LSI, pLSI, LDA and inference methods

Guillaume Obozinski

INRIA - Ecole Normale Supérieure - Paris

RussIR summer school

Yaroslavl, August 6-10th 2012

Latent Semantic Indexing

Latent Semantic Indexing (LSI) (Deerwester et al., 1990)

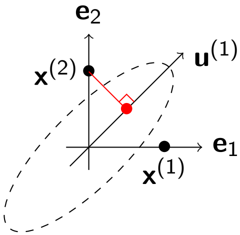
Idea: words that co-occur frequently in documents should be similar.

Let $x_1^{(i)}$ and $x_2^{(i)}$ count resp. the number of occurrences of the words `physician` and `doctor` in the i^{th} document.

Latent Semantic Indexing (LSI) (Deerwester et al., 1990)

Idea: words that co-occur frequently in documents should be similar.

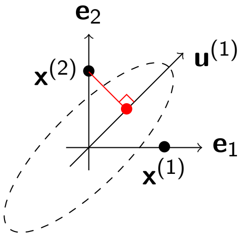
Let $x_1^{(i)}$ and $x_2^{(i)}$ count resp. the number of occurrences of the words `physician` and `doctor` in the i^{th} document.



Latent Semantic Indexing (LSI) (Deerwester et al., 1990)

Idea: words that co-occur frequently in documents should be similar.

Let $x_1^{(i)}$ and $x_2^{(i)}$ count resp. the number of occurrences of the words `physician` and `doctor` in the i^{th} document.



The directions of covariance or **principal directions** are obtained using the **singular value decomposition of** $X \in \mathbb{R}^{d \times N}$

$$X = USV^\top, \quad \text{with} \quad U^\top U = I_d \quad \text{and} \quad V^\top V = I_N$$

and $S \in \mathbb{R}^{d \times N}$ a matrix with non-zero element only on the diagonal: *the singular values* of X, positives and sorted in decreasing order.

LSI: computation of the document representation

$$U = \begin{bmatrix} | & & | \\ \mathbf{u}^{(1)} & \dots & \mathbf{u}^{(d)} \\ | & & | \end{bmatrix}$$: the **principal directions**.

Let $U_K \in \mathbb{R}^{d \times K}$, $V_K \in \mathbb{R}^{N \times K}$ be the matrices retaining the K first columns and $S_K \in \mathbb{R}^{K \times K}$ the top left $K \times K$ corner of S.

LSI: computation of the document representation

$$U = \begin{bmatrix} | & & | \\ \mathbf{u}^{(1)} & \dots & \mathbf{u}^{(d)} \\ | & & | \end{bmatrix}$$: the **principal directions**.

Let $U_K \in \mathbb{R}^{d \times K}$, $V_K \in \mathbb{R}^{N \times K}$ be the matrices retaining the K first columns and $S_K \in \mathbb{R}^{K \times K}$ the top left $K \times K$ corner of S.

LSI: computation of the document representation

$$U = \begin{bmatrix} | & & | \\ \mathbf{u}^{(1)} & \dots & \mathbf{u}^{(d)} \\ | & & | \end{bmatrix}$$: the **principal directions**.

Let $U_K \in \mathbb{R}^{d \times K}$, $V_K \in \mathbb{R}^{N \times K}$ be the matrices retaining the K first columns and $S_K \in \mathbb{R}^{K \times K}$ the top left $K \times K$ corner of S.

The projection of $\mathbf{x}^{(i)}$ on the subspace spanned by U_K yields the

Latent representation: $\tilde{\mathbf{x}}^{(i)} = U_K^\top \mathbf{x}^{(i)}$.

LSI: computation of the document representation

$$U = \begin{bmatrix} | & & | \\ \mathbf{u}^{(1)} & \dots & \mathbf{u}^{(d)} \\ | & & | \end{bmatrix}$$: the **principal directions**.

Let $U_K \in \mathbb{R}^{d\times K}, V_K \in \mathbb{R}^{N\times K}$ be the matrices retaining the K first columns and $S_K \in \mathbb{R}^{K\times K}$ the top left $K \times K$ corner of S.
The projection of $\mathbf{x}^{(i)}$ on the subspace spanned by U_K yields the

Latent representation: $\qquad \tilde{\mathbf{x}}^{(i)} = U_K^\top \mathbf{x}^{(i)}.$

Remarks

- $U_K^\top X = U_K^\top U_K S_K V_K^\top = S_K V_K^\top$

LSI: computation of the document representation

$U = \begin{bmatrix} | & & | \\ \mathbf{u}^{(1)} & \dots & \mathbf{u}^{(d)} \\ | & & | \end{bmatrix}$: the **principal directions**.

