In Proceedings of the 27th International Conference on Machine Learning (ICML 2010) 2010.

Spherical Topic Models

Joseph Reisinger JOERAII@CS.UTEXAS.EDU
Austin Waters AUSTIN@CS.UTEXAS.EDU
Bryan Silverthorn BSILVERT@CS.UTEXAS.EDU
Raymond J. Mooney MOONEY@CS.UTEXAS.EDU
Department of Computer Science, 1 University Station C0500, University of Texas, Austin, TX 78712

Abstract

We introduce the Spherical Admixture Model (SAM), a Bayesian topic model for arbitrary ℓ_2 normalized data. SAM maintains the same hierarchical structure as Latent Dirichlet Allocation (LDA), but models documents as points on a high-dimensional spherical manifold, allowing a natural likelihood parameterization in terms of cosine distance. Furthermore, SAM can model word absence/presence at the document level, and unlike previous models can assign explicit negative weight to topic terms. Performance is evaluated empirically, both through human ratings of topic quality and through diverse classification tasks from natural language processing and computer vision. In these experiments, SAM consistently outperforms existing models.

1. Introduction

Unsupervised admixture, or *topic models*, such as Latent Dirichlet Allocation (LDA; Blei et al., 2003) build compact descriptions of document collections in terms of a small set of semantically coherent topics. Individual documents are decomposed as mixtures over the topic set, with each document maintaining its own set of mixture parameters. Unlike standard mixture models, where each mixture component (topic) is responsible for explaining all of the variation in a subset of the corpus, admixture models allow mixture components to share responsibility, often resulting in a significantly better generative model of the data.

LDA is a fully Bayesian extension of Latent Semantic Analysis (LSA; Hofmann, 1999), representing documents directly as word counts, modeling them implicitly as weighted averages on the multinomial simplex. Unlike similar methods such as the Aspect-Bernoulli Model (ABM; Bingham et al., 2009), LDA is unable to directly model the *absence* of words, only their presence, as document likelihood is based on the multinomial distribution. In contrast, a multivariate Bernoulli likelihood can model word absence, e.g., distinguishing between "true absences" and "missing presences," but is incapable of modeling frequency (McCallum & Nigam, 1998). In this paper, we introduce the Spherical Admixture Model (SAM), a class of topic models that represent data using directional distributions on the unit hypersphere (Mardia & Jupp, 2000), modeling both word frequency and word presence/absence. Specifically, we derive an admixture model with a *von Mises-Fisher* likelihood, which has been demonstrated to model sparse data such as text more accurately than corresponding multinomial models (Banerjee et al., 2005; Zhong & Ghosh, 2005).

SAM offers several other major benefits over LDA. First, documents are modeled as arbitrary unit vectors, allowing for richer feature representations (e.g. *tf-idf* or *t-test* feature weighting). Second, document-topic similarity is measured in terms of weighted cosine distance, defining similarity in terms of the *directions* of their word frequency vectors, which provides significant robustness to feature noise. Third, by exploiting the entire support of the von Mises-Fisher distribution, topics are able to assign *negative* weights to words: for example, seeing "neurons" in a NIPS abstract might imply that we should expect to see "SVM" significantly *less* often on average. Finally, despite its increased complexity, SAM admits an efficient variational Bayesian inference procedure.

We evaluate SAM along two dimensions: (1) as a topic model using the human evaluation methods described in Chang et al. (2009) and (2) as a dimensionality reduction method on three real-world tasks, classifying Usenet posts from the CMU **news-20** collection, detecting thematic shifts in the Italian text of Niccolò Machiavelli's *Il Principe*, and classifying natural scenes in the **13-scene** database (Fei-Fei & Perona, 2005). We find that SAM significantly outperforms LDA both in terms of human interpretability of topics and in its ability to capture salient semantic variation across all three corpora.

This paper is divided into six sections: Section 2 covers re-

Appearing in *Proceedings of the 27^{th} International Conference on Machine Learning*, Haifa, Israel, 2010.

lated models, Section 3 introduces SAM and its variational approximation, Section 4 gives experimental results, Section 5 discusses future work, and Section 6 concludes.

2. Background

2.1. Spherical Mixture Models

In this section and those subsequent, we adopt the terminology of topic models: data consists of D individual "documents," where each document is a set of "words" from a known vocabulary V. Probabilistic models of text have been built around the multinomial distribution and the von Mises-Fisher (vMF) distribution (Mardia & Jupp, 2000), and these distributions are associated with different representations of textual data.

The multinomial distribution is the most straightforward model of discrete data. It assigns probabilities to integer vectors of event counts, which, for textual data, are typically raw non-normalized word counts in $\mathbb{N}^{|V|}$.

The vMF distribution instead has its support on $\mathbb{S}^{d-1}$, the unit $(d-1)$-sphere 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, $c_d(\kappa) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\kappa)}$ is a normalization factor, and $I_r(\cdot)$ is the modified Bessel function of the first kind and order r. vMF distributions have been used to model *tf* and *tf-idf* representations of text documents ℓ_2-normalized onto $\mathbb{S}^{|V|-1}$ (Banerjee et al., 2005), and other directional data (Mardia & Jupp, 2000).

