{\boldsymbol{u}}_k$. In (24), the latent variable vectors $\{\tilde{\boldsymbol{u}}_k\}$, which are independently modeled *a priori* (19), are coupled, thus no more independent *a posteriori* in $\mathsf{P}\left(\{\tilde{\boldsymbol{u}}_k\}|\{\tilde{\boldsymbol{h}}_i\},\theta\right)$. To capture the dependency we treat them in a single vector $\tilde{\boldsymbol{u}} = [\tilde{\boldsymbol{u}}_{(1)}\dots\tilde{\boldsymbol{u}}_{(K)}]^T$. Without loss of generality, *a priori* we assume independence among the words for the same event, i.e., $\tilde{\boldsymbol{\Sigma}}_k = \boldsymbol{I}_{D-1},\ \forall k$. Prior probability distribution reflects our initial belief about the unknown entity; and *a priori* we do not know anything about the inter-word dependencies of the hidden events. Hence, this is a quite reasonable assumption. In any case (under any prior assumption), we learn the posterior distribution for $\tilde{\boldsymbol{u}}$. For the same reason, without loss of generality, we also assume $\tilde{\boldsymbol{\mu}}_k = \boldsymbol{0}_{D-1},\ \forall k$, i.e.,

$$\tilde{\boldsymbol{u}} \sim \mathcal{N}(\boldsymbol{0}_{K(D-1)}, \boldsymbol{I}_{K(D-1)}). \qquad (25)$$

To rewrite (24) in terms of $\tilde{\boldsymbol{u}}$ we note that $\boldsymbol{\eta}_i = \tilde{\boldsymbol{C}}_i^T\tilde{\boldsymbol{u}}$ where

$$\tilde{\boldsymbol{C}}_i = \boldsymbol{I}_{D-1}\otimes\boldsymbol{c}_i, \qquad (26)$$

and ⊗ denotes the Kronecker product.

Then, from (24) and (25), we approximate the complete-data likelihood with the following lower bound

$$
\begin{aligned}
\mathsf{P}\left(\{\tilde{\boldsymbol{h}}_i\}, \tilde{\boldsymbol{u}}|\theta\right) &\geq Q_m(\tilde{\boldsymbol{u}}, \{\boldsymbol{c}_i\}) \\
&= \prod_{i=1}^{P} \frac{f_{\tilde{\psi}_i}}{(2\pi)^{K(D-1)/2}} \\
&\quad \exp\left(-\frac{1}{2}\left[\left(\tilde{\boldsymbol{C}}_i^T \tilde{\boldsymbol{u}} - \tilde{\boldsymbol{h}}_i\right)^T M_i \boldsymbol{A} \left(\tilde{\boldsymbol{C}}_i^T \tilde{\boldsymbol{u}} - \tilde{\boldsymbol{h}}_i\right) + \tilde{\boldsymbol{u}}^T \tilde{\boldsymbol{u}}\right]\right) \\
&= \prod_{i=1}^{P} f_{\tilde{\psi}_i} \exp\left(\left[\boldsymbol{\phi}^T \boldsymbol{\Phi}^{-1} \boldsymbol{\phi} - \tilde{\boldsymbol{h}}_i^T M_i \boldsymbol{A} \tilde{\boldsymbol{h}}_i\right]/2\right) \ |\boldsymbol{\Phi}|^{1/2} \\
&\quad \underbrace{\exp\left(-\frac{1}{2}(\tilde{\boldsymbol{u}} - \boldsymbol{\phi})^T \boldsymbol{\Phi}^{-1} (\tilde{\boldsymbol{u}} - \boldsymbol{\phi})\right) \Big/ (2\pi)^{K(D-1)/2} |\boldsymbol{\Phi}|^{1/2}}_{q_m(\tilde{\boldsymbol{u}})}, \qquad (27)
\end{aligned}
$$

where using (26)

$$
\boldsymbol{\phi} = \boldsymbol{\Phi} \sum_{i=1}^{P} M_i \tilde{\boldsymbol{C}}_i \boldsymbol{A} \tilde{\boldsymbol{h}}_i = \boldsymbol{\Phi} \sum_{i=1}^{P} \left(M_i \boldsymbol{A} \tilde{\boldsymbol{h}}_i\right) \otimes \boldsymbol{c}_i, \qquad (28)
$$

$$
\boldsymbol{\Phi} = \left(\sum_{i=1}^{P} M_i \tilde{\boldsymbol{C}}_i \boldsymbol{A} \tilde{\boldsymbol{C}}_i^T + \boldsymbol{I}_{K(D-1)}\right)^{-1} = \left(\sum_{i=1}^{P} M_i \boldsymbol{A} \otimes \boldsymbol{c}_i \boldsymbol{c}_i^T + \boldsymbol{I}_{K(D-1)}\right)^{-1}. \qquad (29)
$$

Using the lower bound in (27) we approximate the posterior $\mathsf{P}\left(\tilde{\boldsymbol{u}}|\{\tilde{\boldsymbol{h}}_i\}, \theta\right)$ with $q_m(\tilde{\boldsymbol{u}})$, which is $\mathcal{N}(\boldsymbol{\phi}, \boldsymbol{\Phi})$.

