Admixture of Poisson MRFs: A Topic Model with Word Dependencies

David I. Inouye DINOUYE@CS.UTEXAS.EDU
Pradeep Ravikumar PRADEEPR@CS.UTEXAS.EDU
Inderjit S. Dhillon INDERJIT@CS.UTEXAS.EDU
Dept. of Computer Science, University of Texas, Austin, TX 78712, USA

Abstract

This paper introduces a new topic model based on an admixture of Poisson Markov Random Fields (APM), which can model dependencies between words as opposed to previous independent topic models such as PLSA (Hofmann, 1999), LDA (Blei et al., 2003) or SAM (Reisinger et al., 2010). We propose a class of admixture models that generalizes previous topic models and show an equivalence between the conditional distribution of LDA and independent Poissons—suggesting that APM subsumes the modeling power of LDA. We present a tractable method for estimating the parameters of an APM based on the pseudo log-likelihood and demonstrate the benefits of APM over previous models by preliminary qualitative and quantitative experiments.

1. Introduction

Topic models can be understood as a class of statistical models for document collections that model documents as admixtures over *topics*. Specifically, each topic is modeled as a distribution over words, and each document is a separate mixture of such topics (or specifically, the word distributions comprising the topics). Such an admixture can be contrasted with a vanilla mixture of topics, where each document would be drawn from a single topic.

A popular set of topic models is PLSA (Hofmann, 1999), which uses the multinomial distribution as the word distribution for any topic, and its Bayesian counterpart, LDA (Blei et al., 2003), which adds Dirichlet priors. While these topic models have proved enormously useful in modeling varied document collections and have attracted a long line of work with numerous extensions (see (Blei, 2012) for a review of LDA applications and trends), it has some crucial lacunae that arise from its basic use of the multinomial distribution to model word distributions for topics. There are several reasons which make the multinomial distribution an inadequate distribution for documents and topics. The primary issue is that it does not model dependencies between words: if the word "kernels" appears in a document (specifically, a machine learning paper), the appearance of the word "graphs" might be less likely. Alternatively, if the word "classification" appears, "supervised" is more likely to appear than in general documents. Indeed, typical coherence metrics that quantitatively measure the goodness of fit of various topic models primarily *test* for such dependence among estimated top words for the topics (see Sec. 6.2). A second caveat is that the multinomial distribution does not model absences of words. Lastly, the multinomial word distribution does not leverage varying document lengths. For instance, with large counts of other words, some specific word might become less likely.

To address the issue of modeling word absences, Reisinger et al. (2010) proposed the use of von Mises-Fisher distribution for topic distributions. But while this addresses one issue with multinomials, it does not model word dependen-

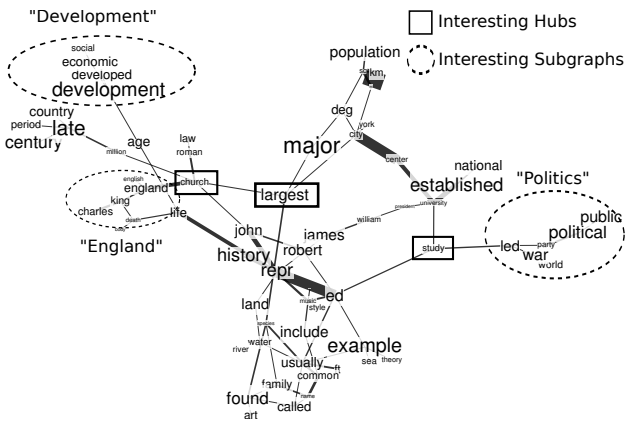


Figure 1: A Poisson MRF can provide interesting insights into a text corpus including multiple word senses (hubs of graph) and semantic concepts (coherent subgraphs).

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

cies, nor does it leverage document lengths in any substantative way.

In this paper, we propose using Poisson MRFs (Yang et al., 2012) for topic distributions and using the resulting admixture of Poisson MRFs (APM) for modeling document collections. These Poisson MRFs allow modeling multivariate count data and use dependencies among the count variables to represent the joint distribution compactly. Moreover, since these are graphical model distributions, the dependencies are Markov with respect to an undirected graph, which thus provides a visually appealing representation of any topic—see Fig. 1—in contrast to just a list of words as in PLSA/LDA.

We position the Poisson MRF in context of topic models by showing that the conditional distributions of the classical LDA model can be written as a Poisson MRF where the underlying graph has no edges and hence no dependencies between words; this connection—which was only recently discovered in the context of matrix factorization (Gopalan et al., 2013)—not only puts into relief the assumptions made by LDA but also opens the door to other approximate inference schemes for LDA (which however we do not explore here). In other contributions of this paper, we define a new class of models called admixtures and show that this class generalizes previous topic models, which thus opens the door to other topic models based on non-Poisson distributions. Finally, we provide qualitative as well as quantitative evidence for the benefits of APM by training the APM model on both a subset of the Grolier encyclopedia and the CMU 20 Newsgroup dataset.

2. Poisson MRFs (PMRFs)

First, we review the Poisson MRF model (PMRF) as proposed by Yang et al. (2012). Second, we contextualize the independent PMRF model by showing an equivalence with the conditional distributions of LDA. Finally, we propose a novel prior distribution for PMRFs that can be viewed as a generalization of the Gamma distribution.

