# The Discrete Infinite Logistic Normal Distribution for Mixed-Membership Modeling

John Paisley, Chong Wang and David Blei

Department of Computer Science
Princeton University, Princeton, NJ

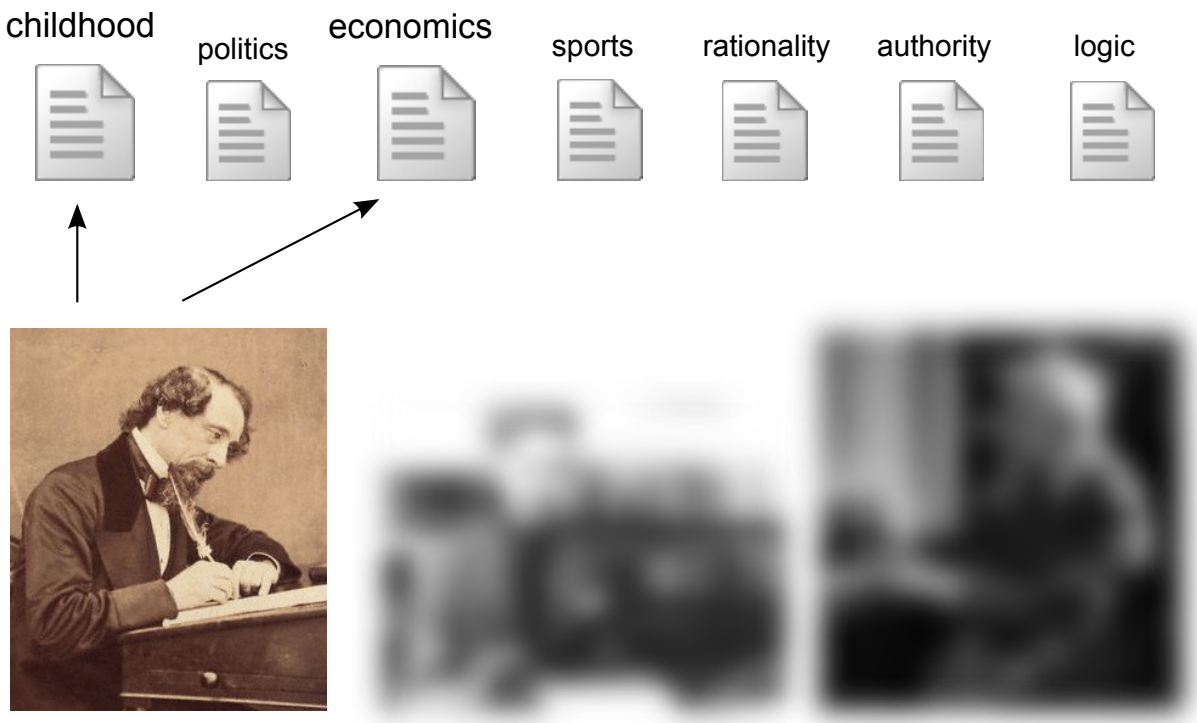
AISTATS 2011

# Motivation

- We develop a mixed-membership model that combines the advantages of the hierarchical Dirichlet process and the correlated topic model.
- We call the resulting prior the discrete infinite logistic normal. We use it as a topic model and derive a VB algorithm.

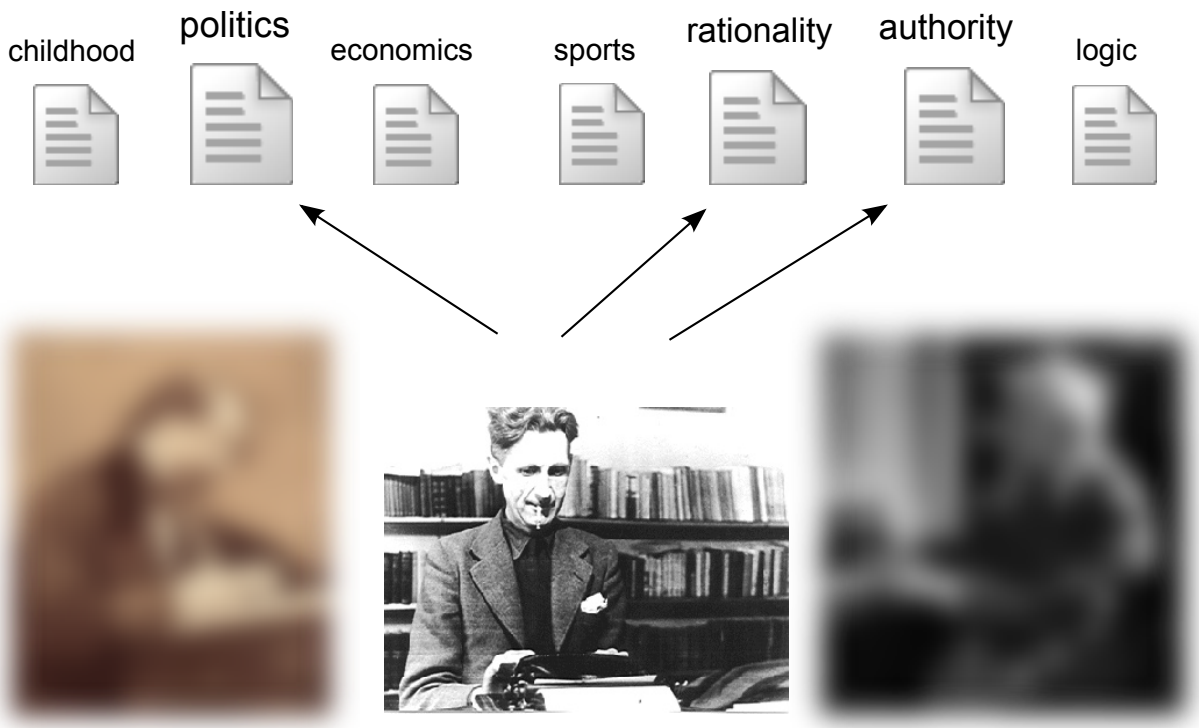

# Motivation

- We develop a mixed-membership model that combines the advantages of the hierarchical Dirichlet process and the correlated topic model.
- We call the resulting prior the discrete infinite logistic normal. We use it as a topic model and derive a VB algorithm.



