Multi-Task Learning of Component Strengths in Non-Conjugate Admixture Models

Bryan Feeney

Read this presentation online at:

http://tinyurl.com/bryansphd2014

bryan.feeney@ucl.ac.uk

October , 2014

Research

Three Components

1. Multi-Task Learning
2. Admixture Modelling ("Topic Models")
3. Local Variational Bounds

Research

Three Components

1. Multi-Task Learning
2. Admixture Modelling ("Topic Models")
3. Local Variational Bounds

http://tinyurl.com/bryansphd2014

Research

Three Components

1. **Multi-Task Learning**
2. Admixture Modelling ("Topic Models")
3. Local Variational Bounds

Multi-Task Learning

Situation: Making many predictions from the same data

1. Predict exam scores in L subjects for several children[12]
2. Predict customers affinity to L observed aspects of a product[1]
3. Propose image captions by predicting $p(word|image)$ from image features for L words (in this case $L > 10,000$)[2]

Multi-Task Learning

Situation: Making many predictions from the same data

1. Predict exam scores in L subjects for several children[12]
2. Predict customers affinity to L observed aspects of a product[1]
3. Propose image captions by predicting $p(word|image)$ from image features for L words (in this case $L > 10,000$)[2]

Problem: Can we improve performance by transferring knowledge between tasks[15]

1. Learn correlations between tasks.
2. Learn a low-rank projection of the tasks themselves
3. Learn structure of features by how we use them - via regularization

Regularization

Learn L vectors w_l. How to *transfer* knowledge from inferring $w_1, \ldots, w_{l-1}$ to the task of inferring w_l

Regularization

Learn L vectors w_l. How to *transfer* knowledge from inferring $w_1, \ldots, w_{l-1}$ to the task of inferring w_l

$$y_{nl}|w_l \sim \mathcal{N}\left(w_l^\top x_n, \sigma^2 I\right) \qquad w_l \sim \mathcal{N}\left(\mathbf{0}, \alpha^2 I\right)$$

Regularization

Learn L vectors w_l. How to *transfer* knowledge from inferring $w_1, \ldots, w_{l-1}$ to the task of inferring w_l

Bayesian approach - learn the prior[1]

$$y_{nl}|w_l \sim \mathcal{N}\left(w_l^\top x_n, \sigma^2 I\right) \qquad w_l \sim \mathcal{N}(m, \Sigma)$$

$$m|\Sigma \sim \mathcal{N}\left(m_0, \frac{1}{\lambda}\Sigma\right) \qquad \Sigma \sim \mathcal{W}^{-1}(\Sigma_0, \nu)$$

Regularization

Learn L vectors w_l. How to *transfer* knowledge from inferring $w_1, \ldots, w_{l-1}$ to the task of inferring w_l

Low-Rank Projections of the Feature Space

$$y_{nl} | w_l \sim \mathcal{N}\left(w_l^\top x_n, \sigma^2 I\right)$$

$$w_l | z_l \sim \mathcal{N}\left(U z_l + m, \alpha^2 I\right)$$

$$z_l \sim \mathcal{N}\left(\mathbf{0}, I\right)$$

$$\implies w_l \sim \mathcal{N}\left(m, \alpha^2 I + U U^\top\right)$$

Regularization

Learn L vectors w_l. How to *transfer* knowledge from inferring $w_1, \ldots, w_{l-1}$ to the task of inferring w_l

Exploit sparsity using ARD priors

$$y_{nl} | w_l \sim \mathcal{N}\left(w_l^\top x_n, \sigma^2 I\right) \qquad w_l \sim \mathcal{N}\left(\mathbf{0}, \mathrm{diag}\left(\boldsymbol{\alpha}\right)\right)$$

$$\alpha_f \sim \mathcal{G}\left(a, b\right)$$

Regularization

Learn L vectors w_l. How to *transfer* knowledge from inferring $w_1, \ldots, w_{l-1}$ to the task of inferring w_l

