Modeling *Science*

David M. Blei

Department of Computer Science
Princeton University

April 17, 2008

Joint work with John Lafferty (CMU)
Poisoning by ice-cream.

No chemist certainly would suppose that the same poison exists in all samples of ice-cream which have produced untoward symptoms in man. Mineral poisons, copper, lead, arsenic, and mercury, have all been found in ice cream. In some instances these have been used with criminal intent. In other cases their presence has been accidental. Likewise, that vanilla is sometimes the bearer, at least, of the poison, is well known to all chemists. Dr. Bartley's idea that the poisonous properties of the cream which he examined were due to putrid gelatine is certainly a rational theory. The poisonous principle might in this ease arise from the decomposition of the gelatine; or with the gelatine there may be introduced into the milk a ferment, by the growth of which a poison is produced.

But in the cream which I examined, none of the above sources of the poisoning existed. There were no mineral poisons present. No gelatine of any kind had been used in making the cream. The vanilla used was shown to be not poisonous. This showing was made, not by a chemical analysis, which might not have been conclusive, but Mr. Davie and I drank of the vanilla extract which was used, and no ill results followed. Still, from this cream we isolated the same poison which I had before found in poisonous cheese (Zeitschrift für physiologische Chemie, x, 26, 97–204). Not that the vanilla extract is poisonous. Everything that we have had to do with vanilla is the result of a very different experience.

---

**RNA Editing and the Evolution of Parasites**

Larry Simpson and Dmitri A. Maslov

The intracellular flagellates, together with their euglenoid and algal ancestry, represent the earliest known lineage of parasitic organisms containing mitochondria (1). Within the diplomonads, there are two major groups, the poorly studied bodonale-cryptophytes, which consist of both branching and nonbranching, and the better known reichardiales, which are obligate parasites (2).

Perhaps because of the ancestry of the diplomonads, these cells possess several unique genetic features (3–5) that give them the potential to shed new light on the evolution of intracellular parasites. This RNA editing function (6–7) creates open reading frames in "introns" by insertion of adenine (or occasional deletion) of uracil; the reading frame is expanded at a few specific sites within the coding region of an mRNA (5–6) editing or at multiple specific sites throughout the mRNA (8–9).

---

**Chaotic Beetles**

Charles Godfrey and Michael Hassell

Eulophis have been used as a pioneering new test system for the study of genetic and evolutionary processes (1–2). They are insect parasites of insects which can be easily cultured in the laboratory. Their reproductive biology is well understood, and they have been extensively used in studies of population genetics and evolution (3–6). Eulophis in this way can serve as a model for the study of genetic and evolutionary processes in natural populations of insects (7). The life cycle of Eulophis is relatively simple, involving one generation of the larva, which matures in about 3 weeks, and then pupates in the soil. Upon emergence, the adult female Eulophis oviposits on the host insect, and the larva develops within the host, feeding on the host's tissues and eventually emerging as an adult.

---

**On-line archives of document collections require better organization. Manual organization is not practical.**

**Our goal: To discover the hidden thematic structure with hierarchical probabilistic models called topic models.**

**Use this structure for browsing, search, and similarity.**
Poisoning by ice-cream.

No chemist certainly would suppose that the same poison exists in all samples of ice-cream which have produced untoward symptoms in man. Mineral poisons, copper, lead, arsenic, and mercury, have all been found in ice-cream. In some instances these have been used with criminal intent. In other cases their presence has been accidental. Likewise, that vanilla is sometimes the bearer, at least, of the poison, is well known to all chemists. Dr. Bartley's idea that the poisonous properties of the cream which he examined were due to putrid gelatine is certainly a rational theory. The poisonous principle might in this ease arise from the decomposition of the gelatine; or with the gelatine there may be introduced into the milk a ferment, by the growth of which a poison is produced...

But in the cream which I examined, none of the above sources of the poisoning existed. There were no mineral poisons present. No gelatine of any kind had been used in making the cream. The vanilla used was shown to be not poisonous. This showing was made, not by a chemical analysis, which might not have been conclusive, but by Mr. Navie and I drank of the vanilla extract which was used, and no ill results followed. Still, from this cream we isolated the same poison which I had before found in poisonous cheese (Zeitschrift für physiologische Chemie, x, 92).