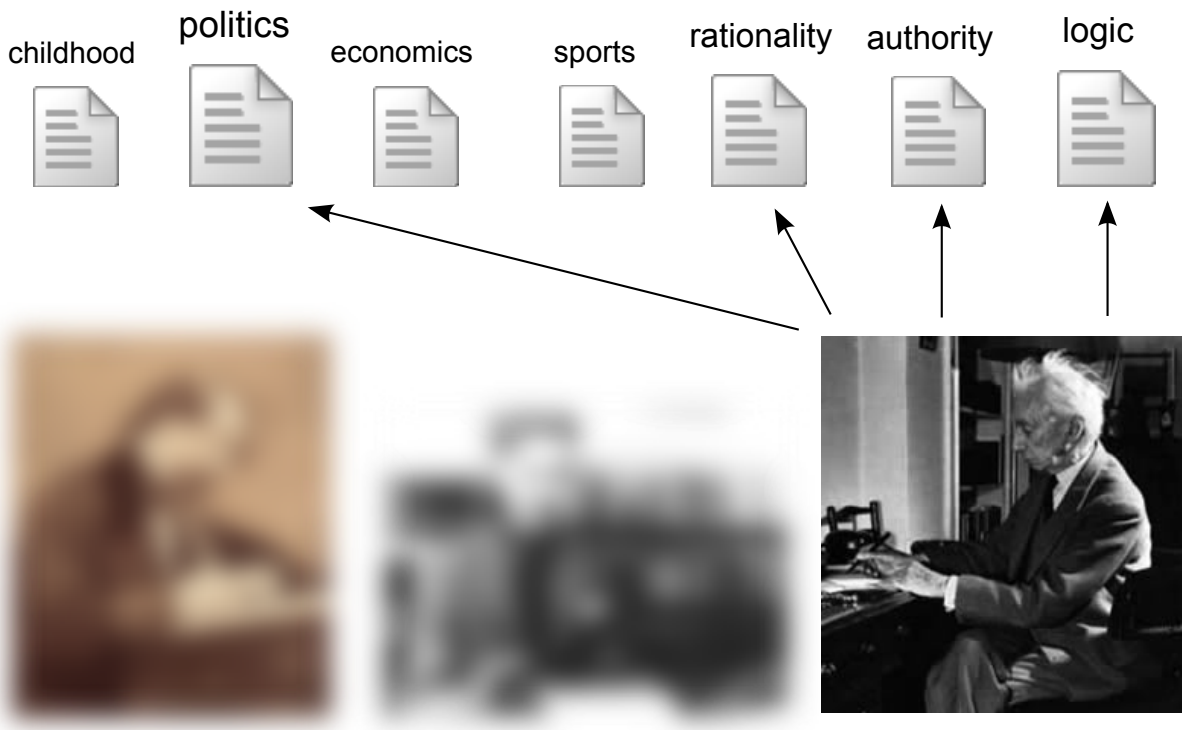
# Motivation

- We develop a mixed-membership model that combines the advantages of the hierarchical Dirichlet process and the correlated topic model.
- We call the resulting prior the discrete infinite logistic normal. We use it as a topic model and derive a VB algorithm.



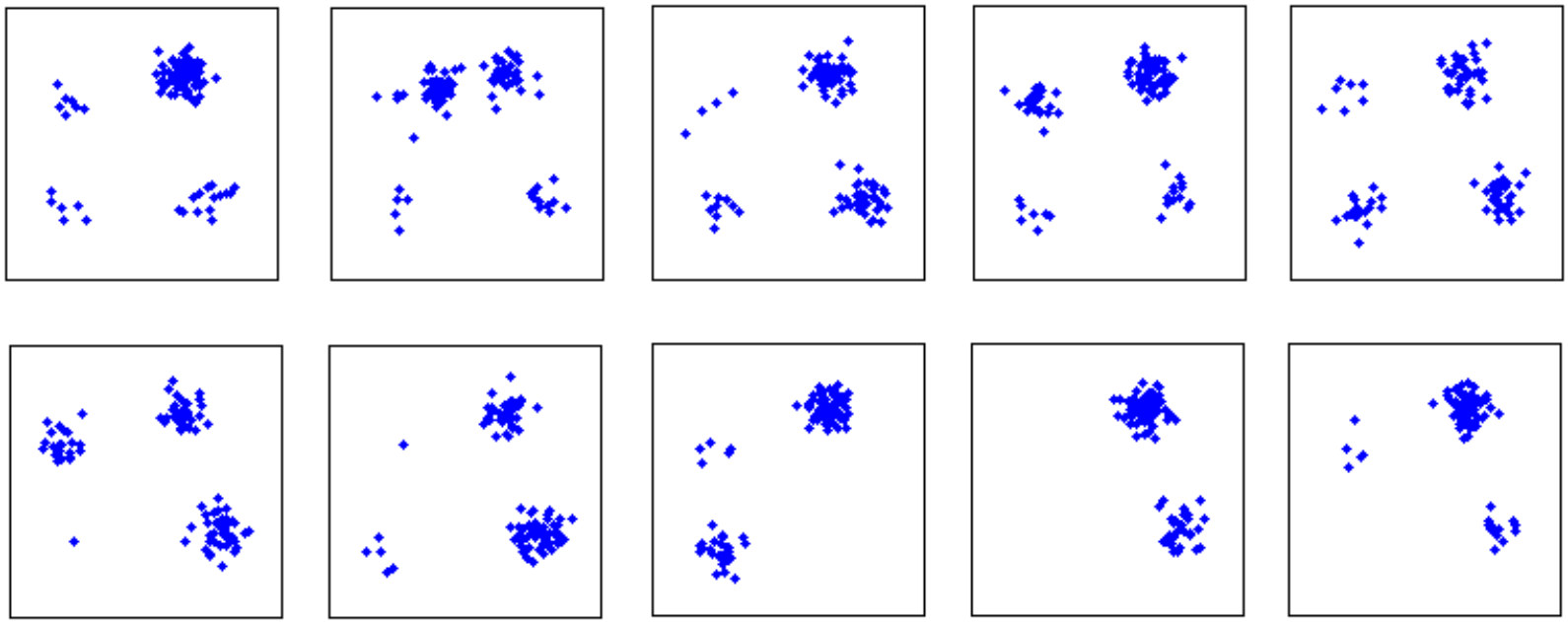
# Motivation

- We develop a mixed-membership model that combines the advantages of the hierarchical Dirichlet process and the correlated topic model.
- We call the resulting prior the discrete infinite logistic normal. We use it as a topic model and derive a VB algorithm.

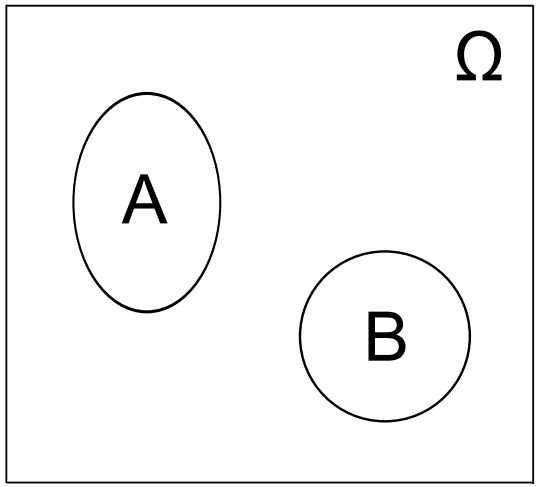


# Introduction: HDP

- The hierarchical Dirichlet process (HDP) is a commonly used BNP prior for grouped data [Teh *et al.*, 2007].
- The HDP extends mixed-membership models (e.g. LDA) to the nonparametric setting.



# Introduction: HDP




H is a completely random measure on Ω.

