5-2014

Von Mises-Fisher Clustering Models

Siddarth Gopal
Carnegie Mellon University

Yiming Yang
Carnegie Mellon University, yiming@cs.cmu.edu

Published In

Journal of Machine Learning Research : Workshop and Conference Proceedings, 32, 154-162.

Von Mises-Fisher Clustering Models

Siddharth Gopal SGOPAL1@CS.CMU.EDU
Carnegie Mellon University, Pittsburgh, PA 15213 USA

Yiming Yang YIMING@CS.CMU.EDU
Carnegie Mellon University, Pittsburgh, PA 15213 USA

Abstract

This paper proposes a suite of models for clustering high-dimensional data on a unit sphere based on von Mises-Fisher (vMF) distribution and for discovering more intuitive clusters than existing approaches. The proposed models include a) A Bayesian formulation of vMF mixture that enables information sharing among clusters, b) a Hierarchical vMF mixture that provides multi-scale shrinkage and tree structured view of the data and c) a Temporal vMF mixture that captures evolution of clusters in temporal data. For posterior inference, we develop fast variational methods as well as collapsed Gibbs sampling techniques for all three models. Our experiments on six datasets provide strong empirical support in favour of vMF based clustering models over other popular tools such as K-means, Multinomial Mixtures and Latent Dirichlet Allocation.

1. Introduction

With the advent of large amounts of unlabeled data, clustering has emerged as an important tool for the end user to obtain a structured view of the data. Probabilistic clustering algorithms such as K-means (Gaussian mixtures), Multinomial Mixtures, Latent Dirichlet allocation (Blei et al., 2003) have emerged as the defacto standard for discovering the latent structures and relations in the data. Such probabilistic models define a generative model for the data by assuming some rigid instance representation, for e.g. Multinomial Mixtures assumes that each instance is represented as discrete feature-counts and is drawn from one of many Multinomial distributions.

However, it is questionable whether such representation of data is appropriate for all domains. For example, in text-mining (classification, retrieval, collaborative filtering etc) documents have typically been represented using a term frequency-Inverse Document frequency Normalized form (Tf-Idf normalization) (Salton & McGill, 1986), where each document is represented as a point on a unit-sphere using a combination of both within-document frequencies and inverse corpus-frequences. Tf-Idf normalization has always shown better performance than feature-counts representation based on several supervised tasks such as classification (Joachims, 2002), retrieval (Robertson, 2004) etc. Similarly, in Image-modeling, unit normalized spatial pyramid vectors is a common representation (Yang et al., 2009). Normalization is often an important step in data analysis because it removes the 'magnitude' of the instances from the picture and places more importance on the directional distribution; in other words, we do not want unduly long documents or big images to distort our inferences, hence we represent the instances as relative contributions of individual features.

Despite the practical successes of normalized data representation, they have not been well studied by Bayesian Graphical models. Since the data lies on a unit-sphere manifolds, popular clustering assumptions such as Gaussian (O'Hagan et al., 2004) or Multinomial (Blei et al., 2003; 2004; Blei & Lafferty, 2006a;b) are not appropriate. On one hand we have empirical success of such normalized representation and on the other hand we have a wide variety of graphical models that model data using a different representation. Can we get the best of both worlds? Can we develop models that are particularly suited to such unit-sphere manifolds but at the same time maintain flexibility like other graphical models? In this paper, we propose a suite of models using von Mises-Fisher (vMF) distributions which have been primarily used to model such directional data. vMF models have been long studied in the directional statistics community (Fisher, 1953; Jupp & Mardia, 1989) and are naturally suited in such scenarios as they model distances between instances using the angle of separation i.e. cosine similarity. Works in vMF have typically focused on low-dimensional problems (2D or 3D spaces) to maintain tractability and relying on Gibbs sampling for inference which is generally difficult to scale in higher dimensions (Hasnat et al., 2013) (Mardia & El-Atoum, 1976) (Guttorp & Lockhart, 1988) (Bangert et al., 2010). More

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

popular works include text clustering work by (Banerjee et al., 2006) (Banerjee et al., 2003) where an EM-based algorithm without Bayesian inference was used, and spherical topic models by (Reisinger et al., 2010) which mimic LDA with a Bayesian inference for learning the mean parameters in vMF, but leave the crucial concentration parameters (the variance parts) of the models to be set manually.

In this paper, using the vMF distribution as the building block, we propose three increasingly powerful models for cluster analysis

1. *The Bayesian vMF Mixture Model* (**B-vMFmix**): A fully Bayesian formulation of mixture of vMF distributions where each instance represented as a point on a unit-sphere is assumed to be drawn from one of many vMF distributions. The parameters of the clusters are themselves drawn from a common prior which helps the clusters to share information among each other.

2. *The Hierarchical vMF Mixture Model* (**H-vMFmix**): When the data we want to analyze is huge, one of the efficient means of browsing is by means of a hierarchy (Cutting et al., 1992). We extend B-vMFmix to H-vMFmix, to enable partitioning the data into increasing levels of specificity as defined by a given input hierarchy. To our knowledge, this is the first hierarchical Bayesian model for vMF-based clustering.

3. *The Temporal vMF Mixture Model* (**T-vMFmix**): For temporal data streams, analyzing how the latent clusters in the data evolve over time is naturally desirable. We augment B-vMFmix to the first temporal vMF-based model that accommodates changes in cluster parameters between adjacent time-points. For example, in a corpus of documents, this could reflect the changing vocabulary within a cluster of documents.

We develop fast variational inference schemes for all the methods and equally fast collapsed Gibbs sampling techniques for the simpler models. Our empirical comparison on several datasets conclusively establishes that vMF distributions are better alternatives to standard Gaussian, Multinomial Mixtures and can be successfully used for cluster analysis. To our knowledge, this is the first work that provides a thorough treatment of vMF distribution in the context of a wide variety of graphical models for data analysis.