The vMF distribution can be thought of as an $\mathbb{S}^{d-1}$ analog of the multivariate Gaussian with spherical covariance, parameterized by *cosine distance* rather than Euclidean distance. Cosine distance computes similarity in terms of the *directions* of ℓ_2-normalized feature vectors and corresponds to the normalized correlation coefficient. Evidence suggests that this type of directional measure is often superior to Euclidean distance in high dimensions (Manning & Schütze, 2000; Zhong & Ghosh, 2005).

The vMF is sensitive to word absence in a way that the multinomial is not: For example, let $\theta = [1/3, 1/3, 1/3]$ be a multinomial parameter vector. Documents $D_1 = [1, 1, 1]$ and $D_2 = [3, 0, 0]$ are equiprobable under Mult(θ); in fact, all three-word documents have equal probability in this example. However, because D_1 and D_2 have different cosine distances from θ, the documents have different densities under a corresponding vMF. Although this is a simple example, it represents a larger issue with the multinomial.

Inspired by the role of cosine distance in information retrieval, Banerjee et al. (2005) introduced the *mixture of von Mises-Fisher* distributions (movMF). The movMF model treats each normalized document *tf* or *tf-idf* vector as drawn from a single vMF distribution centered on one cluster mean, selected by a common multinomial distribution. The likelihood of a document d is $f(d|\Theta) = \sum_{t=1}^{T} \alpha_t \text{vMF}(d|\boldsymbol{\mu}_t, \kappa_t)$, where $\Theta = (\boldsymbol{\alpha}, \boldsymbol{\mu}_1, \kappa_1, \ldots, \boldsymbol{\mu}_T, \kappa_T)$, $\boldsymbol{\alpha}$ is the parameterization of the multinomial over topics, and each $\boldsymbol{\mu}$ and κ parameterizes the vMF distribution for a cluster. movMF generalizes classic clustering methods parameterized by cosine distance: when each cluster concentration κ is taken to infinity, movMF becomes equivalent to spherical k-means (Banerjee et al., 2005).

The movMF model successfully integrates a directional measure of similarity into a probabilistic setting, but its mixture model assumption—that each document is associated with a single cluster—is fundamentally restrictive.

2.2. Latent Dirichlet Allocation

Admixture models such as LDA relax the assumption that each document is drawn exclusively from a single mixture component; instead, documents are drawn from a weighted average over all components. In LDA, this weighted average is implicit in the model structure (Blei et al., 2003). Each document $\mathbf{w}_d$ maintains a separate multinomial distribution $\boldsymbol{\theta}_d$ over topics $\boldsymbol{\phi}$. For each word token $w_{i,d}$ a *topic index* $z_{i,d}$ is drawn from $\boldsymbol{\theta}_d$ and then $w_{i,d}$ is drawn from the corresponding topic multinomial $\boldsymbol{\phi}_{z_{i,d}}$. The generative model is given by:

$$
\begin{array}{llll}
\boldsymbol{\theta}_d|\boldsymbol{\alpha} & \sim \text{Dirichlet}(\boldsymbol{\alpha}), & d \in D, & \text{(topic proportions)} \\
\boldsymbol{\phi}_t|\boldsymbol{\beta} & \sim \text{Dirichlet}(\boldsymbol{\beta}), & t \in T, & \text{(topics)} \\
z_{i,d}|\boldsymbol{\theta}_d & \sim \text{Mult}(\boldsymbol{\theta}_d), & i \in |\mathbf{w}_d|, & \text{(topic indicators)} \\
w_{i,d}|\boldsymbol{\phi}_{z_{i,d}} & \sim \text{Mult}(\boldsymbol{\phi}_{z_{i,d}}), & i \in |\mathbf{w}_d|, & \text{(words)}
\end{array}
$$

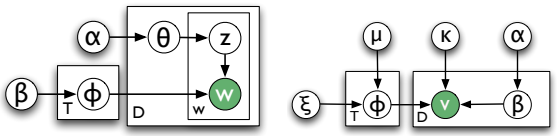
where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are hyperparameters smoothing the per-document topic distributions and per-topic word distributions respectively. As an admixture model, LDA relaxes the assumption that each document is drawn exclusively from a single mixture component. This flexibility allows it to uncover more fine-grained document structure than traditional mixture models. Furthermore, by marginalizing the topic indicators $z_{i,d}$ out of the model, LDA can be shown to draw each document from a multinomial whose parameters are a weighted average of the topics. The same intuition will be used to develop SAM as a weighted average over ℓ_2-normalized topic means.

3. The Spherical Admixture Model

The Spherical Admixture Model (SAM), developed below, is a topic model for arbitrary ℓ_2-normalized data. Like the movMF model, it is built on a probability distribution parameterized by cosine distance and capable of taking into account the absence of words; like LDA, it decomposes in-

Table 1. Top positive and negative term weights learned by SAM on the NIPS corpus and Wikipedia. $(+)$ 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. These correlations appear meaningful: In the case of Wikipedia, negatively weighted words are often related but not directly relevant to the topic.

NIPS				Wikipedia					
$(+)$	$(-)$	$(+)$	$(-)$	$(+)$	$(-)$	$(+)$	$(-)$	$(+)$	$(-)$
svm	network	genetic	mlp	navy	airport	album	opera	india	germany
kernel	experts	fitness	tree	ships	airlines	label	actor	temple	borough
margin	units	crossover	matrix	naval	flights	singles	films	dynasty	england
machines	target	population	discriminant	submarines	bus	chart	players	indian	france
support	clusters	search	lemma	aircraft	satellites	song	conservatory	khan	parish



(a) LDA (b) SAM

Figure 1. Graphical models for LDA and SAM.

dividual documents over multiple topics.