---

RNA Editing and the Evolution of Parasites

Larry Simpson and Dmitri A. Maslov

The karyocellular flagellates, together with their single group of flagellates, represent the earliest known origin of eukaryotic organisms containing mitochondria (1). Within the karyocellulars, there are two major groups, the poorly studied bodonids, and the better known ciliates, which are obligate intracellular parasites of bacteria (2).

Perhaps because of this, the karyocellulars have been thought to have some unusual features. One of the most striking is the ability of some of their species to change their genome through RNA editing. This phenomenon is thought to occur in at least 10% of all bacteria, and is a form of horizontal gene transfer that occurs by reverse transcription.

The RNA editing process involves the insertion or deletion of nucleotides into the RNA transcript, resulting in a change in the corresponding DNA sequence. This process is thought to be important for the evolution of new protein-coding sequences, as well as for the regulation of gene expression.

Furthermore, RNA editing has been shown to be involved in the control of gene expression in many different organisms, including bacteria, plants, and animals. In some cases, RNA editing has been shown to cause changes in the levels of gene expression, leading to changes in the phenotypes of the organisms that carry these changes.

In conclusion, RNA editing is a fascinating process that has important implications for the evolution of new genes and the regulation of gene expression in a wide range of organisms.
Discover topics from a corpus

<table>
<thead>
<tr>
<th>human</th>
<th>evolution</th>
<th>disease</th>
<th>computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>genome</td>
<td>evolutionary</td>
<td>host</td>
<td>models</td>
</tr>
<tr>
<td>dna</td>
<td>species</td>
<td>bacteria</td>
<td>information</td>
</tr>
<tr>
<td>genetic</td>
<td>organisms</td>
<td>diseases</td>
<td>data</td>
</tr>
<tr>
<td>genes</td>
<td>life</td>
<td>resistance</td>
<td>computers</td>
</tr>
<tr>
<td>sequence</td>
<td>origin</td>
<td>bacterial</td>
<td>system</td>
</tr>
<tr>
<td>gene</td>
<td>biology</td>
<td>new</td>
<td>network</td>
</tr>
<tr>
<td>molecular</td>
<td>groups</td>
<td>strains</td>
<td>systems</td>
</tr>
<tr>
<td>sequencing</td>
<td>phylogenetic</td>
<td>control</td>
<td>model</td>
</tr>
<tr>
<td>map</td>
<td>living</td>
<td>infectious</td>
<td>parallel</td>
</tr>
<tr>
<td>information</td>
<td>diversity</td>
<td>malaria</td>
<td>methods</td>
</tr>
<tr>
<td>genetics</td>
<td>group</td>
<td>parasite</td>
<td>networks</td>
</tr>
<tr>
<td>mapping</td>
<td>new</td>
<td>parasites</td>
<td>software</td>
</tr>
<tr>
<td>project</td>
<td>two</td>
<td>united</td>
<td>new</td>
</tr>
<tr>
<td>sequences</td>
<td>common</td>
<td>tuberculosis</td>
<td>simulations</td>
</tr>
</tbody>
</table>
Model the evolution of topics over time

"Theoretical Physics"

FORCEx
RELATIVITY
LASER

"Neuroscience"

OXYGEN
NERVE
NEURON

1880 1900 1920 1940 1960 1980 2000

Model connections between topics
1 Introduction

2 Latent Dirichlet allocation

3 Dynamic topic models

4 Correlated topic models
1. Treat data as observations that arise from a generative probabilistic process that includes hidden variables
   - For documents, the hidden variables reflect the thematic structure of the collection.
2. Infer the hidden structure using posterior inference
   - What are the topics that describe this collection?
3. Situate new data into the estimated model.
   - How does this query or new document fit into the estimated topic structure?
Intuition behind LDA

Simple intuition: Documents exhibit multiple topics.
• Cast these intuitions into a generative probabilistic process

• Each document is a random mixture of corpus-wide topics

• Each word is drawn from one of those topics
In reality, we only observe the documents

Our goal is to infer the underlying topic structure

- What are the topics?
- How are the documents divided according to those topics?
Graphical models (Aside)

- Nodes are random variables
- Edges denote possible dependence
- Observed variables are shaded
- Plates denote replicated structure
Graphical models (Aside)

- Structure of the graph defines the pattern of conditional dependence between the ensemble of random variables
- E.g., this graph corresponds to

\[
p(y, x_1, \ldots, x_N) = p(y) \prod_{n=1}^{N} p(x_n \mid y)
\]
Each piece of the structure is a random variable.
Latent Dirichlet allocation

1. Draw each topic $\beta_i \sim \text{Dir}(\eta)$, for $i \in \{1, \ldots, K\}$.

2. For each document:
   1. Draw topic proportions $\theta_d \sim \text{Dir}(\alpha)$.
   2. For each word:
      1. Draw $Z_{d,n} \sim \text{Mult}(\theta_d)$.
      2. Draw $W_{d,n} \sim \text{Mult}(\beta_{Z_{d,n}})$. 