2.1. PMRF Definition

By assuming that the conditionals of the joint distribution are univariate Poisson, Yang et al. (2012) recently proposed a PMRF model that provides a joint distribution over multivariate count data. They also provided a tractable way to estimate the parameters of such a PMRF using ℓ_1 regularization and proved that the estimator is guaranteed to recover the underlying dependency structure with some assumptions including sparsity of the parameters. The model PMRF$(\boldsymbol{\theta}, \Theta)$ is defined as follows:

$$\Pr_{\text{PMRF}}(\boldsymbol{x} \mid \boldsymbol{\theta}, \Theta) \propto \exp\left\{\boldsymbol{\theta}^T \boldsymbol{x} + \boldsymbol{x}^T \Theta \boldsymbol{x} - \sum_{s=1}^{p} \ln(x_s!)\right\},$$

where $\boldsymbol{\theta} \in \mathbb{R}^p$ and $\Theta \in \{\mathbb{R}^{p \times p}_{-} : \text{diag}(\Theta) = 0\}$. By construction, the conditional distribution of a variable x_s given all other variables $\boldsymbol{x}_{\setminus s}$ is a univariate Poisson with canonical parameter $\eta_s = \theta_s + \Theta_s^T \boldsymbol{x}$ and mean (standard) parameter $\lambda_s = \exp(\eta_s)$. An illustration of the density of a 2D Poisson with negative, zero and positive dependency can be seen in Fig. 2. Other observations about a PMRF:

- The dependency parameter Θ is analogous to the Gaussian precision matrix Σ^{-1} in a Gaussian MRF.
- If $\Theta = 0$, then the PMRF reduces to an independent multivariate Poisson distribution.
- Negative dependencies can help model sparse data (i.e. data with many 0's) because the density can be concentrated on the axes as seen seen in Fig. 2 (left).

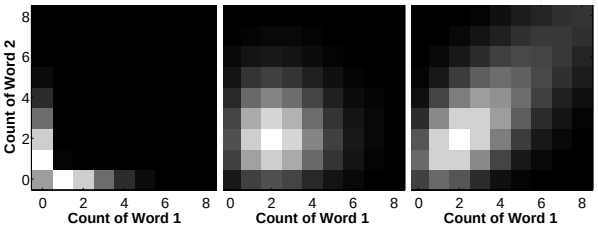


Figure 2: The densities of three 2D Poisson MRFs that show possible dependency structures between two words. Negative dependencies (left) suggest that two words *rarely* co-occur whereas positive dependencies (right) suggest that two words *often* co-occur.

The negativity constraint on Θ is required under the formulation above to ensure that the distribution is normalizable (Yang et al., 2012). However, Yang et al. (2013) propose a slight modification to the sufficient statistics of the PMRF that removes this constraint and allows positive parameter values—see reference for details. For simplicity of notation and wording, throughout the rest of the paper, the basic PMRF notation will be used for derivations though the APM model uses this slightly modified PMRF.

2.2. LDA Conditionals Equivalent to Independent Poissons

In this section, we place Poisson MRFs in the context of topic models by showing the equivalence between the conditionals of LDA and an independent Poisson MRF.[1] LDA assumes the following generative process for a new document given that the topic weights for the document

[1] Gopalan et al. (2013) recently introduced the connection between LDA and Poisson models in the context of matrix factorization.

$\boldsymbol{w}$ and the topic distribution parameters $\boldsymbol{\phi}_{1\ldots k}$ are known: 1) Draw $\tilde{x} \sim \text{Poisson}(\tilde{\lambda})$ 2) For each of the $\tilde{x}$ words: (a) Draw topic index $z \sim \text{Categorical}(\boldsymbol{w})$ (b) Draw word $v \sim \text{Categorical}(\boldsymbol{\phi}_z)$. Notice that because $\tilde{x}$ is independent of the other variables in LDA, it is often simply ignored when estimating the model parameters. In our model, however, $\tilde{x}$ cannot be ignored because words can be dependent. By marginalizing out the topic variable z, step 2 can be collapsed into a draw from a Multinomial with a single parameter $\tilde{\boldsymbol{\phi}}$, which is simply a weighted average over the topic distribution parameters $\boldsymbol{\phi}_z$. This yields the following modified step: 2') Draw document $\boldsymbol{x} \sim \text{Mult}(\tilde{\boldsymbol{\phi}} = \sum_{j=1}^{k} w_j \boldsymbol{\phi}_j \,|\, N = \tilde{x})$. Therefore, the probability of a document $\boldsymbol{x}$ given $\boldsymbol{w}$ and $\boldsymbol{\phi}_{1\ldots k}$ is: $\Pr_{\text{Poiss}}(\tilde{x} \,|\, \tilde{\lambda}) \Pr_{\text{Mult}}(\boldsymbol{x} \,|\, \tilde{\boldsymbol{\phi}} = \sum_{j=1}^{k} w_j \boldsymbol{\phi}_j, N = \tilde{x})$.

Amazingly, this Poisson-Multinomial joint distribution is equivalent to p independent Poissons (Bishop et al., 2007):

$$\begin{aligned}
\Pr_{\text{Ind. Poiss}}(\boldsymbol{x} \,|\, \lambda_1, \cdots, \lambda_p) &= \prod_{s=1}^{p} \frac{e^{-\lambda_s}}{x_s!} \lambda_s^{x_s} \\
&= \frac{\tilde{x}!}{\tilde{x}!} \frac{e^{-\tilde{\lambda}}}{\prod_{s=1}^{p} x_s!} \prod_{s=1}^{p} \left(\frac{\tilde{\lambda} \lambda_s}{\tilde{\lambda}} \right)^{x_s} \\
&= \frac{e^{-\tilde{\lambda}}}{\tilde{x}!} \tilde{\lambda}^{\tilde{x}} \frac{\tilde{x}!}{\prod_{s=1}^{p} x_s!} \prod_{s=1}^{p} \left(\frac{\lambda_s}{\tilde{\lambda}} \right)^{x_s} \\
&= \Pr_{\text{Poiss}}\left(\tilde{x} \,|\, \tilde{\lambda} \right) \Pr_{\text{Mult}}\left(\boldsymbol{x} \,|\, \theta = (\lambda_1, \cdots, \lambda_p) / \tilde{\lambda}, \mathrm{N} = \tilde{x} \right)
\end{aligned}$$

where $\tilde{\lambda} = \sum_{s=1}^{p} \lambda_s$ and $\tilde{x} = \sum_{s=1}^{p} x_s$. Therefore, a PMRF directly generalizes the conditional distribution of PLSA/LDA by relaxing the independence assumption. To more fully generalize LDA, priors must be added to a PMRF as proposed next.