3.1. Model Definition

SAM is a Bayesian admixture model of normalized vectors on $\mathbb{S}^{|V|-1}$. It is therefore not 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 combine topics. To draw a collection of documents in SAM,

1. Draw a set of T topics $\boldsymbol{\phi}$ on the unit hypersphere;
2. For each document d, draw topic weights $\boldsymbol{\theta}_d$ from a Dirichlet with hyperparameter $\boldsymbol{\alpha}$;
3. Draw a document vector $\mathbf{v}_d$ from a vMF with mean $\bar{\boldsymbol{\phi}}_d = \text{Avg}(\boldsymbol{\phi}, \boldsymbol{\theta}_d)$ and concentration κ.

Representing the T topics as columns of matrix $\boldsymbol{\phi}$, and $\boldsymbol{\theta}_d$ as a column vector, the weighted directional average is written as $\bar{\boldsymbol{\phi}}_d \stackrel{\text{def}}{=} \text{Avg}(\boldsymbol{\phi}, \boldsymbol{\theta}_d) = \frac{\boldsymbol{\phi}\boldsymbol{\theta}_d}{\|\boldsymbol{\phi}\boldsymbol{\theta}_d\|}$.[1] The complete generative model for SAM is given by:

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

where $\boldsymbol{\mu}$ is the corpus mean direction, ξ controls the concentration of topics around $\boldsymbol{\mu}$, the elements of $\boldsymbol{\theta}_d$ 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 an arbitrary vector on the unit hypersphere $\mathbb{S}^{|V|-1}$. Topics are equally capable of making words more or less likely: positive entries in a topic mean vector increase the weights of corresponding words in each per-document mean, and negative entries reduce those weights (see Table 1 for an example from the NIPS and Wikipedia datasets). The empirical results in Section 4 demonstrate that this flexibility can help capture useful structure in data.

3.2. Variational Inference

Given a document corpus, we are interested in inferring the posterior distribution of the topic means, topics, and per-document topic proportions: $p(\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\mu}|\mathbf{v}, \xi, \mathbf{m}, \alpha, \kappa_0, \kappa)$. Computing the exact posterior is intractable, thus we develop an efficient *variational mean-field* method to perform approximate inference in SAM. In variational mean-field methods, the true posterior is approximated by another distribution with a simpler, factored parametric form. An EM procedure is used to update the parameters of the approximate posterior and the model hyperparameters so that a lower bound on the log likelihood increases with each iteration (Jordan et al., 1999).

We approximate the posterior as the factored distribution

$$q(\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\mu}|\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\alpha}}, \xi) = q(\boldsymbol{\phi}|\tilde{\boldsymbol{\mu}}, \xi) q(\boldsymbol{\theta}|\tilde{\boldsymbol{\alpha}}) q(\boldsymbol{\mu}|\tilde{\boldsymbol{m}}, \kappa_0),$$

and assume the factors have the parametric forms $q(\phi_t)$=vMF$(\phi_t|\tilde{\boldsymbol{\mu}}, \xi)$, $q(\theta_d)$=Dir$(\theta_d|\tilde{\boldsymbol{\alpha}})$, and $q(\mu_t)$=vMF$(\mu_t|\tilde{\boldsymbol{m}}_t, \kappa_0)$. Here, $\tilde{\boldsymbol{\mu}}$, $\tilde{\boldsymbol{m}}$, and $\tilde{\boldsymbol{\alpha}}$ are the free variational parameters. Given this factorization, a lower bound $L(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{m}})$ on the log likelihood is given by:

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

[1] (**Buss-Fillmore spherical average**) This procedure does not yield the vector that minimizes the weighted sum of geodesic distances to the mean. Buss & Fillmore (2001) introduce the spherical average $\text{Avg}_{BF}(\boldsymbol{\phi}, \boldsymbol{\theta}) \stackrel{\text{def}}{=} \arg\min_q \sum_i \boldsymbol{\theta}_i d_{\mathbb{S}}(\boldsymbol{\phi}_i, q)$, where $d_{\mathbb{S}}(p, q)$ is the geodesic distance between $p, q \in \mathbb{S}^d$. This definition is desirable, but must be computed iteratively.

Note that the expectations in this expression are taken over the *variational* distribution q.

The E step of variational EM consists of optimizing the expression for the log-likelihood lower bound (1) with respect to each of the free parameters $\tilde{\alpha}_{d,i}$, $\tilde{\boldsymbol{\mu}}_t$, and $\tilde{\boldsymbol{m}}$. Similarly, in the M step, eq. (1) is optimized with respect to each of the hyperparameters ξ, m, α, κ_0, and κ. The EM procedure consists of alternating E and M steps until some suitable convergence criterion is reached.

