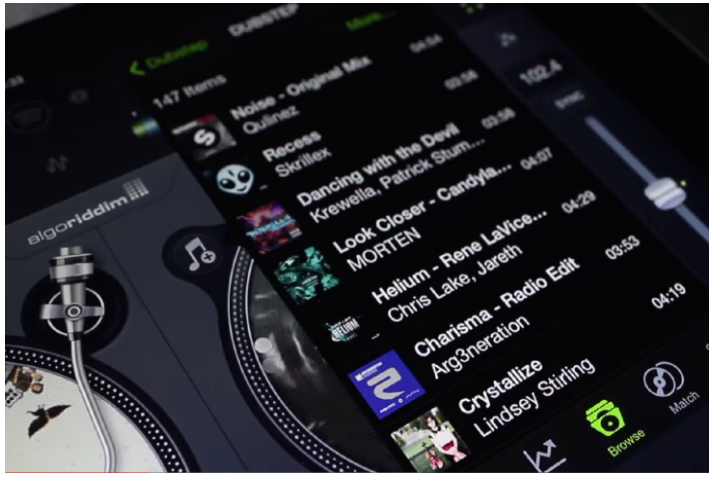
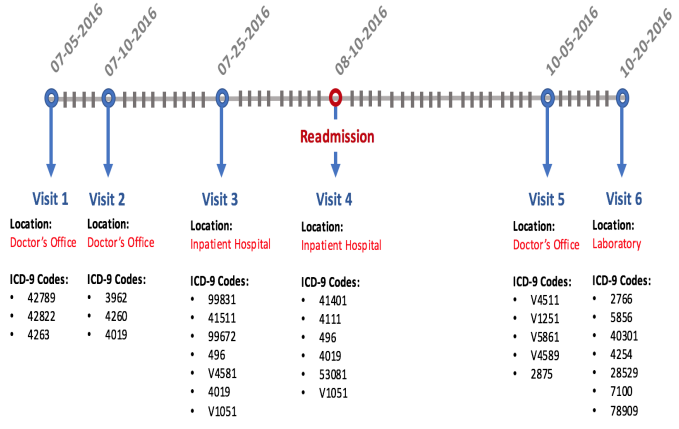
Structured Deep Generative Models

Adji Bousso Dieng

IPAM Workshop on Interpretable Learning in Physical Sciences
University of California Los Angeles, Los Angeles, CA
October, 2019

Structure


Noise - Original Mix
Qulinez
Recess
Skrillex
Dancing with the Devil
Krewella, Patrick Stum...
Look Closer - Candyla...
MORTEN
Helium - Rene LaVice...
Chris Lake, Jareth
Charisma - Radio Edit
Arg3neration
Crystallize
Lindsey Stirling
Browse
Match


07-05-2016
07-10-2016
07-25-2016
08-10-2016
10-05-2016
10-20-2016
Readmission
Visit 1
Location: Doctor's Office
ICD-9 Codes:
• 42789
• 42822
• 4263
Visit 2
Location: Doctor's Office
ICD-9 Codes:
• 3962
• 4260
• 4019
Visit 3
Location: Inpatient Hospital
ICD-9 Codes:
• 99831
• 41511
• 99672
• 496
• V4581
• 4019
• V1051
Visit 4
Location: Inpatient Hospital
ICD-9 Codes:
• 41401
• 4111
• 496
• 4019
• 53081
• V1051
Visit 5
Location: Doctor's Office
ICD-9 Codes:
• V4511
• V1251
• V5861
• V4589
• 2875
Visit 6
Location: Laboratory
ICD-9 Codes:
• 2766
• 5856
• 40301
• 4254
• 28529
• 7100
• 78909

Bayesian Hierarchical Models

- Model is a joint over data $\mathbf{x}$, global latent variables θ, and local latent variables $\mathbf{z}$