Exploit sparsity using Clustered ARD priors[6]

$$y_{nl} | w_l \sim \mathcal{N}\left(w_l^\top x_n, \sigma^2 I\right) \qquad w_l | z_l \sim \mathcal{N}\left(\mathbf{0}, \mathrm{diag}\left(\boldsymbol{\alpha}_{z_l}\right)\right)$$

$$z_l \sim \mathcal{M}\left(\boldsymbol{\theta}, 1\right) \qquad \alpha_{fk} \sim \mathcal{G}\left(a, b\right)$$

$$\boldsymbol{\theta} \sim \mathcal{D}\left(\beta\right)$$

Regularization

Learn L vectors w_l. How to *transfer* knowledge from inferring $w_1, \ldots, w_{l-1}$ to the task of inferring w_l

This is all just hierarchical Bayeisan modelling

However analogous methods exist in error-optimisation approaches to machine learning which learn regularization functions instead of priors.

- Low-Rank projections in [3]
- Heterogeneous sparsity in [4]

Research

Three Components

1. Multi-Task Learning
2. **Admixture Modelling ("Topic Models")**
3. Local Variational Bounds

Mixture Models versus Admixture Models

Classical Mixture Model of Text - One topic per document

$$\boldsymbol{\theta} \sim \mathcal{D}(\alpha) \qquad z_d \sim \mathcal{M}(\boldsymbol{\theta}, 1) \qquad w_{dn} \sim \mathcal{M}(\phi_{z_d}, 1)$$

Where each of the component vocabularies is drawn $\phi_k \sim \mathcal{D}(\beta)$.

Mixture Models versus Admixture Models

Classical Mixture Model of Text - One topic per document

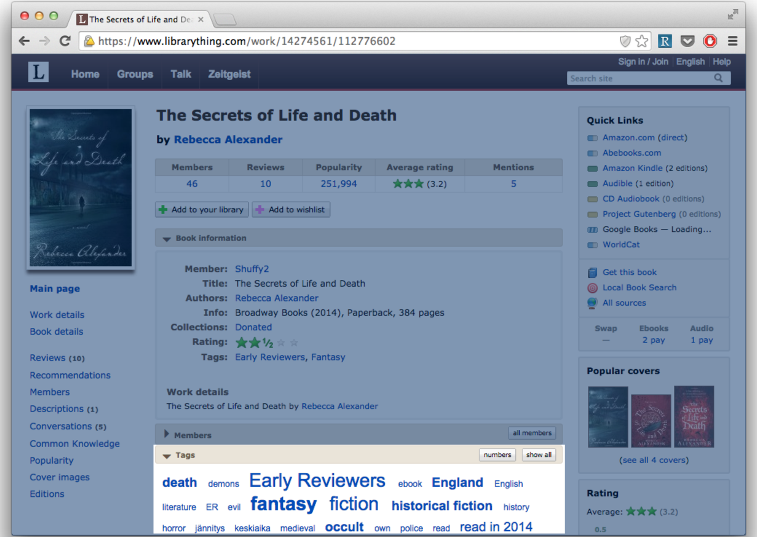
$$\boldsymbol{\theta} \sim \mathcal{D}(\alpha) \qquad z_d \sim \mathcal{M}(\boldsymbol{\theta}, 1) \qquad w_{dn} \sim \mathcal{M}(\phi_{z_d}, 1)$$

Where each of the component vocabularies is drawn $\phi_k \sim \mathcal{D}(\beta)$.