2. Related Background

The von Mises-Fisher (vMF) distribution defines a probability density over points on a unit-sphere. It is parameterized by mean parameter μ and concentration parameter κ - the former defines the direction of the mean and the latter determines the spread of the probability mass around the mean. The density function for $x \in \mathcal{R}^D, \|x\| = 1, \|\mu\| = 1, \kappa > 0$ is given by,

$$f(x|\mu,\kappa) = C_D(\kappa)\exp(\kappa\mu^\top x);\ C_D(\kappa) = \frac{\kappa^{.5D-1}}{(2\pi)^{.5D} I_{.5D-1}(\kappa)}$$

where $I_\nu(a)$ is the modified bessel function of first kind with order ν and argument a. Note that $\mu^\top x$ is the cosine similarity between x and mean μ and that κ plays the role of the inverse of variance.

The simplest vMF mixture model (Banerjee et al., 2006) assumes that each instance is drawn from one of the K vMF distributions with a mixing distribution π, where K is a known constant, and the parameters of the vMF distributions correspond to the underlying themes (clusters) in the data. The cluster assignment variable for instance x_i is denoted by $z_i \in \{1, 2, ..K\}$ in the probabilistic generative model given below,

$$\begin{aligned} z_i &\sim \text{Categorical}(.|\pi) \quad i = 1, 2..N \\ x_i &\sim \text{vMF}(.|\mu_{z_i}, \kappa) \quad i = 1, 2..N \end{aligned}$$

The k'th cluster is defined by mean parameter μ_k and concentration parameter κ. The parameters $\mathcal{P} = \{\boldsymbol{\mu}, \boldsymbol{\kappa}, \pi\}$ are treated as fixed unknown constants and $\mathbf{Z} = \{z_i\}_{i=1}^N$ are treated as a latent variables. To train the model, we can use the familiar EM algorithm, to efficiently iterate between calculating the $E[\mathbf{Z}]$ in the E-step and optimizing $\mathcal{P}$ to maximize the likelihood in the M-step. The update equations are, [1]

$$E[z_{ik}] = \frac{\pi_k \text{vMF}(x_i|\mu_k,\kappa)}{\sum_{j=1}^K \pi_j \text{vMF}(x_i|\mu_j,\kappa)},\ R_k = \sum_{i=1}^N E[z_{ik}]x_i$$

$$\pi_k = \sum_{i=1}^N \frac{E[z_{ik}]}{N}, \mu_k = \frac{R_k}{\|R_k\|},\ \bar{r} = \sum_{k=1}^K \frac{\|R_k\|}{N},\ \kappa = \frac{\bar{r}D - \bar{r}^3}{1-\bar{r}^2}$$

3. Proposed Bayesian vMF models

We improve upon the basic vMF mixture model in the above section by adding bayesian components in three significant ways a) the fully Bayesian vMF mixture model (B-vMFmix) b) A hierarchical extension of B-vMFmix and c) A temporal extension of B-vMFmix.

3.1. Bayesian vMF mixtures (B-vMFmix)

The bayesian vMF mixture model is the fundamental building block for the hierarchical and temporal vMF mixtures. This model enables sharing of information between the clusters by shrinking the cluster parameters towards each other using a prior. The generative model for given data $\mathcal{D} = \{x_i\}_{i=1}^N$ and a fixed number of clusters K is given by,

$$\begin{aligned} \pi &\sim \text{Dirichlet}(.|\alpha) \\ \mu_k &\sim \text{vMF}(.|\mu_0, C_0) \quad k = 1, 2, ..K \\ \kappa_k &\sim \text{logNormal}(.|m, \sigma^2) \quad k = 1, 2, ..K \end{aligned}$$

[1] The supplementary material thoroughly describes all the inference schemes (EM, variational, collapsed gibbs sampling) as well as complete set results only a part of which is presented in the paper

$$z_i \sim \text{Categorical}(.|\pi) \quad i = 1, 2...N$$
$$x_i \sim \text{vMF}(.|\mu_{z_i}, \kappa_{z_i})\ i = 1, 2, ..N$$

The prior parameters are $\{\alpha, \mu_0, C_0, m, \sigma^2\}$. The cluster mean parameters $\boldsymbol{\mu} = \{\mu_k\}_{k=1}^K$ are commonly drawn from a prior vMF distribution with parameters $\{\mu_0, C_0\}$. The cluster concentration parameters $\boldsymbol{\kappa} = \{\kappa_k\}_{k=1}^K$ are commonly drawn from a log-normal prior with mean m and variance σ^2. The mixing distribution π is drawn from a symmetric dirichlet with parameter α. This bayesian model improves over the simple vMF mixture in multiple ways; firstly we share statistical strength by shrinking the cluster mean parameters towards a common μ_0, secondly there is flexibility to learn cluster-specific concentration parameters κ_k without the risk of overfitting if the priors are appropriately set, thirdly the posterior distribution over the parameters gives a measure of uncertainty of the parameters unlike point estimates in simple vMF mixtures. These advantages are evident in our experimental section (in section 4).

The likelihood of the data and the posterior of the parameter is given by,

$$P(\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\kappa}, \pi|.) \propto P(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\kappa}, \pi) P(\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\kappa}, \pi|m, \sigma^2, \mu_0, C_0, \alpha)$$

Since the posterior distribution of the parameters cannot be calculated in closed form, we need to resort to approximate inference using variational methods or sampling techniques.

