In NIPS'09 workshop: Applications for Topic Models: Text and Beyond 2009.

Spherical Topic Models

Joseph Reisinger, Austin Waters, Bryan Silverthorn & Raymond Mooney
Department of Computer Science
The University of Texas at Austin
{joeraii, austin, bsilvert, mooney}@cs.utexas.edu

Abstract

We introduce the Spherical Admixture Model (SAM), a Bayesian topic model over arbitrary ℓ_2 normalized data. SAM models documents as points on a high-dimensional spherical manifold, and is capable of representing negative word-topic correlations and word presence/absence, unlike models with multinomial document likelihood, such as LDA. In this paper, we evaluate SAM as a topic browser, focusing on its ability to model "negative" topic features, and also as a dimensionality reduction method, using topic proportions as features for difficult classification tasks in natural language processing and computer vision.

1 Introduction

Unsupervised admixture, or *topic models*, such as Latent Dirichlet Allocation (LDA) [3] build compact descriptions of document collections in terms of a small set of semantically coherent topics. This paper introduces the Spherical Admixture Model (SAM), a topic models that represent documents using distributions on the unit hypersphere [8], modeling both word frequency and word presence/absence. Specifically, we derive a variant of LDA, replacing the multinomial document likelihood with the *von Mises-Fisher* (vMF) distribution, which has been found to often model sparse data such as text more accurately [2, 1].

SAM offers several major benefits over LDA: documents can be represented as arbitrary unit vectors; document-topic similarity is measured in terms of weighted cosine distance in the generative model; and, by exploiting the entire support of the von Mises-Fisher distribution, topics can model negative correlations between words within each topic.

2 The Spherical Admixture Model

2.1 Mixtures of von Mises-Fisher Distributions

The vMF distribution has its support on $\mathbb{S}^{d-1}$, the $(d-1)$-dimensional unit hypersphere embedded in $\mathbb{R}^d$. Its density is $f(\mathbf{x}; \boldsymbol{\mu}, \kappa) = c_d(\kappa) \exp\left(\kappa \boldsymbol{\mu}^\top \mathbf{x}\right)$, where $\boldsymbol{\mu}$ is the mean direction with $||\boldsymbol{\mu}|| = 1$, $\kappa \geq 0$ is the concentration parameter, $c_d(\kappa) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\kappa)}$, and $I_r(\cdot)$ is the modified Bessel function of the first kind and order r [8].

Motivated by the success of cosine distance in information retrieval, Banerjee *et al*. introduce the *mixture of von Mises-Fisher* distributions (movMF) [2], which treats each normalized document *tf* or *tf-idf* vector as drawn from a vMF distribution centered on a one topic mean. Although movMF outperforms mixture models with multinomial likelihood in several clustering benchmarks [2], the single-topic mixture-model assumption is too restrictive for document modeling.

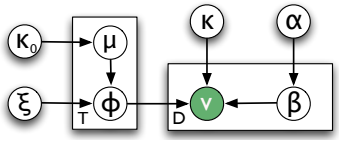
1		2		3	
(+)	(-)	(+)	(-)	(+)	(-)
model	neurons	state	visual	learning	algorithms
gaussian	theorem	policy	distance	state	space
mixture	error	time	feature	time	matrix
image	input	optimal	recognition	system	problem
noise	training	value	model	neurons	data
models	network	number	images	cells	bound
mean	learning	training	image	visual	algorithm

Figure 1: Word frequency correlations learned by SAM on the NIPS corpus. (+) shows the highest weighted words and (-) shows lowest weighted within each topic. Unlike LDA, SAM is able to represent words that are anti-correlated with the topic, rather than just unrelated.

2.2 SAM Definition

SAM is a Bayesian admixture model that operates on normalized vectors on $\mathbb{S}^{|V|-1}$. It is not therefore possible to define the admixture in terms of topic indicators for individual words in each document, as is done by LDA. SAM instead uses a *weighted directional average* to achieve the same goal. Representing the T topics as columns of matrix $\boldsymbol{\phi}$ and $\boldsymbol{\beta}_d$ as a column vector, the weighted directional average is written as: $\bar{\boldsymbol{\phi}}_d \stackrel{\text{def}}{=} \text{Avg}(\boldsymbol{\phi}, \boldsymbol{\beta}_d) = \frac{\boldsymbol{\phi}\boldsymbol{\beta}_d}{\|\boldsymbol{\phi}\boldsymbol{\beta}_d\|}$. The generative model for SAM is given by