2.3. Adding Priors to a PMRF

Similar to LDA's prior, a conjugate prior on the parameters of a PMRF can be defined as being proportional to:

$$\exp\{\boldsymbol{\beta}^T \boldsymbol{\theta} + \boldsymbol{\beta}^T \Theta \boldsymbol{\beta} - \gamma \mathrm{A}(\boldsymbol{\theta}, \Theta) - \lambda_{\boldsymbol{\theta}} \|\boldsymbol{\theta}\|_2^2 - \lambda \|\text{vec}(\Theta)\|_1\},$$

where $\forall s,\ \beta_s > 0$, $\gamma \geq 0$, $\lambda_{\boldsymbol{\theta}} > 0$ and $\lambda > \max_{i,j} \beta_i \beta_j$.[2] One observation is that when $\Theta = 0$, $\exp(\theta_s)$ is essentially $\text{Gam}(\text{shape} = \boldsymbol{\beta}_s; \text{scale} = 1)$. Therefore, for independent Poissons, this is similar to using Gamma priors. The posterior distribution merely modifies the hyperparameters to be $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta} + \boldsymbol{x}$ and $\tilde{\gamma} = \gamma + 1$. Therefore, because this prior adds pseudo-counts $\boldsymbol{\beta}$ to the observations for parameter estimation, this prior for PMRFs is analogous to a Dirichlet prior for multinomials as in LDA.

[2]The conditions on the hyperparameters are needed for normalization. In practice, $\lambda_{\boldsymbol{\theta}}$ can be set arbitrarily small and is thus ignored in subsequent discussions. The λ hyperparameter is used for ℓ_1 regularization as discussed in Sec. 5.1.

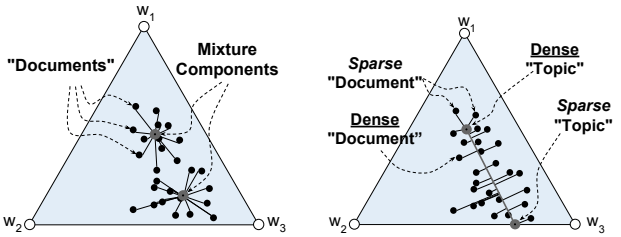


Figure 3: (Left) In *mixtures*, documents are drawn from exactly one component distribution. (Right) In *admixtures*, documents are drawn from a distribution whose parameters are a convex combination of component parameters.

3. Generalized Admixtures

In a simple *mixture model*, an observation is assumed to come from exactly one of k possible components. An illustration of this type of model is shown in Fig. 3 (left) in which documents are drawn from exactly one of two component distributions—the "topics" in the case of document modeling. On the other hand, for *admixtures* each document is drawn from a distribution whose parameters can be any convex combination of the component parameters, allowing each document to be explained by multiple components as illustrated in Fig. 3 (right).[3]

Given this intuition about admixtures, the probability of a single observation $\boldsymbol{x}$ from an *admixture* of some base distribution (e.g. multinomial, von Mises-Fisher, PMRF)—assuming that the admixture weights $\boldsymbol{w}$ and component canonical parameters $\Phi = \boldsymbol{\phi}_{1\ldots k}$ are given—is defined as:

$$\Pr_{\text{Admix.}}(\boldsymbol{x} \,|\, \boldsymbol{w}, \Phi) = \Pr_{\text{Base}}\left(\boldsymbol{x} \,\middle|\, \bar{\boldsymbol{\phi}} = \Psi^{-1}\left(\sum_{j=1}^{k} w_j \Psi(\boldsymbol{\phi}_j) \right) \right), \tag{1}$$

where Ψ allows for the mixing to occur in a suitable transformation of the parameter space. In the context of exponential families, the mixing could occur either in the canonical parameter space in which case Ψ would be the identity function, or it could occur in the mean parameter space (such as the mean μ and covariance Σ for a multivariate Gaussian). For this paper, unless otherwise specified, we will assume that Ψ is equal to the identity function.

If priors are given for the admixture weights $\boldsymbol{w}$ and the topic parameters $\boldsymbol{\phi}_{1\ldots k}$ with parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ hyperparameters respectively, the joint distribution of a single observation and the parameters is:

$$\Pr_{\text{Base}}\left(\boldsymbol{x} \,\middle|\, \bar{\boldsymbol{\phi}} = \sum_{j=1}^{k} w_j \boldsymbol{\phi}_j \right) \Pr(\boldsymbol{w}) \prod_{j=1}^{k} \Pr(\boldsymbol{\phi}_j)$$

[3]Fig. 3 is only meant as an illustration and not as a rigorous visualization of mixtures or admixtures. It should be noted that, in general, the KL-divergence should be minimized rather than ℓ_2 distance as suggested by the figure.