3.1.1. Variational Inference

Using variational inference, we try to find a distribution that is closest in KL-divergence to the true posterior. We assume the following factored form for the approximate posterior.

$$q(\pi) \sim \text{Dirichlet}(.|\rho)$$
$$q(\mu_k) \sim \text{vMF}(.|\psi_k, \gamma_k)\ \ k = 1, 2..K$$
$$q(z_i) \sim \text{Categorical}(.|\lambda_i)\ \ i = 1, 2..N$$

For the concentration parameters, the inference is not straight-forward because of the presence of log-bessel function. We present two different ways of estimating the posterior of the concentration parameters - (a) Sampling scheme and (b) Bounding scheme. Each scheme assumes a different form for the posterior distribution of κ_k's.

$$q(\kappa_k) \equiv \text{No-specific form}\ \ k = 1, 2..K\ \ \text{[Sampling]}$$
$$q(\kappa_k) \sim \text{logNormal}(.|a_k, b_k)\ \ k = 1, 2..K\ \ \text{[Bounding]}$$

We develop a variational algorithm where the posterior parameters - $\rho, \boldsymbol{\psi}, \boldsymbol{\gamma}, \boldsymbol{\lambda}$ are iteratively optimized to maximize the variational lower bound (VLB)[1] .

$$\text{VLB} = E_q[\log P(\mathcal{D}, \mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\kappa}, \pi|m, \sigma^2, \mu_0, C_0, \alpha)] - H(q)$$

The closed form updates for posterior parameters for μ_k and z_{ik} are given by,

$$\psi_k = \frac{R_k}{\|R_k\|}, \quad \gamma_k = \|R_k\|,$$
$$\text{where } R_k = E_q[\kappa_k] \sum_{i=1}^N E_q[z_{ik}] x_i + C_0 \mu_0$$
$$\lambda_{ik} \propto \exp(E_q[\log \text{vMF}(x_i|\mu_k, \kappa_k)] + E_q[\log(\pi_k)])$$
$$\log \text{vMF}(x_i|\mu_k, \kappa_k) = E_q[\log C_D(\kappa_k)] + E_q[\kappa_k] x_i^\top E_q[\mu_k]$$

Next we present the posterior estimation of the concentration parameters.

Sampling: In the sampling scheme, we rely on estimating κ_k's (and related quantities such as $\log C_D(\kappa_k)$) by drawing samples from the posterior distribution. Sampling requires the computation of the conditional distribution for κ_k. However, variational inference (for the other parameters) does not maintain samples but instead maintains only posterior distributions. To overcome this issue, we rewrite the conditional distribution for κ_k in terms of the expectation of posterior parameters rather than samples. Using the VLB and Jensen's inequality on the conditional of κ_k,

$$\begin{aligned}
&P(\kappa_k|\mathbf{X}, m, \sigma^2, \mu_0, C_0, \alpha) \propto P(\kappa_k, \mathbf{X}|m, \sigma^2, \mu_0, C_0, \alpha) \\
&\approx E_q\left[P(\kappa_k, \mathbf{X}, \mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\kappa}^{-k}, \pi|m, \sigma^2, \mu_0, C_0, \alpha)\right] \\
&\geq \exp\left(E_q\left[\log P(\kappa_k, \mathbf{X}, \mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\kappa}^{-k}|m, \sigma^2, \mu_0, C_0, \pi)\right]\right) \\
&\propto \exp\left(\sum_{i=1}^N E_q[z_{ik}] \log C_D(\kappa_k) + \kappa_k \sum_{i=1}^N E_q[z_{ik}] x_i^\top E_q[\mu_k]\right) \\
&\quad \times \text{logNormal}(\kappa_k|m, \sigma^2) \qquad (1)
\end{aligned}$$

Having identified the proportionality of the conditional distribution, we can use MCMC sampling with a suitable proposal distribution to draw samples. We used a log-normal distribution around the current iterate as the proposal distribution.

The MCMC sampling step for the posterior of κ introduces flexibility into the model as the samples are now from the true posterior (given the rest of the parameters). The downside is that the variational bound is no longer guaranteed to be a valid lower-bound since $E_q[\log C_D(\kappa_k)]$ and $E_q[\kappa_k]$ are estimated through the samples. However, we observed that the posterior distribution of the κ_k's was highly concentrated around the mode and the estimates from samples gave good empirical performance in terms of predictive power.

This partial MCMC sampling does not increase computational costs much since there are only K variables to be estimated and the major computational bottleneck is only in updating $\boldsymbol{\lambda}$. Another alternative to repeated sampling is to use grid search, where we break down the continuous posterior distribution of κ_k across a finite set of points, and use this discrete distribution to estimate the expectation of various quantities.Refer the last part of supplementary material for a thorough discussion of the computational issues.

Note that the same model with an unnormalized prior for κ

was used in . However since κ is not a natural parameter of the vMF distribution, it is not clear whether the unnormalized prior results in a valid posterior distribution.

Bounding: The core problem in doing full variational inference for the model is the computation of the $E_q[\log C_D(\kappa_k)]$ in the VLB. We are not aware of any distribution $q(\kappa_k)$ for which is there is a closed form expression for $E_q[\log C_D(\kappa_k)]$ (due to the presence of the log bessel function in $C_D(\kappa_k)$). To overcome this issue, we first upper-bound the log bessel function using a Turan type inequality (Baricz et al., 2011), followed by an approximation using the delta method. More specifically, the growth of the modified bessel function of first kind with order v and argument u i.e. $I_v(u)$ can be lower-bounded by (Baricz et al., 2011),

$$u\frac{I_v(u)'}{\sqrt{u^2+v^2}} \leq I_v(u) \quad \Rightarrow \frac{I_v(u)'}{I_v(u)} \leq \sqrt{1+\frac{v}{u}^2}$$

