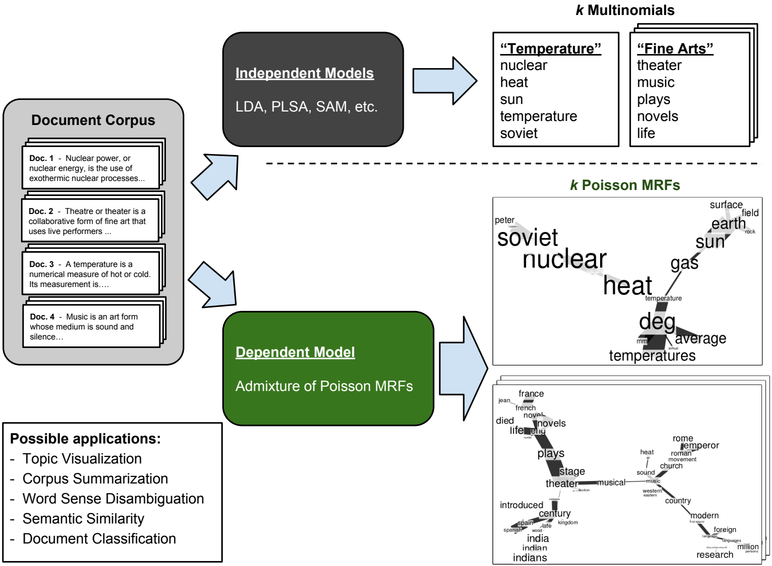
Admixture of Poisson MRFs (APM): A Topic Model with Word Dependencies

David Inouye*, Pradeep Ravikumar, Inderjit Dhillon

Tuesday, June 24, 2014

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

* Presenter



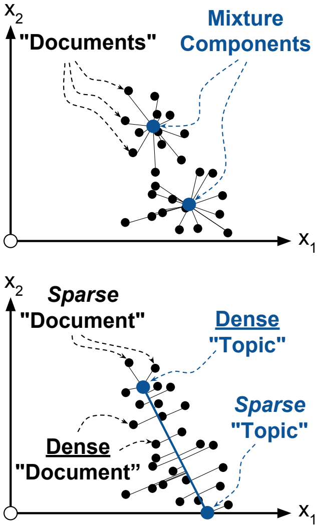
- Previous topic models assume independence between words.
- An Admixture of Poisson MRFs (APM), however, **explicitly models word dependencies**.

Main Contributions

1. Generalized Admixtures
2. (Background) Poisson MRF [Yang et al. 2012])
   - Poisson MRFs in the context of LDA
   - Novel conjugate prior for a Poisson MRF
3. Admixture of Poisson MRFs (APM)
4. Tractable MAP parameter estimation

Formalizing Generalized Admixtures

- **Mixtures** - Draws from *single* component distribution. (Top)
- **Admixtures** - Draws from a distribution whose parameters are a *convex combination* of component parameters. (Bottom)



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

- Examples of different Ψ

$$\Pr_{\text{Admix.}}(x \,|\, \mathbf{w}, \lambda^{1\ldots k}) = \Pr_{\text{Poiss.}}\left(x \,\middle|\, \bar{\lambda} = \sum_{j=1}^{k} w_j \lambda^j\right)$$

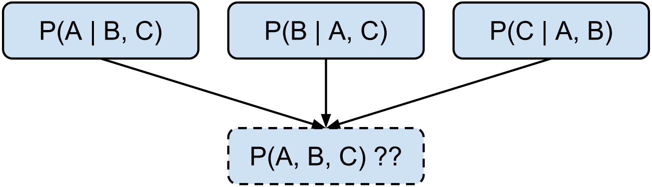
$$\Pr_{\text{Admix.}}(x \,|\, \mathbf{w}, \lambda^{1\ldots k}) = \Pr_{\text{Poiss.}}\left(x \,\middle|\, \bar{\lambda} = \exp\Big(\sum_{j=1}^{k} w_j \ln(\lambda^j)\Big)\right)$$

Examples of Admixture Models

1. LDA [Blei et al. 2003]
   - LDA is an admixture of Multinomials
     (i.e. Mult(p^1), Mult(p^2), $\cdots$, Mult(p^k))
   - Dirichlet prior over $p^{1\ldots k}$
2. Population Admixtures
   - Equivalent model to LDA in genetics [Pritchard et al. 2000]
   - *Admixture* term comes from genetics literature
   - Original ancestors of population correspond to "topics"
   - Individuals of a population correspond to "documents"
3. Spherical Admixture Model [Reisinger et al. 2010]
   - Von Mises-Fisher base distribution
     (an independent Gaussian analog on unit hypersphere)
   - Von Mises-Fisher priors

Background: Poisson MRFs [Yang et al., 2012]

If we assume the node conditional distributions are Poisson,



does there exist a **joint MRF distribution**
that has these conditionals?

