THE UNIVERSITY OF TEXAS AT AUSTIN
Department of Computer Science
College of Natural Sciences

Admixture of Poisson MRFs (APM): A Topic Model with Word Dependencies

David Inouye, Pradeep Ravikumar, Inderjit Dhillon

Motivation

- Previous topic models cannot model intuitive **dependencies between words**. (e.g. if the word "classification" occurs, "supervised" is more likely to occur.)
- Several **topic coherence metrics** that correlate with human judgment primarily *test* for word dependence [Minmo et al. 2011, Newman et al. 2010].

Contributions

1. Introduce **Admixture of Poisson MRFs (APM)** (a new topic model that considers *word dependencies*)
2. Formalize **admixtures** (a *generalization* of previous topic models)
3. Define a **novel conjugate prior** for a Poisson MRF
4. Develop **APM parameter estimation** method using an approximate MAP estimator
5. Show some preliminary qualitative and topic coherence results

Overview

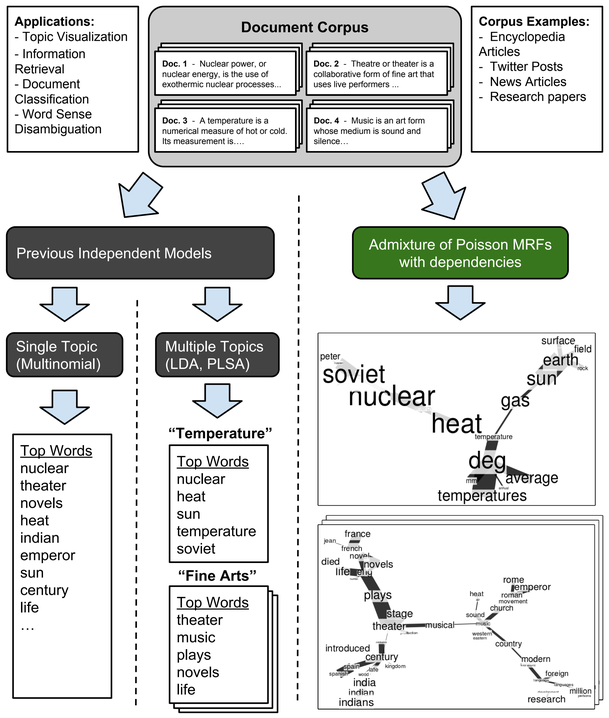


Figure: Previous topic models assume words are *independent* of each other and thus these previous models can only represent topics as a *list of words* ordered by frequency. However, our model, an admixture of Poisson MRFs, can model dependencies between words and hence can represent topics as a *graph over words*.

Generalized Admixtures

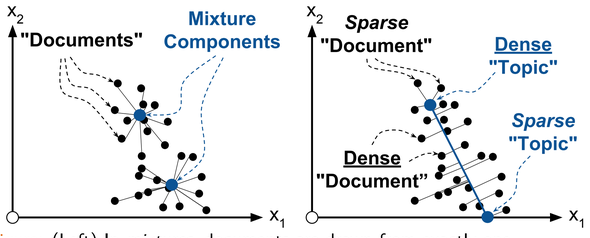


Figure: (Left) In *mixtures*, documents are drawn from exactly one component distribution. (Right) In *admixtures*, documents are drawn from a distribution whose parameters are a convex combination of component parameters.