Note that $K(D-1)$ can be very large due to the dictionary size D. As a result, it is, in general, not practical to perform the matrix inversion in (29). From the Matrix Inversion Lemma, it can be shown that

$$
\begin{aligned}
\boldsymbol{\Phi} &= \boldsymbol{I}_{D-1} \otimes \boldsymbol{F}^{-1} - \boldsymbol{1}_{D-1} \boldsymbol{1}_{D-1}^T \otimes \boldsymbol{\Delta}, \qquad (30) \\
\boldsymbol{F} &= \frac{1}{2} \boldsymbol{C} \boldsymbol{\Lambda}_{M_i} \boldsymbol{C}^T + \boldsymbol{I}_K, \\
\boldsymbol{\Delta} &= \boldsymbol{F}^{-1} \boldsymbol{C} \boldsymbol{Y} \boldsymbol{C}^T \boldsymbol{F}^{-1} \\
\boldsymbol{Y} &= -\frac{\boldsymbol{\Lambda}_{M_i}}{2D} - \frac{\boldsymbol{\Lambda}_{M_i}}{2D} \boldsymbol{C}^T \left(\frac{\boldsymbol{F}}{D-1} - \boldsymbol{C} \frac{\boldsymbol{\Lambda}_{M_i}}{2D} \boldsymbol{C}^T\right)^{-1} \boldsymbol{C} \frac{\boldsymbol{\Lambda}_{M_i}}{2D},
\end{aligned}
$$

where $\boldsymbol{C} = [\boldsymbol{c}_1 \dots \boldsymbol{c}_P]$, and $\boldsymbol{\Lambda}_{M_i}$ is the diagonal matrix whose entries are $M_1, \dots, M_P$. Using (30) we efficiently compute $\boldsymbol{\Phi}$ by only inverting $K \times K$ matrices. Since, typically, the number of events is selected a small number, the proposed algorithm is now feasible for big datasets with large P and D (see Theorem 1 and Section 4).

We can similarly simplify the computation of $\boldsymbol{\phi}$, given in (28). Define

$$
\boldsymbol{z}_i = M_i \boldsymbol{A} \tilde{\boldsymbol{h}}_i,
$$

and partition the posterior mean $\boldsymbol{\phi}$ of the $K(D-1)$ event-word scores into $D-1$ vectors of size K

$$
\boldsymbol{\phi} = [\boldsymbol{x}_1^T \dots \boldsymbol{x}_{D-1}^T]^T, \quad \boldsymbol{X} = [\boldsymbol{x}_1 \dots \boldsymbol{x}_{D-1}]. \qquad (31)
$$

We can efficiently compute $\boldsymbol{X}$, which is nothing but a reorganized version of $\boldsymbol{\phi}$, as

$$\begin{aligned}\boldsymbol{X} &= \boldsymbol{F}^{-1}\boldsymbol{C}\boldsymbol{Z} - \boldsymbol{\Delta}\boldsymbol{C}\tilde{\boldsymbol{Z}}, \\ \boldsymbol{Z} &= [\boldsymbol{z}_1 \dots \boldsymbol{z}_P]^T, \\ \tilde{\boldsymbol{Z}} &= \boldsymbol{Z}\ \mathbf{1}_{D-1}\mathbf{1}_{D-1}^T.\end{aligned} \tag{32}$$

3.2.2 M-step

The mean and covariance of $\tilde{\boldsymbol{u}}$ are estimated using (28) and (29). From [20], the optimum value of $\boldsymbol{\psi}_i$ is given by

$$\boldsymbol{\psi}_i = \tilde{\boldsymbol{C}}_i^T\boldsymbol{\phi} = \boldsymbol{X}^T\boldsymbol{c}_i. \tag{33}$$

For the estimation of $\{\boldsymbol{c}_i\}$, which is considered in the next section, we use the expected value of the lower bound to the complete-data log-likelihood, given in (27),

$$\begin{aligned}\mathsf{E}_{q_m(\tilde{\boldsymbol{u}})}\left[\log Q_m(\tilde{\boldsymbol{u}}, \{\boldsymbol{c}_i\})\right] &= -\frac{1}{2}\mathsf{E}_{q_m(\tilde{\boldsymbol{u}})}\left[\tilde{\boldsymbol{u}}^T\left(\sum_{i=1}^P M_i\tilde{\boldsymbol{C}}_i\boldsymbol{A}\tilde{\boldsymbol{C}}_i^T + \boldsymbol{I}_{K(D-1)}\right)\tilde{\boldsymbol{u}}\right] \\ &\quad + \mathsf{E}_{q_m(\tilde{\boldsymbol{u}})}\left[\tilde{\boldsymbol{u}}\right]^T\left(\sum_{i=1}^P M_i\tilde{\boldsymbol{C}}_i\boldsymbol{A}\tilde{\boldsymbol{h}}_i\right) + \text{Const.} \\ &= -\frac{1}{2}\mathsf{Tr}\left[\left(\sum_{i=1}^P M_i\tilde{\boldsymbol{C}}_i\boldsymbol{A}\tilde{\boldsymbol{C}}_i^T + \boldsymbol{I}_{K(D-1)}\right)\left(\boldsymbol{\Phi} + \boldsymbol{\phi}\boldsymbol{\phi}^T\right)\right] \\ &\quad + \boldsymbol{\phi}^T\sum_{i=1}^P M_i\tilde{\boldsymbol{C}}_i\boldsymbol{A}\tilde{\boldsymbol{h}}_i + \text{Const.},\end{aligned} \tag{34}$$

where $\mathsf{Tr}(\cdot)$ is the trace of a matrix, and the expectation is taken with respect to $q_m(\tilde{\boldsymbol{u}})$ (see (27)). To compute the expectation of the quadratic term we use the fact that $\mathsf{E}\left[\tilde{\boldsymbol{u}}^T\boldsymbol{X}\tilde{\boldsymbol{u}}\right] = \mathsf{E}\left[\mathsf{Tr}(\tilde{\boldsymbol{u}}^T\boldsymbol{X}\tilde{\boldsymbol{u}})\right] = \mathsf{E}\left[\mathsf{Tr}(\boldsymbol{X}\tilde{\boldsymbol{u}}\tilde{\boldsymbol{u}}^T)\right] = \mathsf{Tr}\left(\boldsymbol{X}\mathsf{E}[\tilde{\boldsymbol{u}}\tilde{\boldsymbol{u}}^T]\right)$.