Integrating over $u > 0$,

$$log(I_v(u)) \leq \sqrt{u^2+v^2} - v\log(u) - v\log\left(v\sqrt{v^2+u^2}+v^2\right)+v^2\Big) - \sqrt{u^2+v^2}$$

$$E_q[\log I_v(u)] \leq E_q[\sqrt{u^2+v^2}] - vE_q[\log(u)] - vE_q[\log(v\sqrt{v^2+u^2}+v^2)+v^2)] - E_q[\sqrt{u^2+v^2}] \quad (2)$$

Since all the expectations on the RHS of eq (2) are twice differentiable, we can use the delta approximation method. The expectation of a function g over a distribution q is given by

$$E_q[g(x)] \approx g(E_q[x]) + g''(E_q[x])\frac{Var_q[x]}{2} \quad (3)$$

Applying the equations (2), (3) to the VLB, we can estimate the posterior parameters a_k, b_k by optimizing VLB using gradient descent[1] . Although it is tempting to directly apply the delta method to calculate $E_q[\log I_v(u)]$, this leads to expressions that are not computable directly as well as numerically unstable.

3.1.2. Collapsed Gibbs Sampling

We can develop efficient sampling techniques for the model by using the fact that vMF distributions are conjugate w.r.t each other. This enables us to completely integrate out $\{\mu_k\}_{k=1}^K$ and π and update the model only by maintaining the cluster assignment variables $\{z_i\}_{i=1}^N$ and the concentration parameters $\{\kappa_k\}_{k=1}^K$. The conditional distributions [1] are given by,

$$P(z_i = k|\mathbf{Z}^{-i},..) \propto (\alpha + \sum_{j=1,j\neq i}^{N} I(z_j = k))C_D(\kappa_k) \frac{C_D(\|\kappa_k \sum_{j\neq i, z_j=k} x_j + C_0\mu_0\|)}{C_D(\|\kappa_k \sum_{j:z_j=k} x_j + C_0\mu_0\|)}$$

$$P(\kappa_k|\boldsymbol{\kappa}^{-k},..) \propto \frac{C_D(C_0)C_D(\kappa_k)^{\sum_j I(z_j=k)}}{C_D(\|\kappa_k \sum_{j:z_j=k} x_j + C_0\mu_0\|)} \times \text{logNormal}(\kappa_k|m,\sigma^2)$$

The conditional distribution of κ_k is again not of a standard form and as before, we use a step of MCMC sampling (with log-normal proposal distribution around the current iterate) to sample κ_k. The advantage of sampling is that the distribution of samples eventually converge to the true posterior, however, the downside is that it is not clear how many samples must be drawn.

3.1.3. Empirical Bayes

When the user does not have enough information to set the prior parameters, it is useful to be able to learn them directly from the data. The prior parameters $\{\alpha, \mu_0, C_0, m, \sigma^2\}$ are estimated by maximizing the variational lower-bound (which acts as a proxy to the true marginal likelihood). The details are discussed in the supplementary material.

3.2. Hierarchical vMF Mixtures (H-vMFmix)

Often the data that the user wants to analyze is large and manually inspecting a flat layer of several clusters is harder than hierarchically browsing the data (Cutting et al., 1992). For such cases, we develop a hierarchical vMF mixture model that enables a hierarchically nested organization of the data.

We assume that the user wants to organize the data into a given fixed hierarchy of nodes $\mathcal{N}$. The hierarchy is defined by the parent function $pa(x) : \mathcal{N} \to \mathcal{N}$ which denotes the parent of the node x in the hierarchy. The generative model for the H-vMFmix is given by,

$$\begin{aligned}
&\pi \sim \text{Dirichlet}(.|\alpha)\\
&\mu_n \sim \text{vMF}(.|\mu_{pa(k)}, \kappa_{pa(k)})\ n \in \mathcal{N}\\
&\kappa_n \sim \text{lgNorm}(.|m,\sigma^2)\ n \in \mathcal{N}\\
&z_i \sim \text{Categorical}(.|\pi), z_i \in \{\text{Leaf nodes of hierarchy }\}\\
&x_i \sim \text{vMF}(.|\mu_{z_i}, \kappa_{z_i}) \quad i = 1, 2..N
\end{aligned}$$

The user specified parameters are $\{\mu_0, C_0\}$ - parameter of the root-node and $\{m, \sigma^2, \alpha\}$. Each node n is equipped with a vMF distribution with parameters μ_n, κ_n. The mean parameters of the siblings nodes are drawn from a common prior defined by their parent vMF distribution. The concentration parameters for all the nodes are commonly drawn from a log-normal distribution with mean m and variance σ^2. The instance x_i is drawn from the one of the leaf-nodes z_i (Note that the data $\mathbf{X}$ resides on the leaf-nodes). One can also formulate slightly different models for e.g. by letting all the sibling nodes share the same concentration parameter; or by forming a hierarchy of concentration parameters etc - we leave such models for future research. By drawing the parameters of siblings from a common parent node, we

Table 1. Dataset Statistics

Dataset	#Instances	#Training	#Testing	#True Clusters	#Features
TDT4	622	311	311	34	8895
TDT5	6366	3183	3183	126	20733
CNAE	1079	539	540	9	1079
K9	2340	1170	1170	20	21839
NEWS20	18744	11269	7505	20	53975
NIPS	2483	1241	1242	-	14036

are enforcing the constraint that nodes which are closer to each other in the hierarchy share similar parameters. Our hope is that this would enable data to be organized into the appropriate levels of granularity as defined by the hierarchy. For inference, we can develop a similar variational inference methods with Empirical Bayes step to estimate the prior parameters. The details are presented in the supplementary material.