---

D. Blei  Modeling Science 13 / 53
From a collection of documents, infer
- Per-word topic assignment $z_{d,n}$
- Per-document topic proportions $\theta_d$
- Per-corpus topic distributions $\beta_k$

Use posterior expectations to perform the task at hand, e.g., information retrieval, document similarity, etc.
• Computing the posterior is intractable:

\[
\frac{p(\theta | \alpha) \prod_{n=1}^{N} p(z_n | \theta)p(w_n | z_n, \beta_{1:K})}{\int_\theta p(\theta | \alpha) \prod_{n=1}^{N} \sum_{z=1}^{K} p(z_n | \theta)p(w_n | z_n, \beta_{1:K})}
\]

• Several approximation techniques have been developed.
Latent Dirichlet allocation

- Mean field variational methods (Blei et al., 2001, 2003)
- Expectation propagation (Minka and Lafferty, 2002)
- Collapsed Gibbs sampling (Griffiths and Steyvers, 2002)
- Collapsed variational inference (Teh et al., 2006)
Example inference

**Data**: The OCR’ed collection of *Science* from 1990–2000
- 17K documents
- 11M words
- 20K unique terms (stop words and rare words removed)

**Model**: 100-topic LDA model using variational inference.
Seeking Life’s Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here, two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today’s organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a single parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn’t be enough.

Although the numbers don’t match precisely, those predictions “are not all that far apart,” especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. “It may be a way of organizing any newly sequenced genome,” explains Aracdy Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an


SCIENCE • VOL. 272 • 24 MAY 1996
Example topics

human
genome
dna
genetic
genes
sequence
gene
molecular
sequencing
map
information
genetics
mapping
project
sequences

evolution
evolutionary
species
organisms
life
origin
biology
groups
phylogenetic
living
diversity
group
new
two

disease
host
bacteria
diseases
resistance
bacterial
new
strains
control
infectious
malaria
parasite
parasites
united
	
tuberculosis

computer
models
information
data
computers
system
network
systems
model
parallel
methods
networks
software
new
simulations

D. Blei  Modeling Science  16 / 53
LDA summary

- LDA is a powerful model for
  - Visualizing the hidden thematic structure in large corpora
  - Generalizing new data to fit into that structure
- LDA is a mixed membership model (Erosheva, 2004) that builds on the work of Deerwester et al. (1990) and Hofmann (1999).
- For document collections and other grouped data, this might be more appropriate than a simple finite mixture
LDA summary

- **Modular**: It can be embedded in more complicated models.
  - E.g., syntax and semantics; authorship; word sense
- **General**: The data generating distribution can be changed.
  - E.g., images; social networks; population genetics data
- Variational inference is fast; lets us to analyze large data sets.

- See Blei et al., 2003 for details and a quantitative comparison.
- Code to play with LDA is freely available on my web-site, http://www.cs.princeton.edu/~blei.
LDA summary

- But, LDA makes certain assumptions about the data.
- When are they appropriate?
Outline

1 Introduction

2 Latent Dirichlet allocation

3 Dynamic topic models

4 Correlated topic models
LDA and exchangeability

- LDA assumes that documents are exchangeable.
- I.e., their joint probability is invariant to permutation.
- This is too restrictive.
"Instantaneous Photography" (1890)

"Infrared Reflectance in Leaf-Sitting Neotropical Frogs" (1977)

Documents are not exchangeable

- Documents about the same topic are not exchangeable.
- Topics evolve over time.
Dynamic topic model