This gives the joint distribution over a set of n independent observations as:

$$\Pr_{\text{Admix.}}(X, \mathrm{W}, \Phi) = \prod_{i=1}^{n} \Pr_{\text{Base}}\left(\boldsymbol{x}_i \,\middle|\, \bar{\boldsymbol{\phi}} = \sum_{j=1}^{k} w_{i,j}\boldsymbol{\phi}_j\right) \Pr(\boldsymbol{w}_i) \prod_{j=1}^{k} \Pr(\boldsymbol{\phi}_j) \tag{2}$$

Intuitively, this admixture model formulation means that each observation can be explained by a mixture of a relatively small number of component distributions parameterized by $\boldsymbol{\phi}_j$. In the special case where w is an indicator vector, this distribution becomes a standard mixture model where each observation is explained by only one component. In the special case where $k = 1$, the admixture simply reduces to every observation being drawn from a single base distribution. Therefore, this admixture formulation generalizes both single and mixture distributions.

This admixture model also generalizes previous topic models and provides a general framework for defining new admixture models based on any parametric distribution. In the next sections, several examples of previous admixture models are given followed by the formulation of this paper's main model—an admixture of Poisson MRFs which, to the authors' best knowledge, is the first admixture model to allow dependencies between words.

Example 1 - LDA As shown in Sec. 2.2, LDA assumes that each document is draw from an admixture of multinomials. The admixture weights and the parameters for each topic multinomial are drawn from a Dirichlet prior. It is important to notice that LDA mixes in the standard multinomial mean parameter space (i.e. Ψ_{LDA} is the canonical to mean parameter transformation).

Example 2 - Population Admixtures In the genetic community, the term *admixture* has been used to describe a population produced by interbreeding several previously-isolated populations into a new *admixed* population. Pritchard et al. (2000) use a model equivalent to LDA to explore this concept. Under this population model, the original ancestors of a population correspond to *topics* and individuals correspond to *documents*.

Example 3 - SAM The Spherical Admixture Model (SAM) as proposed by Reisinger et al. (2010) is an admixture model where the base distribution is a Von Mises-Fisher distribution—the independent Gaussian analog defined on the unit hypersphere. The model, which is motivated by the observation that cosine distance is an important document similarity, assumes Dirichlet and Von Mises-Fisher priors on the admixture weights and component parameters respectively.

4. Admixture of Poisson MRFs

With the background on PMRFs and the development of admixtures, the main model of this paper—an admixture of Poisson MRFs (APM)—can be developed. Relaxing the independence assumption of previous admixture models such as LDA, APM assumes that the base distribution is a PMRF. This yields the following joint distribution:

$$\begin{aligned}\Pr_{\text{APM}}(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{\theta}_{1\ldots k}, \Theta_{1\ldots k}) \\ = \Pr_{\text{PMRF}}\left(\boldsymbol{x} \,\middle|\, \bar{\boldsymbol{\theta}} = \sum_{j=1}^{k} w_j\boldsymbol{\theta}_j, \bar{\Theta} = \sum_{j=1}^{k} w_j\Theta_j\right) \\ \times \Pr_{\text{Dir}}(\boldsymbol{w}) \prod_{j=1}^{k} \Pr(\boldsymbol{\theta}_j, \Theta_j)\end{aligned} \tag{3}$$

where $\Pr_{\text{Dir}}(\boldsymbol{w} \mid \boldsymbol{\alpha})$ is a Dirichlet prior on the admixture weights (similar to LDA) and $\Pr(\boldsymbol{\theta}_j, \Theta_j)$ is the PMRF prior defined in Sec. 2.3. Because of the equivalence described in Sec. 2.2, APM subsumes the expressive power of LDA. The primary difference between an independent APM and LDA is that LDA mixes in the standard Multinomial parameter space whereas APM mixes in the canonical parameter space. An interesting open area for future research could be admixing the component PMRFs in a different parameter space such as the mean parameter space.[4] Fundamentally, however, this model is much more expressive than all previous admixture models because it allows for dependencies between words.

4.1. Topic Representation

In the APM model, topics are represented as PMRFs, and therefore, each topic provides a full graph over words showing word dependencies rather than just a list of words as in independent models such as LDA (see Fig. 4 in Sec. 6 for example topic graphs). This representation opens up a whole new area for interpreting, exploring and visualizing topics using a graph. In addition, all the metrics and algorithms on graphs such as tree width or shortest path could be used to explore each topic.

4.2. Document Representation

Documents could be represented in at least two different ways under the APM model. First, they could be represented by their admixture weights, and therefore, APM could be used as a type of dimensionality reduction technique. Second, each document can be represented as a full graph over words just like a topic because each document is associated with an admixed PMRF. This graph represen-

[4] For more information on the relationship between the mean and canonical parameter spaces, see (Wainwright & Jordan, 2008).

tation provides a powerful new way to visualize and summarize a document that was not possible with independent models like LDA.

5. Parameter Estimation by Optimizing Approximate Posterior

The parameters of an admixture of Poisson MRFs can be estimated by minimizing the negative log posterior. Because the true log-likelihood of a Poisson MRF is computationally intractable for complex multivariate distributions (Wainwright & Jordan, 2008), the pseudo log-likelihood—which approximates the joint distribution as a product of node conditionals—will be used instead. With the Dirichlet prior on $\boldsymbol{w}$ and the prior described in Sec. 2.3 on the component parameters, the approximate posterior is:

$$\mathcal{P} \approx \hat{\mathcal{P}}(\mathrm{W}, \boldsymbol{\theta}_{1\ldots k}, \Theta_{1\ldots k} \mid X) \tag{4}$$
$$\propto \sum_{i=1}^{n} \left\{ \left[\sum_{s=1}^{p} \eta_{s,i} \hat{x}_{s,i} - (\gamma+1)\mathrm{A}\left(\eta_{s,i}\right) \right] + (\boldsymbol{\alpha}-1)^T \ln(\boldsymbol{w}_i) \right\},$$

where $\hat{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{\beta}$ and $\eta_{s,i} = \sum_{j=1}^{k} w_{i,j}(\boldsymbol{\theta}_{j,s} + \Theta_{j,s}\hat{\boldsymbol{x}}_{i,\setminus s})$ is the canonical parameter of a univariate Poisson.