The conditional distribution given the admixture weights and component distributions is merely the base distribution with parameters that are instance-specific mixtures of the component parameters:

$$\Pr_{\text{Admix.}}(\mathbf{x} \mid \mathbf{w}, \Phi) = \Pr_{\text{Base}}\left(\mathbf{x} \,\middle|\, \bar{\phi} = \Psi^{-1}\Big[\sum_{j=1}^{k} w_j \Psi(\phi^j)\Big]\right)$$

Examples of admixtures/topic models:

- PLSA [Hofmann, 1999] - An admixture of Multinomials
- LDA [Blei et al. 2003] - An admixture of Multinomials with Dirichlet priors
- Spherical Admixture Model (SAM) [Reisinger et al., 2010] - An admixture of Von-Mises Fisher distributions

Background: Poisson MRF (Multivariate Poisson)

By assuming that the conditional distribution of a variable x_s given all other variables $\mathbf{x}_{\setminus s}$ is a univariate Poisson, a joint Poisson distribution can defined (Yang 2012, 2013):

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

where $\boldsymbol{\theta} \in \mathbb{R}^p$ and $\Theta \in \{\mathbb{R}^{p\times p} : \text{diag}(\Theta) = 0\}$.
Node conditionals (i.e. the distribution of one word given all other words) are 1-D Poissons:

$$\Pr(x_s \mid \mathbf{x}_{-s}, \theta_s, \Theta_s) \propto \exp\{\underbrace{(\theta_s + \mathbf{x}^T\Theta_s)}_{\eta_{ls}} x_s - \ln(x_s!)\}.$$

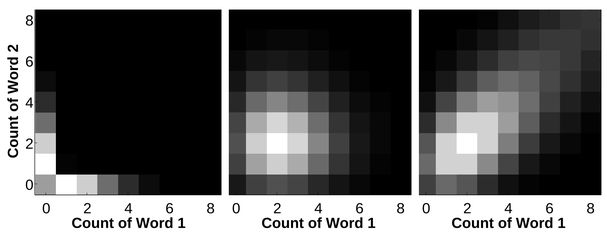


Figure: The densities of three 2D Poisson MRFs that show possible dependency structures between two words. Negative dependencies (left) suggest that two words *rarely* co-occur whereas positive dependencies (right) suggest that two words *often* co-occur.

Poisson MRFs in Context of LDA

LDA uses Multinomial distributions but if the parameter $N \sim \text{Poisson}(\tilde{x} = \sum_{s=1}^{p} x_s | \tilde{\lambda} = \sum_{s=1}^{p}\lambda)$, then the joint distribution is an independent Poisson model:[a]

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

Therefore, the topic-word distributions of LDA can be viewed as special cases of Poisson MRFs.

Novel Conjugate Prior for PMRF

Form of a conjugate prior:

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

where $\mathrm{A}(\boldsymbol{\theta}, \Theta)$ is the log partition function of a PMRF.[b]

- $\lambda\|\text{vec}(\Theta)\|_1$ term encourages **sparsity** in Θ and is similar to adding a Laplace prior on Θ.
- $\boldsymbol{\beta}$ can be viewed as adding **pseudo-counts** to the observations similar to a Dirichlet prior for a Multinomial.

Admixture of Poisson MRFs

An Admixture of Poisson MRFs (APM) is an *admixture* with Poisson MRFs as the component distributions:

$$\Pr_{\text{APM}}(\mathbf{x}, \mathbf{w}, \boldsymbol{\theta}^{1\ldots k}, \Theta^{1\ldots k}) = \Pr_{\text{PMRF}}\left(\mathbf{x} \,\middle|\, \bar{\boldsymbol{\theta}} = \sum_{j=1}^{k} w_j\boldsymbol{\theta}^j, \bar{\Theta} = \sum_{j=1}^{k} w_j\Theta^j\right)\Pr_{\text{Dir}}(\mathbf{w})\prod_{j=1}^{k}\Pr(\boldsymbol{\theta}^j, \Theta^j)$$

Parameter Estimation

Parameter estimation is done by optimizing the approximate posterior (i.e. using pseudo log-likelihood) which has an ℓ_1 constraint, which enforces sparse parameters:

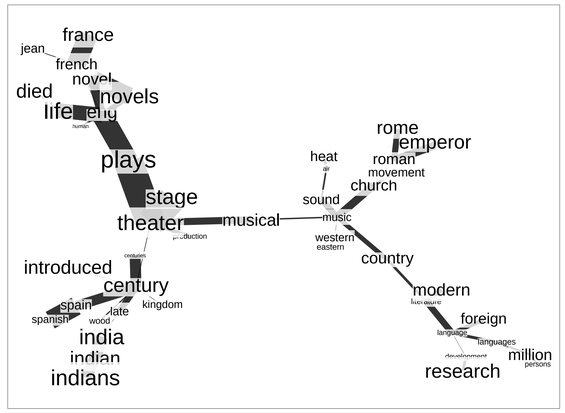
$$\underset{\mathrm{W}, \boldsymbol{\theta}^{1\ldots k}, \Theta^{1\ldots k}}{\arg\min} \underbrace{-\hat{\mathcal{L}}(\mathrm{W}, \boldsymbol{\theta}^{1\ldots k}, \Theta^{1\ldots k})}_{\text{differentiable}} + \underbrace{\delta_{\mathbb{W}}(\mathrm{W}) + \lambda\sum_{j=1}^{k}\|\Theta^j\|_1}_{\text{nonsmooth but convex}},$$

where $\hat{\mathcal{L}}$ is the pseudo log-likelihood and $\delta_{\mathbb{W}}(\mathrm{W})$ ensures that the weights are on the simplex. A proximal gradient method can be used to find a local minimum.

[a] Gopalan et al. (2013) recently introduced the connection between LDA and independent Poissons in the context of matrix factorization.

[b] $\lambda_{\boldsymbol{\theta}}\|\boldsymbol{\theta}\|_2^2$ and $\lambda\|\text{vec}(\Theta)\|_1$ needed for normalization of this prior distribution. In practice, $\lambda_{\boldsymbol{\theta}}$ can be set arbitrarily small and is thus ignored in subsequent discussion.

Qualitative Experiment



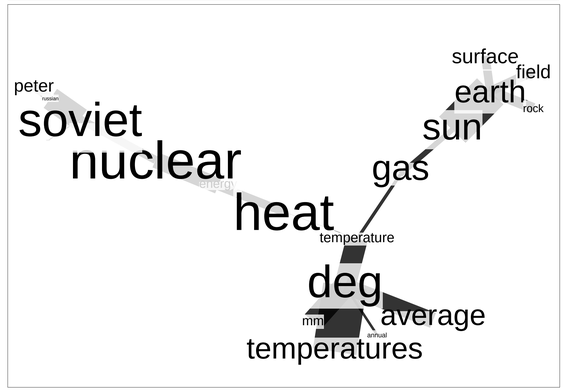


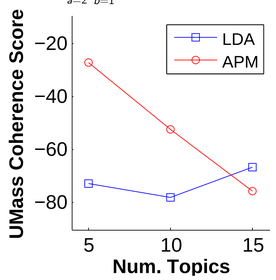
Figure: These APM topic visualizations ("Music and Fine Arts" and "Temperature") illustrate that PMRFs are much more intuitive than multinomials (as in LDA/PLSA), which can only be represented as a list of words. Word size signifies relative word frequency and edge width signifies the strength of word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

Topic Coherence Experiments

Dataset	# of Words	# of Documents
CMU 20 Newsgroup	200	18,846

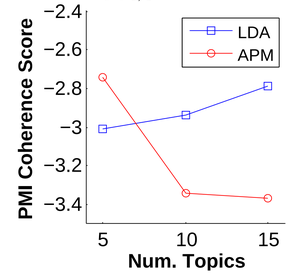
- UMass Coherence Metric [Minmo et al. 2011]

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



- Pointwise Mutual Info. [Newman et al. 2010]

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



For this preliminary experiment, APM seems to outperform LDA when the number of topics is small but is only comparable to LDA for a larger number of topics (Median scores shown).

{dinouye,pradeepr,inderjit}@cs.utexas.edu