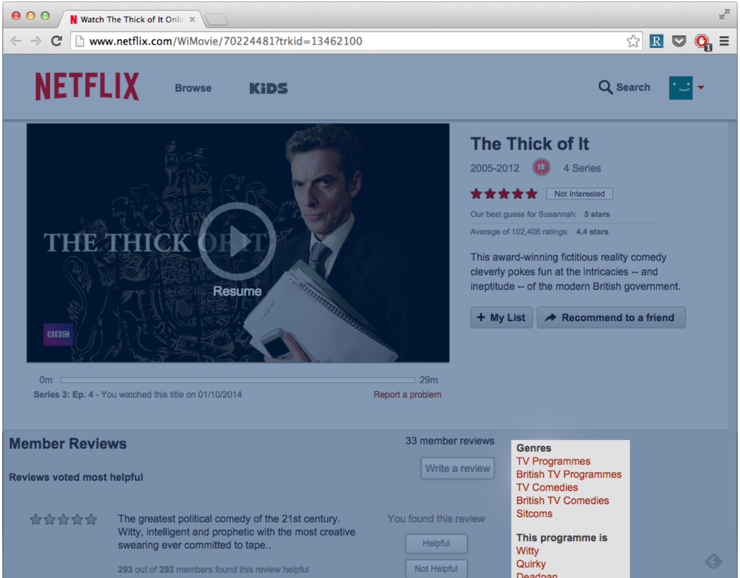
Mixture models struggle to generalise.

- Fewer clusters mean coarser estimates of cluster centroids
- But more clusters mean fewer datapoints per cluster, and thus sparser estimates of cluster centroids (for text), due to the assumption of one cluster per document

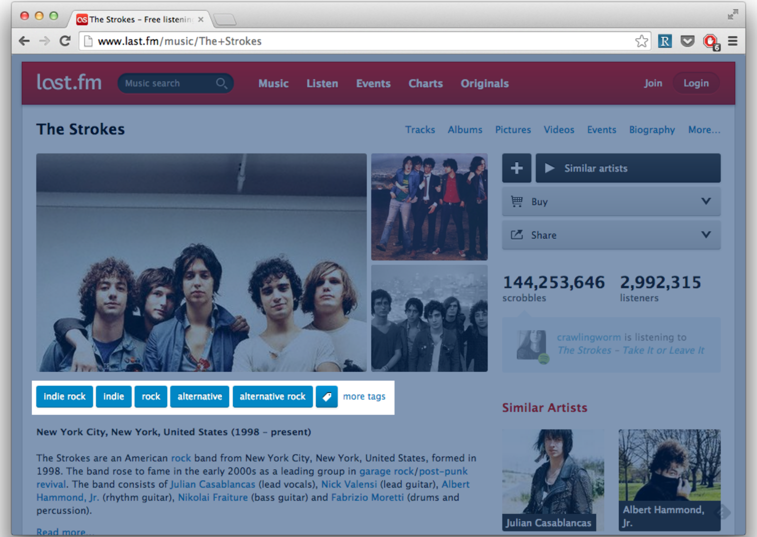
Mixture Models versus Admixture Models



Mixture Models versus Admixture Models



Mixture Models versus Admixture Models



Mixture Models versus Admixture Models

Classical Mixture Model of Text - One topic per document

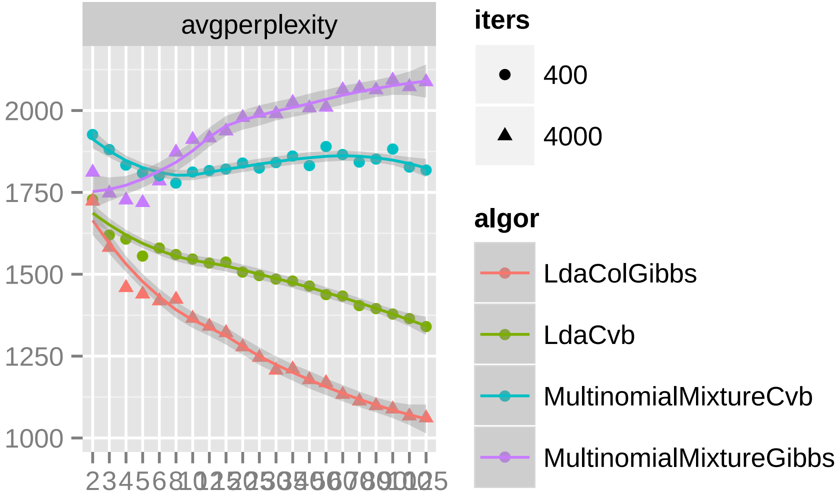
$$\boldsymbol{\theta} \sim \mathcal{D}(\alpha) \qquad z_d \sim \mathcal{M}(\boldsymbol{\theta}, 1) \qquad w_{dn} \sim \mathcal{M}(\phi_{z_d}, 1)$$