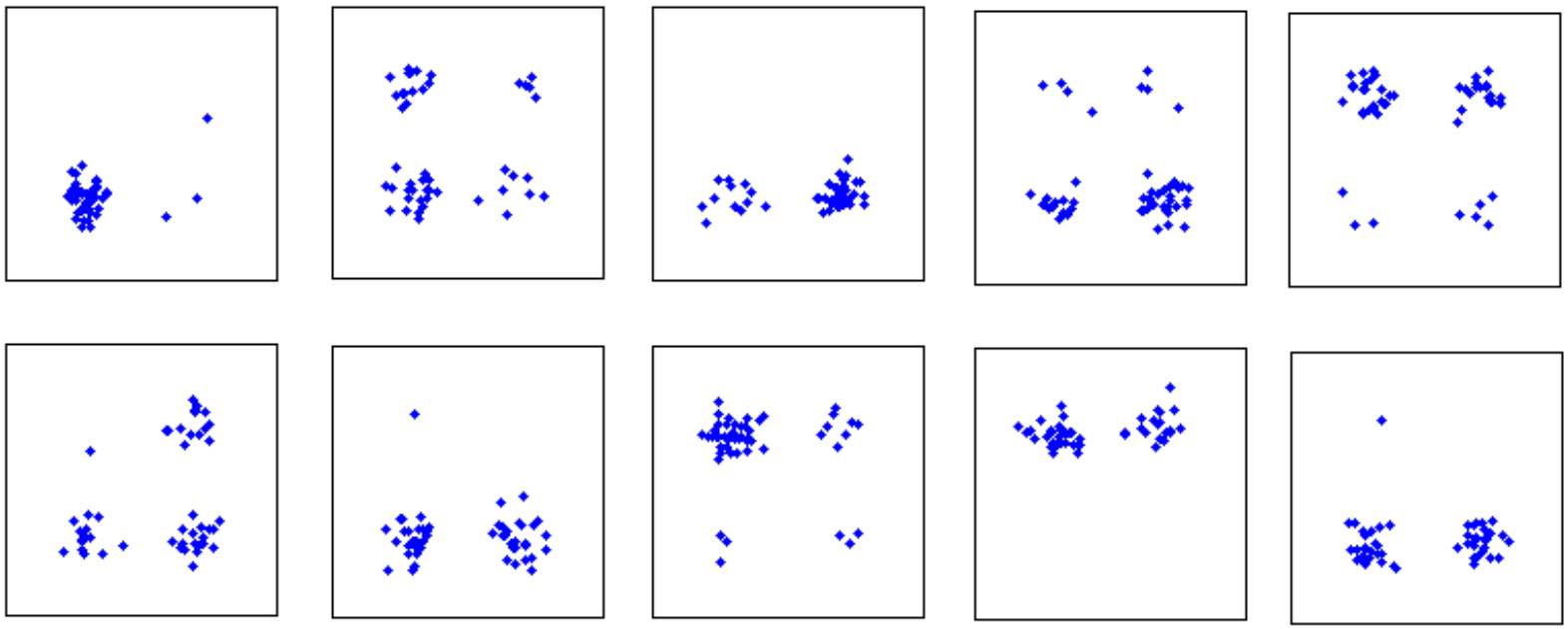
$$H(A) \perp\!\!\!\perp H(B)$$

- The HDP does not explicitly model correlations between the mixing weights of any group-level distribution.

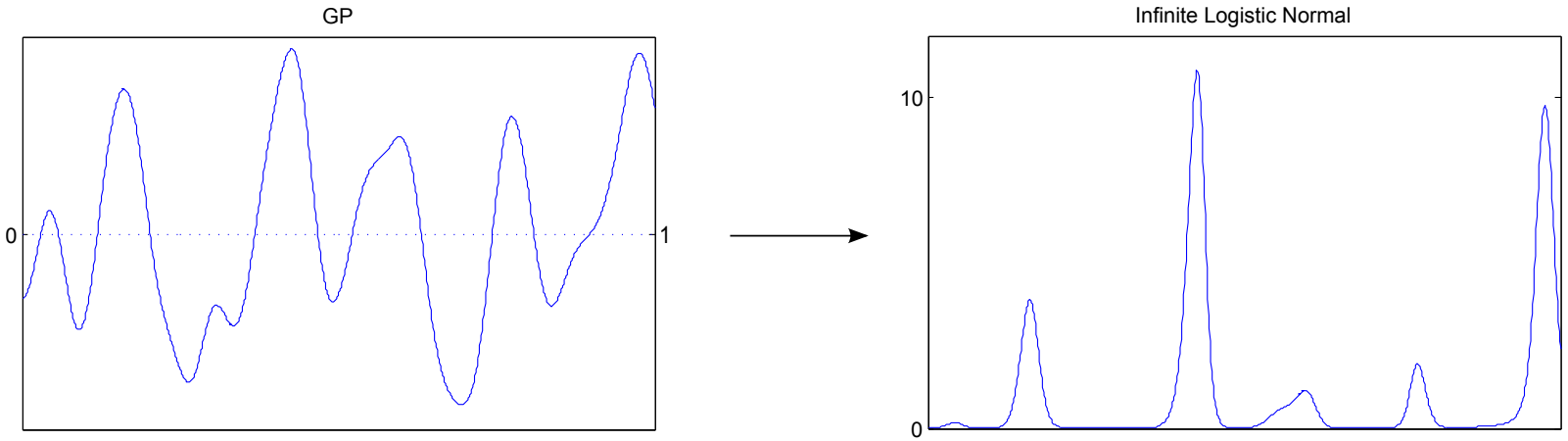
# Introduction: CTM

- The correlated topic model (CTM) addresses this in the finite setting by replacing the Dirichlet prior with a logistic normal prior [Blei & Lafferty, 2007; Aitchison, 1982].
- The generative process of a logistic normal vector is

$$\textit{Gaussian vector} \rightarrow \textit{exponentiation} \rightarrow \textit{normalization}.$$



# Introduction: CTM




- One idea would be to exponentiate and normalize a Gaussian process on the parameter space.
- However, the infinite logistic normal is not discrete and so cannot be used in the mixed-membership setting [Lenk, 1988].

# Introduction: DILN

- Goal: Define a prior for an "infinite correlated topic model".

# Introduction: DILN

- Goal: Define a prior for an "infinite correlated topic model".

  Objective 1 We want this prior to be discrete, so that groups of data use the same set of atoms.

# Introduction: DILN

- Goal: Define a prior for an "infinite correlated topic model".

  Objective 1 We want this prior to be discrete, so that groups of data use the same set of atoms.

  Objective 2 We want to explicitly model correlations between the mixing weights of each group.

# Introduction: DILN

- We define the discrete infinite logistic normal distribution (DILN) to achieve these two goals.

# Introduction: DILN

- We define the discrete infinite logistic normal distribution (DILN) to achieve these two goals.
- We pronounce DILN as "Dylan".

# Review: The Dirichlet process and HDP

- The Dirichlet process [Ferguson, 1973] is a useful BNP prior for mixture models,

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\eta_k}.$$

Atoms $\eta_k \overset{iid}{\sim} G_0$ and $\boldsymbol{\pi}$ depends on $\alpha > 0 \rightarrow G \sim DP(\alpha G_0)$.