Let $U_K \in \mathbb{R}^{d \times K}, V_K \in \mathbb{R}^{N \times K}$ be the matrices retaining the K first columns and $S_K \in \mathbb{R}^{K \times K}$ the top left $K \times K$ corner of S.
The projection of $\mathbf{x}^{(i)}$ on the subspace spanned by U_K yields the

$$\textbf{Latent representation:} \qquad \tilde{\mathbf{x}}^{(i)} = U_K^\top \mathbf{x}^{(i)}.$$

Remarks

- $U_K^\top X = U_K^\top U_K S_K V_K^\top = S_K V_K^\top$
- $\mathbf{u}^{(k)}$ is somehow like a **topic** and $\tilde{\mathbf{x}}^{(i)}$ is the vector of **coefficients of decomposition** of a document on the K "topics".

LSI: computation of the document representation

$$U = \begin{bmatrix} | & & | \\ \mathbf{u}^{(1)} & \dots & \mathbf{u}^{(d)} \\ | & & | \end{bmatrix}$$: the **principal directions**.

Let $U_K \in \mathbb{R}^{d \times K}$, $V_K \in \mathbb{R}^{N \times K}$ be the matrices retaining the K first columns and $S_K \in \mathbb{R}^{K \times K}$ the top left $K \times K$ corner of S.
The projection of $\mathbf{x}^{(i)}$ on the subspace spanned by U_K yields the

Latent representation: $\quad \tilde{\mathbf{x}}^{(i)} = U_K^\top \mathbf{x}^{(i)}$.

Remarks

- $U_K^\top X = U_K^\top U_K S_K V_K^\top = S_K V_K^\top$
- $\mathbf{u}^{(k)}$ is somehow like a **topic** and $\tilde{\mathbf{x}}^{(i)}$ is the vector of **coefficients of decomposition** of a document on the K "topics".
- The similarity between two documents can now be measured by

$$\cos(\angle(\tilde{\mathbf{x}}^{(i)}, \tilde{\mathbf{x}}^{(j)})) = \frac{\tilde{\mathbf{x}}^{(i)}}{\|\tilde{\mathbf{x}}^{(i)}\|} \cdot \frac{\tilde{\mathbf{x}}^{(j)}}{\|\tilde{\mathbf{x}}^{(j)}\|}$$

LSI vs PCA

LSI is almost identical to Principal Component Analysis (PCA) proposed by Karl Pearson in 1901.

LSI vs PCA

LSI is almost identical to Principal Component Analysis (PCA) proposed by Karl Pearson in 1901.

- Like PCA, LSI aims at finding the directions of high correlations between words called principal directions.
- Like PCA, it retains the projection of the data on a number k of these principal directions, which are called the **principal components**.

LSI vs PCA

LSI is almost identical to Principal Component Analysis (PCA) proposed by Karl Pearson in 1901.

- Like PCA, LSI aims at finding the directions of high correlations between words called principal directions.
- Like PCA, it retains the projection of the data on a number k of these principal directions, which are called the **principal components**.
- Difference between LSI and PCA
  - LSI does not center the data (no specific reason).
  - LSI is typically combined with TF-IDF

Limitations and shortcomings of LSI

- The generative model of the data underlying PCA is a Gaussian cloud which does not match the structure of the data.

Limitations and shortcomings of LSI

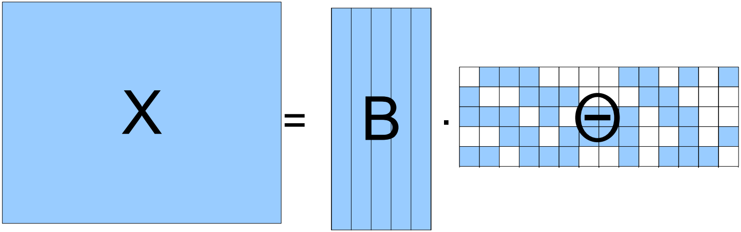
- The generative model of the data underlying PCA is a Gaussian cloud which does not match the structure of the data.
- In particular: LSI ignores
  - That the data are counts, frequencies or tf-idf scores.
  - The data is positive ($\mathbf{u}_k$ typically has negative coefficients)