3.3. Temporal vMF mixtures (T-vMFmix)

Sometimes, a data collection can evolve over time. In such cases, it will be useful to develop models that can capture the evolution of clusters over time. We present Temporal vMF mixture (T-vMFmix), a state-space model based on the vMF distribution where the parameters of a cluster at a given time point have evolved from the previous time point. Given data across T time-points, $\mathbf{X} = \{\{x_{t,i}\}_{i=1}^{N_t}\}_{t=1}^{T}$ and a fixed number of clusters K, the generative model is given by,

$$\pi \sim \text{Dirichlet}(.|\alpha), \quad \kappa_k \sim \text{logNormal}(.|m, \sigma^2) \quad k = 1, 2..K$$
$$\mu_{1,k} \sim \text{vMF}(.|\mu_{-1}, C_0)\ k = 1, 2, 3..K$$
$$\mu_{t,k} \sim \text{vMF}(.|\mu_{t-1,k}, C_0)\ \ t = 2...T,\ k = 1, 2, 3..K$$
$$z_{t,i} \sim \text{Mult}(.|\pi) \quad t = 1, 2...T; i = 1, 2, ..N_t,$$
$$x_{t,i} \sim \text{vMF}(.|\mu_{z_{t,i}}, \kappa_{z_{t,i}}) \quad t = 1, 2..T;\ i = 1, 2, ...N_t$$

The prior parameters are $\{\mu_{-1}, C_0, \alpha, m, \sigma^2\}$. The cluster-specific concentration parameters κ_k's are commonly drawn from a log-Normal distribution with mean m and variance σ^2. The mean parameters of the clusters at time t are drawn from a vMF distribution centered around the previous time $t-1$ with a concentration C_0. This time-evolution of the cluster parameters introduces a layer of flexibility and enables T-vMFmix to accommodate changes in the mean parameter within a given cluster. The C_0 parameter controls the sharpness of time-evolution; having a large value of C_0 ensures that cluster parameters are more or less the same over time whereas a low value of C_0 enables the cluster parameter to fluctuate wildly between adjacent time points. Note that it is also possible to incorporate time-evolution of the mixing distribution i.e. Dirichlet$(.|\alpha)$ (similar to (Blei & Lafferty, 2006a))

$$\eta_t \sim \mathcal{N}(.|\eta_{t-1}, \Sigma^{-1}), \qquad \alpha_{t,k} = \exp(\eta_{t,k}) / \sum_{j=1}^{K} \exp(\eta_{t,j})$$
$$\pi_t \sim \text{Dirichlet}(.|\alpha_t)$$

For inference, we develop a mean-field variational approach with Empirical Bayes step for estimating the prior parameters[1] . The prior parameter C_0 in some sense acts as a regularization term forcing the parameters of the next time-step to be similar to the previous one. We recommend setting the parameter manually than directly learning it from data[1] .

4. Empirical Evaluation

Throughout our experiments we used several popular benchmark datasets (Table 1) - TDT-{4,5} (Allan et al., 1998) , CNAE [2] , K9 [3], NEWS20 [4] , and NIPS (Globerson et al., 2007). All datasets (except NIPS) have associated class-labels and are single-labeled i.e. each instance is assigned to exactly one class-label. For TDT-{4,5}, we used only those subset of documents for which relevance judgements were available.

4.1. Metrics and Baselines

First and the most natural question that we would like to answer is 'are vMF mixtures any better than standard Gaussian or Multinomial mixtures?'. Generally, comparing two clustering models is in itself a hard problem. Conclusive comparisons often involve detailed user-studies (Boyd-Graber et al., 2009) which are time-consuming and not always feasible. Therefore likelihood based comparisons have been commonly used as an alternative (Blei & Lafferty, 2006a),(Blei & Lafferty, 2006b),(Teh et al., 2006). Likelihood measures the predictive power of the model on unseen data i.e. the generalization ability of the model - a metric widely used for model selection.

However, since the support of the models are different - vMF models are defined on unit-spheres, Multinomial models are defined for non-negative integers etc, we cannot use likelihood on held-out test set to compare vMF and other non-vMF models (numerically, the likelihood of the vMF models is around 5 orders of magnitude larger than Multinomial or Gaussian mixtures). To address this issue, we compare vMF and non-vMF clustering models based on how well they are able to recover the ground-truth clusters - the human assigned class-labels are assumed to be the true ground truth clusters. We use six widely used eval-

[2] http://archive.ics.uci.edu/ml/datasets/CNAE-9

[3] http://www-users.cs.umn.edu/boley/ftp/PDDPdata/

[4] http://people.csail.mit.edu/jrennie/20Newsgroups/

Table 2. Comparison of vMFmix vs other clustering models models with 30 clusters using NMI and ARI metrics[1]. Each result is averaged over 10 different starting values for the algorithms. Bold face numbers indicate best performing method.
The results of the significance tests against B-vMFmix are denoted by a * for significance at 5% level, † for significance at 1% level.

Dataset	TDT4		TDT5		CNAE		K9		NEWS20	
Method/Metric	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI
B-vMFmix	**.900**	**.799**	**.860**	**.710**	**.748**	**.669**	**.551**	**.352**	**.567**	**.397**
vMFmix	.880†	.729†	.851*	.676†	.650†	.426†	.543*	.350	.565	.386†
K-means	.842†	.675†	.827†	.591†	.605†	.311†	.487†	.264†	.492†	.217†
Gaussian Mixtures	.857†	.693†	.833†	.617†	.620†	.357†	.482†	.293†	.518†	.267†
Multinomial Mixtures	.720†	.467†	.796†	.580†	.726†	.528†	.487†	.321†	.454†	.200†
LDA	.841†	.692†	.776†	.501†	.320†	.169†	.269†	.110†	.190†	.091†

Table 3. Average Likelihood on held-out test-set of B-vMFmix and vMFmix with 20, 30 clusters [1]. Each result is averaged over 10 different starting values for the algorithms. Bold face numbers indicate best performing method.