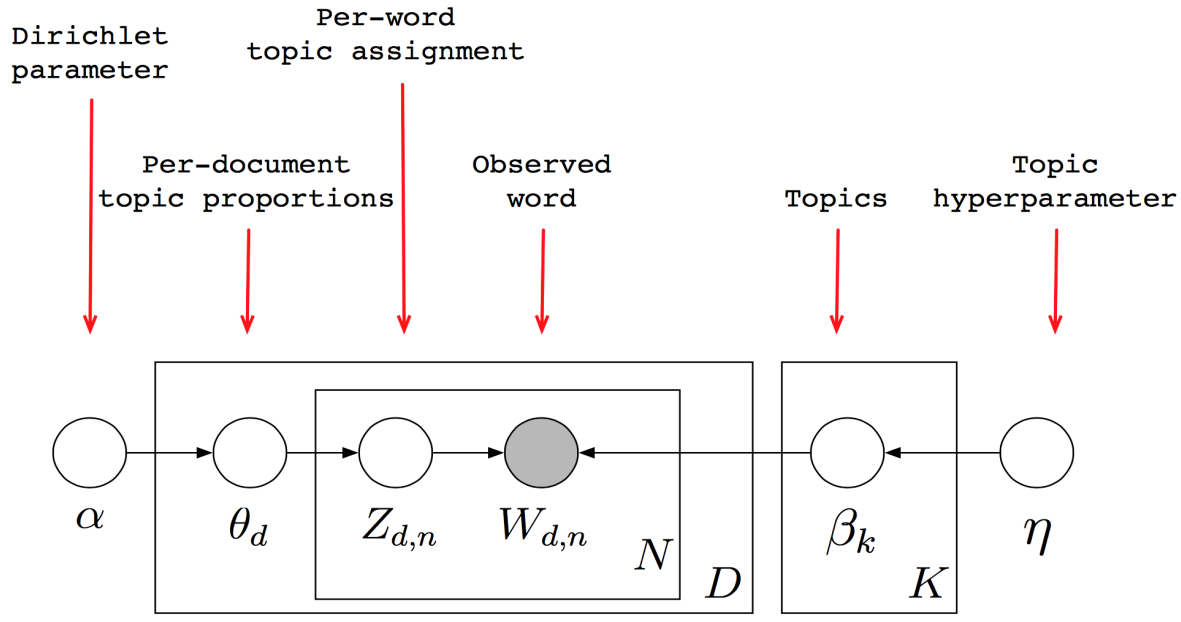
$$p(\mathbf{x}, \mathbf{z}, \theta) = p(\theta) \prod_{n=1}^{N} p(\mathbf{x}_n \mid \mathbf{z}_n, \theta) p(\mathbf{z}_n \mid \theta)$$

- Find structure via posterior inference

$$p(\mathbf{z}, \theta \mid \mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{z}, \theta)}{p(\mathbf{x})}$$

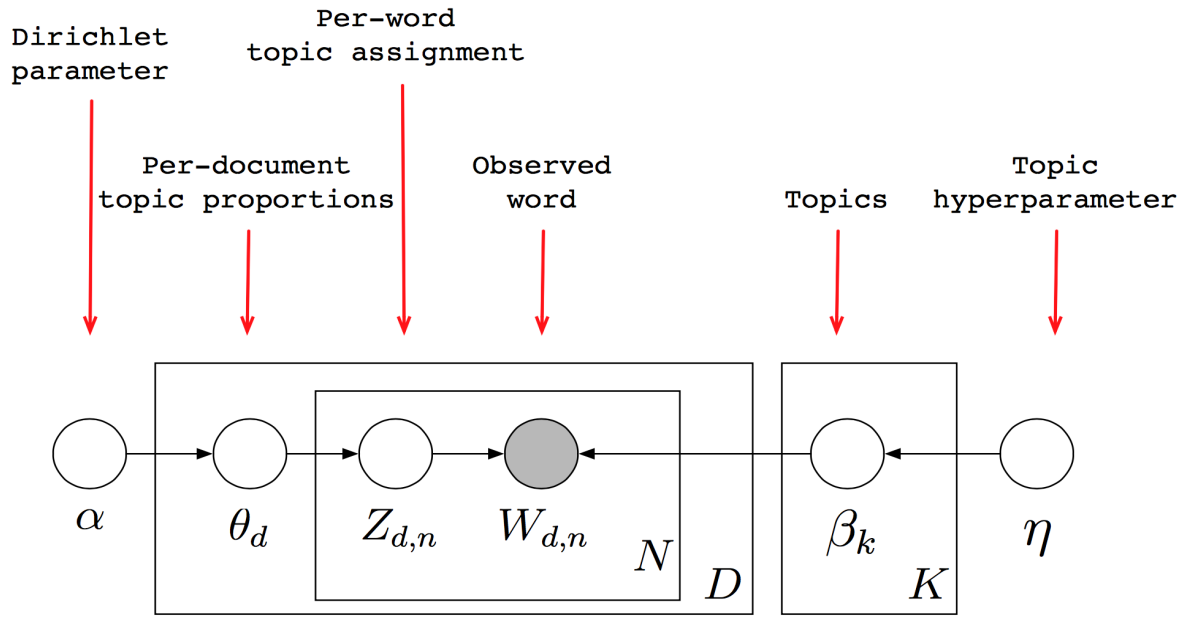
- Interpretable probabilistic structure

Example: Latent Dirichlet Allocation



- Define K shared latent topics $\beta_k \in \text{Dir}(\eta)$
- For each document d
  1. Draw proportions $\theta_d \sim \text{Dir}(\alpha)$
  2. For each word n in the document:
     - Draw assignment $\mathbf{z}_{nd} \sim \text{Cat}(\theta_d)$
     - Draw word $w_{nd} \sim \beta_{\mathbf{z}_{nd}}$

Example: Latent Dirichlet Allocation



- Enjoys conjugacy
- Can be fit using coordinate ascent variational inference
- Potential problem in high dimensions (very large vocabularies)