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

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

Here Ψ is the digamma function and $A_D(c)$ 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. Elkan (2006). Because ρ_d itself does not have a closed form, we use the approximation:

$$\begin{aligned}
\mathbb{E}[\text{Avg}(\boldsymbol{\phi}, \theta_d)] &\approx \mathbb{E}[\boldsymbol{\phi}\theta_d] \, \mathbb{E}\left[\sqrt{\theta_d^\top \boldsymbol{\phi}^\top \boldsymbol{\phi} \theta_d} \right]^{-1} \quad (2) \\
&\approx \mathbb{E}[\boldsymbol{\phi}\theta_d] \, \mathbb{E}[\theta_d^\top \boldsymbol{\phi}^\top \boldsymbol{\phi} \theta_d]^{-1/2}. \quad (3)
\end{aligned}$$

The last factor is the *expected squared norm* of the random vector $\boldsymbol{\phi}\theta_d$, which we will refer to as S_d. These expectations can be computed in closed form using known properties of the Dirichlet and vMF distributions, yielding:

$$\rho_d \approx A_V(\xi) \tilde{\alpha}_{d,0}^{-1} S_d^{-1/2} (\tilde{\boldsymbol{\mu}} \tilde{\alpha}_d)^\top v_d,$$

where

$$S_d = \frac{\tilde{\alpha}_{d,0} + (1 - A_V(\xi)^2) \sum \tilde{\alpha}_{di}^2 + A_v(\xi)^2 ||\tilde{\boldsymbol{\mu}} \tilde{\alpha}_d||^2}{\tilde{\alpha}_{d,0}(\tilde{\alpha}_{d,0} + 1)}.$$

Differentiating with respect to $\tilde{\alpha}_{d,j}$ and $\tilde{\boldsymbol{\mu}}_j$, respectively, yields:

$$\frac{d\rho_d}{d\tilde{\alpha}_{d,j}} = \frac{A_V(\xi)}{\tilde{\alpha}_{d,0}} \left(\frac{\tilde{\boldsymbol{\mu}}_j - \tilde{\boldsymbol{\mu}} \tilde{\alpha}_d / \tilde{\alpha}_{d,0}}{\sqrt{S_d}} - \frac{\tilde{\boldsymbol{\mu}} \tilde{\alpha}_d}{2 S_d^{3/2}} \frac{dS_d}{d\tilde{\alpha}_{d,j}} \right)^\top v_d$$

$$\nabla_{\tilde{\boldsymbol{\mu}}_j} \rho_d = \frac{A_V(\xi)}{\tilde{\alpha}_{d,0}} \left(\frac{\tilde{\alpha}_{d,j} v_d}{\sqrt{S_d}} - \frac{(\tilde{\boldsymbol{\mu}} \tilde{\alpha}_d)^\top v_d}{2 S_d^{3/2}} \cdot \nabla_{\tilde{\boldsymbol{\mu}}_j} S_d \right)$$

The derivatives of S_d are:

$$\begin{aligned}
\frac{dS_d}{d\tilde{\alpha}_{d,j}} &= \frac{1 + 2(1 - A_v(\xi)^2)\tilde{\alpha}_{d,j} + 2A_v(\xi)^2 \tilde{\alpha}_d^\top \tilde{\boldsymbol{\mu}}^\top \tilde{\boldsymbol{\mu}}_j}{\tilde{\alpha}_{d,0}(\tilde{\alpha}_{d,0} + 1)} \\
&\quad - \frac{2\tilde{\alpha}_{d,0} + 1}{\tilde{\alpha}_{d,0}(\tilde{\alpha}_{d,0} + 1)} S_d \\
\nabla_{\tilde{\boldsymbol{\mu}}_j} S_d &= \frac{2(1 - A_V(\xi)^2)\tilde{\boldsymbol{\mu}}_j + 2A_V(\xi)^2 \tilde{\alpha}_{d,j} \tilde{\boldsymbol{\mu}} \tilde{\alpha}_d}{\tilde{\alpha}_{d,0}(\tilde{\alpha}_{d,0} + 1)}
\end{aligned}$$

Unlike the variational topics and topic proportions, the variational corpus mean $\tilde{m}$ has a simple closed-form update rule. The gradient of (1) with respect to $\tilde{\boldsymbol{m}}$ is:

$$\nabla_{\tilde{\boldsymbol{m}}} L = \kappa_0 A_V(\kappa_0) m + A_V(\xi) A_V(\kappa_0) \xi \sum_{t=1}^{T} \tilde{\boldsymbol{\mu}}_t + 2\lambda \tilde{\boldsymbol{m}},$$

where λ is a Lagrange multiplier used to enforce the constraint that $\tilde{\boldsymbol{m}}$ must have unit ℓ_2 norm. Setting the gradient to zero and solving, we attain the closed-form update rule $\tilde{\boldsymbol{m}} \propto \left(\kappa_0 \mathbf{m} + A_V(\xi) \xi \sum_{t=1}^{T} \tilde{\boldsymbol{\mu}}_t \right)$. Update rules for the model hyperparameters can be derived using a process very similar to that above.[2]

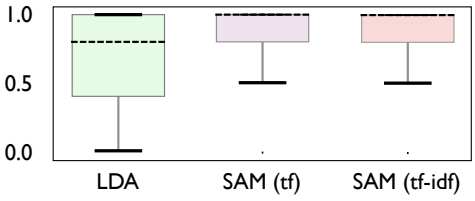
4. Experiments

We evaluate SAM along two dimensions: First, the semantic coherence and relevance of its topics is compared against LDA using the subjective methods described by Chang et al. (2009). Second, its performance as a dimensionality reduction method is evaluated on three real-world tasks: classifying Usenet posts from the CMU **news-20** collection, detecting thematic shifts in the Italian text of Niccolò Machiavelli's *Il Principe*, and classifying natural scenes in the **13-scene** database (Fei-Fei & Perona, 2005). Four models are compared:

- **LDA** – The Latent Dirichlet Allocation model, outlined in Section 2.2.[3]
- **movMF** – The mixture of von-Mises Fisher clustering with soft assignments (Banerjee et al., 2005).
- **SAM** $[\mathbb{S}]$ – SAM with topic means in $\mathbb{S}^{|V|-1}$ that can contain both negative and positive entries.
- **SAM** $[\mathbb{S}_+]$ – SAM with topics and spherical combinations restricted to the positive orthant of the unit hypersphere, ablating the model's ability to assign negative term weights in topics.