Where each of the component vocabularies is drawn $\phi_k \sim \mathcal{D}(\beta)$.

Admixture models assign a *mixture* of topics to each document, In the case of text the “topic-model” implementation[10] assigns a single topic to each word w_{dn}.

$$\boldsymbol{\theta}_d \sim \mathcal{D}(\alpha) \qquad z_{dn} \sim \mathcal{M}(\boldsymbol{\theta}_d, 1) \qquad w_{dn} \sim \mathcal{M}(\phi_{z_{dn}}, 1)$$

Mixture Models versus Admixture Models



Topic Models: Multinomial Admixtures

Five broad areas of research

- Richer Observation Models
  - Language Models[41][44][26]
  - Alternative distrbutions such as logistic-normal[9][17] or von Mises - Fisher[34]

Topic Models: Multinomial Admixtures

Five broad areas of research

- Richer Observation Models
- Alternatives to the Dirchlet prior
  - The "Logistic Normal" prior for the Correlated Topic Model[7]

$$\boldsymbol{\eta}_d \sim \mathcal{N}(\mu, \Sigma)$$
$$\theta_{dk} = \sigma_k(\boldsymbol{\theta}_d) = \frac{\exp(\eta_{dk})}{\sum_j \exp(\eta_{dj})}$$

  - Same quality of model fit with fewer topics

Topic Models: Multinomial Admixtures

Five broad areas of research

- Richer Observation Models
- Alternatives to the Dirchlet prior
- Bayesian Non-Parametrics to estimate Topic Counts
    - Hierarchical Dirichlet Processes[38]
    - Discrete Infinite Logistic Normal Distribution[32]

Topic Models: Multinomial Admixtures

Five broad areas of research

- Richer Observation Models
- Alternatives to the Dirchlet prior
- Bayesian Non-Parametrics to estimate Topic Counts
- Scalable Implementations:
    - Gibbs Sampling[33] and Collapsed Gibbs Sampling[19]
    - Variational[10] inference and Collapsed Variational[39][21] inference
    - MAP and other approximations[5]
    - Distributed "Big Data" approaches[37][31][16]
    - Optimised online approaches[22][23][29]

Topic Models: Multinomial Admixtures

Five broad areas of research

- Richer Observation Models
- Alternatives to the Dirchlet prior
- Bayesian Non-Parametrics to estimate Topic Counts
- Scalable Implementations:
- Use of Covariates x_d
  - Ad-hoc: Time[43], author[28], region[17]
  - "Downstream" Models[8][35][40]

$$p(w_d, x_d) = \int_{\boldsymbol{\theta}_d} p(w_d|\boldsymbol{\theta}_d)p(x_d|\boldsymbol{\theta}_d)p(\boldsymbol{\theta}_d) \tag{1}$$

  - "Upstream" Models:[30]

$$p(w_d|x_d) = \int_{\boldsymbol{\theta}_d} p(w_d|\boldsymbol{\theta}_d)p(\boldsymbol{\theta}_d|x_d) \tag{2}$$

.