	Dataset	TDT4	TDT5	CNAE	K9	NEWS20	NIPS
#Clusters	Method						
20	B-vMFmix	**8628821.5**	**235115133.1**	**942719.5**	**91683824.8**	**1637772501.4**	**58811701.0**
	vMFmix	8387289.5	232542058.5	936675.9	91476580.9	1637630340.0	58671832.3
30	B-vMFmix	**8638797.7**	**235145615.9**	**944888.3**	**91694902.0**	**1638015418.8**	**58825835.7**
	vMFmix	8300513.7	231946214.7	939730.9	91376189.0	1637866392.5	58600285.2

uation metrics for this comparison - Normalized Mutual Information (NMI) , Mutual Information (MI) , Rand Index (RI) , Adjusted Rand Index (ARI) , Purity and Macro-F1 (ma-F1). The definitions of the metrics can be found in (Banerjee et al., 2006), (Steinley, 2004) and (Manning et al., 2008). We compare the following 5 clustering methods,

- **B-vMFmix**: Our proposed Bayesian vMF mixture model that extracts flat clusters.
- **vMFmix**: A mixture of vMF distributions described in Section 2. Note that this is similar to the model developed in (Banerjee et al., 2006), except that all clusters share the same κ. Using the same κ for all clusters performed significantly better than allowing cluster-specific κ_k's - this may be due to the absence of Bayesian prior.
- **K-means (KM)**: The standard k-means with euclidean distance and hard-assignments.
- **Gaussian Mixtures (GM)**: A mixture of Gaussian distributions with the means commonly drawn from a Normal prior and a single variance parameter for all clusters - using cluster specific variances performed worse (even with an Inverse gamma prior).
- **Multinomial Mixture (MM)**: A graphical model where each instance is represented as a vector of feature counts, drawn from a mixture of K Multinomial distributions.
- **Latent Dirichlet Allocation (LDA)** (Blei et al., 2003) An admixture model where the words in an instance are drawn from an instance-specific mixture of K multinomials. Although LDA may not be most suitable, we still present the results for the sake of completeness.

We used the Tf-Idf normalized data representation for vMFmix, KM and GM, and feature counts representation (without normalization) for MM. For all the methods, every instance is assigned to the cluster with largest probability before evaluation. Also, for the sake of a fair comparison, we set the same *random* initial values of the cluster-assignment variables in all the algorithms; the results of each method is averaged over 10 different starting values.

4.2. Results

Table 2 summarizes the main results of B-vMFmix, vMFmix, KM, GM and MM in the ground-truth based evaluation on five data sets. Due to the lack space we only include the results for 30 clusters and two metrics NMI and ARI [5]. Note that no separate test set is needed for this type of evaluation. The parameters of all the probabilistic models are estimated using *MAP Inference*. Among the six methods, B-vMFmix achieves the best performance on all the datasets. In fact, both B-vMFmix as well as vMFmix show a consistently strong performance against other methods, suggesting that the clusters generated by vMF models are indeed better aligned with the ground truth than the non-vMF methods. To further validate our findings, we conducted two-sided significance tests using paired t-test between every method and B-vMFmix for both the metrics. The results of the 10 different runs are considered as observed samples. The null hypothesis is that there is no significant difference between the methods. Our experiments almost all the observed results are statistically significant. Table 3 compares the performance of the basic vMF mix-

[5] The complete set of results for varying number of clusters and metrics is given in the supplementary material

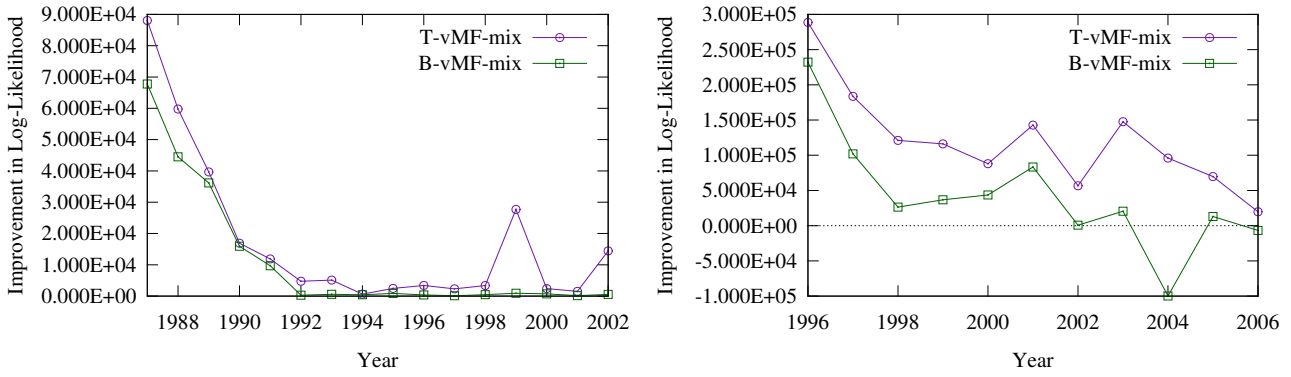

Figure 1. Relative improvement in likelihood of T-vMFmix, B-vMFmix models over vMFmix across time (with 30 clusters) - NIPS dataset (left), NYTC-elections dataset (right)