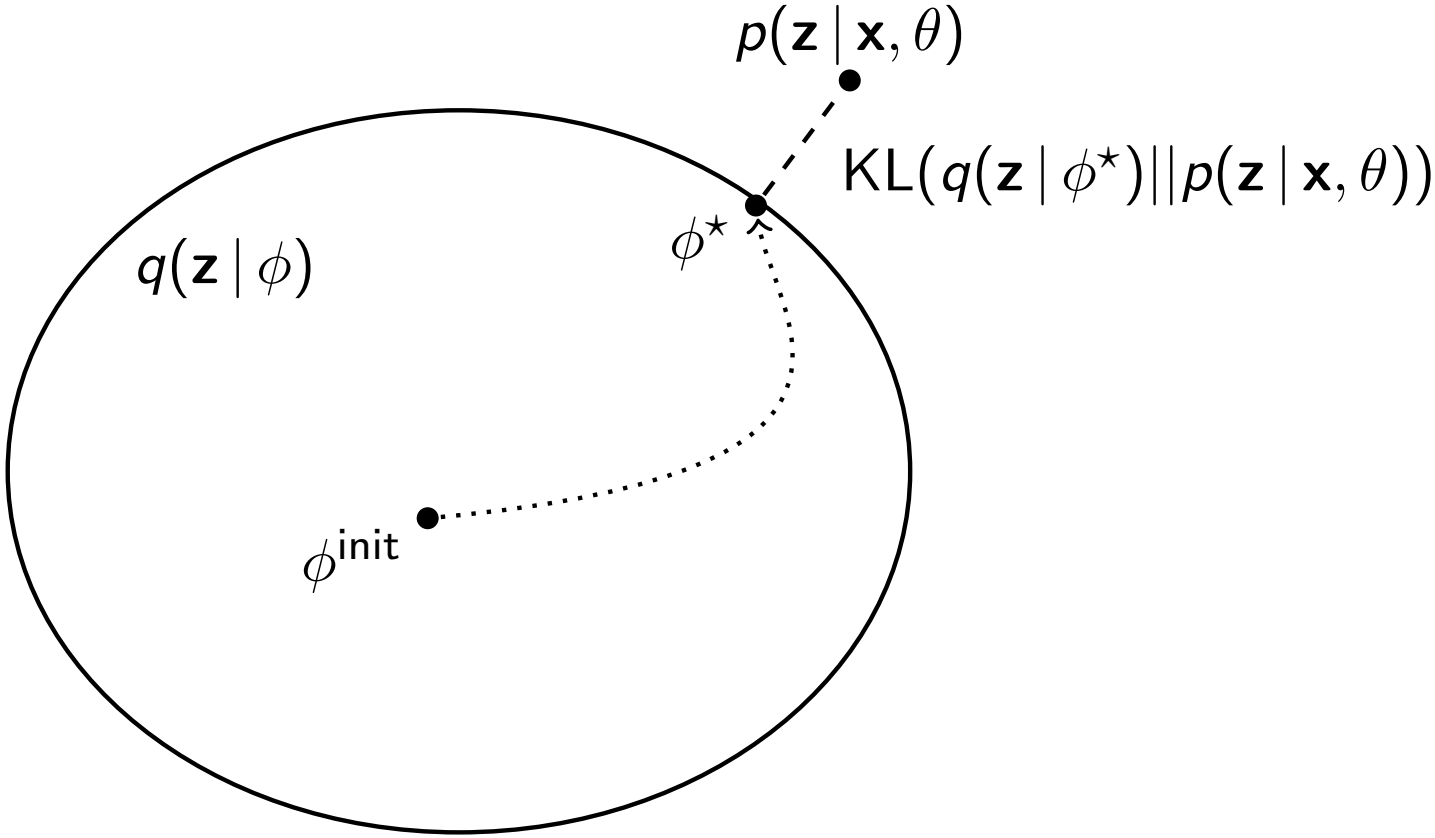
"Example": Deep Latent Gaussian Model

- Replace prior over θ with a neural network
- Model is a joint over data $\mathbf{x}$ and local latent variables $\mathbf{z}$

$$p(\mathbf{x}, \mathbf{z} \mid \theta) = \prod_{n=1}^{N} p(\mathbf{x}_n \mid \mathbf{z}_n, \theta) p(\mathbf{z}_n \mid \theta)$$

- Often $p(\mathbf{z}_n \mid \theta) = p(\mathbf{z}_n) = \mathcal{N}(\mathbf{0}, \mathbf{I})$
- The conditional $p(\mathbf{x}_n \mid \mathbf{z}_n, \theta)$ is a neural network with parameters θ that takes $\mathbf{z}$ as input
- We lost the probabilistic structure over shared global variables θ