- Divide corpus into sequential slices (e.g., by year).
- Assume each slice’s documents exchangeable.
  - Drawn from an LDA model.
- Allow topic distributions evolve from slice to slice.
Dynamic topic models

\begin{align*}
D & \quad \theta_d \\
\alpha & \quad \theta_d \\
Z_{d,n} & \quad \theta_d \\
W_{d,n} & \quad \theta_d \\
N & \quad \theta_d \\
K & \quad \beta_{k,1} \\
\alpha & \quad \beta_{k,1} \\
Z_{d,n} & \quad \beta_{k,1} \\
W_{d,n} & \quad \beta_{k,1} \\
N & \quad \beta_{k,1} \\
D & \quad \beta_{k,1} \\
K & \quad \beta_{k,2} \\
\alpha & \quad \beta_{k,2} \\
Z_{d,n} & \quad \beta_{k,2} \\
W_{d,n} & \quad \beta_{k,2} \\
N & \quad \beta_{k,2} \\
D & \quad \beta_{k,2} \\
K & \quad \beta_{k,T} \\
\alpha & \quad \beta_{k,T} \\
Z_{d,n} & \quad \beta_{k,T} \\
W_{d,n} & \quad \beta_{k,T} \\
N & \quad \beta_{k,T} \\
D & \quad \beta_{k,T} \\
\end{align*}
Modeling evolving topics

- Use a logistic normal distribution to model evolving topics (Aitchison, 1980)
- A state-space model on the natural parameter of the topic multinomial (West and Harrison, 1997)

\[
\beta_{t,k} \mid \beta_{t-1,k} \sim \mathcal{N}(\beta_{t-1,k}, I\sigma^2)
\]

\[
p(w \mid \beta_{t,k}) = \exp \left\{ \beta_{t,k} - \log(1 + \sum_{v=1}^{V-1} \exp\{\beta_{t,k,v}\}) \right\}
\]
Posterior inference

• Our goal is to compute the posterior distribution,

\[ p(\beta_{1:T,1:K}, \theta_{1:T,1:D}, z_{1:T,1:D} \mid w_{1:T,1:D}) \].

• Exact inference is impossible
  • Per-document mixed-membership model
  • Non-conjugacy between \( p(w \mid \beta_{t,k}) \) and \( p(\beta_{t,k}) \)

• MCMC is not practical for the amount of data.

• Solution: Variational inference
Genome sequencing projects reveal the genetic makeup of an organism by reading off the sequence of the DNA bases, which encodes all of the information necessary for the life of the organism. The base sequence contains four nucleotides—adenine, thymidine, guanosine, and cytosine—which are linked together into long double-helical chains. Over the last two decades, automated DNA sequencers have made the process of obtaining the base-by-base sequence of DNA...

- Analyze JSTOR’s entire collection from *Science* (1880-2002)
- Restrict to 30K terms that occur more than ten times
- The data are 76M words in 130K documents
Sequencing the Genome, Fast
James C. Malin and Amanda A. Murray

D. Blei
Modeling Science
Analyzing a document

Original article

**TECH VIEW: DNA SEQUENCING**

**Sequencing the Genome, Fast**

Jawneice C. Hadfield and Amanda A. McMurray

DNA sequencing projects reveal the genetic makeup of an organism. These projects involve sequencing DNA bases, which encode all of the information necessary for the life of the organism. The base sequence contains four nucleotides—adenine, thymine, guanine, and cytosine—which are linked together into long double-stranded chains. The process involves identifying the sequence of DNA bases, which makes up the genome of an organism.

The process of sequencing DNA involves several steps. First, the DNA is isolated from the organism and fragmented into small pieces. These fragments are then amplified using PCR (Polymerase Chain Reaction) to create a large number of copies. The amplified DNA is then subjected to a series of reactions to separate the DNA fragments into their constituent nucleotides. The nucleotides are then identified using a variety of methods, such as electrophoresis or mass spectrometry.

The results of these reactions are then analyzed to determine the sequence of the DNA. This information can be used to understand the function of genes and the regulation of gene expression. Additionally, it can be used to identify new genes and to study the evolution of species. DNA sequencing has revolutionized many fields, including medicine, biology, and genetics.

**Most likely words from top topics**

- sequence
- genome
- genes
- sequences
- human
- gene
dna
- sequencing
- chromosome
- regions
- analysis
- data
- genomic
- number
- devices
- device
- materials
- current
- high
- gate
- light
- silicon
- material
- technology
- electrical
- fiber
- power
- based
- data
- information
- network
- web
- computer
- language
- networks
- time
- software
- system
- words
- algorithm
- number
- internet
Visualizing trends within a topic

"Theoretical Physics"

FORCE
RELATIVITY
LASER

"Neuroscience"

OXYGEN
NERVE
NEURON
Time-corrected document similarity

- Consider the expected Hellinger distance between the topic proportions of two documents,

\[ d_{ij} = E \left[ \sum_{k=1}^{K} (\sqrt{\theta_{i,k}} - \sqrt{\theta_{j,k}})^2 \mid w_i, w_j \right] \]

- Uses the latent structure to define similarity
- Time has been factored out because the topics associated to the components are different from year to year.
- Similarity based only on topic proportions
The Brain of the Orang (1880)
Representation of the Visual Field on the Medial Wall of Occipital-Parietal Cortex in the Owl Monkey (1976)
Automatic Analysis, Theme Generation, and Summarization of Machine-Readable Texts

Gerard Salton, James Allan, Chris Buckley.