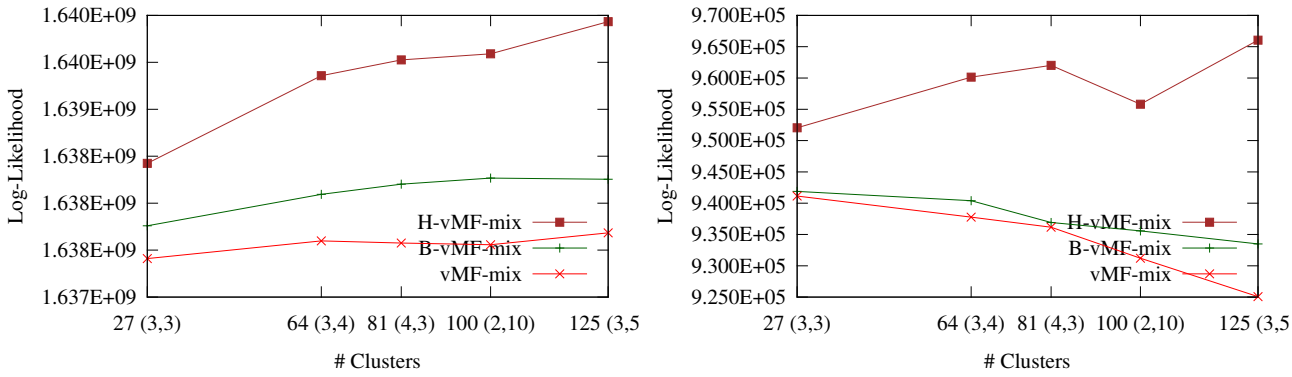

Figure 2. Average Likelihood of H-vMFmix, B-vMFmix and vMFmix for different input hierarchies. The number of clusters for B-vMFmix, vMFmix is set to number of leaf nodes in the hierarchy - NEWS20 (left) , CNAE (right)

ture (vMFmix) and the fully Bayesian vMF mixture model (B-vMFmix) on six data sets. Since the support of both models are the same, i.e. points on a unit-sphere, we can directly compare them by looking at the predictive power on a held out test-set. The B-vMFmix is trained through variational inference (using *Sampling method* for the concentration parameters) and vMFmix is trained through the EM algorithm. In our experiments, the *sampling method* performed better than the *bounding method* with an insignificant difference in running time, we therefore report all results using the sampling based method. Due to lack space we only include the results for $K = 20, 30$ clusters (refer supplementary material for full results). The results show that B-vMFmix is able to assign a higher likelihood for unseen data for all six datasets.

Figure 2 compares the performance of the hierarchical vMF mixture model (H-vMFmix) with vMFmix and B-vMFmix. We test H-vMFmix on 5 different hierarchies with varying depth of hierarchy and branch factor (branch factor is the #children under each node). For example (height h=3,branching factor b=4) is a hierarchy 3 levels deep where each internal node in the hierarchy has 4 children each. We also plot the performance of the corresponding vMFmix and B-vMFmix; the number of clusters was set equal to the number of leaf nodes in the hierarchy. Due to the lack of space we plot the results for only two datasets - NEWS20 and CNAE (refer supplementary material for full results). The superior performance of H-vMFmix in terms of predictive power strongly supports the rationale for hierarchically shrinking the model parameters.

Finally, we compare the predictive power of T-vMFmix with vMFmix and B-vMFmix on the following datasets which are temporal in nature,

1. **NIPS**: This dataset outlined in table 1 has 2483 research papers spread over a span of 17 years. The collection was divided into 17 time-points based on year of publication of each paper.
2. **NYTC-elections**: This is a collection of New York Times news articles from the period 1996-2007 which have been tagged 'elections'. The collection was divided into 11 time-points based on year of news release.

We test the models by predicting the likelihood of data in the next time-point given all the articles from the beginning

(1987)	(1991)	(1994)	(1997)	(2000)	(2003)
excitatory	activity	activity	cortical	scene	scene
inhibitor	synaptic	cortical	visual	spikes	accessible
activity	firing	cortex	neuronal	spike	quantitative
synaptic	inhibitory	stimulus	orientation	quantitative	dissimilarity
firing	excitatory	neuronal	spatial	visual	fold
cells	neuron	response	tuning	stimuli	distinguish

Figure 3. The change in vocabulary over the years for one of the topics from the NIPS dataset

Table 4. Comparison of different data representations using vMFmix and K-means using 30 clusters. Bold face numbers indicate best performing data representation.

	Dataset	NEWS20		TDT5	
Method	Representation/Metric	NMI	ARI	NMI	ARI
vMFmix	Tf-Idf normalization	**.565**	**.386**	**.851**	**.676**
	Tf normalization	.084	.027	.827	.657
K-means	Tf-Idf normalization	**.492**	**.213**	**.827**	**.591**
	Tf normalization	.089	.0284	.7993	.551

to the current time-point, similar to the setup used in (Blei & Lafferty, 2006a). For simplicity, we fix the number of clusters to 30. For ease of visualization, we plot the relative improvement in log-likelihood of the B-vMFmix and T-vMFmix models over the simple vMFmix model (this is because the log-likelihood between adjacent time-points are not comparable and fluctuate wildly). The results are plotted in Fig 1 (the supplementary material contains the results for 2 additional datasets). The results suggest that T-vMFmix by taking the temporal nature into account, is able to always assign a higher likelihood to the next time-point than B-vMFmix and vMFmix. To visualize the the cluster evolutionion over time, Fig 3 shows the progress of the mean parameter for one the the clusters in NIPS. The six words with largest weights in the mean parameter are shown over time.