Variational Inference



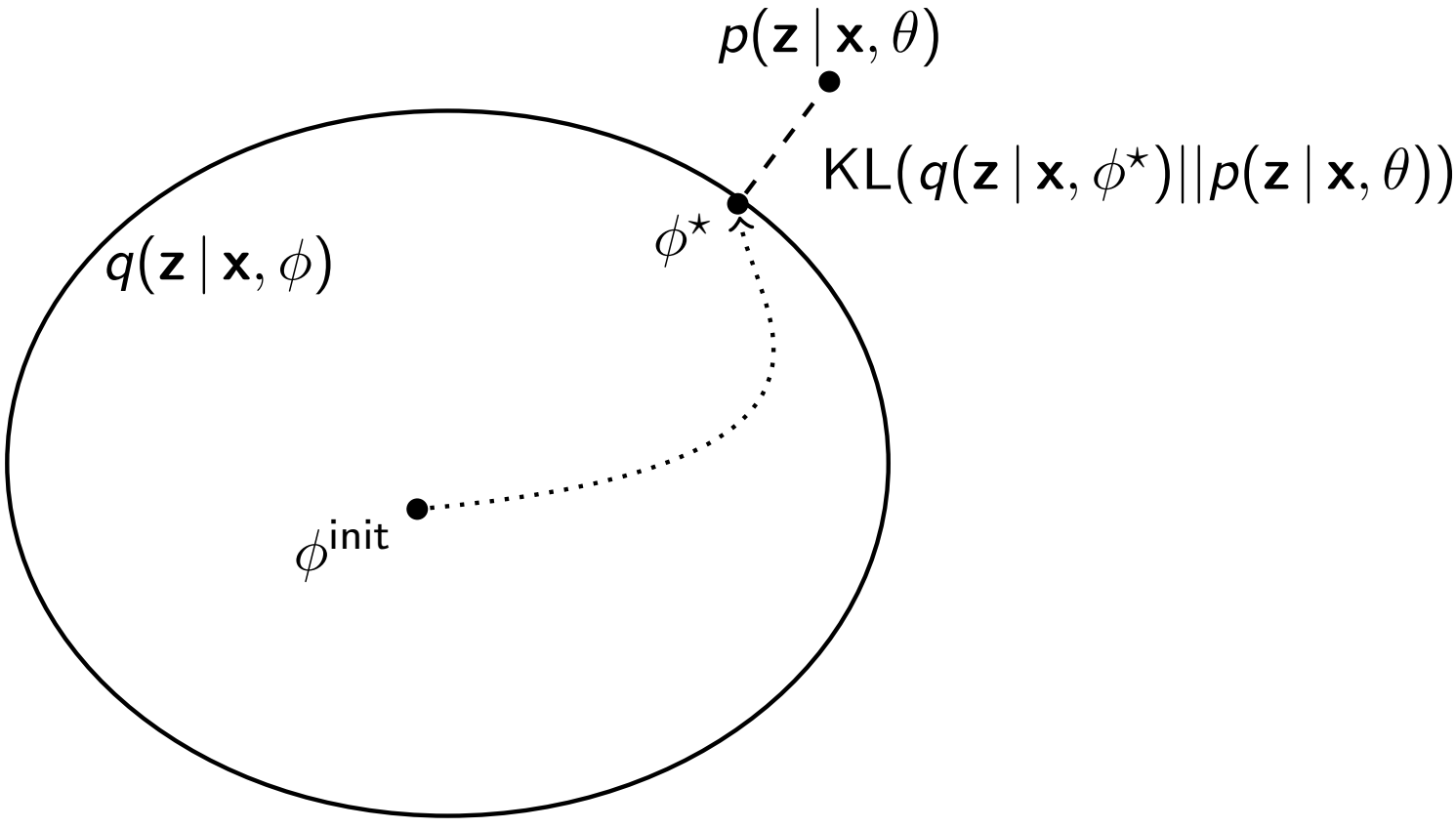
Minimizing the KL divergence

$$\mathrm{KL}(q(\mathbf{z} \mid \phi) || p(\mathbf{z} \mid \mathbf{x}, \theta)) = \log p(\mathbf{x} \mid \theta) - \mathrm{ELBO}$$

is equivalent to maximizing the ELBO,

$$\mathrm{ELBO} = \mathbb{E}_{q(\mathbf{z} \mid \phi)} \left[\log p(\mathbf{x}, \mathbf{z} \mid \theta) - \log q(\mathbf{z} \mid \phi) \right]$$

Amortized Variational Inference

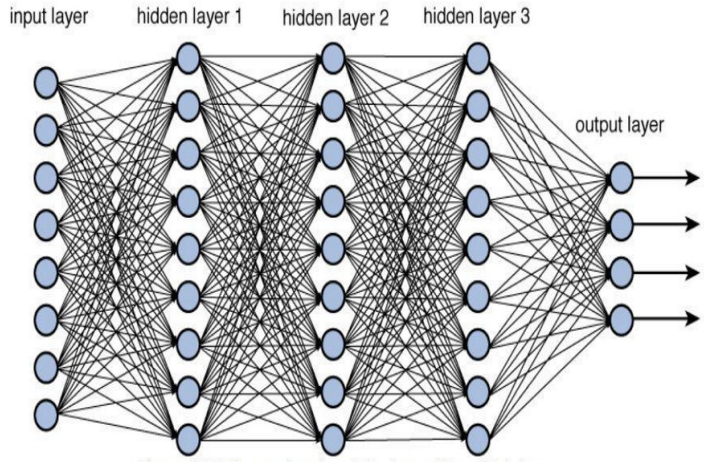


Parameterize q with a neural network that takes data $\mathbf{x}$ as input and maximize ELBO as before:

$$\text{ELBO} = \mathbb{E}_{q(\mathbf{z} \mid \mathbf{x}, \phi)} \left[\log p(\mathbf{x}, \mathbf{z} \mid \theta) - \log q(\mathbf{z} \mid \mathbf{x}, \phi) \right]$$