Vast amounts of text material are now available in machine-readable form and are amenable to automatic processing. Because the available databases are large and cover many different subject areas, automatic aids must be provided to users interested in accessing the data. It has been suggested that links be placed between related pieces of text, connecting, for example, particular text paragraphs or other paragraphs covering related subject matter. Such a linked text structure, often called hypertext, makes it feasible for the reader to start with particular text passages and use the linked structure to find related text elements. Unfortunately, until now, viable methods for automatically building large hypertext structures and for using such structures in a sophisticated way have not been available. Here we give methods for constructing text relation maps and for using text relations to access and use text databases. In particular, we outline procedures for determining text themes, traversing texts selectively, and extracting summary statements that reflect text content.

Text Analysis and Retrieval: The Smart System

The Smart system is a sophisticated text retrieval tool, developed over the past 20 years, that is based on the vector space model of retrieval. In this model, all inform. as well as inform. sensed by texts, or as it is typically a word, associated with the text. In principle, a text consists of a set of text elements, but by constructing such an uncorrected text to derive the terms under consideration, terms assigned to a text content.

Because the text for consent represents a term-weighted, high weights to and lower weights to A powerful technique is the well-known term frequency x inverse document frequency (TF-IDF). This technique is used in many text similarity measures. Thus, p


<table>
<thead>
<tr>
<th>TOPIC</th>
<th>PROB</th>
</tr>
</thead>
<tbody>
<tr>
<td>data computer system information network</td>
<td>0.30</td>
</tr>
<tr>
<td>information library text index libraries</td>
<td>0.19</td>
</tr>
<tr>
<td>two three four different single</td>
<td>0.16</td>
</tr>
</tbody>
</table>

D. Blei Modeling Science

Global Text Matching for Information Retrieval

Gerard Salton and Chris Buckley

An approach is outlined for the retrieval of natural language texts in response to available search requests and for the recognition of common similarities between text excerpts. The proposed retrieval system is based on a text retrieval procedure carried out in a number of different text environments and is applicable to large text collections covering a wide range of subject matter. For unstructured text environments this system appears to perform better than currently available methods.
Quantitative comparison

- Compute the probability of each year’s documents conditional on all the previous year’s documents,

\[ p(w_t \mid w_1, \ldots, w_{t-1}) \]

- Compare exchangeable and dynamic topic models
Quantitative comparison

![Graph showing Per-word negative log likelihood over years with lines for LDA and DTM]
Outline

1 Introduction

2 Latent Dirichlet allocation

3 Dynamic topic models

4 Correlated topic models
The Dirichlet is an exponential family distribution on the simplex, positive vectors that sum to one.

However, the near independence of components makes it a poor choice for modeling topic proportions.

An article about fossil fuels is more likely to also be about geology than about genetics.
The logistic normal is a distribution on the simplex that can model dependence between components.

- The natural parameters of the multinomial are drawn from a multivariate Gaussian distribution.

\[
X \sim \mathcal{N}_{K-1}(\mu, \Sigma)
\]

\[
\theta_i = \exp\{x_i - \log(1 + \sum_{j=1}^{K-1} \exp\{x_j\})\}
\]
Correlated topic model (CTM)

- Draw topic proportions from a logistic normal, where topic occurrences can exhibit correlation.
- Use for:
  - Providing a “map” of topics and how they are related
  - Better prediction via correlated topics
Summary

• Topic models provide useful descriptive statistics for analyzing and understanding the latent structure of large text collections.

• Probabilistic graphical models are a useful way to express assumptions about the hidden structure of complicated data.

• Variational methods allow us to perform posterior inference to automatically infer that structure from large data sets.

• Current research
  • Choosing the number of topics
  • Continuous time dynamic topic models
  • Topic models for prediction
  • Inferring the impact of a document
“We should seek out unfamiliar summaries of observational material, and establish their useful properties... And still more novelty can come from finding, and evading, still deeper lying constraints.” (John Tukey, *The Future of Data Analysis*, 1962)
Most topic models are *unsupervised*. They are fit by maximizing the likelihood of a collection of documents.