5.1. Enforcing Sparsity of Θ_j by ℓ_1 Regularization

For interpretability, generalizability and computational tractability, the parameters of high-dimensional MRFs are often assumed to be sparse (i.e. a small number of non-zeros compared to zeros). This sparsity assumption is usually incorporated into the problem by adding an ℓ_1 regularization term to the objective function. This ℓ_1 regularized estimator has been shown to have theoretical guarantees on structural recovery for Bernoulli MRFs/Ising Models (Ravikumar et al., 2010), Gaussian MRFs (Ravikumar et al., 2011) and, more recently, Poisson MRFs (Yang et al., 2012; 2013). For similar reasons, APM assumes that the parameter matrices (Θ_j) for each topic PMRF are sparse and, like the aforementioned methods, estimates this sparse solution by using an ℓ_1 regularization term. Intuitively, this sparsity assumption makes sense because most words are only directly related to a small subset of other relevant words.

5.2. Unconstrained Optimization

Along with the regularization of the Θ_j parameter matrices, APM requires that the columns of the admixture weights matrix W be probability vectors (i.e. properly defined mixture weights that lie on the k-dimensional simplex). This leads to the following unconstrained optimization problem:

$$\underset{\mathrm{W}, \boldsymbol{\theta}_{1\ldots k}, \Theta_{1\ldots k}}{\arg\min} \; -\hat{\mathcal{P}} + \delta_{\mathbb{W}}(\mathrm{W}) + \lambda \sum_{j=1}^{k} \|\mathrm{vec}(\Theta_j)\|_1 \tag{5}$$

where λ is the ℓ_1 regularization parameter, $\mathbb{W}$ is the set of all possible matrices such that the columns are probability vectors and $\delta_{\mathbb{W}}(\mathrm{W}) = \{0, \text{if } \mathrm{W} \in \mathbb{W};\ \infty, \text{otherwise}\}$.

5.3. Proximal Optimization Algorithms

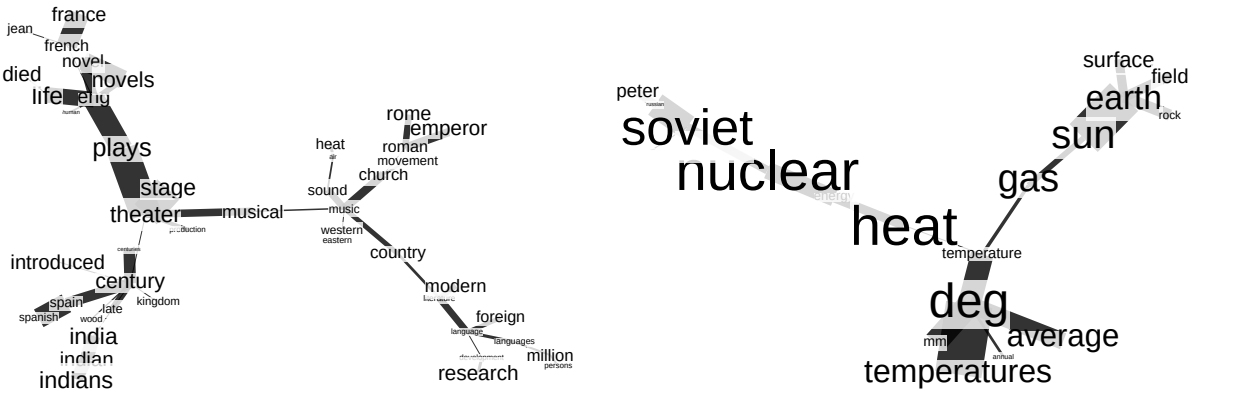
Because the objective in (5) is composed of a differentiable term (i.e. $\hat{\mathcal{P}}$) and two non-differentiable terms (i.e. $\delta_{\mathbb{W}}(\mathrm{W})$ and $\lambda \sum_{j=1}^{k} \|\mathrm{vec}(\Theta_j)\|_1$), a simple gradient descent algorithm cannot be used to solve the problem. Therefore, in this paper, we use a proximal optimization algorithm (Parikh & Boyd, 2013). Essentially, a proximal algorithm is an iterative algorithm that computes each new parameter estimate using only the previous estimate and the differentiable term. After finding a new estimate, a proximal algorithm applies the prox operator(s) to this estimate to incorporate the non-differentiable terms of the objective function and iterates until convergence.

The prox operator for the ℓ_1 regularization term in (5) is the simple soft thresholding operator: $\mathcal{S}_\lambda(z) = \mathrm{sign}(z)\max(|z| - \lambda, 0)$. The prox operator for $\delta_{\mathbb{W}}(\mathrm{W})$ is simply the best Euclidean projection onto the simplex which can be computed using the algorithm from (Chen & Ye, 2011). With these operators, any proximal algorithm can optimize (5) but, in this paper, a FISTA-like algorithm was used (Beck & Teboulle, 2009). Because APM is a model with many parameters and FISTA-like algorithms are only first-order optimization methods, the models trained for this paper required many iterations to converge (> 5000). Using faster and more complex proximal optimization algorithms such as a proximal Newton method would be an excellent area for future work.

6. Preliminary Experiments

Because previous admixture models have been independent, it is difficult to directly compare APM to previous models. Therefore, first, an experiment was conducted by running APM (with $k = 5$ and $p = 500$) on approximately 31,000 articles of the Grolier encyclopedia.[5] Visualizations of the topics were constructed using the graph visualization program Gephi[6] in order to show some qualitative results on the model output and suggest that APM can provide a more interesting, intuitive and visually appealing representation of topics than merely a list of words as in standard topic models. Two topics of this run can be seen in Fig. 4 and other topic graph examples are given in the appendix. Also, a simple experiment was conducted to give some evidence, though inconclusive, that the APM model subsumes the power of the LDA model because of the model equivalence described in Sec. 2.2.