$$
\begin{array}{lcll}
\boldsymbol{\mu}_t|\kappa_0 & \sim & \text{vMF}(\mathbf{m}, \kappa_0), & t \in T, \quad \text{(topic means)} \\
\boldsymbol{\phi}_t|\boldsymbol{\mu}_t, \xi & \sim & \text{vMF}(\boldsymbol{\mu}_t, \xi), & t \in T, \quad \text{(topics)} \\
\boldsymbol{\beta}_d|\boldsymbol{\alpha} & \sim & \text{Dirichlet}(\boldsymbol{\alpha}), & d \in D, \quad \text{(topic proportions)} \\
\bar{\boldsymbol{\phi}}_d|\boldsymbol{\phi}, \boldsymbol{\beta}_d & = & \text{Avg}(\boldsymbol{\phi}, \boldsymbol{\beta}_d), & d \in D, \quad \text{(spherical average)} \\
\mathbf{v}_d|\bar{\boldsymbol{\phi}}_d, \kappa & \sim & \text{vMF}\left(\bar{\boldsymbol{\phi}}_d, \kappa\right), & d \in D, \quad \text{(documents)}
\end{array}
$$



where $\boldsymbol{\mu}$ is the corpus mean direction, ξ controls the concentration of topics around $\boldsymbol{\mu}$, the elements of $\boldsymbol{\beta}_d$ are the mixing proportions for document d, $\boldsymbol{\phi}_t$ is the mean of topic t, and $\mathbf{v}_d$ is the observed vector for document d.

Each topic $\boldsymbol{\phi}_t$ is vector on the unit hypersphere $\mathbb{S}^{|V|-1}$. Negative entries in a topic mean vector reduce the frequency of corresponding words in the resulting mean. This expressive power does not exist in admixture models with multinomial likelihood, and the empirical results in Section 3 demonstrate that this flexibility captures useful structure in real data.

2.3 Variational Approximation

We employ a *variational mean-field* method to perform approximate inference on SAM [7]. The posterior is approximated as the factored distribution $q(\boldsymbol{\phi}|\tilde{\boldsymbol{\mu}}, \xi)q(\boldsymbol{\beta}|\tilde{\boldsymbol{\alpha}})q(\boldsymbol{\mu}|\tilde{m}, \kappa_0)$ and the factors are assumed to have the parametric forms $q(\phi_t) = \text{vMF}(\phi_t|\tilde{\boldsymbol{\mu}}, \xi)$, $q(\beta_d) = \text{Dir}(\beta_d|\tilde{\boldsymbol{\alpha}})$, and $q(\mu_t) = \text{vMF}(\mu_t|\tilde{m}_t, \kappa_0)$. Here, $\tilde{\boldsymbol{\mu}}$, $\tilde{m}$, and $\tilde{\boldsymbol{\alpha}}$ are the free variational parameters. Given this factorization, it can be shown that a lower bound on the log likelihood is given by the expression

$$
\begin{aligned}
L(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\alpha}}, \tilde{m}) &= \mathbb{E}_q[\log p(\mathbf{v}, \boldsymbol{\phi}, \boldsymbol{\beta}, \boldsymbol{\mu})] - \mathbb{E}_q[\log q(\boldsymbol{\phi}, \boldsymbol{\beta}, \boldsymbol{\mu}; \tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\phi}}, \tilde{m})] \qquad (1) \\
&= \mathbb{E}_q[\log p(\mathbf{v}|\boldsymbol{\phi}, \boldsymbol{\beta})] + \mathbb{E}_q[\log p(\boldsymbol{\phi}|\boldsymbol{\mu}, \boldsymbol{\xi})] + \mathbb{E}_q[\log p(\boldsymbol{\beta})] + \mathbb{E}_q[\log p(\boldsymbol{\mu})] \\
&\quad -\mathbb{E}_q[\log q(\boldsymbol{\phi}|\tilde{\boldsymbol{\mu}}, \xi)] - \mathbb{E}_q[\log q(\boldsymbol{\beta}|\tilde{\alpha})] - \mathbb{E}_q[\log q(\boldsymbol{\mu}|\tilde{m}, \kappa_0)]
\end{aligned}
$$