3.3 Mixture Coefficients

From (11) and (34), we estimate the mixture coefficients of the i-th hashtag as

$$\begin{aligned}\hat{\boldsymbol{c}}_i &= \arg\max_{\boldsymbol{c}_i}\ \mathsf{E}_{q_v(\boldsymbol{v}_k)}\left[\log Q_v(\{\boldsymbol{v}_k\}, \{\boldsymbol{c}_i\})\right] + \mathsf{E}_{q_m(\tilde{\boldsymbol{u}})}\left[\log Q_m(\tilde{\boldsymbol{u}}, \{\boldsymbol{c}_i\})\right] \\ &= \left(\boldsymbol{B}\frac{\boldsymbol{c}_i}{\sum_{k=1}^K c_{ik}}\right)^T \kappa_i\sum_{n=1}^{N_i}\boldsymbol{w}_{in} + \boldsymbol{\phi}^T M_i\tilde{\boldsymbol{C}}_i\boldsymbol{A}\tilde{\boldsymbol{h}}_i - \frac{1}{2}\mathsf{Tr}\left[M_i\tilde{\boldsymbol{C}}_i\boldsymbol{A}\tilde{\boldsymbol{C}}_i^T\left(\boldsymbol{\Phi} + \boldsymbol{\phi}\boldsymbol{\phi}^T\right)\right],\end{aligned} \tag{35}$$

where $\boldsymbol{B} = [\boldsymbol{b}_1 \dots \boldsymbol{b}_K]$ holds the posterior mean directions of the event geolocations (see (9)). From (28),

$$\begin{aligned}\boldsymbol{\phi}^T M_i\tilde{\boldsymbol{C}}_i\boldsymbol{A}\tilde{\boldsymbol{h}}_i &= \boldsymbol{\phi}^T\left(\boldsymbol{z}_i \otimes \boldsymbol{c}_i\right) \\ &= \boldsymbol{c}_i^T\boldsymbol{X}\boldsymbol{z}_i.\end{aligned} \tag{36}$$

Using the definitions of $\boldsymbol{A}$ and $\tilde{\boldsymbol{C}}_i$, given in (22) and (26), we write

$$\begin{aligned} M_i\tilde{\boldsymbol{C}}_i\boldsymbol{A}\tilde{\boldsymbol{C}}_i^T &= \frac{M_i}{2}\tilde{\boldsymbol{C}}_i\tilde{\boldsymbol{C}}_i^T - \frac{M_i}{2D}(\tilde{\boldsymbol{C}}_i\boldsymbol{1}_{D-1})(\tilde{\boldsymbol{C}}_i\boldsymbol{1}_{D-1})^T \\ &= \boldsymbol{I}_{D-1}\otimes\frac{M_i}{2}\boldsymbol{c}_i\boldsymbol{c}_i^T - \boldsymbol{1}_{D-1}\boldsymbol{1}_{D-1}^T\otimes\frac{M_i}{2D}\boldsymbol{c}_i\boldsymbol{c}_i^T. \end{aligned}$$

As a result, from (30),

$$\begin{aligned} \mathsf{Tr}\Big(M_i\tilde{\boldsymbol{C}}_i\boldsymbol{A}\tilde{\boldsymbol{C}}_i^T\boldsymbol{\Phi}\Big) &= \mathsf{Tr}\left(\boldsymbol{I}_{D-1}\otimes\frac{M_i}{2}\boldsymbol{c}_i\boldsymbol{c}_i^T\boldsymbol{F}^{-1} - \boldsymbol{1}_{D-1}\boldsymbol{1}_{D-1}^T\otimes\frac{M_i}{2D}\boldsymbol{c}_i\boldsymbol{c}_i^T\boldsymbol{F}^{-1}\right) \\ &\quad - \mathsf{Tr}\left(\boldsymbol{1}_{D-1}\boldsymbol{1}_{D-1}^T\otimes\frac{M_i}{2}\boldsymbol{c}_i\boldsymbol{c}_i^T\boldsymbol{\Delta} - \boldsymbol{1}_{D-1}\boldsymbol{1}_{D-1}^T\otimes\frac{M_i(D-1)}{2D}\boldsymbol{c}_i\boldsymbol{c}_i^T\boldsymbol{\Delta}\right) \\ &= \boldsymbol{c}_i^T\left[\frac{M_i(D-1)^2}{2D}\boldsymbol{F}^{-1} - \frac{M_i(D-1)}{2D}\boldsymbol{\Delta}\right]\boldsymbol{c}_i. \end{aligned} \tag{37}$$