[5] www.cs.nyu.edu/~roweis/data.html
[6] www.gephi.org



(a) "Music and Fine Arts"

(b) "Temperature"

Figure 4: These APM topic visualizations illustrate that PMRFs are much more intuitive than multinomials (as in LDA/PLSA), which can only be represented as a list of words. Word size signifies relative word frequency and edge width signifies the strength of word dependency (only positive dependencies shown).

6.1. Qualitative Experiment

The graphs as seen in Fig. 4 have many interesting structural features that can be interpreted.[7] In the first example (Fig. 4a), the word "musical" is a hub word that connects the two concepts of "theatre" and "music". A similar idea happens in the second example (Fig. 4b) where "temperature" connects the concepts of "heat", "gas" and "deg".

Another interesting feature is chains of words whose endpoints are not directly related but only related through other words. For example, the chain "music" $\leftrightarrow$ "musical" $\leftrightarrow$ "theater" $\leftrightarrow$ "plays" suggests that "music" and "plays" are related, albeit indirectly. The chain "sun" $\leftrightarrow$ "gas" $\leftrightarrow$ "temperature" $\leftrightarrow$ "heat" $\leftrightarrow$ "nuclear" in Fig. 4b shows the connection that the sun is related to nuclear reactions through the words "heat", "gas" and "temperature". For other chains, the endpoints are not related even though each edge seems reasonable. For example, the chain "novel" $\leftrightarrow$ "eng" $\leftrightarrow$ "plays" $\leftrightarrow$ "theater" $\leftrightarrow$ "musical" $\leftrightarrow$ "music" $\leftrightarrow$ "church" has logical connections for each edge but "novel" is not usually associated with "church".

Though these features give evidence for the usefulness and power of APM, they do not capture any of the negative dependencies between words—words that do *not* tend to co-occur. For example, the words "novel" and "math" would not tend to co-occur. This might be helpful in excluding documents from certain categories in document categorization. For example, if the words "history", "war" and "politics" appear in a document, the document is unlikely to be science literature. Though it may be more difficult to visualize these negative dependencies, negative dependencies can provide interesting structural information of the underlying dataset.

[7] These graphs were manually filtered to simplify the whole graph. A future area of research could be to automatically filter the graph to important clusters.

6.2. Coherence Experiment

Though the APM model gives significantly more information than simply a list of topic words as in LDA, another experiment was conducted to give evidence that APM can be used to produce a list of topic words similar to other topic models. The APM model was applied to the CMU 20 Newsgroup dataset evaluated with the two metrics explained next. Because of the complexity of APM, only the top 200 words were used. Please see Sec. 8 for some possible future research that could make APM more scalable. This experiment is meant only to be a preliminary experiment. Extensive experiments on larger datasets and exploration of the parameters of the model are significant areas of future work but are outside the scope of this paper since this paper focuses on model definition, contextualization and parameter estimation.

Two topic coherence metrics that have been shown to correlate with human annotators were used in this experiment. First, the UMass coherence metric introduced by Mimno et al. (2011) and further explored by Stevens & Kegelmeyer (2012), evaluates the intrinsic coherence of the generated topics by computing co-occurrence statistics from the training data. Letting each topic t be an *ordered* list of top m words $t = (v_1, \cdots, v_m)$, the UMass coherence metric is defined as follows:

$$\text{coh}_{\text{UMass}}(t) = \sum_{a=2}^{m} \sum_{b=1}^{a-1} \ln \left(\frac{D(v_a, v_b) + \epsilon}{D(v_b)} \right)$$

where $D(v_a, v_b)$ and $D(v_b)$ are the co-occurrence and marginal co-occurrence statistics in the training corpus and

ϵ is introduced to avoid taking the log of zero. Loosely, this measures how well the model fits the training data.

The second coherence metric, Pointwise Mutual Information (PMI), was introduced by Newman et al. (2010) and has also been shown to correlate with human judgments of topic coherence. The PMI metric is defined as follows:

$$\text{coh}_{\text{PMI}}(t) = \frac{1}{m(m-1)} \sum_{a=1}^{m} \sum_{b \neq a} \ln \left(\frac{\Pr(v_a, v_b) + \epsilon}{\Pr(v_a) \Pr(v_b)} \right)$$

where $\Pr(v_a, v_b)$ and $\Pr(v_a)$ are computed from the local co-occurrence statistics in a sliding window of an external corpus. To compute the probabilities for the PMI metric, a recent dump of Wikipedia was used with a sliding window of 20 words.

Stevens & Kegelmeyer (2012) explored the importance of ϵ in both coherence metrics and, in light of this, ϵ was set to 10^{-12} as it was in (Stevens & Kegelmeyer, 2012). For simplicity, a set of topic words was chosen based on the θ parameter of the PMRF. In general, because a PMRF contains information about word dependencies, the best words could be chosen using some sort of graph density algorithm such as the one described in (Yuan & Zhang, 2013).

LDA was trained using the MATLAB Topic Modeling Toolbox,[8] which uses the Gibbs sampling method described in (Steyvers & Griffiths, 2007) and was run for 5000 iterations. For both LDA and APM, the hyperparameters α and β were set to $\alpha = 200/p$ and $\beta = 50/k$ respectively as suggested by the documentation of the toolbox. For APM, the parameter λ was set near 10^{-7}, which was chosen so that there would be some edges in the initial iterations—however, as discussed in the following section, the final converged APM solution did not have any edges. Because LDA and APM might perform differently with different number of topics k, three values for $k = \{5, 10, 15\}$ were evaluated for both LDA and APM. We might expect that LDA will need more topics to model the data because LDA assumes independence and hence has less parameters.