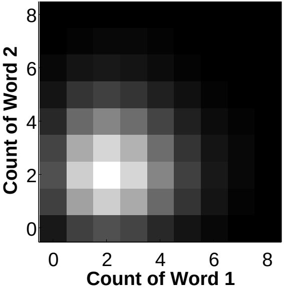
- Poisson MRF joint distribution:

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

- Node conditionals are 1-D Poissons:

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

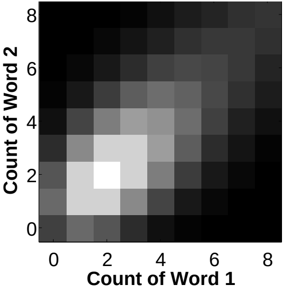
Independent PMRF



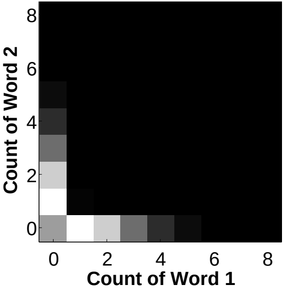
1. Each **conditional ("slice")** of a PMRF is 1-D Poisson.
2. **Distinct** from Gaussian MRF
3. Positive dependencies can model **word co-occurence**.[a]

[a] See [Yang et al. 2013] for SPMRF model that allows for positive dependencies.

Positive Dependency PMRF



Negative Dependency PMRF



Poisson MRFs in the 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:[1]

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

- Therefore, the topic-word distribution of LDA can be viewed as **a special case of a Poisson MRF**.

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

Novel conjugate prior for a Poisson MRF

- Form of a conjugate prior:

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

where $\mathrm{A}(\boldsymbol{\theta}, \Theta)$ is the log partition function of a PMRF.[2]

- $\lambda \|\mathrm{vec}(\Theta)\|_1$ term encourages **sparsity** in Θ (i.e. a Laplace prior on Θ).

- $\boldsymbol{\beta}$ can be viewed as adding **pseudo-counts** (similar to a Dirichlet prior for a Multinomial)

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

Admixture of Poisson MRFs (APM)

- Poisson MRF base distribution
- Priors
  - Dirichlet prior on admixture weights
  - Conjugate prior on component PMRFs

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

- Topics $\rightarrow$ graphs over words (from PMRF parameters)
- Documents $\rightarrow$ weights over topics (dimensionality reduction)

Parameter Estimation using Approximate Posterior

- Because the Poisson MRF likelihood does not have a closed-form solution, we approximate the likelihood with the **pseudo log-likelihood**:

$$\mathcal{L} \approx \hat{\mathcal{L}}(X \mid \mathrm{W}, \boldsymbol{\theta}^{1\ldots k}, \Theta^{1\ldots k})$$
$$= \sum_{i=1}^{n} \left[\sum_{s=1}^{p} \underbrace{\eta_{is} x_{is} - \ln(x_{is}!) - \mathrm{A}\left(\eta_{is}\right)}_{\text{Conditional Poisson log-likelihood}} \right],$$

where $\eta_{is} = \sum_{j=1}^{k} w_{ij} (\boldsymbol{\theta}_s^j + \mathbf{x}_i^T \Theta_s^j)$ is the canonical parameter of a univariate Poisson (i.e. $\lambda_{is} = \exp(\eta_{is})$).

Tractable MAP Parameter Estimation

- The approximate log posterior is:

$$\mathcal{P}(\mathrm{W}, \boldsymbol{\theta}^{1\ldots k}, \Theta^{1\ldots k} \mid X) \approx \hat{\mathcal{L}} \times \ln(\text{priors})$$

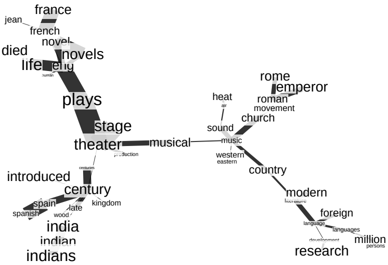
$$\propto \sum_{i=1}^{n} \left\{ \left[\sum_{s=1}^{p} \eta_{is} \underbrace{(x_{is} + \beta_s)}_{\text{psuedo-counts}} - (\gamma+1)\mathrm{A}(\eta_{is}) \right] + \underbrace{(\boldsymbol{\alpha} - 1)^T \ln(\mathbf{w}_i)}_{\text{Dirichlet prior}} \right\} - \underbrace{\sum_{j=1}^{k} \lambda \|\Theta^j\|_1}_{\substack{\ell_1 \text{ penalty} \\ \text{for sparsity}}}$$