4.2.1. Effect of Tf-Idf Normalization

In order to fully understand the benefits of data representation using Tf-Idf normalization, we performed controlled experiments on the two largest datasets - NEWS20 and TDT5. We compared the performance of representing the data using plain Tf normalization against Tf-Idf normalization with 'ltc' [6] based term weighting. The results of experiments using vMFmix and K-means (as reported in Table 4) show that on both the datasets, representing the data using Tf-Idf normalization gives significant performance benefits.

4.3. Experimental Settings

All our experiments were run on 48 core AMD opteron 6168 @ 1.92Ghz with 66GB RAM with full parallelization wherever possible. The main computational bottleneck in all our variational inference algorithm is the computation of $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$. The updates for both these parameters can be rewritten in terms of matrix products which can be efficiently computed using state-of-art parallel matrix multiplication tools. Refer section 10 of the supplementary material for a thorough discussion of the computational issues.

[6] http://nlp.stanford.edu/IR-book/html/htmledition/document-and-query-weighting-schemes-1.html

5. Conclusion

In this paper we proposed a suite of powerful Bayesian vMF models on unit-sphere manifolds as an alternative to the popular approaches based on multinomial or Gaussian distributions. Our models enable full Bayesian inference with vMF models in flat, hierarchical and temporal clustering. Our fast variational/sampling algorithms make the methods scalable to reasonable data volumes with high-dimensional feature spaces. The experiments provide strong empirical support for the effectiveness of our approaches - all our models outperformed strong baselines (k-means, Multinomial Mixtures and Latent Dirichlet Allocation) by a large margin on most data sets. For future work we would develop non-parametric versions of our vMF models as well as handle multi-field information.

Acknowledgements

This work is supported in part by the National Science Foundation (NSF) under grant IIS_1216282. We thank the reviewers for their excellent feedback and Guy Blelloch for sharing his computational resources.

References

Allan, James, Carbonell, Jaime G, Doddington, George, Yamron, Jonathan, and Yang, Yiming. Topic detection and tracking pilot study final report. 1998.

Banerjee, Arindam, Dhillon, Inderjit, Ghosh, Joydeep, and

Sra, Suvrit. Generative model-based clustering of directional data. In *SIGKDD*, pp. 19–28. ACM, 2003.

Banerjee, Arindam, Dhillon, Inderjit S, Ghosh, Joydeep, and Sra, Suvrit. Clustering on the unit hypersphere using von mises-fisher distributions. *JMLR*, 6(2):1345, 2006.

Bangert, Mark, Hennig, Philipp, and Oelfke, Uwe. Using an infinite von mises -fisher mixture model to cluster treatment beam directions in external radiation therapy. In *ICMLA*, 2010.

Baricz, Árpád, Ponnusamy, Saminathan, and Vuorinen, Matti. Functional inequalities for modified bessel functions. *Expositiones Mathematicae*, 29(4):399–414, 2011.

Blei, David M and Lafferty, John D. Dynamic topic models. In *ICML*, 2006a.

Blei, David M, Ng, Andrew Y, and Jordan, Michael I. Latent dirichlet allocation. *JMLR*, 2003.

Blei, David M, Griffiths, T, Jordan, M, and Tenenbaum, J. Hierarchical topic models and the nested chinese restaurant process. *NIPS*, 16:106–114, 2004.

Blei, MD and Lafferty, JD. Correlated topic models. In *NIPS*, pp. 147–155. Citeseer, 2006b.

Boyd-Graber, Jonathan, Chang, Jordan, Gerrish, Sean, Wang, Chong, and Blei, David. Reading tea leaves: How humans interpret topic models. In *NIPS*, 2009.

Cutting, Douglass R, Karger, David R, Pedersen, Jan O, and Tukey, John W. Scatter/gather: A cluster-based approach to browsing large document collections. In *SIGIR*, 1992.

Fisher, Ronald. Dispersion on a sphere. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 1953.

Globerson, A., Chechik, G., Pereira, F., and Tishby, N. Euclidean Embedding of Co-occurrence Data. *JMLR*, 8: 2265–2295, 2007.

Guttorp, Peter and Lockhart, Richard A. Finding the location of a signal: A bayesian analysis. *Journal of the American Statistical Association*, 83(402):322–330, 1988.

Hasnat, Md Abul, Alata, Olivier, and Trémeau, Alain. Hierarchical 3-d von mises-fisher mixture model. *Workshop on Divergences and Divergence Learning, ICML*, 2013.

Joachims, Thorsten. *Learning to classify text using support vector machines: Methods, theory and algorithms*. Kluwer Academic Publishers, 2002.

Jupp, PE and Mardia, KV. A unified view of the theory of directional statistics, 1975-1988. *International Statistical Review/Revue Internationale de Statistique*, 1989.

Manning, Christopher D, Raghavan, Prabhakar, and Schütze, Hinrich. *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge, 2008.

Mardia, KV and El-Atoum, SAM. Bayesian inference for the von mises-fisher distribution. *Biometrika*, 63(1): 203–206, 1976.

O'Hagan, Anthony, Forster, Jonathan, and Kendall, Maurice George. *Bayesian inference*. Arnold London, 2004.

Reisinger, Joseph, Waters, Austin, Silverthorn, Bryan, and Mooney, Raymond. Spherical topic models. In *ICML*, 2010.

Robertson, Stephen. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of documentation*, 60(5):503–520, 2004.

Salton, Gerard and McGill, Michael J. Introduction to modern information retrieval. 1986.

Steinley, Douglas. Properties of the hubert-arable adjusted rand index. *Psychological methods*, 2004.

Teh, Yee Whye, Jordan, Michael I, Beal, Matthew J, and Blei, David M. Hierarchical dirichlet processes. *JASA*, 101(476), 2006.

Yang, Jianchao, Yu, Kai, Gong, Yihong, and Huang, Thomas. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, pp. 1794–1801. IEEE, 2009.