[2] A MATLAB implementation is available at `http://www.cs.utexas.edu/~austin`

[3] Collapsed Gibbs sampler with an asymmetric Dirichlet prior on the topic proportions, cf. Wallach et al. (2009); implemented in HBC (Daumé III, 2009).



SAM (easy)	1: vishnu, tamil, kerala, singh, nadu, **meteorologist** 2: oxidation, protein, potassium, **footballers**, hydrogen, symptoms
SAM (hard)	1: **saloon**, huron, burlington, county, mississippi, wl 2: tang, hong, **howe**, wu, kong, leone
LDA (easy)	1: male, mammals, **empire**, plants, species, birds 2: court, crimes, police, law, security, **jazz**
LDA (hard)	1: brother, sister, manga, anime, ride, **orchestra** 2: water, **earth**, power, energy, production, oil

Figure 2. (**top**) Boxplot showing summarizing human rater accuracy in the word-intruder task. (**bottom**) Examples of word intrusion questions that human raters found easy or difficult. Intruder words are shown in bold.

Quantitative evaluation measures common in clustering, such as normalized mutual information (Banerjee & Basu, 2007), 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.[4] Instead, we focus our evaluation on qualitative corpus exploration and classifier accuracy, comparing topic proportion features derived from SAM and LDA to standard bag-of-words features (Blei et al., 2003).

4.1. Topic Interpretability

Since SAM and LDA are incomparable in terms of perplexity, we instead evaluate the coherence and relevance of topics generated by both methods directly with human raters, adapting the procedure described by Chang et al. (2009). All experiments described in this section use Amazon's *Mechanical Turk*. Both LDA and SAM are trained on a random 10K document subset of Wikipedia.[5] Unlike Chang et al. (2009), we perform significantly less aggressive post-processing, and named entities are kept intact, making the inference task more difficult. In both experiments, each topic model is used to infer 50 topics. Responses are averaged over 8 human raters.

Topic *coherence* is evaluated using a simple word intrusion task: the top five words from a topic are shuffled and a single intruder word is added to the set, drawn from the high probability words in a different topic. The rater is then asked to identify the intruder. As the semantic coherence and distinctness from other topics increases, this task becomes easier.

Using LDA, raters were able to correctly identify the intruder words in 67.1% of cases (50 per model). Using SAM topics, raters were able to identify the intruders in 82.7% of cases with tf-idf features and 80.4% of cases with tf features (Figure 2). Both SAM results differ significantly from LDA ($p<0.05$; Student's *t-test*), indicating that SAM topics, when represented as the top weighted terms, are more semantically coherent than LDA topics.

Topic *relevance* is evaluated through a forced-choice experiment: evaluators are presented with one of 100 randomly selected articles from Wikipedia and are asked to judge which of two topics is most relevant. Topics are ranked from both models and paired together for presentation: i.e. the highest weighted topic from SAM is paired with the highest weighted topic from LDA, etc. After discarding trials with low inter-rater agreement ($\kappa < 0.4$; 47 trials), topics drawn from SAM are preferred roughly 3:2 over topics from LDA (0.616 ± 0.08), indicating that SAM topics are more relevant on average.

4.2. Classification Tasks

In this section we compare the models through their performance as dimensionality reduction methods. The topic or cluster proportions inferred by each model are evaluated as features in several multiclass classification tasks.

In all experiments in this paper, LDA is run with an asymmetric α prior and symmetric β prior, optimized using a hybrid Gibbs-EM empirical Bayes procedure. SAM uses $\kappa = 1500$, ℓ_2-normalized *tf* or *tf-idf* document representations and inference is performed with the Variational EM (VEM) procedure discussed in section 3.2. A simple Adaptive Metropolis-Hastings (MH) sampler for SAM is also evaluated, with $\alpha = \eta = 0.1$.[6] The total number of topics is fixed at 50.[7] All results reported use Logistic Regression with a ridge estimator (le Cessie & van Houwelingen, 1992) and use 10×10-fold cross-validation.

4.2.1. CMU 20 NEWSGROUPS

This first classification task is derived from the CMU **news-20** data set. Each news post is treated as a document

[4] (**Measurability**) vMF distributions are continuous, while multinomials are discrete; hence, neither perplexity nor the likelihood ratio test are applicable.

[5] (**Wikipedia dataset**) Snapshot taken on 9/29/09; wikitext markup is removed, as are articles with fewer than 100 words. The 10K document subset has a vocabulary size of 16552 unique words and a total of ∼2M tokens.

[6] (**Adaptive Metropolis-Hastings**) Proposals for ϕ_t are drawn from $\mathcal{N}(\phi_t, \text{diag}(\sigma))$ and projected onto the unit hypersphere.