- A MAP parameter estimate can be computed by the following:

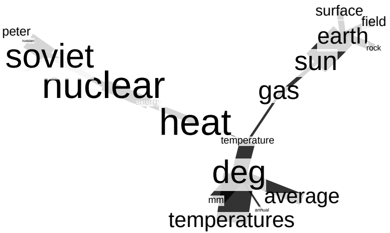
$$\underset{\mathrm{W}, \boldsymbol{\theta}^{1\ldots k}, \Theta^{1\ldots k}}{\arg\min} \underbrace{- f(\mathrm{W}, \boldsymbol{\theta}^{1\ldots k}, \Theta^{1\ldots k})}_{\text{differentiable}} + \underbrace{\delta_{\mathbb{W}}(\mathrm{W}) + \lambda \sum_{j=1}^{k} \|\Theta^j\|_1}_{\text{nonsmooth but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

- A proximal gradient method can used

Qualitative Experiment



(a) "Music and Fine Arts"



(b) "Temperature"

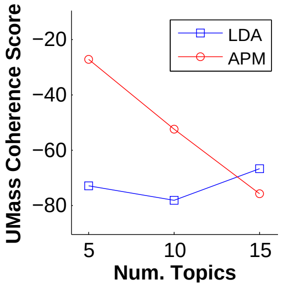
- APM topic visualizations are intuitive
  (versus list of word representations in LDA/PLSA)
- Explicit structure such as word chains
  (e.g. [*plays* – *theater* – *musical* – *music*])
- Interesting central words
  (e.g. [*theater*], [*music*], [*temperature*])

Preliminary Coherence Experiments

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

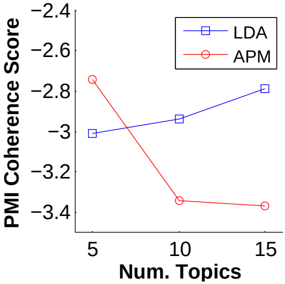
- UMass Coherence Metric [Minmo et al. 2011]

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



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

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



Summary

- Introduced Admixture of Poisson MRFs that explictily models word dependencies
- Formalized a class of models called admixtures that generalizes previous topic models
- Provided tractable MAP parameter estimation
- Showed preliminary results on datasets

Future Work

- *Scalability*
  - Obvious concern since number of parameters is $O(p^2)$
  - Faster, parallel parameter estimation algorithm (promising initial work on this)
- *Empirical Experiments*
  - Evaluate semantic meaningfulness of edges in PMRF graph (promising initial work on this)
  - Word Sense Disambiguation (WSD)
  - Document classification
- *Visualization*
  - Visualize topics
  - Visualize documents
  - Visual information retrieval

Thanks for listening!

- Blei, D., Ng, A., and Jordan, M. Latent dirichlet allocation. *JMLR*, 3:993-1022, 2003.
- Mimno, D., Wallach, H. M., Talley, E., Leenders, M., and McCallum, A. Optimizing semantic coherence in topic models. In *EMNLP*, pp. 262-272, 2011.
- Newman, D., Noh, Y., Talley, E., Karimi, S., and Baldwin, T. Evaluating topic models for digital libraries. In *Proc. of ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pp. 215-224, 2010.
- Pritchard, J. K., Stephens, M., and Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945-59, June 2000.
- Reisinger, J., Waters, A., Silverthorn, B., and Mooney, R. J. Spherical topic models. In *ICML*, pp. 903-910, 2010.
- Yang, E., Ravikumar, 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.

Thanks for listening!

Document Corpus

Doc. 1 - Nuclear power, or nuclear energy, is the use of exothermic nuclear processes...

Doc. 2 - Theatre or theater is a collaborative form of fine art that uses live performers ...

Doc. 3 - A temperature is a numerical measure of hot or cold. Its measurement is….

Doc. 4 - Music is an art form whose medium is sound and silence…

Independent Models

LDA, PLSA, SAM, etc.

***k* Multinomials**

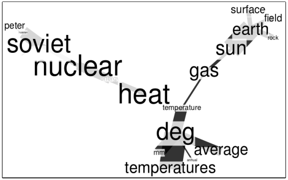
"Temperature"
nuclear
heat
sun
temperature
soviet

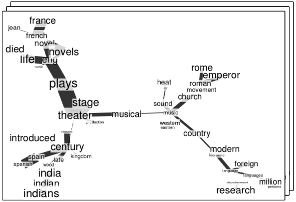
"Fine Arts"
theater
music
plays
novels
life

Dependent Model

Admixture of Poisson MRFs

***k* Poisson MRFs**





Possible applications:
- Topic Visualization
- Corpus Summarization
- Word Sense Disambiguation
- Semantic Similarity
- Document Classification