# Review: The Dirichlet process and HDP

- The Dirichlet process [Ferguson, 1973] is a useful BNP prior for mixture models,

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\eta_k}.$$

Atoms $\eta_k \overset{iid}{\sim} G_0$ and $\boldsymbol{\pi}$ depends on $\alpha > 0 \rightarrow G \sim DP(\alpha G_0)$.

- The HDP allows for multiple DPs to share the same atoms,

$$G \sim DP(\alpha G_0), \quad G'_m \overset{iid}{\sim} DP(\beta G).$$

Because $G$ is a discrete distribution, $G'_m$ and $G'_{m'}$ share atoms.

# The discrete infinite logistic normal

- We formulate DILN as a scaled HDP.

# The discrete infinite logistic normal

- We formulate DILN as a scaled HDP.
- Let $\alpha, \beta > 0$ and $G_0 \times L_0$ be a base distribution, where $G_0$ covers some parameter space and $L_0$ some location space.

# The discrete infinite logistic normal

- We formulate DILN as a scaled HDP.
- Let $\alpha, \beta > 0$ and $G_0 \times L_0$ be a base distribution, where $G_0$ covers some parameter space and $L_0$ some location space.
    1. A top level DP gives a distribution on atoms and their corresponding locations: $G \sim DP(\alpha G_0 \times L_0)$.

# The discrete infinite logistic normal

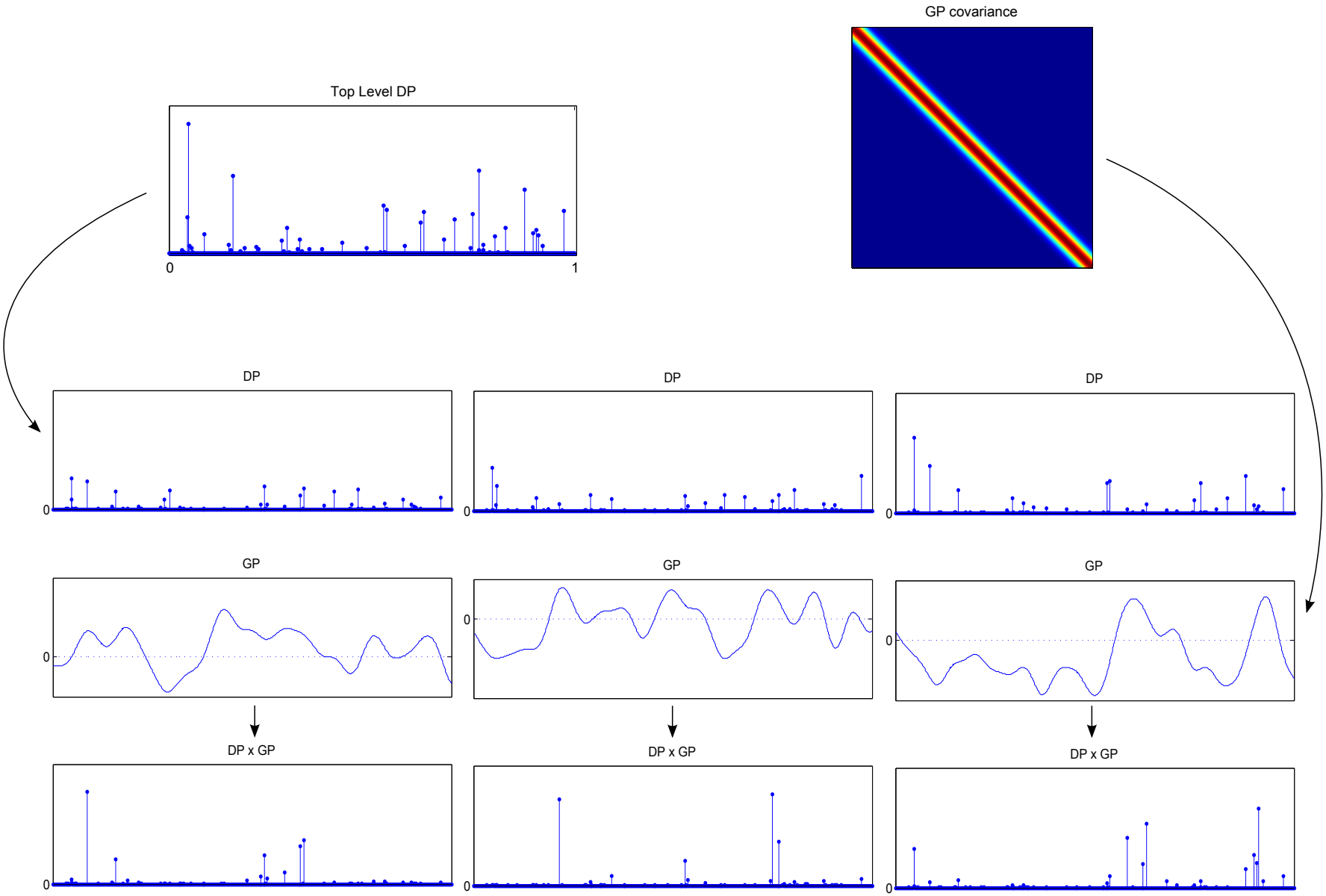
- We formulate DILN as a scaled HDP.
- Let $\alpha, \beta > 0$ and $G_0 \times L_0$ be a base distribution, where $G_0$ covers some parameter space and $L_0$ some location space.
  1. A top level DP gives a distribution on atoms and their corresponding locations: $G \sim DP(\alpha G_0 \times L_0)$.
  2. We then draw a group-level DP and GP,

$$G_m^{DP} \sim DP(\beta G), \quad w^{(m)}(\ell) \sim GP(\mathbf{m}(\ell), \mathbf{K}(\ell, \ell')).$$

# The discrete infinite logistic normal

- We formulate DILN as a scaled HDP.
- Let $\alpha, \beta > 0$ and $G_0 \times L_0$ be a base distribution, where $G_0$ covers some parameter space and $L_0$ some location space.
  1. A top level DP gives a distribution on atoms and their corresponding locations: $G \sim DP(\alpha G_0 \times L_0)$.
  2. We then draw a group-level DP and GP,
  $$G_m^{DP} \sim DP(\beta G), \quad w^{(m)}(\ell) \sim GP(\mathbf{m}(\ell), \mathbf{K}(\ell, \ell')).$$
  3. Finally, we scale the group-level DP by the exponentiated GP,
  $$G'_m(\{\eta, \ell\}) \propto G_m^{DP}(\{\eta, \ell\}) \exp\{w^{(m)}(\ell)\}.$$

# Rough and Intuitive Example



# Constructing DILN

- How do we construct draws from DILN?

# Constructing DILN

- How do we construct draws from DILN?