Similarly, using (31) we write

$$\begin{aligned} \mathsf{Tr}\Big(M_i\tilde{\boldsymbol{C}}_i\boldsymbol{A}\tilde{\boldsymbol{C}}_i^T\boldsymbol{\phi}\boldsymbol{\phi}^T\Big) &= \sum_{d=1}^{D-1}\mathsf{Tr}\left(\frac{M_i}{2}\boldsymbol{c}_i\boldsymbol{c}_i^T\boldsymbol{x}_d\boldsymbol{x}_d^T\right) - \sum_{d=1}^{D-1}\sum_{j=1}^{D-1}\mathsf{Tr}\left(\frac{M_i}{2D}\boldsymbol{c}_i\boldsymbol{c}_i^T\boldsymbol{x}_d\boldsymbol{x}_j^T\right) \\ &= \boldsymbol{c}_i^T\left[\frac{M_i}{2}\sum_{d=1}^{D-1}\boldsymbol{x}_d\boldsymbol{x}_d^T - \frac{M_i}{2D}\sum_{d=1}^{D-1}\sum_{j=1}^{D-1}\boldsymbol{x}_d\boldsymbol{x}_j^T\right]\boldsymbol{c}_i. \end{aligned} \tag{38}$$

Substituting (36), (37) and (38) in (35) we have the following quadratic program

$$\begin{aligned} &\hat{\boldsymbol{c}}_i = \arg\max_{\boldsymbol{c}_i}\; -\frac{1}{2}\boldsymbol{c}_i^T\boldsymbol{\Gamma}_i\boldsymbol{c}_i + \boldsymbol{c}_i^T\boldsymbol{\gamma}_i \\ &\qquad\quad \text{subject to } \; c_{ik}\geq 0,\; k=1,\ldots,K, \\ &\boldsymbol{\Gamma}_i = \frac{M_i(D-1)^2}{2D}\boldsymbol{F}^{-1} - \frac{M_i(D-1)}{2D}\boldsymbol{\Delta} + \frac{M_i}{2}\sum_{d=1}^{D-1}\boldsymbol{x}_d\boldsymbol{x}_d^T - \frac{M_i}{2D}\sum_{d=1}^{D-1}\sum_{j=1}^{D-1}\boldsymbol{x}_d\boldsymbol{x}_j^T \\ &\boldsymbol{\gamma}_i = \boldsymbol{B}^T\kappa_i\sum_{n=1}^{N_i}\boldsymbol{w}_{in} + \boldsymbol{X}\boldsymbol{z}_i, \end{aligned} \tag{39}$$

which can be efficiently solved using the interior point method.

The resulting algorithm is summarized as Algorithm 1.

3.4 Computational Complexity

In the following theorem, we show that the computational complexity of Algorithm 1 scales linearly with each dimension of the problem. As a result, the proposed algorithm can be efficiently used for large datasets, as demonstrated in the following section.

Theorem 1 *At each iteration of the proposed EM algorithm, given by Algorithm 1, the computational complexity linearly scales with the number of factors K, the number of hashtags P, and the number of words in the dictionary D, i.e., $O(KPD)$.*

Algorithm 1 The proposed EM algorithm

```
1: Input $\{\bar{\boldsymbol{w}}_i^{3\times 1} = \sum_{n=1}^{N_i} \boldsymbol{w}_{in}, \boldsymbol{h}_i^{D\times 1}\},\ i = 1, \ldots, P$
2: Initialize $\{\boldsymbol{c}_i^{K\times 1}, \kappa_i, s_k, \beta_k^{3\times 1}\},\ k = 1, \ldots, K$
3: while not converged do
4:     Compute posterior parameters $\{\boldsymbol{b}_k, r_k, \tau_i\}$ for vMF as in (9), (10), (12)
5:     Update vMF parameters:
       $\kappa_i = \max\left\{0, \frac{3\tau_i - \tau_i^3}{1-\tau_i^2}\right\},\ s_k = \max\left\{0, \frac{3\boldsymbol{\beta}_k^T \boldsymbol{b}_k - (\boldsymbol{\beta}_k^T \boldsymbol{b}_k)^3}{1-(\boldsymbol{\beta}_k^T \boldsymbol{b}_k)^2}\right\},\ \boldsymbol{\beta}_k = \boldsymbol{b}_k$
6:     Compute posterior parameters $\{\boldsymbol{\Phi}, \boldsymbol{X}\}$ for multinomial as in (30), (32)
7:     Update multinomial parameter $\boldsymbol{\psi}_i = \boldsymbol{X}^T \boldsymbol{c}_i$
8:     Update mixture coefficients $\{\boldsymbol{c}_i\}$ by solving (39)
9: end while
```

Proof First of all, note that typically $K \ll P \ll D$. We start with the vMF E-step (line 4 in Algorithm 1). The most expensive computation in the vMF E-step is the posterior mean direction $\boldsymbol{b}_k$, given by (9). Note that the sum of geolocation vectors $\sum_{n=1}^{N_i} \boldsymbol{w}_{in}$ is computed offline once for each hashtag i; hence the number of geotagged tweets N_i does not contribute to the computational complexity. Each $\boldsymbol{b}_k$ has a computational complexity of $O(P)$. As a result, the computational complexity for the vMF E-Step is $O(KP)$. There is no expensive computation in the vMF M-step (line 5 in Algorithm 1).

In the multinomial E-step (line 6 in Algorithm 1), the computational complexity of the posterior covariance $\boldsymbol{\Phi}$ is $O(K^2P)$ due to the computation of $\boldsymbol{F}$ and $\boldsymbol{\Delta}$ (see (30)). The computation of the posterior mean $\boldsymbol{X}$ in (32) has a complexity of $O(KPD)$ due to the multiplication of the matrices $\boldsymbol{C}^{K\times P}$ and $\boldsymbol{Z}^{P\times D}$. The computational complexity of the multinomial M-step (line 7 in Algorithm 1) is $O(KD)$.