Topic Models: Multinomial Admixtures

Five broad areas of research

- Richer Observation Models
- Alternatives to the Dirchlet prior
- Bayesian Non-Parametrics to estimate Topic Counts
- Scalable Implementations:
- Use of Covariates x_d

LDA also linked to multinomial PCA[14] and Non-Negative Matrix Factorization[18]

Research

Three Components

1. Multi-Task Learning
2. Admixture Modelling ("Topic Models")
3. **Local Variational Bounds**

Local Variational Bounds

The softmax transformation

$$\eta_d \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma) \qquad z_{dn} \sim \mathcal{M}(\boldsymbol{\sigma}(\boldsymbol{\eta}_d), 1) \tag{1}$$

Problems

- Prior (a Gaussian) is not conjugate to the likelihood (a multinomial mixture)
- No analytic form for the posterior.

Approach is to approximate the likelihood using bounds

- Classic approach is a global bound over the entire likelihood (e.g. the Laplace approximation)
- A better fit, and more flexibility, can be obtained by using *local* bounds over the likelihood of each document.

Local Variational Bounds

- Global Bounds: Laplace, Delta Method[42]
- Quadratic Bounds: Bohning[11], Bouchard[13]
- Other more recent bounds
    - Piecewise Quadratic Bounds[27]
    - Tilted Bound[25]
    - Stick-breaking bound[24]

Research

Use-Cases: Microtexts

- Tweets: Predict text given user features (username, time tweet was posted). Notably predict absent words such as hashtags.
- Image captions: Predict caption given image features.

Incorporate multi-task learning into an upstream topic model

Partial Model

The Model

$$\eta_d \sim \mathcal{N}(Ax_d, \Sigma) \qquad \boldsymbol{\theta}_d = \boldsymbol{\sigma}(\eta_d)$$

$$z_{dn} \sim \mathcal{M}(\boldsymbol{\theta}_d, 1) \qquad w_{dn} \sim \mathcal{M}(\phi_{z_{dn}}, 1)$$

Partial Model

The Model

$$\eta_d \sim \mathcal{N}(Ax_d, \Sigma) \qquad \theta_d = \sigma(\eta_d)$$

$$z_{dn} \sim \mathcal{M}(\theta_d, 1) \qquad w_{dn} \sim \mathcal{M}(\phi_{z_{dn}}, 1)$$

Prior over A?

Matrix-Variate Priors

Matrix-Variate Normal Distribution [20] :

$$A \sim \mathcal{N}(M, \Omega, \Sigma) \implies \mathrm{vec}(A) \sim \mathcal{N}(\mathrm{vec}(M), \Sigma \otimes \Omega)$$

Matrix-Variate Priors

Matrix-Variate Normal Distribution [20] :

$$A \sim \mathcal{N}(M, \Omega, \Sigma) \implies \mathrm{vec}(A) \sim \mathcal{N}(\mathrm{vec}(M), \Sigma \otimes \Omega)$$

$$\ln p(W) = -\frac{FL}{2}\ln 2\pi - \frac{F}{2}\ln|\Omega| - \frac{L}{2}\ln|\Sigma| - \frac{1}{2}\mathrm{etr}\left(\Sigma^{-1}(W-M)\Omega^{-1}(W-M)^\top\right)$$

where $\mathrm{etr}(X) = \exp(\mathrm{tr}(X))$

Matrix-Variate Priors

Matrix-Variate Normal Distribution [20] :

$$A \sim \mathcal{N}(M, \Omega, \Sigma) \implies \mathrm{vec}(A) \sim \mathcal{N}(\mathrm{vec}(M), \Sigma \otimes \Omega)$$

$$V \sim \mathcal{N}(0, I, I) \qquad A|V \sim \mathcal{N}(UZ, I, \Sigma)$$

$$A \sim \mathcal{N}\left(0, I + UU^\top, \Sigma\right)$$

See also [2] for examples of matrix-variate normal priors.