Hierarchical Bayes + Neural Networks

+



- to learn models with interpretable probabilistic structure
- to deal well with the high dimensionality of the data
- for efficient inference

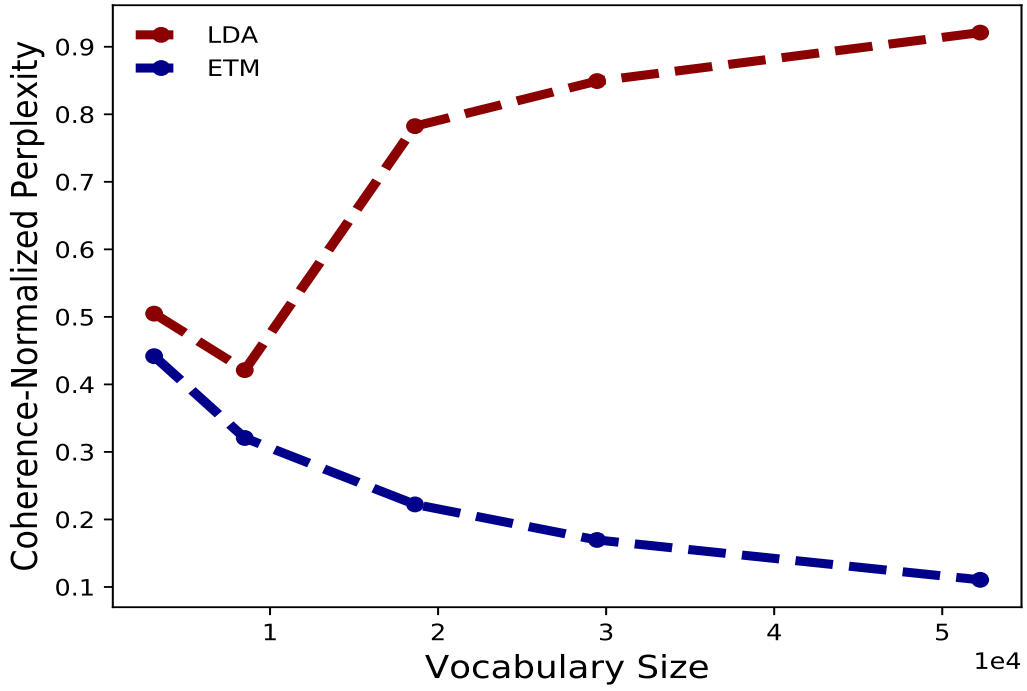
I will next describe three instances of a family of models for high dimensional data that (1) have an interpretable probabilistic structure and (2) are fit efficiently using amortized variational inference. The key underlying idea is to work in the *meaning* space.

Embedded Topic Model

- Define a deterministic per-word embedding $\rho \in \mathbb{R}^{V \times E}$
- Define a deterministic shared global embedding matrix $\alpha \in \mathbb{R}^{K \times E}$
- For each document d
  1. Draw proportions $\theta_d \sim \mathcal{LN}(\mathbf{0}, \mathbf{I})$
  2. For each word n in the document:
     - Draw assignment $\mathbf{z}_{nd} \sim \text{Cat}(\theta_d)$
     - Draw word $w_{nd} \sim \text{Cat}(\text{softmax}(\rho^\top \alpha_{\mathbf{z}_{nd}}))$
- Deal with high dimensions by working on the embedding space
- Same interpretable probabilistic structure as LDA
- Fit model using amortized variational inference