In the variational EM procedure, we use gradient ascent to update the variational topic means $\tilde{\mu}$ and per-document topic proportions $\tilde{\alpha}_d$. For convenience, we define $\tilde{\alpha}_{d,0} = \sum_{j=1}^{k} \tilde{\alpha}_{d,j}$ and $\rho_d = \mathbb{E}_q[\text{Avg}(\boldsymbol{\phi}, \beta_d)]^\top \mathbf{v}_d$, where $d \in \{1 \dots D\}$ ranges over the documents. Taking gradients of eq. (1) with respect to the variational parameters, we have

$$
\begin{aligned}
\frac{dL}{d\tilde{\alpha}_{d,i}} &= \kappa \left(\frac{d}{d\tilde{\alpha}_{d,i}} \rho_d \right) + \Psi'(\tilde{\alpha}_{d,0})(\tilde{\alpha}_{d,0} - \alpha_0) - \Psi'(\tilde{\alpha}_{d,i})(\tilde{\alpha}_{d,i} - \alpha_i) \\
\nabla_{\tilde{\mu}_t} L &= A_V(\xi) A_V(\kappa_0) \xi \tilde{m}_t + \kappa \sum_{d=1}^{D} \nabla_{\tilde{\mu}_t} (\rho_d)
\end{aligned}
$$

Model	Accuracy (%)		
	different	similar	same
Bag-of-words	75.1 ± 3.2	68.7 ± 2.1	57.2 ± 2.6
LDA	75.1 ± 2.4	67.5 ± 1.3	52.2 ± 5.7
movMF	65.3 ± 0.8	62.0 ± 3.2	49.1 ± 0.4
MH SAM $[\mathbb{S}]$	80.9 ± 1.1	69.5 ± 1.3	62.2 ± 1.6
VEM SAM $[\mathbb{S}_+]$	72.0 ± 1.3	69.6 ± 0.9	62.2 ± 0.5
VEM SAM $[\mathbb{S}]$	80.2 ± 1.0	72.4 ± 0.9	65.5 ± 0.4

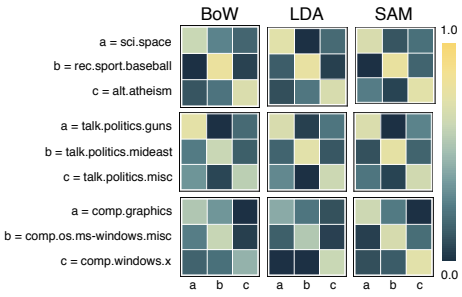


Figure 2: (**left**) Classification performance and standard deviations on the three **news-20** tasks. SAM topic proportions make better features, particularly in more difficult domains. SAM is implemented using both adaptive rejection sampling (MH) and variational EM (VEM). (**left**) Confusion matrices for each feature set (bag-of-words, LDA, and VEM SAM) and task.

$A_D(c)$ denotes the *mean resultant length* of a vMF distribution of dimension D with concentration c. This quantity can be approximated stably in high dimension using the approach of Abramowitz and Stegum, cf. [5]. Because ρ_d itself does not have a simple closed form, we use the approximations

$$\mathbb{E}[\text{Avg}(\boldsymbol{\phi}, \beta_d)] \approx \mathbb{E}[\boldsymbol{\phi}\beta_d]\ \mathbb{E}\left[\sqrt{\beta^\top \boldsymbol{\phi}^\top \boldsymbol{\phi}\beta}\right]^{-1} \approx \mathbb{E}[\boldsymbol{\phi}\beta_d]\ \mathbb{E}[\beta^\top \boldsymbol{\phi}^\top \boldsymbol{\phi}\beta]^{-1/2} \tag{2}$$

Closed-form expressions for these expectations can be derived from well-known properties of the Dirichlet and vMF distributions.

3 Experiments

SAM's performance is evaluated empirically on Usenet post and natural scene image classification. Four models are compared: (i) **LDA**[1]; (ii) **movMF**, the mixture of von-Mises Fisher clustering algorithm with soft assignments [2]; (iii) **SAM** $[\mathbb{S}]$, SAM with topic means in $\mathbb{S}^{|V|-1}$; and (iv) **SAM** $[\mathbb{S}_+]$, SAM with topics and spherical combinations restricted to the positive orthant of the unit hypersphere, ablating the ability to model negative correlations between word and topics. In all models we use *only* term-frequency information (counts for LDA and normalized counts for SAM), despite SAM's ability to handle e.g., *tf-idf* vectors.