Full Model

The Model

$$V \sim \mathcal{N}(0, I, I) \qquad A|V \sim \mathcal{N}(UV, I, \Sigma) \tag{2}$$

$$\boldsymbol{\eta}_d \sim \mathcal{N}(Ax_d, \Sigma) \qquad \boldsymbol{\theta}_d = \boldsymbol{\sigma}(\boldsymbol{\eta}_d) \tag{3}$$

$$z_{dn} \sim \mathcal{M}(\boldsymbol{\theta}_d, 1) \qquad w_{dn} \sim \mathcal{M}(\phi_{z_{dn}}, 1) \tag{4}$$

Inference

- Bound the likelihood using quadratic bounds
- Use the variational bound with the mean-field approximation $q(V, A, \Theta, Z, \Phi) = q(V)q(A)q(\Theta)q(Z)q(\Phi)$
- Numerically eliminate $Z \in \mathbb{R}^{D \times N \times K}$ as it consumes huge amounts of computer memory.

Model

Two Datasets NIPS Papers from 1987 to 1999

- 682 Documents. Vocabulary of 12503 words
- Median document length of 1,532 words, total of 1,075,323 words across the corpus
- Features are: authors; citations; the year

Tweets from April to September 2013 (inclusive)

- 735,868 tweets from 572 users. Vocabulary of 82,698 words
- Median tweet length is 10 words, total of 7,272,228 word observations across the corpus
- Features are: authors; time at various granularities (hour, day, week, month)

Images are ongoing

Twitter Hashtags

- Our model predicts words according to $p(w_d|x_d)$
- So given a tweets features, we can generate words ourselves, instead of using the observed words
- We can generate hashtags
  - So given all tweets that do not have a given hashtag (like `#eurozone`)
  - In which tweet is it most likely to occur given the features

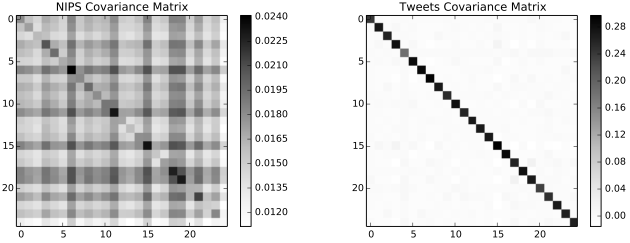
Twitter Hashtags

Hashtag	Tweet Words
#eurozone	#German finmin #Schaeuble argues #Karlsruhe may have no jurisdiction over #ECB measures.1st question by court's judges also focuses on this.
#usopen	Essa foi apenas a 2a vitoria de Gasquet nas oitavas de um Grand Slam. O frances alcanca sua primeira QFs em GS desde Wimbledon 2007 —
#f1	MT @f1paddockpass: ...a big shout out to the marshalls & volunteers here in Singapore. To all of you, heartfelt thanks @F1NightRace

Twitter Hashtags

Hashtag	**Tweet Words**
#tcot	#Delaware Senate rejects bill to keep guns away from unstable people deemed danger to others http://hrld.us/110xnGt #guncontrol #NRA
#beer	No-Li Brewhouse on track for 150% growth, expanding annual capacity to 10,000 barrels ...
#obamacare	@michellemalkin #feded supporters simply believe you won't challenge them. #stopcommoncore @afpne

Covariances



(a) NIPS (b) Tweets

Figure: Covariances over topics inferred using the Bohning implementation of the model for $K = 25$

What Next?

- Look at more complex matrix-variate priors, with low-rank approximations to both row and column covariances (the obvious approach doesn't work well)
- Large Scale inference (to 33 million tweets), potentially via SGD[23]
- Look at alternatives to the low-rank decomposition such as the use of matrix-variate Gaussian Scale models for sparsity[46]
- Look at richer downstream models for covariates such as collective-matrix[36] and tensor factorization[45].

Questions?

Reference I

[1] Greg M Allenby and Peter E Rossi.
Marketing models of consumer heterogeneity.
Journal of Econometrics, 89:57–59, 1999.

[2] C Archambeau, Shengbo Guo, and Onno Zoeter.
Sparse Bayesian multi-task learning.
NIPS, (1):41, 2011.