Top-Level DP

$$V_k \overset{iid}{\sim} Beta(1, \alpha), \quad \eta_k \overset{iid}{\sim} G_0, \quad \ell_k \overset{iid}{\sim} L_0$$

$$\downarrow$$

$$G = \sum_{k=1}^{\infty} V_k \prod_{j=1}^{k-1} (1 - V_j) \delta_{\{\eta_k, \ell_k\}}$$

- We think of $\ell_k$ as the location of atom $k$.

# Constructing DILN

- We use gamma r.v.'s to construct group-level distributions.

# Constructing DILN

- We use gamma r.v.'s to construct group-level distributions.

Group-Level Distributions

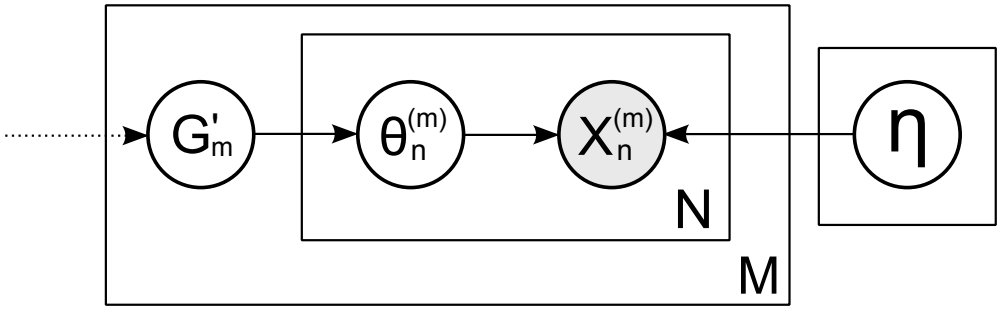
$$Z_k^{(m)} \sim Gamma(\beta p_k, e^{-w_k^{(m)}}), \quad w^{(m)} \overset{iid}{\sim} GP(\mathbf{m}, \mathbf{K})$$

$$\downarrow$$

$$G'_m = \sum_{k=1}^{\infty} \frac{Z_k^{(m)}}{\sum_{j=1}^{\infty} Z_j^{(m)}} \delta_{\eta_k}$$

- If $w^{(m)} = 0$, then this is a representation of the HDP.

# DILN Topic Model




- We use DILN as a topic modeling prior.

- Given $G'_m$, words in document $m$ are generated according to

$$X_n^{(m)} \sim Mult(\theta_n^{(m)}), \quad \theta_n^{(m)} \overset{iid}{\sim} G'_m \,.$$

- For inference, we introduce a latent indicator $C_n^{(m)}$, such that

$$\theta_n^{(m)} = \eta_{C_n^{(m)}}.$$

# Variational inference for DILN

- We use variational inference to learn the approximate posterior of the DILN model [Jordan *et al.*, 1999].

# Variational inference for DILN

- We use variational inference to learn the approximate posterior of the DILN model [Jordan *et al.*, 1999].
  - Mean-field variational inference uses a factorized $q$ distribution to approximate the true posterior of a model's parameters.
  - Searches for the parameters of $q$ that minimize the KL divergence between $q$ and the true posterior.

# Variational inference for DILN

- We use variational inference to learn the approximate posterior of the DILN model [Jordan *et al.*, 1999].
  - Mean-field variational inference uses a factorized $q$ distribution to approximate the true posterior of a model's parameters.
  - Searches for the parameters of $q$ that minimize the KL divergence between $q$ and the true posterior.
- In a DILN topic model, the hidden variables are

  *Document level*: $\mathbf{Z}$, $\mathbf{w}$, $\mathbf{C}$

  *Corpus level*: $\boldsymbol{\eta}$, $\mathbf{V}$, $\mathbf{m}$, $\mathbf{K}$, $\alpha$, $\beta$

# Variational inference for DILN

- We use variational inference to learn the approximate posterior of the DILN model [Jordan *et al.*, 1999].
  - Mean-field variational inference uses a factorized $q$ distribution to approximate the true posterior of a model's parameters.
  - Searches for the parameters of $q$ that minimize the KL divergence between $q$ and the true posterior.
- In a DILN topic model, the hidden variables are

  *Document level*: $\mathbf{Z}$, $\mathbf{w}$, $\mathbf{C}$

  *Corpus level*: $\boldsymbol{\eta}$, $\mathbf{V}$, $\mathbf{m}$, $\mathbf{K}$, $\alpha$, $\beta$
- Note: We learn $\mathbf{K}$ directly, rather than latent locations, $\ell_k$. This leads to a fast, closed-form update.

# Variational inference for DILN

- Inference note: For each group, we use the lower bound

$$-\mathbb{E}_Q\left[\ln\sum_{k=1}^{T} Z_k\right] \geq -\ln\xi - \frac{\sum_{k=1}^{T}\mathbb{E}_Q[Z_k]-\xi}{\xi}.$$

- Results in analytical updates for $q(Z_k) = Gamma(Z_k|a_k, b_k)$,

$$\begin{aligned} a_k &= \beta p_k + \sum_{n=1}^{N}\phi_n(k), \\ b_k &= \mathbb{E}_Q[\exp\{-w_k\}] + \frac{N}{\xi}. \end{aligned}$$

- If $w_k = 0$, this is a new inference algorithm for HDPs.

# Experiments

- We test on four text corpora: Huffington Post, New York Times, Science and Wikipedia.
- We compare with the HDP and CTM. For the HDP, we use the algorithm previously discussed.
- For the CTM, we vary the number of topics. For the HDP and DILN we truncated $T = 200$.
- We use $\mathit{Dirichlet}(\gamma \mathbf{1}_{|W|})$ for the base distribution, and vary $\gamma$.

# Experiments

- We test on four text corpora: Huffington Post, New York Times, Science and Wikipedia.
- We compare with the HDP and CTM. For the HDP, we use the algorithm previously discussed.
- For the CTM, we vary the number of topics. For the HDP and DILN we truncated $T = 200$.
- We use $\mathit{Dirichlet}(\gamma \mathbf{1}_{|W|})$ for the base distribution, and vary $\gamma$.

# Experiments

- We test on four text corpora: Huffington Post, New York Times, Science and Wikipedia.
- We compare with the HDP and CTM. For the HDP, we use the algorithm previously discussed.
- For the CTM, we vary the number of topics. For the HDP and DILN we truncated $T = 200$.
- We use $\mathit{Dirichlet}(\gamma \mathbf{1}_{|W|})$ for the base distribution, and vary $\gamma$.

# Experiments

- We test on four text corpora: Huffington Post, New York Times, Science and Wikipedia.
- We compare with the HDP and CTM. For the HDP, we use the algorithm previously discussed.
- For the CTM, we vary the number of topics. For the HDP and DILN we truncated $T = 200$.
- We use $\mathit{Dirichlet}(\gamma \mathbf{1}_{|W|})$ for the base distribution, and vary $\gamma$.