[7] Accuracy increases with T, but the main results here do not change significantly for $T > 50$.

Table 2. Classification accuracy and 95% confidence intervals on the three **news-20** tasks. SAM topic proportions make better features, particularly in more semantically tight domains. Since no significant difference was found between SAM $[\mathbb{S}]$ and SAM $[\mathbb{S}_+]$, only SAM $[\mathbb{S}]$ is shown.

Model	Accuracy (%)		
	different	**similar**	**same**
Bag-of-Words (tf)	91.3 ± 0.4	85.3 ± 0.7	75.9 ± 0.6
Bag-of-Words (tf-idf)	91.7 ± 0.3	85.9 ± 0.5	77.5 ± 0.8
Topic Only			
LDA	87.8 ± 0.6	78.5 ± 2.7	66.3 ± 2.6
movMF (tf)	71.4 ± 0.3	64.5 ± 0.6	59.4 ± 0.4
movMF (tf-idf)	71.9 ± 0.3	74.2 ± 0.4	56.0 ± 0.6
SAM (tf)	88.6 ± 0.4	81.2 ± 0.4	70.5 ± 0.5
SAM (tf-idf)	93.3 ± 0.3	85.9 ± 0.3	75.0 ± 0.4
Topic + Bag-of-Words			
LDA	91.8 ± 0.4	85.7 ± 0.7	75.6 ± 0.8
movMF (tf)	91.1 ± 0.3	84.9 ± 0.5	75.8 ± 0.8
movMF (tf-idf)	91.4 ± 0.5	84.9 ± 0.5	75.3 ± 0.6
SAM (tf)	91.9 ± 0.4	86.3 ± 0.5	75.6 ± 0.6
SAM (tf-idf)	94.1 ± 0.3	88.1 ± 0.5	78.1 ± 0.6

and labeled with its newsgroup. Following Banerjee & Basu (2007), three subsets of **news-20** are used for evaluation: (1) **news-20-different**, with posts from the unrelated groups *rec.sport.baseball*, *sci.space* and *alt.atheism*; (2) **news-20-similar**, with posts from the more similar groups *rec.sport.baseball*, *talk.politics.guns* and *talk.politics.misc*; and (3) **news-20-same**, with posts from the highly related groups *comp.os.ms-windows.misc*, *comp.windows.x* and *comp.graphics*. These domains span corpora with varying degrees of subject similarity, making it possible to measure how well SAM and LDA identify meaningful topics that capture fine-grained semantic structure. Each model is evaluated based on the performance of its topic proportions as features for classification, using raw bag-of-words features as the baseline.

Table 2 summarizes the experimental results. In general, SAM finds better features than the other models, performing about as well as raw bag-of-words. The difference between SAM and LDA persists even as the task becomes more semantically tight, indicating that it finds more meaningful distinctions between finer-grained topics.[8] Furthermore, features derived from tf-idf SAM significantly improve classification accuracy in **news-20-different** and **news-20-similar** when combined with raw bag-of-words features, unlike LDA (**news-20-different**: 94.1% accuracy vs. 91.8% accuracy for LDA and 91.7% accuracy for bag-of-words only; **news-20-similar**: 88.1% accuracy vs. 85.7% accuracy for LDA and 85.9% accuracy for bag-of-words only).

The bag-of-words representation is best for the semantically tight **news-20-same** dataset, but SAM nearly matches its performance (75.0% accuracy vs. 77.5% accuracy) despite the information lost in the reduction from $\sim$3000 features to only 50. Neither LDA nor movMF can match this accuracy. The performance gap between SAM and LDA suggests that generative models based on vMF distributions are better suited to capturing fine-grained semantic variation in text than are multinomial models.

[8] **(Classifier robustness)** The results do not change significantly when replacing Logistic Regression with an SVM or Naive Bayes; implementations from Weka (Witten & Frank, 2005).

Table 3. Logistic regression accuracy using inferred features on the four-class *Il Principe* thematic shift detection task. Standard *tf* representations are used in all models. SAM infers significantly better features overall.

Model	Accuracy (%)				
	Overall	*prin.*	*war*	*cond.*	*Italy*
Bag-of-Words	57.9 ± 3.4	60.5	71.3	55.3	45.1
LDA	57.3 ± 3.0	59.4	63.9	58.1	34.9
movMF	49.6 ± 8.3	47.6	11.7	55.8	0.0
MH SAM $[\mathbb{S}_+]$	46.1 ± 6.9	46.5	31.8	54.4	8.3
MH SAM $[\mathbb{S}]$	59.4 ± 5.4	60.9	51.7	64.8	31.4
VEM SAM $[\mathbb{S}_+]$	58.7 ± 0.6	64.9	71.1	60.8	13.9
VEM SAM $[\mathbb{S}]$	$\mathbf{65.2 \pm 0.3}$	71.3	65.1	62.5	50.6