[3] A. Argyriou, C.A. Micchelli, M. Pontil, and Y. Ying.
A spectral regularization framework for multi-task structure learning.
NIPS, 2007.

Reference II

[4] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil.
Multi-Task feature learning.
In *Advances in neural information processing systems*, volume 19, page 41, January 2005.

[5] A Asuncion, M Welling, P Smyth, and YW Teh.
On smoothing and inference for topic models.
In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 27–34. AUAI Press, 2012.

[6] Bart Bakker and Tom Heskes.
Task Clustering and Gating for Bayesian Multitask Learning.
Journal of Machi, 4:83–99, 2003.

Reference III

[7] D Blei and J Lafferty.
Correlated topic models.
Advances in neural information processing systems, 18:147, 2006.

[8] David M Blei and Michael I Jordan.
Modeling annotated data.
In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval.*, pages 127–134. ACM, 2003.

[9] David M Blei and John D Lafferty.
Dynamic Topic Models.
In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120, 2006.

Reference IV

[10] David M. Blei, Andrew Y. Ng, and Michael I Jordan.
Latent dirichlet allocation.
The Journal of machine Learning research, 3:993–1022, 2003.

[11] Dankmar Bohning and Bruce G Lindsay.
Monotonicity of Quadratic-Approximation Algorithms.
Annals of the Institute of Statistical Mathematics, 40(4):641–663, 1988.

[12] E. Bonilla, K.M. Chai, and C. Williams.
Multi-task Gaussian process prediction.
Advances in Neural Information Processing Systems, 20:153–160, 2008.

Reference V

[13] Guillaume Bouchard.
Efficient Bounds for the Softmax Function and Applications to Approximate Inference in Hybrid models.
In *NIPS*, pages 1–9, 2007.

[14] Wray Buntine.
Variational Extensions to EM and Multinomial PCA.
Machine Learning: ECML 2002, pages 23–34, 2002.

[15] Rich Caruana.
Multitask Learning.
Machine Learning, 28:41–75, 1997.

[16] Jianfei Chen, Jun Zhu, Zi Wang, Xun Zheng, and Bo Zhang.
Scalable Inference for Logistic-Normal Topic Models.
In *NIPS*, pages 1–9, 2013.

Reference VI

[17] J Eisenstein and B O'Connor.
A latent variable model for geographic lexical variation.
Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 1277–1287, 2010.

[18] Prem Gopalan, Jake M Hofman, and David M Blei.
Scalable Recommendation with Poisson Factorization.
Technical report, ArXiv:1311.1704, 2013.

[19] TL Griffiths and M Steyvers.
Finding scientific topics.
Proceedings of the National Academy of Sciences of the United States of America, 101:5228–5235, 2004.

[20] A. K. Gupta and D. K Nagar.
Matrix Variate Distributions.
CRC Press, 1999.

Reference VII

[21] James Hensman, Magnus Rattray, and Neil D Lawrence.
Fast Variational Inference in the Conjugate Exponential Family.
pages 1–20, 2012.

[22] M Hoffman, Francis Bach, and DM Blei.
Online learning for latent dirichlet allocation.
Advances in Neural Information Processing Systems, pages 1–9, 2010.

[23] Matt Hoffman, DM Blei, Chong Wang, and John Paisley.
Stochastic variational inference.
Journal of Machine Learning Research, (2), 2012.

Reference VIII

[24] Mohamed Emtiyaz Khan, Shakir Mohamed, Benjamin M Marlin, and Kevin P Murphy.
A Stick-Breaking Likelihood for Categorical Data Analysis with Latent Gaussian Models.
In *15th International Conference on Artificial Intelligence and Statistics*, volume XX, 2012.

[25] David A Knowles and Thomas P Minka.
Non-conjugate Variational Message Passing for Multinomial and Binary Regression.
Advances in Neural Information Processing Systems, pages 1701–1709, 2011.