# Experiments

- For testing, we partition a test document into two halves.
- We learn document-specific parameters on one half and predict the other half.
- We then calculate the per-word perplexity

$$\text{perplexity} = \exp\left\{\frac{-\ln p(X_{\text{half2}}|X_{\text{half1}})}{N}\right\}.$$

# Experiments

- For testing, we partition a test document into two halves.
- We learn document-specific parameters on one half and predict the other half.
- We then calculate the per-word perplexity

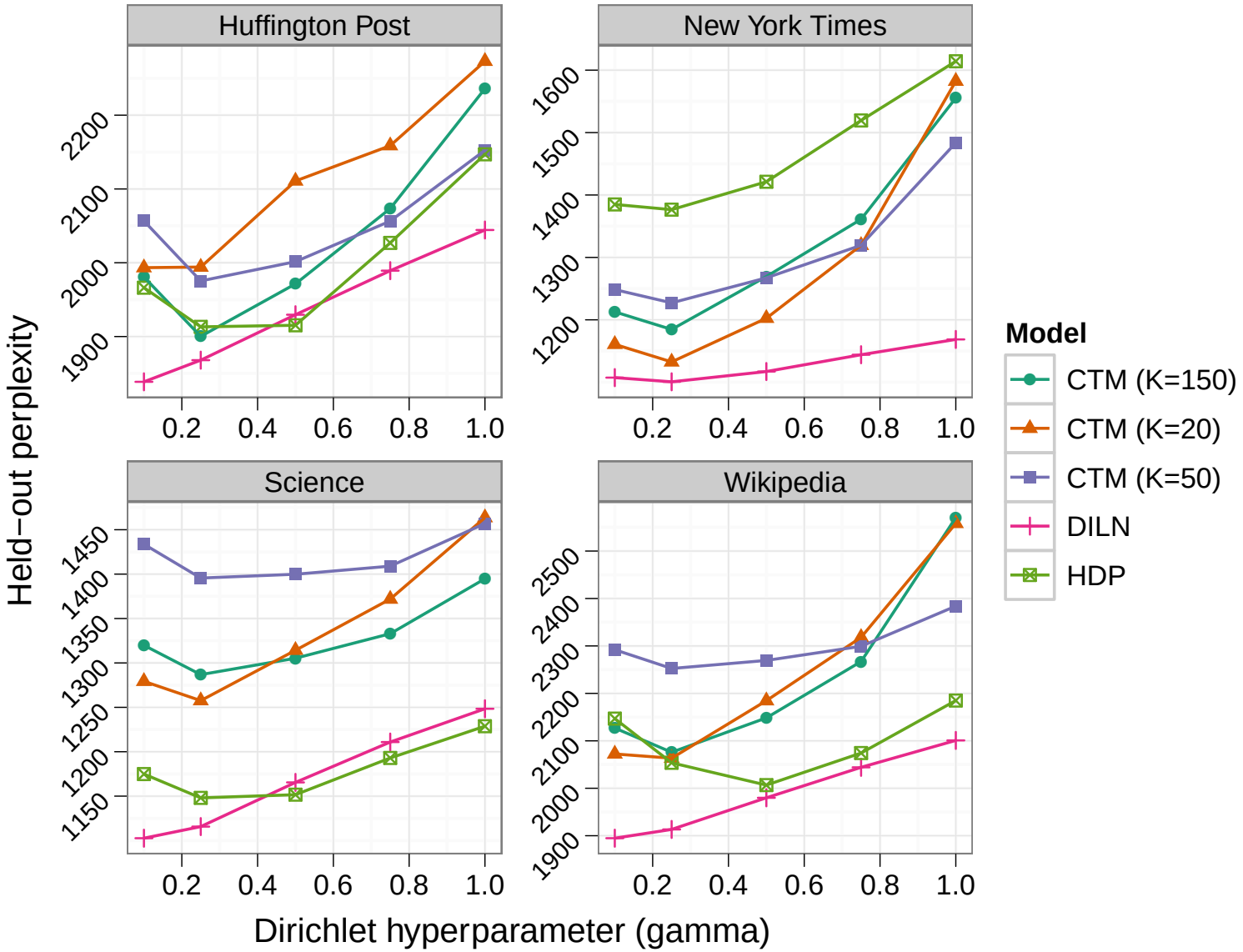
$$\text{perplexity} = \exp\left\{\frac{-\ln p(X_{\text{half2}}|X_{\text{half1}})}{N}\right\}.$$

# Experiments

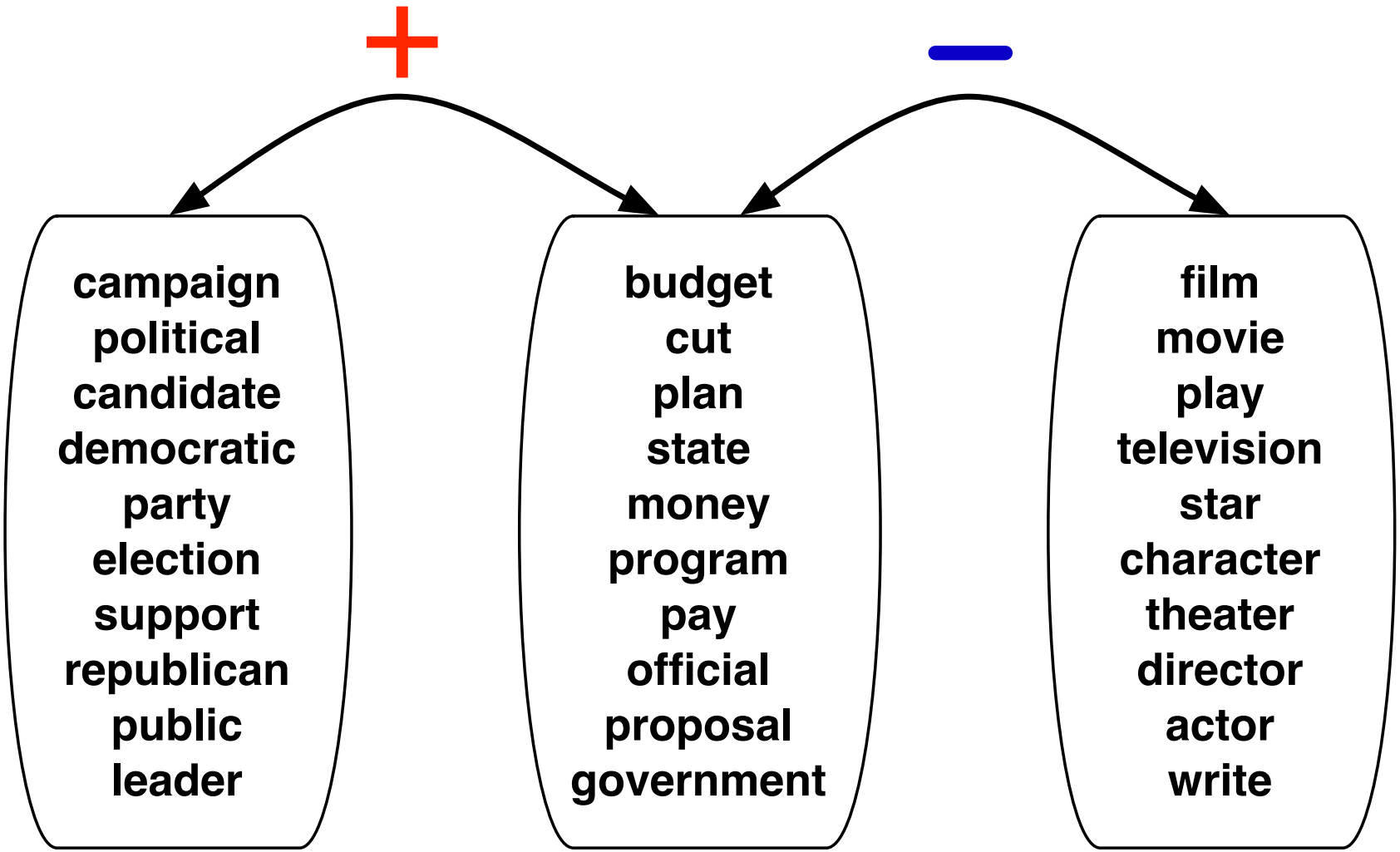
- For testing, we partition a test document into two halves.
- We learn document-specific parameters on one half and predict the other half.
- We then calculate the per-word perplexity