Finally, the complexity for updating the coefficients $\{\boldsymbol{c}_i\}$ (line 8 in Algorithm 1) is $O(KPD)$ since for each $\boldsymbol{\gamma}_i$, given by (39), the complexity is $O(KD)$ due to the multiplication $\boldsymbol{X}\boldsymbol{z}_i$. In (39), the matrices $\{\boldsymbol{\Gamma}_i\}$ entail the complexity of $O(K^2D)$ due to the computation of $\sum_{d=1}^{D-1} \boldsymbol{x}_d \boldsymbol{x}_d^T = \boldsymbol{X}\boldsymbol{X}^T$, which is computed once and used for each $\boldsymbol{\Gamma}_i$. Note that $\sum_{d=1}^{D-1} \sum_{j=1}^{D-1} \boldsymbol{x}_d \boldsymbol{x}_j^T = \left(\sum_{d=1}^{D-1} \boldsymbol{x}_d\right) \left(\sum_{d=1}^{D-1} \boldsymbol{x}_d\right)^T$ is nothing but the outer product of the sum vector $\sum_{d=1}^{D-1} \boldsymbol{x}_d$, requiring $O(KD)$ computations for the sum and $O(K^2)$ computations for the outer product. In solving the constrained quadratic program given in (39) for each coefficient vector $\boldsymbol{c}_i$, the number of iterations, in practice, is bounded by a constant; and in each iteration linear algebra operations in the K-dimensional space are performed. Hence, the overall complexity does not exceed $O(KPD)$. Note also that each $\boldsymbol{c}_i$ can be updated in parallel.

4 Experiments

We have tested the proposed algorithm on a Twitter dataset from August 2014 obtained from the Twitter stream API at gardenhose level access. It spans the whole month, and includes a random sample of 10 % of all tweets from all over the world. We consider about 30 million geotagged tweets, among which around 3 million use approximately 1 million unique hashtags. We have organized the data in terms of hashtags. That is, each unique hashtag is an instance with bag-of-words and geolocation features. The rarely used hashtags and the hashtags with small geographical distribution are filtered out, leaving us with 13000 hashtags ($P = 13000$), and a dictionary of 67000 significant words ($D = 67000$). The number of geotags, N_i, for hashtags varies from 2 to 71658; and the number of words, M_i, varies from 10 to 426892.

We run the algorithm in a hierarchical manner with a small number of events in each round (e.g., $K = 10$). In each round, the hashtags that localize well in an event with a dominant mixture coefficient are pruned, and the remaining hashtags are further processed in the next round. In other words, in the next round, we zoom into the previously unexplored sections of the data to discover new events. We also zoom into the broadly discovered events to find specific events. For example, in the first round, we have discovered popular events such as the Ice Bucket Challenge and the Robin Williams' death, and also generic events for the British and Asian hashtags (top figure in Fig. 6). In the following round, separately processing the generic events and the non-localized data we have identified further specific events such as the Ferguson unrest and the USA national basketball team for the FIBA world championship (bottom figure in Fig. 6). Specifically, we have identified an Indian event about a Hindu religious leader in jail, and a British event about the Commonwealth Games in the generic Asian and British events, respectively. In Fig. 6, it is seen also that the previously found Ice Bucket Challenge event has decomposed into a local event and a global event in the second round. It is seen in Fig. 6 that the proposed algorithm successfully finds interesting events in terms of both topic and geolocation. The geographical distribution of the tweets that use hashtags associated with the events about the Commonwealth Games and the Hindu religious leader are depicted in Fig. 7. Similarly, Fig. 8 illustrates the geographical distributions of the tweets that use hashtags about the death of Robin Williams. The geolocations of the tweets that are shown in Fig. 7 and Fig. 8 are consistent with the corresponding events. As expected, the tweets that mention Robin Williams are well distributed around the world with a center in the USA, whereas the tweets about the Commonwealth Games are sent only from the Commonwealth countries, and the tweets about the Hindu leader are only from India.

Next, as an application, we cluster the hashtags based on the mixture coefficients $\boldsymbol{c}_i$. A sample result using k-means and multidimensional scaling (MDS) is shown in Fig. 9. For this example, we have chosen a small subset of the dataset with 314 hashtags and annotated each hashtag with an event name, which resulted in 7 events as shown in Fig. 9. This hand annotation provides us a ground truth (color and shape coded by markers), which we compare with the clustering result from the proposed algorithm (shown by ellipses). Running the algorithm with $K = 7$ we get a 7-dimensional coefficient vector $\boldsymbol{c}_i$ for each hashtag i [1]. We then cluster the vectors $\{\boldsymbol{c}_i\}$ into 7 groups using the k-means technique. The rand index and the adjusted rand index between the clustering result and the ground truth are 0.9807 and 0.9539, respectively. The rand index is a similarity measure between two clusterings. It takes a value between 0 and 1, where 0 indicates no agreement, and 1 indicates perfect agreement. Hence, from the rand index result

[1] The issue of unknown number of events can be handled using the silhouette values.