Limitations and shortcomings of LSI

- The generative model of the data underlying PCA is a Gaussian cloud which does not match the structure of the data.
- In particular: LSI ignores
  - That the data are counts, frequencies or tf-idf scores.
  - The data is positive ($\mathbf{u}_k$ typically has negative coefficients)
- The singular value decomposition is expensive to compute

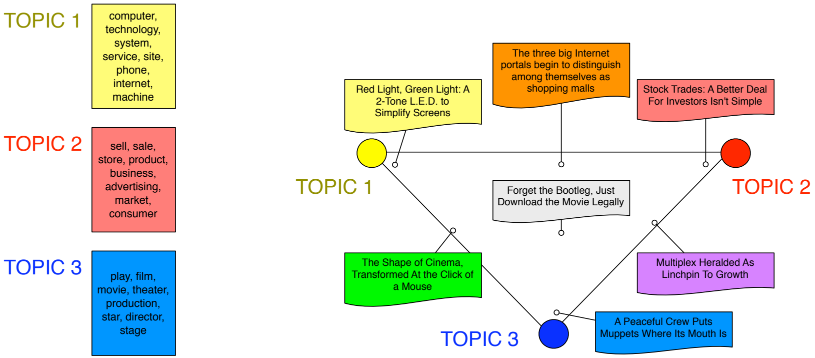
Topic models and matrix factorization

- $\mathbf{X} \in \mathbb{R}^{d \times M}$ with columns $\mathbf{x}_i$ corresponding to documents
- $\mathbf{B}$ the matrix whose columns correspond to different topics
- Θ the matrix of decomposition coefficients with columns $\boldsymbol{\theta}_i$ associated each to one document and which encodes its "topic content".



Probabilistic LSI

Probabilistic Latent Semantic Indexing (Hofmann, 2001)



(a) Topics

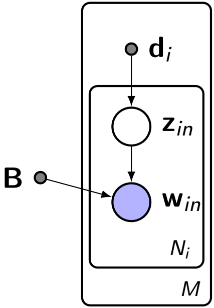
(b) Document Assignments to Topics

Probabilistic Latent Semantic Indexing (Hofmann, 2001)

Obtain a more expressive model by allowing several topics per document in various proportions so that each word $\mathbf{w}_{in}$ gets its own topic $\mathbf{z}_{in}$ drawn from the multinomial distribution $\mathbf{d}_i$ unique to the i^{th} document.

Probabilistic Latent Semantic Indexing (Hofmann, 2001)

Obtain a more expressive model by allowing several topics per document in various proportions so that each word $\mathbf{w}_{in}$ gets its own topic $\mathbf{z}_{in}$ drawn from the multinomial distribution $\mathbf{d}_i$ unique to the i^{th} document.



- $\mathbf{d}_i$ topic proportions in document i
- $\mathbf{z}_{in} \sim \mathcal{M}(1, \mathbf{d}_i)$
- $(\mathbf{w}_{in} | z_{ink} = 1) \sim \mathcal{M}(1, (b_{1k}, \ldots, b_{dk}))$

EM algorithm for pLSI

Denote j^*_{in} the index in the dictionary of the word appearing in document i as the nth word.

EM algorithm for pLSI

Denote j^*_{in} the index in the dictionary of the word appearing in document i as the nth word.

Expectation step

$$q_{ink}^{(t)} = p(z_{ink} = 1 \mid \mathbf{w}_{in}\,;\, \mathbf{d}_i^{(t-1)}, \mathbf{B}^{(t-1)}) = \frac{d_{ik}^{(t-1)}\, b_{j^*_{in}k}^{(t-1)}}{\sum_{k'=1}^{K} d_{ik'}^{(t-1)}\, b_{j^*_{in}k'}^{(t-1)}}$$

Maximization step

$$d_{ik}^{(t)} = \frac{\sum_{n=1}^{N^{(i)}} q_{ink}^{(t)}}{\sum_{n}^{N^{(i)}} \sum_{k'=1}^{K} q_{ink'}^{(t)}} = \frac{\tilde{N}_k^{(i)}}{N^{(i)}} \qquad \text{and} \qquad b_{jk}^{(t)} = \frac{\sum_{i=1}^{M} \sum_{n=1}^{N^{(i)}} q_{ink}^{(t)}\, w_{inj}}{\sum_{i=1}^{M} \sum_{n=1}^{N^{(i)}} q_{ink}^{(t)}}$$

Issues with pLSI

Too many parameters $\rightarrow$ overfitting !