$$\text{perplexity} = \exp\left\{\frac{-\ln p(X_{\text{half2}}|X_{\text{half1}})}{N}\right\}.$$

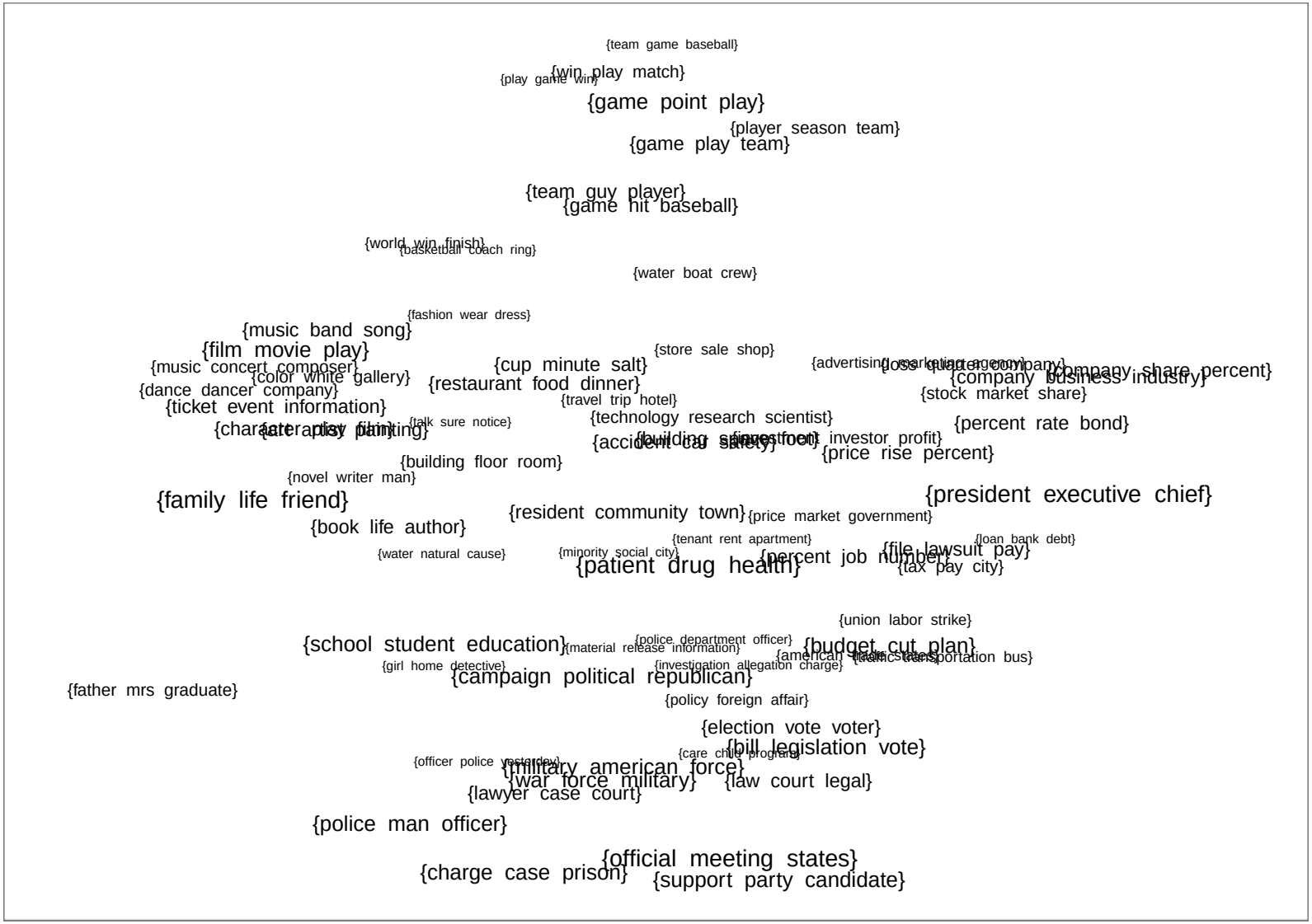
# Experiments: Perplexity Results



# Experiments: New York Times Topic Correlations



# Experiments: New York Times Topic Mapping



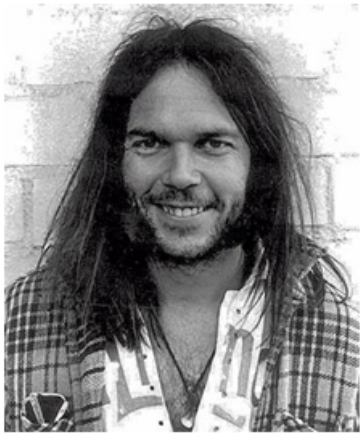
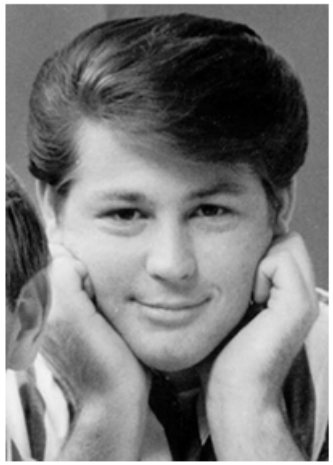
# Conclusion

- DILN: A nonparametric mixed membership model that models correlations across mixing weights at the group level.
- Close relationship between DILN and HDP results in new variational inference algorithm for HDP.
- Demonstrated performance of DILN in topic modeling setting.

# Future Work






# References


Aitchison, J. (1982). The statistical analysis of compositional data. Journal of the Royal Statistical Society, Series B 44, 139-177.

Blei, D. & Lafferty, J. (2007). A correlated topic model of Science. Annals of Applied Statistics 1, 17-35.

Blei, D., Ng, A. & Jordan, M. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research 3, 993-1022.

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. The Annals of Statistics 1, 209-230.

Jordan, M., Ghahramani, Z., Jaakkola, T. & Saul, L. (1999). An introduction to variational methods for graphical models. Machine Learning 37, 183-233.