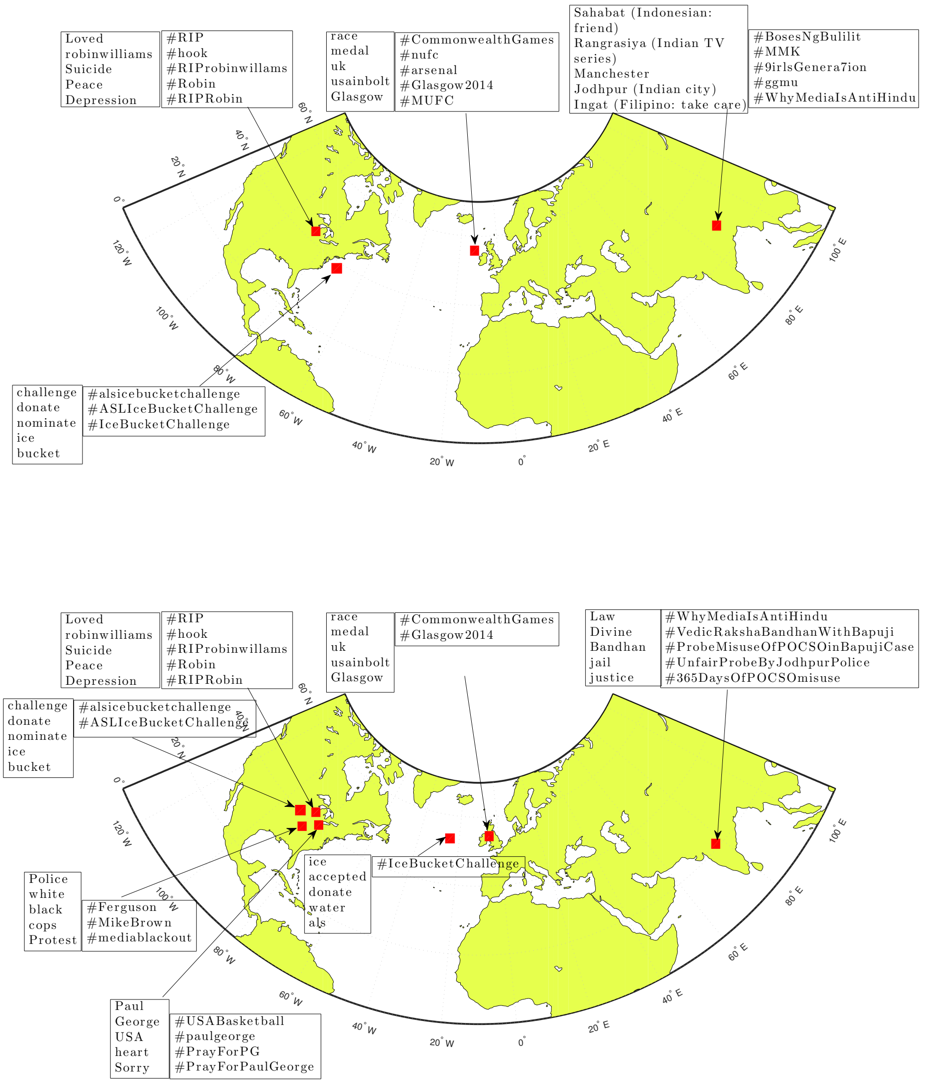

Fig. 6 Some events discovered in the first round of the algorithm (top). Some specific events discovered after two rounds (bottom). Dominant hashtags and words used for the events, as well as their mean geolocations are displayed.

and also Fig. 9, it is seen that the proposed algorithm can be used to effectively cluster multimodal big datasets.

Finally, for the same subset used in Fig. 9, we measure, in Fig. 10, how well our model fits the data by comparing the likelihood values of the geolocation data under the vMF model given by the proposed algorithm with the likelihood values under the individual vMF models separately fitted for each hashtag. The individual

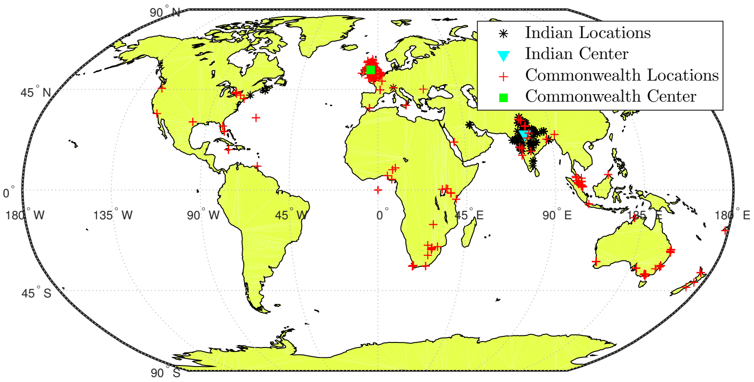

Fig. 7 Geographical distribution of the tweets the algorithm associates with the Commonwealth Games and the Hindu religious leader in jail.

vMF models for each hashtag provide us a baseline for comparison since this is what one would typically do to fit the geolocation data for each hashtag without the goal of event detection. Fig. 10 shows that the proposed event detection algorithm fits the geolocation data as well as the baseline data fitting technique.

5 Conclusion

We have treated the event detection problem in a multimodal Twitter hashtag network. Utilizing the bag-of-words and the geotags from related tweets as the features for hashtags we have developed a variational EM algorithm to detect events according to a generative model. The computational complexity of the proposed algorithm has been simplified such that it is viable for big datasets. A hierarchical version of the proposed algorithm has been tested on a Twitter dataset with 13000 hashtags from August 2014. By pruning data in each round multiresolution events (higher in each round) have been learned. Significant events, such as Robin Williams' death, and the Ice Bucket Challenge, as well as some generic events, such as the British and the Asian hashtags, have been learned in the first round. Later in the second round, new specific events have been discovered within the generic events. We have also successfully clustered a set of hashtags using the detected events. In addition to event detection, we have shown that the proposed algorithm fits the geolocation data as well as the baseline data fitting technique which separately models each hashtag. The proposed algorithm is justified by the remarkable clustering and goodness of fit results, and the low computational complexity that linearly scales with the number of events, number of hashtags, number of tweets, and number of words in the dictionary.