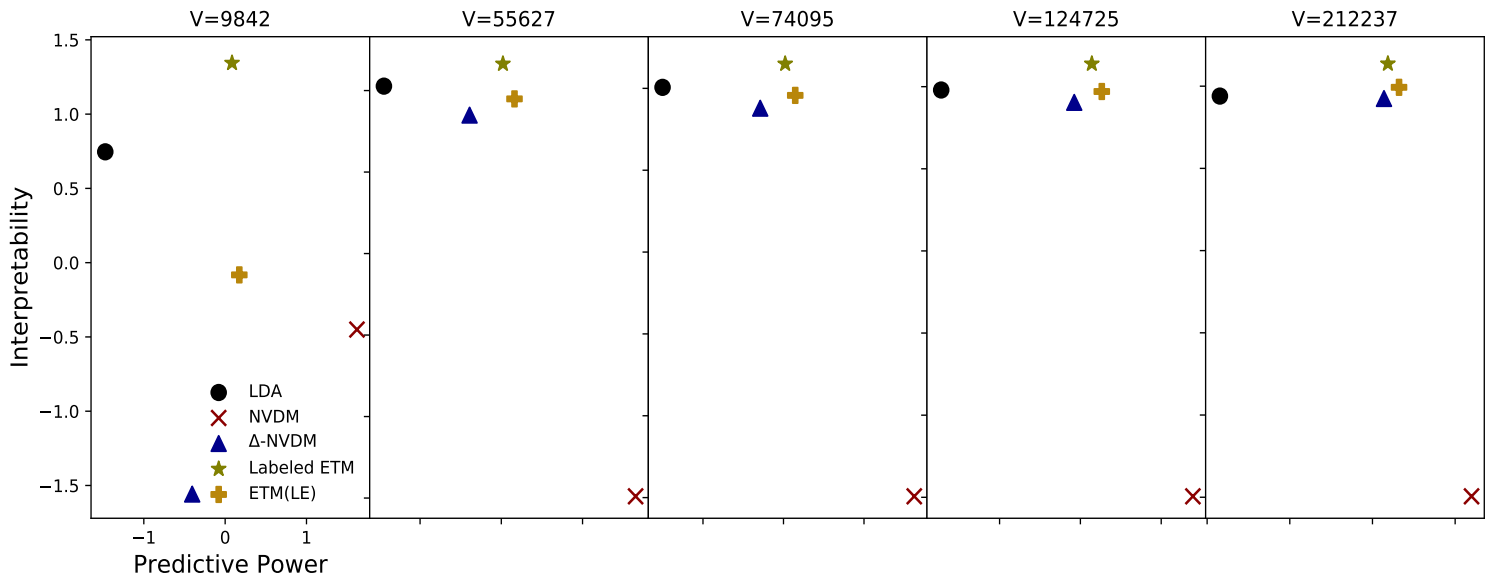
Embedded Topic Model

- Comparing ETM to LDA on the 20Newsgroup corpus.
- LDA's performance degrades as the dimensionality increases
- ETM deals well with high dimensionality



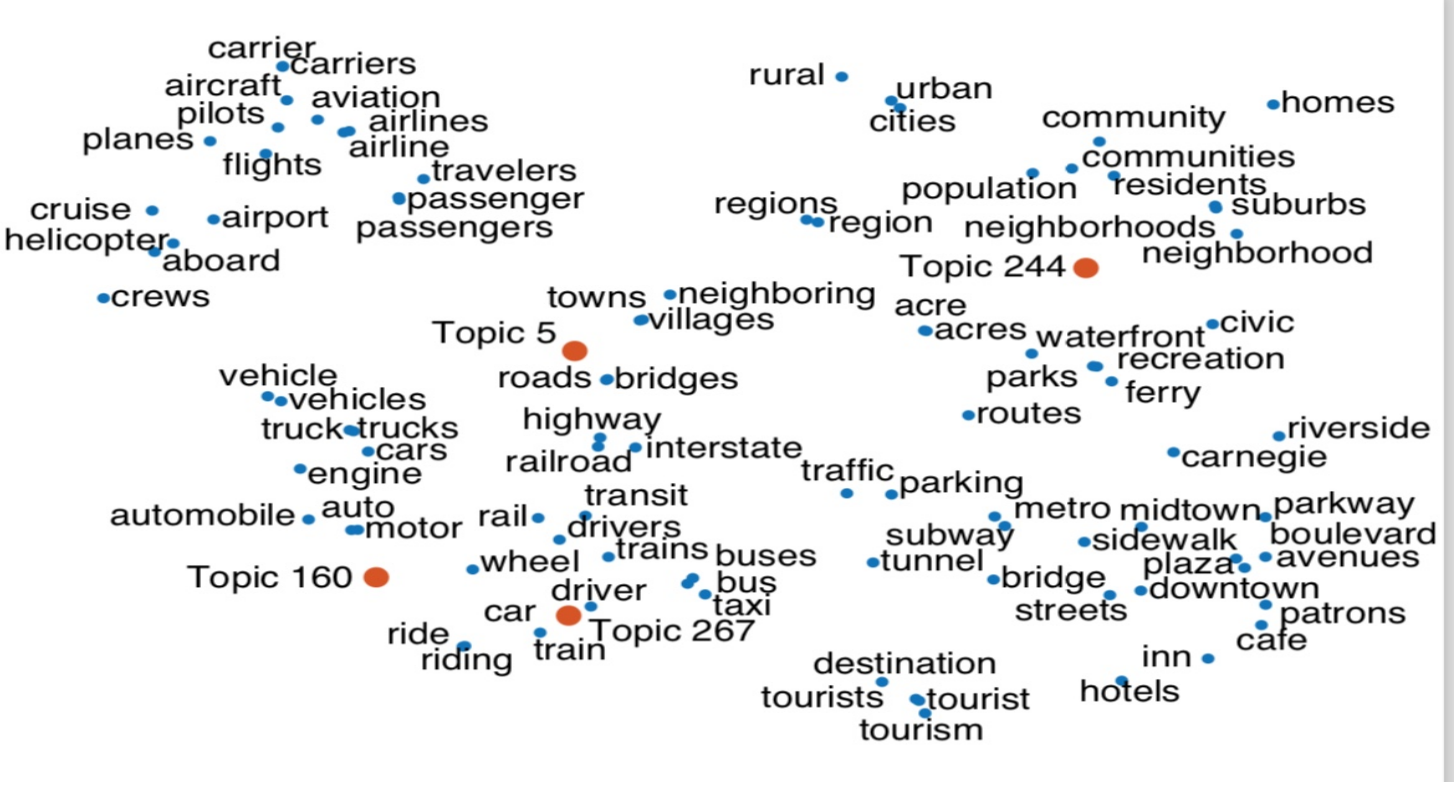
Embedded Topic Model

- Comparing ETM to several document models on the New York Times corpus as the vocabulary (V) increases.
- A good document model is on the top right; interpretable with high predictive power.