4.2.2. Detecting Thematic Shifts in *Il Principe*

Both SAM and LDA perform well when the corpus covers a wide variety of topics. To more precisely illustrate their differences, then, it is instructive to compare them in classification tasks where small semantic distinctions are important. In this section we perform supervised textual segmentation (identifying *thematic shifts* in discourse; cf. Hearst (1994)) on Niccolò Machiavelli's *Il Principe*. Since the book is short, singly-authored, and thematically tight, topics must be fine-grained to be helpful. For training the topic models, documents are taken to be individual paragraphs of text. For classification, each paragraph is assigned one of four labels corresponding to the main themes of the book: (1) the types of *principalities* (chapters I-XI), (2) the types of *armies* (chapters XII-XIV), (3) the character and *conduct* of Princes (chapters XV-XXIII), and (4) the current political situation in *Italy* (circa 1505; chapters XXIV-XXVI). This split yields a challenging 4-way classification problem.[9]

SAM $[\mathbb{S}]$ discovers the best features in all settings; SAM significantly outperforms LDA and movMF, cutting relative classification error by by 18.5% (Table 3; significance determined using Fisher's Least Significant Difference test). Broken down by class, SAM sees the largest relative reductions in error for *Italy*, the most thematically ambiguous section. SAM $[\mathbb{S}]$ also outperforms SAM $[\mathbb{S}_+]$ by a significant margin, highlighting the utility of explicitly representing negative term weights. Finally, since the Adaptive MH and VEM versions of SAM were run using the

[9] **(*Il Principe* dataset)** The base text is the original Italian version, converted to lowercase with stopwords removed. A total of 128 paragraphs are extracted; 39.8% are labeled *principalities*, 37.5% are labeled *conduct*, 12.5% *armies* and 9.3% *Italy*.

same convergence criterion, the results indicate that the approximations made in VEM SAM do not significantly affect performance, despite the fact that MH sampling takes significantly longer ($\sim$10 hours as opposed to $\sim$20 minutes). The number of topics chosen has a large impact on performance for $T < 50$, but performance is relatively stable for $T > 50$. Hence selecting T using a nonparametric prior may be justifiable (Teh et al., 2006).

4.2.3. 13 NATURAL SCENE CATEGORIES

The final task involves classifying visual images according to their natural scene type, e.g. living room, coast, forest, etc, using the 13 scene database proposed by Fei-Fei & Perona (2005). We divide the full 13 class visual scene recognition task, **13-scene-full**, into four separate 4-class problems: **13-scene-different** (including *livingroom*, *MITstreet*, *CALsuburb*, and *MITopencountry*), **13-scene-similar** (*MITinsidecity*, *MITstreet*, *CALsuburb*, *MITtallbuilding*), **13-scene-outdoor** (*MITcoast*, *MITforest*, *MITmountain*, *MITopencountry*), and **13-scene-indoor** (*bedroom*, *kitchen*, *livingroom*, *PARoffice*), ordered by their classification difficulty. We follow Fei-Fei and Perona's preprocessing steps: densely sampling patches, computing 128-dimensional SIFT descriptors, then clustering the descriptors using spherical k-means. The resulting clusters are treated as *visual words*, and each image is represented by its visual word counts. Note that despite the fact that a similar visual-bag-of-words representation is employed in this task, it differs fundamentally from the previous tasks in terms of sparsity: Figure 3 indicates that most visual words tend to occur in most scenes, leading to denser document representations. 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 significantly outperforms LDA across all scene recognition tasks (Figure 4) when 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, indicating that SAM may perform better relative to LDA with less data. We find that neither topic model significantly outperforms a simple visual-bag-of-words representation for $|V| = 200$; for $|V| = 1500$, however, both significantly outperform visual-bag-of-words. With dense features, SAM provides a smaller benefit relative to LDA, as cosine distance and KL-divergence correspond more closely.

5. Future Work

SAM opens up a new class of admixture models: those based on spherical distributions. From that class originate three important avenues for future work. First, most extensions to LDA proposed in recent literature can easily be used with SAM, such as modeling infinite-capacity (Teh et al., 2006) or correlated topics (Blei & Lafferty, 2005). Furthermore, SAM can be extended to model word burstiness explicitly, as in DCM-LDA (Doyle & Elkan, 2009), by exploiting the conjugate prior of Nuñez-Antonio & Gutiérrez-Peña (2005). Second, although sparse document vectors improve the efficiency of the inference methods presented here, the topic vectors themselves are not sparse, leading to storage overhead that scales as $O(|V| \cdot T)$. Such overhead is undesirable with larger corpora. One possible solution is a spherical admixture with sparse topic representations (i.e., each topic only spans a subspace of the full $\mathbb{S}^{|V|-1}$), leading to more efficient inference and lower storage overhead.

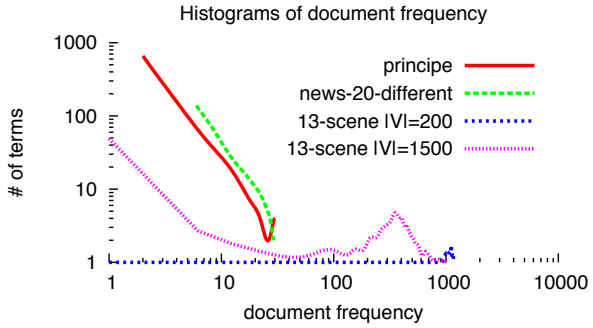


Figure 3. Histograms showing the number of terms with a given document frequency; **13-scene** is significantly more dense in terms of vocabulary coverage than the textual domains.

6. Conclusion

This paper has developed SAM, an admixture model that decomposes spherically distributed data into weighted combinations of component vMF distributions. Unlike previous spherical models, 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 parameterized by cosine distance, can explicitly assign negative weight to topic terms, and models document-level word absence.