6.3. Results from Preliminary Experiments

Results for APM and LDA on the 20 Newsgroup dataset can be seen in Fig. 5. The topic words chosen for both models when $k = 10$ can be found in the appendix. APM seems to outperform LDA in this simple 200-word experiment for the UMass metric whereas APM is only comparable to LDA for the PMI metric. Because APM directly models the co-occurrence of words, it seems reasonable that APM would perform better in the UMass coherence metric, which focuses on model fit and internal coherence. However, on the PMI coherence metric, APM only seems to do at least as well as LDA suggesting APM should probably be studied through an extensive experimental comparison as suggested in the future works section. In addition, as expected because of model complexity, LDA seems to perform better with larger k. APM's performance seems to degrade as the k increases which might be due to the estimation procedure not fully converging given the high model complexity.

Interestingly, though some of the initial iterations of APM included many word dependencies, the final iterations gave admixtures of independent Poisson MRFs. This is likely due to the fact that only a small text collection was used, and therefore, the power of dependencies is not needed to appropriately model the data. This result could also be caused by the choice of the regularization parameter λ. Though λ was chosen by simply trying several values, an important area for future research is how to choose λ appropriately for the application. However, this result shows that APM can effectively model words even in the independent case and can perform competitively with LDA as expected by the equivalence discussion in Sec. 2.2.

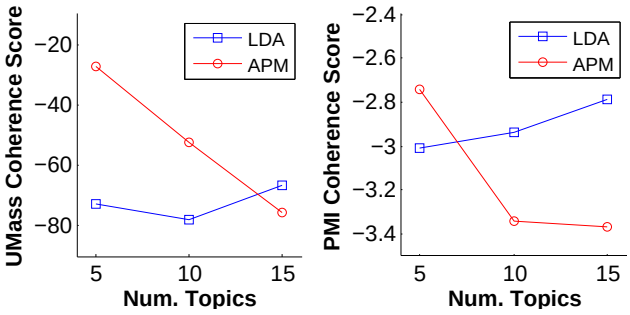


Figure 5: APM seems to outperform LDA in a simple 200-word experiment when the number of topics is small but is only comparable to LDA for a larger number of topics. (Median score is shown to reduce the effect of outliers.)

7. Related Work

Many probabilistic models for documents have been constructed using the multinomials. Nigam et al. (2000) introduced a mixture of multinomials to model document collections, and later, Hofmann (1999) proposed an admixture of multinomials called Probabilistic Latent Semantic Analysis (PLSA). This model was followed by the very successful Latent Dirichlet Allocation (LDA) topic model proposed by Blei et al. (2003) that added priors to the distributions as well as provided a more coherent framework for extending the model. There have been numerous extensions of LDA that incorporate other knowledge such as author information (Steyvers et al., 2004), time (Blei & Lafferty, 2006) and topic dependency (Blei & Lafferty, 2005). However, none of these models considers dependencies between words since the base distribution is multinomial.

[8] www.psiexp.ss.uci.edu/research/programs_data/toolbox.htm

Replicated Softmax (Hinton & Salakhutdinov, 2009) uses a restricted Boltzmann machine (RBM) with parameter biases to create a generative model for word count vectors. The hidden layer is binary-valued and allows for topic parameters to be mixed in the canonical parameter space (similar to APM). *Wordfish* (Slapin & Proksch, 2008) is a Poisson IRT (Item Response Theory) model that attempts to characterize the latent position of a political party based on political manifestos (e.g. determining left or right wing political views). Though *Wordfish* also adds fixed-effect parameters, this model is similar to an independent APM model with $k = 2$ (i.e. only one latent dimension). Both *Replicated Softmax* and *Wordfish* significantly differ from APM because they do not consider word dependencies.

Sparse Word Graphs (Nallapati et al., 2007) attempts to create graph visualizations of the topics by combining LDA and Bernoulli MRFs (Ising model) in a two-stage approach. First, LDA is used to estimate the topic assignments for every word in the corpus. Then, these topic assignments are used to train k independent Bernoulli MRFs for each topic. To transform the LDA output into the input for the Bernoulli MRF estimation algorithm, binary word-document matrices are constructed for each topic based on the LDA topic assignments. Though this leads to a graph over words for each topic, one major difference with APM is that this two-stage method is not a unified probabilistic model but rather two separate probability models. Another significant difference is that *Sparse Word Graphs* estimates simpler Bernoulli MRFs instead of PMRFs as in APM.

In (Hu et al., 2011), users can interactively add soft constraints to LDA so that the probability of the words in the constraint set will tend to be similar (e.g. either all low or all high probability). The soft dependency is added through a latent constraint variable and only provides indirect dependence of words rather than direct dependence between words as in APM. Another difference is that these constraints can only be supplied as user-specified disjoint groups of words rather than automatically-discovered arbitrary structure as in APM.

Collins et al. (2001) develop a generalization of PCA by using the likelihood of exponential families as the loss function instead of squared loss—which would correspond to Gaussian errors. While exponential PCA is related to admixtures, it does not place constraints on the admixture weights but rather allows them to be arbitrary real numbers. This is analagous to the difference between SVD and constrained non-negative matrix factorization (NMF).

8. Future Work

Scalability Because APM allows for dependencies between all words, the model is quadratic in the number of words p. Therefore, scalability could be an significant obstacle to overcome in future research. However, since sparsity is assumed on the dependencies, the effective number of parameters can be reduced significantly. For Gaussian MRFs, this fact was recently exploited by Hsieh et al. (2013) to find the dependency parameters—the precision matrix in this case—even for very high dimensional data. We believe that some of the intuitions in (Hsieh et al., 2013) could be employed to effectively scale APM.