Issues with pLSI **?**

Too many parameters $\rightarrow$ overfitting ! **Not clear**

Solutions

Issues with pLSI **?**

Too many parameters $\rightarrow$ overfitting ! **Not clear**

Solutions or alternative approaches

- Frequentist approach: regularize + optimize $\rightarrow$ *Dictionary Learning*

$$\min_{\boldsymbol{\theta}_i} -\log p(\mathbf{x}_i|\boldsymbol{\theta}_i) + \lambda\Omega(\boldsymbol{\theta}_i)$$

Issues with pLSI ?

Too many parameters $\rightarrow$ overfitting ! **Not clear**

Solutions or alternative approaches

- Frequentist approach: regularize + optimize $\rightarrow$ *Dictionary Learning*

$$\min_{\boldsymbol{\theta}_i} -\log p(\mathbf{x}_i|\boldsymbol{\theta}_i) + \lambda\Omega(\boldsymbol{\theta}_i)$$

- Bayesian approach: prior + integrate $\rightarrow$ Latent Dirichlet Allocation

$$p(\boldsymbol{\theta}_i|\mathbf{x}_i, \boldsymbol{\alpha}) \propto p(\mathbf{x}_i|\boldsymbol{\theta}_i)\, p(\boldsymbol{\theta}_i|\boldsymbol{\alpha})$$

Issues with pLSI ?

Too many parameters $\rightarrow$ overfitting ! **Not clear**

Solutions or alternative approaches

- Frequentist approach: regularize + optimize $\rightarrow$ *Dictionary Learning*

$$\min_{\boldsymbol{\theta}_i} -\log p(\mathbf{x}_i|\boldsymbol{\theta}_i) + \lambda\Omega(\boldsymbol{\theta}_i)$$

- Bayesian approach: prior + integrate $\rightarrow$ Latent Dirichlet Allocation

$$p(\boldsymbol{\theta}_i|\mathbf{x}_i, \boldsymbol{\alpha}) \propto p(\mathbf{x}_i|\boldsymbol{\theta}_i)\, p(\boldsymbol{\theta}_i|\boldsymbol{\alpha})$$

- "Frequentist + Bayesian" $\rightarrow$ integrate + optimize

$$\max_{\boldsymbol{\alpha}} \prod_{i=1}^{M} \int p(\mathbf{x}_i|\boldsymbol{\theta}_i)\, p(\boldsymbol{\theta}_i|\boldsymbol{\alpha})\, d\boldsymbol{\theta}$$

Issues with pLSI ?

Too many parameters $\rightarrow$ overfitting ! **Not clear**

Solutions or alternative approaches

- Frequentist approach: regularize + optimize $\rightarrow$ *Dictionary Learning*

$$\min_{\boldsymbol{\theta}_i} -\log p(\mathbf{x}_i|\boldsymbol{\theta}_i) + \lambda\Omega(\boldsymbol{\theta}_i)$$

- Bayesian approach: prior + integrate $\rightarrow$ Latent Dirichlet Allocation

$$p(\boldsymbol{\theta}_i|\mathbf{x}_i, \boldsymbol{\alpha}) \propto p(\mathbf{x}_i|\boldsymbol{\theta}_i)\, p(\boldsymbol{\theta}_i|\boldsymbol{\alpha})$$

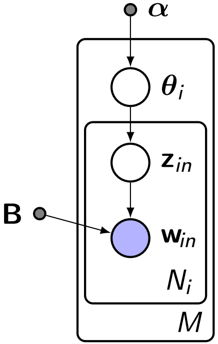
- "Frequentist + Bayesian" $\rightarrow$ integrate + optimize

$$\max_{\boldsymbol{\alpha}} \prod_{i=1}^{M} \int p(\mathbf{x}_i|\boldsymbol{\theta}_i)\, p(\boldsymbol{\theta}_i|\boldsymbol{\alpha})\, d\boldsymbol{\theta}$$