Lenk, P. (1988). The logistic normal distribution for Bayesian, nonparametric, predictive densities. JASA 83, 509-516.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. Statistica Sinica 4, 639-650.

Teh, Y., Jordan, M., Beal, M. & Blei, D. (2007). Hierarchical Dirichlet processes. JASA 101, 1566-1581.

Discussion of

# "The Discrete Infinite Logistic Normal (DILN) Distribution for Mixed-Membership Modeling"

Frank Wood

Columbia University
fwood@stat.columbia.edu

April 12, 2011

# Context

- Mixed-membership models (e.g. “topic models”) are inarguably popular

# Context

- Mixed-membership models (e.g. “topic models”) are inarguably popular
- Key to mixed-membership model popularity : simplicity (e.g. LDA [Blei et al., 2003])

# Context

- Mixed-membership models (e.g. “topic models”) are inarguably popular
- Key to mixed-membership model popularity : simplicity (e.g. LDA [Blei et al., 2003])
  - Easy to describe

# Context

- Mixed-membership models (e.g. “topic models”) are inarguably popular
- Key to mixed-membership model popularity : simplicity (e.g. LDA [Blei et al., 2003])
  - Easy to describe
  - Accessible intuition

# Context

- Mixed-membership models (e.g. "topic models") are inarguably popular
- Key to mixed-membership model popularity : simplicity (e.g. LDA [Blei et al., 2003])
  - Easy to describe
  - Accessible intuition
  - Visibly fascinating output

# Context

- Mixed-membership models (e.g. "topic models") are inarguably popular
- Key to mixed-membership model popularity : simplicity (e.g. LDA [Blei et al., 2003])
  - Easy to describe
  - Accessible intuition
  - Visibly fascinating output
- Promising for browsing and search applications.

# Problem

Simplicity a virtue—also a vice.

- Mixed-membership models posit latent features (topics, objects, etc.)

# Problem

Simplicity a virtue—also a vice.

- Mixed-membership models posit latent features (topics, objects, etc.)
- Latent feature occurrence is correlated in the real world (i.e. objects appearing in visual scenes [Doshi-Velez and Ghahramani, 2009]).

# Problem

Simplicity a virtue—also a vice.

- Mixed-membership models posit latent features (topics, objects, etc.)
- Latent feature occurrence is correlated in the real world (i.e. objects appearing in visual scenes [Doshi-Velez and Ghahramani, 2009]).
- Most mixed-membership models have un-correlated features

# Problem

Simplicity a virtue—also a vice.

- Mixed-membership models posit latent features (topics, objects, etc.)
- Latent feature occurrence is correlated in the real world (i.e. objects appearing in visual scenes [Doshi-Velez and Ghahramani, 2009]).
- Most mixed-membership models have un-correlated features
- Correlated feature models in development [Blei and Lafferty, 2006, Doshi-Velez and Ghahramani, 2009, Rai and Daumé III, 2009]

# DILN: Key Idea

- Associate every feature with a latent vector

# DILN: Key Idea

- Associate every feature with a latent vector
- Use "distance" in a latent space to encode feature co-occurrence
  - "Close" features tend to occur together more often

# DILN: Key Idea

- Associate every feature with a latent vector
- Use "distance" in a latent space to encode feature co-occurrence
  - "Close" features tend to occur together more often
- Gaussian process [Rasmussen and Williams, 2006] prior on feature space ensures that close "locations" have similar prevalences (smoothness).

# DILN: Key Idea

- Associate every feature with a latent vector
- Use "distance" in a latent space to encode feature co-occurrence
  - "Close" features tend to occur together more often
- Gaussian process [Rasmussen and Williams, 2006] prior on feature space ensures that close "locations" have similar prevalences (smoothness).
- Expert practitioners: HDP-LDA [Teh et al., 2006] + Gamma process representation + GP latent space prior + variational inference =

# DILN: Key Idea

- Associate every feature with a latent vector
- Use "distance" in a latent space to encode feature co-occurrence
  - "Close" features tend to occur together more often
- Gaussian process [Rasmussen and Williams, 2006] prior on feature space ensures that close "locations" have similar prevalences (smoothness).
- Expert practitioners: HDP-LDA [Teh et al., 2006] + Gamma process representation + GP latent space prior + variational inference =
  - Mathematically elegant mixed-membership model with correlated features.

# Questions

- What kind topic prevalence correlations can be captured by DILN?

$$Cov\left[Z_i^{(m)}, Z_j^{(m)}|\cdot\right] = \beta^2 p_i p_j e^{\frac{1}{2}(k_{ii}+k_{jj})}(e^{k_{ij}} - 1)$$

# Questions

- What kind topic prevalence correlations can be captured by DILN?

$$Cov\left[Z_i^{(m)}, Z_j^{(m)}|\cdot\right] = \beta^2 p_i p_j e^{\frac{1}{2}(k_{ii}+k_{jj})}(e^{k_{ij}} - 1)$$

- How scalable is the model? (particularly with respect to the number of topics in the VB truncation)

# Questions

- What kind topic prevalence correlations can be captured by DILN?

$$Cov\left[Z_i^{(m)}, Z_j^{(m)}|\cdot\right] = \beta^2 p_i p_j e^{\frac{1}{2}(k_{ii}+k_{jj})}(e^{k_{ij}} - 1)$$

- How scalable is the model? (particularly with respect to the number of topics in the VB truncation)
- How are the latent topic locations represented?

# Questions

- What kind topic prevalence correlations can be captured by DILN?

$$Cov\left[Z_i^{(m)}, Z_j^{(m)}|\cdot\right] = \beta^2 p_i p_j e^{\frac{1}{2}(k_{ii}+k_{jj})}(e^{k_{ij}} - 1)$$

- How scalable is the model? (particularly with respect to the number of topics in the VB truncation)
- How are the latent topic locations represented?
- Why VB? How would sampling work in this model? Incremental inference?

# Questions

- What kind topic prevalence correlations can be captured by DILN?

$$Cov\left[Z_i^{(m)}, Z_j^{(m)}|\cdot\right] = \beta^2 p_i p_j e^{\frac{1}{2}(k_{ii}+k_{jj})}(e^{k_{ij}} - 1)$$

- How scalable is the model? (particularly with respect to the number of topics in the VB truncation)
- How are the latent topic locations represented?
- Why VB? How would sampling work in this model? Incremental inference?

# Big Picture

- What about interpretability?

# Big Picture

- What about interpretability?
  - Distance only? Why not further hierarchy?

# Big Picture

- What about interpretability?
  - Distance only? Why not further hierarchy?
- How much data? (data vs. model complexity)

D. Blei and J. Lafferty. Correlated topic models. *Advances in Neural Information Processing Systems*, 18:147, 2006. ISSN 1049-5258.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

F. Doshi-Velez and Z. Ghahramani. Correlated non-parametric latent feature models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 143–150. AUAI Press, 2009.

P. Rai and H. Daumé III. The infinite hierarchical factor regression model, 2009.

C.E. Rasmussen and C.K.I. Williams. *Gaussian processes for machine learning*. Springer, 2006.

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.