Embedded Topic Model

- ETM topic embeddings found in the New York Times corpus.



Embedded Topic Model

- ETM word embeddings found in the New York Times corpus compared to Skipgram word embeddings.

Skip-gram embeddings				ETM embeddings			
love	**family**	**woman**	**politics**	**love**	**family**	**woman**	**politics**
loved	families	man	political	joy	children	girl	political
passion	grandparents	girl	religion	loves	son	boy	politician
loves	mother	boy	politicking	loved	mother	mother	ideology
affection	friends	teenager	ideology	passion	father	daughter	speeches
adore	relatives	person	partisanship	wonderful	wife	pregnant	ideological

Dynamic Embedded Topic Model

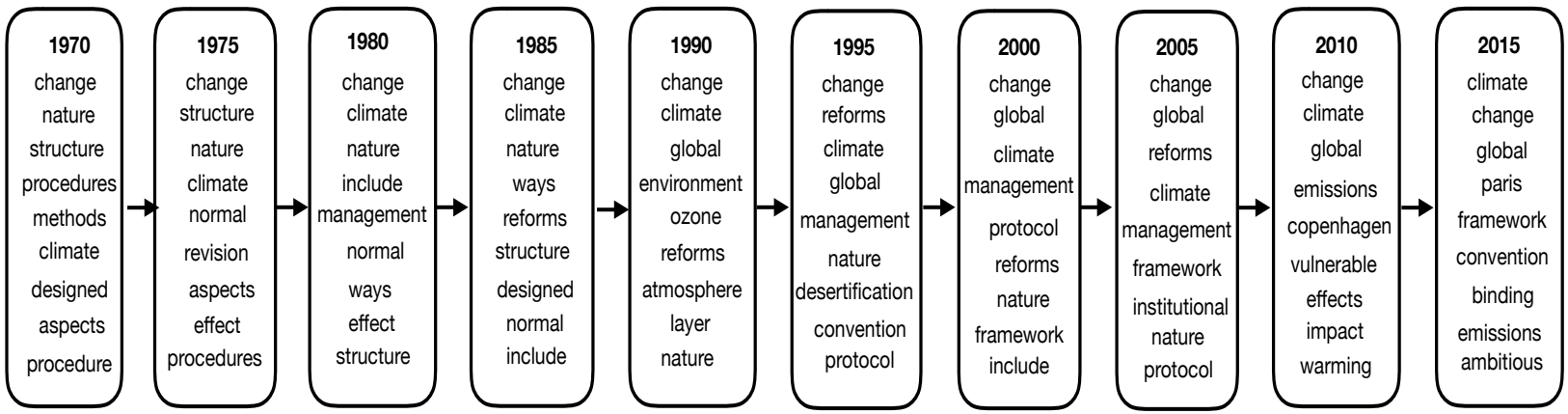
- Define a deterministic per-word embedding $\rho \in \mathbb{R}^{V \times E}$
- Define a latent per-time-step shared global embedding matrix

$$\alpha_t \sim \mathcal{N}(\alpha_{t-1}, \gamma^2 \mathbf{I}) \quad \text{where} \quad \alpha_t \in \mathbb{R}^{K \times E}$$

- For each document d
  1. Draw $\theta_d \sim \mathcal{LN}(\eta_{t_d}, \sigma^2 \mathbf{I})$ where $\eta_t \sim \mathcal{N}(\eta_{t-1}, \delta^2 \mathbf{I})\ \forall t$
  2. For each word n in the document:
     - Draw assignment $\mathbf{z}_{nd} \sim \text{Cat}(\theta_d)$
     - Draw word $w_{nd} \sim \text{Cat}(\text{softmax}(\rho^\top \alpha_{\mathbf{z}_{nd}, t_d}))$
- Same interpretable probabilistic structure as Dynamic LDA
- Fit model using structured amortized variational inference w/ LSTM