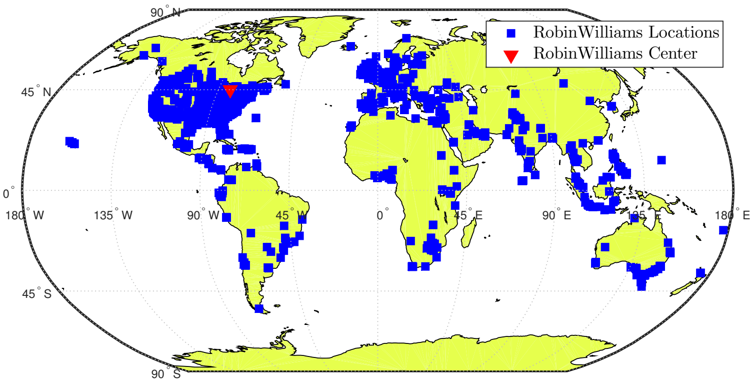

Fig. 8 Geographical distribution of the tweets the algorithm associates with the death of Robin Williams.

References

1. http://www.socialbakers.com/statistics/twitter/
2. Farzindar, A., & Khreich, W. (2015). A Survey of Techniques for Event Detection in Twitter. *Computational Intelligence*, 31(1), 132–164.
3. Liu, K.L., Li, W., & Guo, M. (2012). Emoticon Smoothed Language Models for Twitter Sentiment Analysis. *AAAI Conference on Artificial Intelligence.*
4. Amer-Yahia, S., Anjum, S., Ghenai, A., Siddique, A., Abbar, S., Madden, S., Marcus, A., & El-Haddad, M. (2012). MAQSA: A System for Social Analytics on News. *ACM SIGMOD International Conference on Management of Data.*
5. Zhao, Z., Resnick, P., & Mei, Q. (2015). Enquiring Minds: Early Detection of Rumors in Social Media from Enquiry Posts. *International World Wide Web Conference.*
6. Oselio, B., Kulesza, A., & Hero, A. (2015). Information Extraction from Large Multi-Layer Social Networks. *IEEE International Conference on Acoustics, Speech, and Signal Processing.*
7. Tumasjan, A., Sprenger, T.O., Sandner, P.G., & Welpe, I.M. (2012). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *International Conference on Weblogs and Social Media.*
8. Wang, X., Gerber, M.S., & Brown, D.E. (2012). Automatic Crime Prediction Using Events Extracted from Twitter Posts. *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction.*
9. Yang, Y., Pierce, T., & Carbonell, J. (1998). A Study of Retrospective and On-line Event Detection. *ACM SIGIR Conference on Research and Development in Information Retrieval.*
10. Phuvipadawat, S., & Murata, T. (2010). Breaking News Detection and Tracking in Twitter. *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.*
11. Popescu, A. M., Pennacchiotti, M., & Paranjpe, D. (2011). Extracting Events and Event Descriptions from Twitter. *International Conference Companion on World Wide Web.*
12. Benson, E., Haghighi, A., & Barzilay, R. (2011). Event Discovery in Social Media Feeds. *Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.*
13. Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors. *International Conference on World Wide Web.*

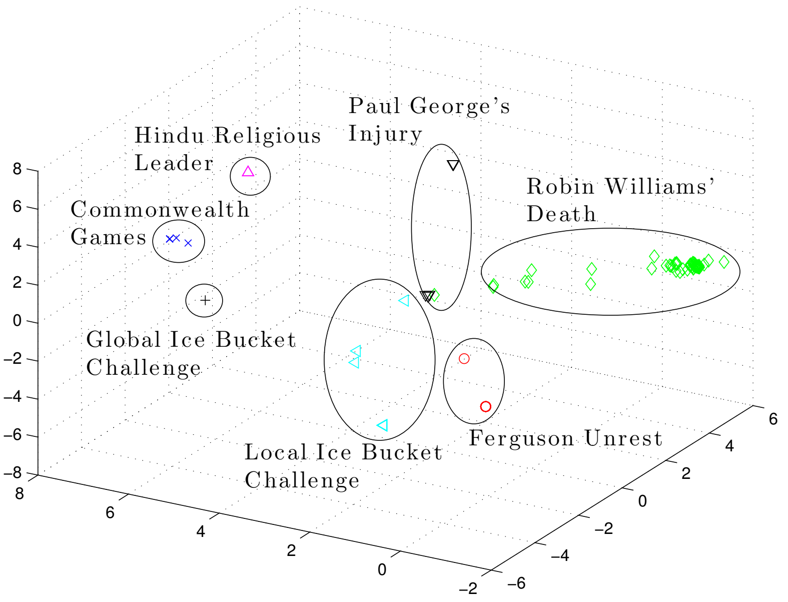

Fig. 9 Clustering via k-means and MDS based on the mixture coefficients. The ground truth is shown with the color and shape of the markers; and the clustering result is shown with the ellipses.

14. Petrovic, S., Osborne, M., & Lavrenko, V. (2010). Streaming First Story Detection with Application to Twitter. *Human Language Technologies: Annual Conference of the North American Chapter of the Association for Computational Linguistics.*

15. Becker, H., Naaman, M., & Gravano, L. (2011). Beyond Trending Topics: Real-World Event Identification on Twitter. *International Conference on Weblogs and Social Media.*

16. Long, R., Wang, H., Chen, Y., Jin, O., & Yu, Y. (2011). Towards Effective Event Detection, Tracking and Summarization on Microblog Data. *Web-Age Information Management, Vol. 6897 of Lecture Notes in Computer Science.* Edited by Wang, H., Li, S., Oyama, S., Hu, X., & Qian, T. Springer: Berlin/Heidelberg, 652663.