In both subjective human studies and dimensionality reduction experiments, SAM was found to produce more relevant topic features than did either the movMF spherical mixture model or LDA, particularly on data where fine-grained topic distinctions are important. Three properties—cosine distance, negative terms, and word absence/presence—were shown to contribute to its performance.

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 par-

Accuracy	diff.	sim.	outdoor	indoor	all
LDA	79.3	68.5	60.9	43.6	43.4
SAM [𝕊]	**85.0**	**74.4**	**68.4**	**50.2**	**50.3**

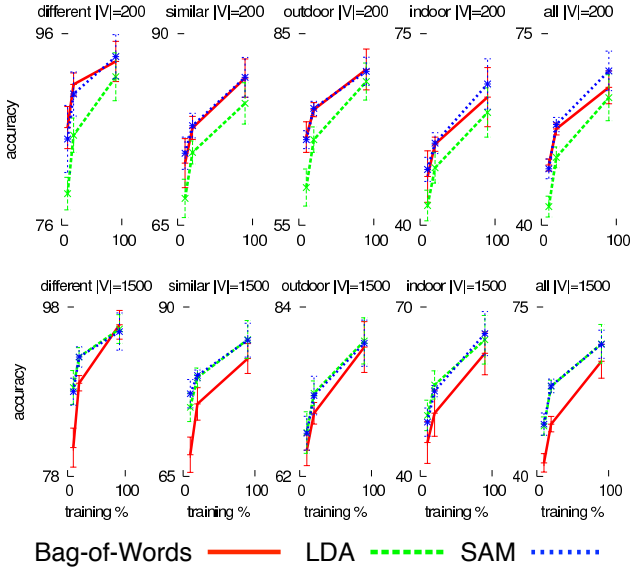


Figure 4. (**top**) Classification accuracy for **13-scene** with $|V| = 200$. (**bottom**) Learning curves for all classes, $|V| = 200$ and $|V| = 1500$.

tially supported by a Google Research Award and an NSF Graduate Research Fellowship to the first author. Experiments were run on the Mastodon Cluster, provided by NSF Grant EIA-0303609.

References

Banerjee, Arindam and Basu, Sugato. Topic models over text streams: A study of batch and online unsupervised learning. In *SDM*, 2007.

Banerjee, Arindam, Dhillon, Inderjit S., Ghosh, Joydeep, and Sra, Suvrit. Clustering on the unit hypersphere using von Mises-Fisher distributions. *JMLR*, 6, 2005.

Bingham, Ella, Kabán, Ata, and Fortelius, Mikael. The aspect Bernoulli model: multiple causes of presences and absences. *Pattern Analysis and Applications*, 12(1), 2009.

Blei, David, Ng, Andrew, and Jordan, Michael. Latent Dirichlet allocation. *JMLR*, 3, 2003.

Blei, David M. and Lafferty, John D. Correlated topic models. In *NIPS*, 2005.

Buss, Samuel R. and Fillmore, Jay P. Spherical averages and applications to spherical splines and interpolation. *ACM Transactions on Graphics*, 20(2), 2001.

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

Daumé III, Hal. HBC: Hierarchical Bayes compiler, 2009. `http://hal3.name/HBC`.

Doyle, Gabriel and Elkan, Charles. Accounting for word burstiness in topic models. In *ICML*, 2009.

Elkan, Charles. Clustering documents with an exponential-family approximation of the Dirichlet compound multinomial distribution. In *ICML*, 2006.

Fei-Fei, Li and Perona, Pietro. A Bayesian hierarchical model for learning natural scene categories. *CVPR*, 2005.

Hearst, Marti A. Multi-paragraph segmentation of expository text. In *ACL*, 1994.

Hofmann, Thomas. Probabilistic latent semantic indexing. In *Research and Development in Information Retrieval*, 1999.

Jordan, Michael I., Ghahramani, Zoubin, Jaakkola, Tommi S., and Saul, Lawrence K. An introduction to variational methods for graphical models. *Machine Learning*, 37(2), 1999.

le Cessie, S. and van Houwelingen, J.C. Ridge estimators in logistic regression. *Applied Statistics*, 41(1), 1992.

Manning, Christopher and Schütze, Hinrich. *Foundations of Statistical Natural Language Processing*. 2000.

Mardia, Kanti V. and Jupp, Peter E. *Directional Statistics*. Wiley, 2000.

McCallum, Andrew and Nigam, Kamal. A comparison of event models for Naive Bayes text classification. In *AAAI Workshop on Learning for Text Categorization*, 1998.

Nuñez-Antonio, Gabriel and Gutiérrez-Peña, Eduardo. A Bayesian analysis of directional data using the von Mises-Fisher distribution. *Communications in Statistics – Simulation and Computation*, 34, 2005.

Teh, Yee W., Jordan, Michael I., Beal, Matthew J., and Blei, David M. Hierarchical Dirichlet processes. *JASA*, 101, 2006.

Wallach, Hanna, Mimno, David, and McCallum, Andrew. Rethinking LDA: Why priors matter. In *NIPS*. 2009.

Witten, Ian H. and Frank, Eibe. *Data Mining: Practical Machine Learning Tools and Techniques*. 2005.

Zhong, Shi and Ghosh, Joydeep. Generative model-based document clustering: A comparative study. *KAIS*, 8(3), 2005.