Reference IX

[26] Robert V Lindsey, William P Headden Iii, and Michael J Stipicevic.
A Phrase-Discovering Topic Model Using Hierarchical Pitman-Yor Processes.
In *EMNLP-CoNLL '12 Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, number July, pages 214–222, 2012.

[27] BM Marlin, ME Khan, and KP Murphy.
Piecewise Bounds for Estimating Bernoulli-Logistic Latent Gaussian Models.
ICML, 2011.

Reference X

[28] Andrew McCallum, Xuerwui Wang, and Andres Corrada-Emmanuel.
Topic and role discovery in social networks with experiments on enron and academic email.
Journal of Artificial Intelligence Research, 30(1):249–272, 2007.

[29] David Mimno, M Hoffman, and D Blei.
Sparse stochastic inference for latent dirichlet allocation.
In *International Conference on Machine Learning*, Edinburgh, Scotland, 2012.

Reference XI

[30] David Mimno and Andrew McCallum.
Topic Models Conditioned on Arbitrary Features with Dirichlet-multinomial Regression.
In *Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, June 2008.

[31] David Newman and Max Welling.
Distributed algorithms for topic models.
The Journal of machine Learning research, 10:1801–1828, 2009.

[32] John Paisley, Chong Wang, and David M. Blei.
The Discrete Infinite Logistic Normal Distribution.
Bayesian Analysis, 7(4):997–1034, December 2012.

Reference XII

[33] J K Pritchard, M Stephens, and P Donnelly.
Inference of population structure using multilocus genotype data.
Genetics, 155(2):945–59, June 2000.

[34] Joseph Reisinger, Austin Waters, and Raymond J Mooney.
Spherical Topic Models.
2010.

[35] Konstantin Salomatin, Y Yang, and A Lad.
Multi-field correlated topic modeling.
In *Siam International Conference on Data Mining (SDM)*, pages 628–637. SIAM, 2009.

Reference XIII

[36] Ajit P Singh and Geoffrey J Gordon.
Relational Learning via Collective Matrix Factorization Categories and Subject Descriptors.
In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 650–658. ACM, 2008.

[37] Alexander Smola and S Narayanamurthy.
An architecture for parallel topic models.
Proceedings of the VLDB Endowment, pages 703–710, 2010.

[38] Whye Teh, I Michael, Matthew J Beal, and David M Blei.
Hierarchical Dirichlet Processes.
Journal of the american statistical association, 101(476):1566–1581, 2006.

Reference XIV

[39] YW Teh, David Newman, and Max Welling.
A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation.
Advances in Neural Information Processing Systems, 19(1353), 2007.

[40] Seppo Virtanen, Yangqing Jia, Arto Klami, and Trevor Darrell.
Factorized multi-modal topic model.
In *Proceedings of the Twenty-Eighth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-12)*, pages 843–851. AUAI Press, 2012.

Reference XV

[41] Hanna M Wallach.
Topic modeling: beyond bag-of-words.
In *Proceedings of the 23rd international conference on Machine learning*, pages 977–984. ACM, 2006.

[42] Chong Wang and David M Blei.
Variational Inference in Nonconjugate Models 1 Introduction.
14:1005–1031, 2013.

[43] X Wang and A McCallum.
Topics over time: a non-Markov continuous-time model of topical trends.
Proceedings of the 12th ACM SIGKDD . . ., 2006.

Reference XVI

[44] Xuerui Wang, Andrew McCallum, and Xing Wei.
Topical n-grams: Phrase and topic discovery, with an application to information retrieval.
In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 697–702. IEEE, 2007.

[45] Liang Xiong, Xi Chen, Jeff Schneider, and Jaime G Carbonell.
Temporal Collaborative Filtering with Bayesian Probabilistic Tensor Factorization.
pages 211–222.

[46] Ming Yang and Yingming Li.
Multi-Task Learning with Gaussian Matrix Generalized Inverse Gaussian Model.
In *ICML*, volume 1, pages 423–431, 2013.