Quantitative evaluation measures common in clustering, such as normalized mutual information [1], are inappropriate in topic modeling because inferred topics do not necessarily correspond to pure partitions of the document collection. Furthermore, SAM and LDA cannot be compared directly in terms of perplexity, as they inhabit fundamentally different base measures. Instead, we focus our evaluation on qualitative corpus exploration (highest weighted positive and negative features in the NIPS corpus; figure 1) and classifier accuracy, comparing topic proportion features derived from SAM and LDA to standard bag-of-words features [3].

3.1 CMU 20 Newsgroups

This section evaluates using the learned topic proportions $\boldsymbol{\beta}$ as features for classification in the CMU **news-20** data set. Each post is treated as a document and labeled with its group. Following Banerjee and Basu [1], three subsets, with posts on different, similar, and the same subject, are used.

Figure 2 summarizes the results. SAM finds better features than the other models, and more meaningful distinctions between finer-grained topics (20% reduction in relative error for **different**; 27.8% reduction for **same**). The differences between SAM $[\mathbb{S}_+]$ and SAM $[\mathbb{S}]$ highlight the utility of allowing topics to encode negative correlations between terms and topics, and the differences between SAM $[\mathbb{S}]$ and LDA suggest that generative models based on vMF distributions are a better match for text than multinomial models.

Model Accuracy (%)	different	similar	outdoor	indoor	all
LDA	79.3 ± 1.7	68.5 ± 2.5	60.9 ± 2.9	43.6 ± 2.8	43.4 ± 1.9
SAM $[\mathbb{S}]$	85 ± 3.5	74.4 ± 2.1	68.4 ± 1.4	50.2 ± 2.2	50.3 ± 1.8

Figure 3: Classification accuracy for **13-scene** with $|V| = 200$.

3.2 13 Natural Scene Categories

We divide the image recognition task of [6] into separate 4-class problems: **13-scene-different**, **13-scene-similar**, **13-scene-outdoor**, and **13-scene-indoor**, ordered by their classification difficulty. We follow Fei-Fei and Perona's preprocessing steps, representing each image with counts of its *visual words*. Note that this task differs fundamentally from the textual tasks in terms of sparsity: most visual words tend to occur in most scenes. Thus the comparative results obtained in this domain can be considered an ablation of SAM's ability to model the lack of features.

Using 200 visual words, we find that SAM outperforms LDA across all scene recognition tasks (Table 3.2); 10% of the data is used for training a Logistic Regression classifier. As more training data is used, the performance of LDA and SAM converge.

4 Discussion

SAM opens up a new class of admixture models based on spherical distributions. Unlike previous spherical mixtures, SAM is a fully Bayesian admixture model that allows multiple component vMFs to explain different aspects of the data; unlike previous admixture models, SAM uses directional distributions that are parameterized by cosine distance and is therefore capable of modeling negative correlations between features as well as word absence/presence.

Using classification performance for evaluation, SAM was found to produce more relevant topic features than the movMF spherical mixture model and LDA, particularly on data where fine-grained topic distinctions are important. Two properties of SAM—its use of directional distributions, and ability to model negative correlations—were found to contribute to its performance. It is an important step in improving generative topic models.

Acknowledgements

We would like to thank Arindam Banerjee for early discussions and the movMF implementation and Kristen Grauman for input on the vision domain. This work was partially supported by a Google Research Award and an NSF Graduate Research Fellowship to the first author.

References

[1] A. Banerjee and S. Basu. Topic models over text streams: A study of batch and online unsupervised learning. In *SDM*. SIAM, 2007.

[2] A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra. Clustering on the unit hypersphere using von Mises-Fisher distributions. *Journal of Machine Learning Research*, 6:1345–1382, 2005.

[3] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 2003.

[4] H. Daumé III. HBC: Hierarchical Bayes compiler, 2009. `http://hal3.name/HBC`.

[5] C. Elkan. Clustering documents with an exponential-family approximation of the Dirichlet compound multinomial distribution. In *Proceedings of the Twenty-Third International Conference on Machine Learning (ICML '06)*, pages 289–296, New York, NY, USA, 2006. ACM.

[6] L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. *CVPR*, pages 524–531, 2005.

[7] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

[8] K. V. Mardia and P. E. Jupp. *Directional Statistics*. John Wiley, 2000.

[1] Collapsed Gibbs sampler with full hyperparameter estimation, implemented in HBC [4]