Consider documents paired with response variables. For example:

- Movie reviews paired with a number of stars
- Web pages paired with a number of “diggs”

We develop *supervised topic models*, models of documents and responses that are fit to find topics predictive of the response.
Supervised LDA

1. Draw topic proportions $\theta | \alpha \sim \text{Dir}(\alpha)$.
2. For each word
   1. Draw topic assignment $z_n | \theta \sim \text{Mult}(\theta)$.
   2. Draw word $w_n | z_n, \beta_{1:K} \sim \text{Mult}(\beta_{z_n})$.
3. Draw response variable $y | z_{1:N}, \eta, \sigma^2 \sim \text{N}(\eta^\top \bar{z}, \sigma^2)$, where

$$
\bar{z} = (1/N) \sum_{n=1}^{N} z_n.
$$
• SLDA is used as follows.
  • Fit coefficients and topics from a collection of document-response pairs.
  • Use the fitted model to predict the responses of previously unseen documents,

\[
E[Y \mid w_{1:N}, \alpha, \beta_{1:K}, \eta, \sigma^2] = \eta^\top E[\bar{Z} \mid w_{1:N}, \alpha, \beta_{1:K}].
\]

• The process enforces that the document is generated first, followed by the response. The response is generated from the particular topics that were realized in generating the document.
• We fit a 10-topic sLDA model to movie review data (Pang and Lee, 2005).
  • The documents are the words of the reviews.
  • The responses are the number of stars associated with each review (modeled as continuous).
• Each component of coefficient vector $\eta$ is associated with a topic.
Simulations

Movie corpus

Digg corpus

sLDA
LDA

D. Blei
Modeling Science
Let $x_{1:N}$ be observations and $z_{1:M}$ be latent variables.

Our goal is to compute the posterior distribution

$$p(z_{1:M} | x_{1:N}) = \frac{p(z_{1:M}, x_{1:N})}{\int p(z_{1:M}, x_{1:N}) \, dz_{1:M}}$$

For many interesting distributions, the marginal likelihood of the observations is difficult to efficiently compute.
Variational inference

- Use Jensen’s inequality to bound the log prob of the observations:

\[
\log p(x_{1:N}) \geq E_{q_\nu} [\log p(z_{1:M}, x_{1:N})] - E_{q_\nu} [\log q_\nu(z_{1:M})].
\]

- We have introduced a distribution of the latent variables with free variational parameters \( \nu \).
- We optimize those parameters to tighten this bound.
- This is the same as finding the member of the family \( q_\nu \) that is closest in KL divergence to \( p(z_{1:M} \mid x_{1:N}) \).
Mean-field variational inference

- Complexity of optimization is determined by factorization of $q_\nu$
- In mean field variational inference $q_\nu$ is fully factored

\[
q_\nu(z_{1:M}) = \prod_{m=1}^{M} q_{\nu m}(z_m).
\]

- The latent variables are independent.
  - Each is governed by its own variational parameter $\nu_m$.
- In the true posterior they can exhibit dependence (often, this is what makes exact inference difficult).
Suppose the distribution of each latent variable conditional on the observations and other latent variables is in the exponential family:

\[ p(z_m | z_{-m}, x) = h_m(z_m) \exp\{g_m(z_{-m}, x)^T z_m - a_m(g_i(z_{-m}, x))\} \]

Assume \( q_\nu \) is fully factorized and each factor is in the same exponential family:

\[ q_{\nu_m}(z_m) = h_m(z_m) \exp\{\nu_m^T z_m - a_m(\nu_m)\} \]
Variational inference is the following coordinate ascent algorithm

$$\nu_m = \mathbb{E}_{q_v}[g_m(z_{-m}, x)]$$

Notice the relationship to Gibbs sampling
Variational family for the DTM

- Distribution of $\theta$ and $z$ is fully-factorized (Blei et al., 2003)
- Distribution of $\{\beta_{1,k}, \ldots, \beta_{T,k}\}$ is a variational Kalman filter
- Gaussian state-space model with free observations $\hat{\beta}_{k,t}$.
- Fit observations such that the corresponding posterior over the chain is close to the true posterior.
Given a document collection, use coordinate ascent on all the variational parameters until the KL converges.

Yields a distribution close to the true posterior of interest

Take expectations w/r/t the simpler variational distribution