... called *Empirical Bayes* approach or Type II Maximum Likelihood

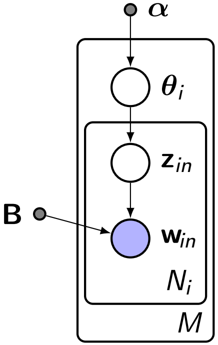
Latent Dirichlet Allocation

Latent Dirichlet Allocation (Blei et al., 2003)



- K topics
- $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ parameter vector
- $\boldsymbol{\theta}_i = (\theta_{1i}, \ldots, \theta_{Ki}) \sim \mathrm{Dir}(\boldsymbol{\alpha})$ topic proportions
- $\mathbf{z}_{in}$ topic indicator vector for n^{th} word of i^{th} document:
  - $\mathbf{z} = (z_{in1}, \ldots, z_{inK})^\top \in \{0, 1\}^K$
  - $\mathbf{z}_{in} \sim \mathcal{M}(1, (\theta_{1i}, \ldots, \theta_{Ki}))$
  - $p(\mathbf{z}_{in}|\boldsymbol{\theta}_i) = \prod_{k=1}^{K} \left[\theta_{ki}\right]^{z_{ink}}$

Latent Dirichlet Allocation (Blei et al., 2003)



- K topics
- $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ parameter vector
- $\boldsymbol{\theta}_i = (\theta_{1i}, \ldots, \theta_{Ki}) \sim \mathrm{Dir}(\boldsymbol{\alpha})$ topic proportions
- $\mathbf{z}_{in}$ topic indicator vector for n^{th} word of i^{th} document:
  - $\mathbf{z} = (z_{in1}, \ldots, z_{inK})^\top \in \{0,1\}^K$
  - $\mathbf{z}_{in} \sim \mathcal{M}(1, (\theta_{1i}, \ldots, \theta_{Ki}))$
  - $p(\mathbf{z}_{in}|\boldsymbol{\theta}_i) = \prod_{k=1}^{K} \left[\theta_{ki}\right]^{z_{ink}}$
- $\mathbf{w}_{in} \,|\, \{z_{ink}=1\} \sim \mathcal{M}(1, (b_{1k}, \ldots, b_{dk}))$
- $p(w_{inj} = 1 \mid z_{ink} = 1) = b_{jk}$

LDA likelihood

$$p\Big(\big(\mathbf{w}_{in}, \mathbf{z}_{in}\big)_{1 \leq m \leq N_i} \mid \boldsymbol{\theta}_i\Big) \quad =$$

LDA likelihood

$$p\Big(\big(\mathbf{w}_{in}, \mathbf{z}_{in}\big)_{1\leq m\leq N_i} \mid \boldsymbol{\theta}_i\Big) \quad = \quad \prod_{n=1}^{N_i} p\big(\mathbf{w}_{in}, \mathbf{z}_{in} \mid \boldsymbol{\theta}_i\big)$$

LDA likelihood

$$
\begin{aligned}
p\Big(\big(\mathbf{w}_{in}, \mathbf{z}_{in}\big)_{1 \leq m \leq N_i} \mid \boldsymbol{\theta}_i\Big) &= \prod_{n=1}^{N_i} p\big(\mathbf{w}_{in}, \mathbf{z}_{in} \mid \boldsymbol{\theta}_i\big) \\
&= \prod_{n=1}^{N_i} \prod_{j=1}^{d} \prod_{k=1}^{K} \big(b_{jk}\, \theta_{ki}\big)^{\,w_{inj}\, z_{ink}}
\end{aligned}
$$

LDA likelihood

$$
\begin{aligned}
p\Big(\big(\mathbf{w}_{in}, \mathbf{z}_{in}\big)_{1 \leq m \leq N_i} \mid \boldsymbol{\theta}_i\Big) &= \prod_{n=1}^{N_i} p\big(\mathbf{w}_{in}, \mathbf{z}_{in} \mid \boldsymbol{\theta}_i\big) \\
&= \prod_{n=1}^{N_i} \prod_{j=1}^{d} \prod_{k=1}^{K} (b_{jk}\, \theta_{ki})^{\,w_{inj}\, z_{ink}}
\end{aligned}
$$

$$
p\Big(\big(\mathbf{w}_{in}\big)_{1 \leq m \leq N_i} \mid \boldsymbol{\theta}_i\Big)
$$