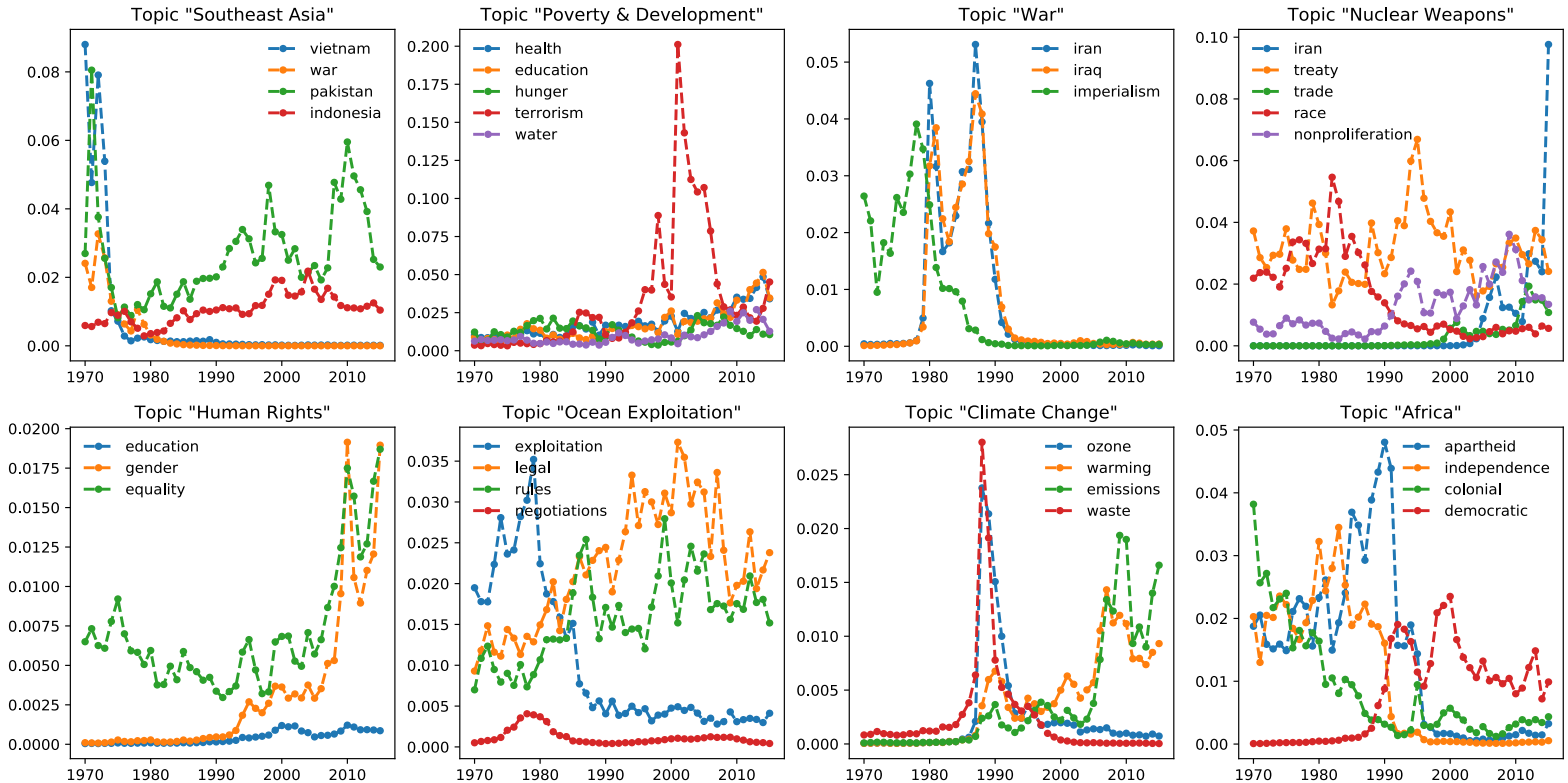
Dynamic Embedded Topic Model

- Trajectory of a topic about climate change found by the model on the United Nations Debates corpus.



Dynamic Embedded Topic Model

- Evolution of word probability across time for eight different topics learned by the model.



Latent Implicit Model Allocator

- Define a deterministic per-observation embedding $\rho \in \mathbb{R}^{V \times E}$
- Define a shared set of K neural networks; each with parameters γ_k
- For each observation d
  1. Draw proportions $\theta_d \sim \mathcal{LN}(\mathbf{0}, \mathbf{I})$
  2. For each element n in d:
     - Draw assignment $\mathbf{z}_{nd} \sim \text{Cat}(\theta_d)$
     - Draw noise $\epsilon_{nd} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
     - Compute landmark $\alpha_{nd} \sim \text{NN}(\epsilon_{nd}; \gamma_{\mathbf{z}_{nd}})$
     - Draw element $x_{nd} \sim \text{ExpFam}(g(\rho^\top \alpha_{nd}))$
- Fit model using amortized variational inference

Latent Implicit Model Allocator

- LIMA vs VAE and other structured deep generative models
- All models have same complexity (#*parameters*)
- Generalization performance as measured by log-likelihood on three benchmark image datasets

Method	MNIST	CIFAR-10	CELEBA
VAE	−85.05	−2121	−6518
SVAE (Johnson et al., 2016)	−85.58	−2144	−6441
kVAE (Locatello et al., 2018)	−87.61	−2217	−7007
LIMA + pretraining	**−84.92**	**−2107**	−6178
LIMA w/o pretraining	−85.20	−2108	**−6073**