Empirical Study Because this paper focuses on model definition, extensive empirical experiments were not conducted. In future work, several parameter settings such as the choice of hyperparameters (α, β) or the regularization parameter λ could be evaluated or automatically fitted in a Bayesian manner. Also, extensive user studies on the effectiveness of visualizing the topics could be conducted to consider the usefulness of this model in real-world settings.

9. Conclusion

This work lays the foundation for a new class of topic models based on an admixture of Poisson MRFs that can model dependencies between words unlike all previous topic models that assume word independence. Independent Poisson MRFs are shown to generalize the conditional distributions of LDA, which thus suggests that APM subsumes the expressive power of LDA and adds significantly greater modeling power than LDA. In addition to APM, a generalized class of admixture models is defined which opens the way for admixtures of any parametric distribution. For parameter estimation of this new model, a tractable method using the approximate posterior is explained. Finally, several experiments give evidence that APM can provide visually appealing and interpretable results as well as subsume the power of the LDA model. The development of APM opens up a whole new area of research with many interesting open questions in both theory (e.g. scalability, other admixtures, hyperparameter choice) and applications (e.g. visualization, user interaction, document exploration).

Acknowledgments

D. Inouye was supported by the NSF Graduate Research Fellowship via DGE-1110007. P. Ravikumar acknowledges support from NSF via IIS-1149803 and DMS-1264033, and ARO via W911NF-12-1-0390. I. Dhillon acknowledges support from NSF via CCF-1117055.

References

Beck, A. and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, January 2009. ISSN 1936-4954.

Bishop, Y., Fienberg, S., and Holland, P. Sampling models for discrete data. In *Discrete Multivariate Analysis: Theory and Practice*, chapter 13, pp. 435–456. Springer, 2007.

Blei, D. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, November 2012. ISSN 1053-5888.

Blei, D., Ng, A., and Jordan, M. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

Blei, D. M. and Lafferty, J. D. Correlated topic models. In *NIPS*, pp. 147–154, 2005.

Blei, D. M. and Lafferty, J. D. Dynamic topic models. In *ICML*, pp. 113–120, 2006. ISBN 1595933832.

Chen, Y. and Ye, X. Projection onto a simplex. *arXiv preprint arXiv:1101.6081*, pp. 1–7, 2011.

Collins, M., Dasgupta, S., and Schapire, R. E. A generalization of principal component analysis to the exponential family. In *NIPS*, pp. 617–624, 2001.

Gopalan, P., Hofman, J., and Blei, D. Scalable recommendation with poisson factorization. *arXiv preprint arXiv:1311.1704*, 2013.

Hinton, G. and Salakhutdinov, R. Replicated softmax: An undirected topic model. *NIPS*, pp. 1607–1614, 2009.

Hofmann, T. Probabilistic latent semantic analysis. In *Uncertainty in Artificial Intelligence (UAI)*, pp. 289–296. Morgan Kaufmann Publishers Inc., 1999.

Hsieh, C., Sustik, M. A., Dhillon, I. S., Ravikumar, P. K., and Poldrack, R. A. BIG & QUIC: Sparse inverse covariance estimation for a million variables. In *NIPS*, pp. 3165–3173, 2013.

Hu, Y., Boyd-Graber, J., and Satinoff, B. Interactive topic modeling. In *ACL*, pp. 248–257, 2011.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., and McCallum, A. Optimizing semantic coherence in topic models. In *EMNLP*, pp. 262–272, 2011.

Nallapati, R., Ahmed, A., Cohen, W., and Xing, E. Sparse word graphs: A scalable algorithm for capturing word correlations in topic models. *ICDM*, pp. 343–348, 2007.

Newman, D., Noh, Y., Talley, E., Karimi, S., and Baldwin, T. Evaluating topic models for digital libraries. In *ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pp. 215–224, 2010. ISBN 9781450300858.

Nigam, K., Mccallum, A. K., Thrun, S., and Mitchell, T. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2-3):103–134, 2000.

Parikh, N. and Boyd, S. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):123–231, 2013.

Pritchard, J. K., Stephens, M., and Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–59, June 2000. ISSN 0016-6731.

Ravikumar, P., Wainwright, M. J., and Lafferty, J. D. High-dimensional Ising model selection using l1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, June 2010. ISSN 0090-5364.

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011. ISSN 1935-7524.

Reisinger, J., Waters, A., Silverthorn, B., and Mooney, R. J. Spherical topic models. In *ICML*, pp. 903–910, 2010.

Slapin, J. and Proksch, S. A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, 52(3):705–722, 2008.

Stevens, K. and Kegelmeyer, P. Exploring topic coherence over many models and many topics. In *EMNLP-CoNLL*, pp. 952–961, 2012.

Steyvers, M. and Griffiths, T. Probabilistic topic models. In *Latent Semantic Analysis: A Road to Meaning*, pp. 424–440. 2007.

Steyvers, M., Smyth, P., Rosen-Zvi, M., and Griffiths, T. Probabilistic author-topic models for information discovery. In *KDD*, pp. 306–315, 2004. ISBN 1581138881.

Wainwright, M. J. and Jordan, M. I. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(12):1–305, 2008. ISSN 1935-8237.

Yang, E., Ravkiumar, P., Allen, G. I., and Liu, Z. Graphical models via generalized linear models. In *NIPS*, pp. 1367–1375, 2012.

Yang, E., Ravikumar, P., Allen, G., and Liu., Z. On poisson graphical models. In *NIPS*, pp. 1718–1726, 2013.

Yuan, X. and Zhang, T. Truncated power method for sparse eigenvalue problems. *JMLR*, 14:899–925, 2013.