17. Weng, J., & Lee, B.-S. (2011). Event Detection in Twitter. *International Conference on Weblogs and Social Media.*

18. Cordeiro, M. (2012). Twitter Event Detection: Combining Wavelet Analysis and Topic Inference Summarization. *Doctoral Symposium on Informatics Engineering.*

19. Yılmaz, Y., & Hero, A. (2015). Multimodal Factor Analysis. *IEEE International Workshop on Machine Learning for Signal Processing.*

20. Khan, M.E., Bouchard, G., Marlin, B.M., & Murphy, K.P. (2010). Variational Bounds for Mixed-Data Factor Analysis. *Neural Information Processing Systems (NIPS) Conference.*

21. Adali, T., Levin-Schwartz, Y., & Calhoun, V.D. (2015). Multimodal Data Fusion Using Source Separation: Two EffectiveModels Based on ICA and IVA and Their Properties. *Proceedings of the IEEE*, 103(9), 1478–1493.

22. Bramon, R., Boada, I., Bardera, A., Rodriguez, J., Feixas, M., Puig, J., & Sbert, M. (2012). Multimodal Data Fusion Based on Mutual Information. *IEEE Transactions on Visualization and Computer Graphics*, 18(9), 1574–1587.

23. Wu, Y., Chang, K.C.-C., Chang, E.Y., & Smith, J.R. (2004). Optimal Multimodal Fusion for Multimedia Data Analysis. *ACM International Conference on Multimedia.*

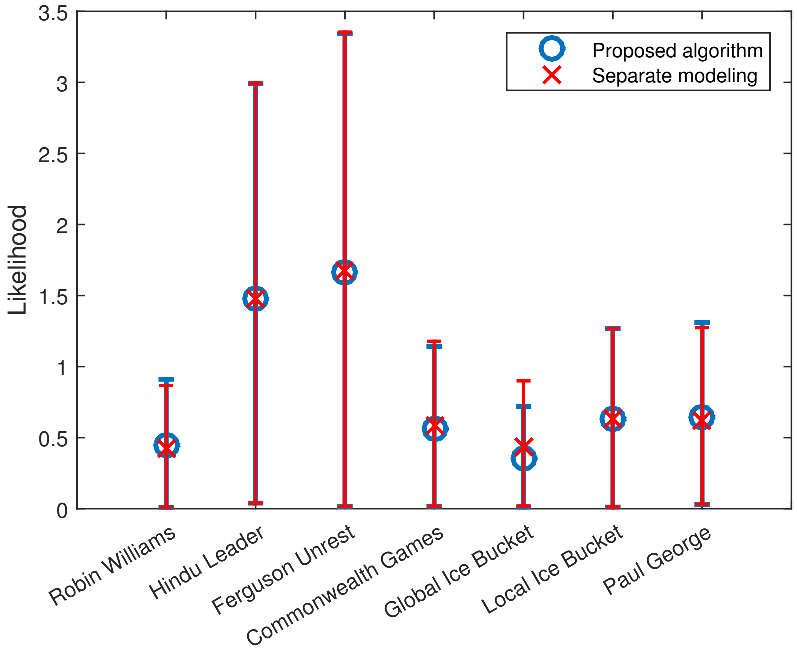

Fig. 10 Goodness of fit plot for the proposed event detection algorithm and the individual vMF models separately fitted to the geolocation data for each hashtag. Note that the separate modeling does not perform event detection; it is the baseline technique for data fitting. It is seen that the proposed algorithm not only detects events (see the previous figures), but also fits the geolocation data as well as the baseline technique. For each event, the mean likelihood values are shown with markers (circle and cross), and the 95% confidence intervals are shown with bars.

24. Sui, J., Adali, T., Yu, Q., Chen, J., & Calhoun, V.D. (2012). A review of Multivariate Methods for Multimodal Fusion of Brain Imaging Data. *Journal of Neuroscience Methods*, 204(1), 68–81.
25. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A.Y. (2011). Multimodal Deep Learning. *International Conference on Machine Learning.*
26. Christoudias, C.M., Urtasun, R., & Darrell, T. (2008). Multi-View Learning in the Presence of View Disagreement. *Conference on Uncertainty in Artificial Intelligence.*
27. He, J., & Lawrence, R. (2011). A Graph-Based Framework for Multi-Task Multi-View Learning. *International Conference on Machine Learning.*
28. Sun, S. (2013). A Survey of Multi-View Machine Learning. *Neural Computing and Applications*, 23(7), 2031–2038.
29. Mardia, K.V., & Jupp, P.E. (2000). Directional Statistics. Chichester: Wiley.
30. Mardia, K.V., & El-Atoum, S.A.M. (1976). Bayesian Inference for the Von Mises-Fisher Distribution. *Biometrika*, 63(1), 203–206.
31. Harman, H.H., (1976). Modern Factor Analysis. University of Chicago Press.
32. Abramowitz, M., & Stegun, I.A. (1972). Handbook of Mathematical Functions. *National Bureau of Standards Applied Mathematics Series*, 55.
33. Banerjee, A., Dhillon, I.J., Ghosh J., & Sra, S. (2005). Clustering on the Unit Hypersphere using von Mises-Fisher Distributions. *Journal of Machine Learning Research*, 6(Sep), 1345–1382.
34. Böhning, D. (1992). Multinomial Logistic Regression Algorithm. *Ann. Inst. Statist. Math.* 44(1), 197–200.