LDA likelihood

$$
\begin{aligned}
p\Big(\big(\mathbf{w}_{in}, \mathbf{z}_{in}\big)_{1 \leq m \leq N_i} \mid \boldsymbol{\theta}_i\Big) &= \prod_{n=1}^{N_i} p\big(\mathbf{w}_{in}, \mathbf{z}_{in} \mid \boldsymbol{\theta}_i\big) \\
&= \prod_{n=1}^{N_i} \prod_{j=1}^{d} \prod_{k=1}^{K} \left(b_{jk}\, \theta_{ki}\right)^{w_{inj}\, z_{ink}} \\
p\Big(\big(\mathbf{w}_{in}\big)_{1 \leq m \leq N_i} \mid \boldsymbol{\theta}_i\Big) &= \prod_{n=1}^{N_i} \sum_{\mathbf{z}_{in}} p\big(\mathbf{w}_{in}, \mathbf{z}_{in} \mid \boldsymbol{\theta}_i\big)
\end{aligned}
$$

LDA likelihood

$$
\begin{aligned}
p\Big(\big(\mathbf{w}_{in}, \mathbf{z}_{in}\big)_{1 \le m \le N_i} \mid \boldsymbol{\theta}_i\Big) &= \prod_{n=1}^{N_i} p\big(\mathbf{w}_{in}, \mathbf{z}_{in} \mid \boldsymbol{\theta}_i\big) \\
&= \prod_{n=1}^{N_i} \prod_{j=1}^{d} \prod_{k=1}^{K} \big(b_{jk}\, \theta_{ki}\big)^{\, w_{inj}\, z_{ink}} \\
p\Big(\big(\mathbf{w}_{in}\big)_{1 \le m \le N_i} \mid \boldsymbol{\theta}_i\Big) &= \prod_{n=1}^{N_i} \sum_{\mathbf{z}_{in}} p\big(\mathbf{w}_{in}, \mathbf{z}_{in} \mid \boldsymbol{\theta}_i\big) \\
&= \prod_{n=1}^{N_i} \prod_{j=1}^{d} \left[\sum_{k=1}^{K} b_{jk} \theta_{ki}\right]^{w_{inj}},
\end{aligned}
$$

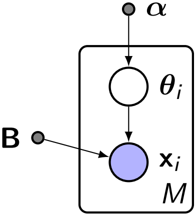
LDA likelihood

$$
\begin{aligned}
p\Big(\big(\mathbf{w}_{in}, \mathbf{z}_{in}\big)_{1 \leq m \leq N_i} \mid \boldsymbol{\theta}_i\Big) &= \prod_{n=1}^{N_i} p\big(\mathbf{w}_{in}, \mathbf{z}_{in} \mid \boldsymbol{\theta}_i\big) \\
&= \prod_{n=1}^{N_i} \prod_{j=1}^{d} \prod_{k=1}^{K} (b_{jk}\, \theta_{ki})^{\, w_{inj}\, z_{ink}} \\
p\Big(\big(\mathbf{w}_{in}\big)_{1 \leq m \leq N_i} \mid \boldsymbol{\theta}_i\Big) &= \prod_{n=1}^{N_i} \sum_{\mathbf{z}_{in}} p\big(\mathbf{w}_{in}, \mathbf{z}_{in} \mid \boldsymbol{\theta}_i\big) \\
&= \prod_{n=1}^{N_i} \prod_{j=1}^{d} \left[\sum_{k=1}^{K} b_{jk} \theta_{ki} \right]^{w_{inj}},
\end{aligned}
$$

so that $\quad \mathbf{w}_{in} \mid \boldsymbol{\theta}_i \overset{\text{i.i.d.}}{\sim} \mathcal{M}(1, \mathbf{B}\boldsymbol{\theta}_i) \quad$ or $\quad \mathbf{x}_i \mid \boldsymbol{\theta}_i \overset{\text{i.i.d.}}{\sim} \mathcal{M}(N_i, \mathbf{B}\boldsymbol{\theta}_i)$.

LDA as Multinomial Factorial Analysis

Eliminating **z**s from the model yields a conceptually simpler model in which $\boldsymbol{\theta}_i$ can be interpreted as latent factors as in *factorial analysis*.



- Topic proportions for document i: $\boldsymbol{\theta}_i \in \mathbb{R}^K$

$$\boldsymbol{\theta}_i \sim \mathrm{Dir}(\boldsymbol{\alpha})$$

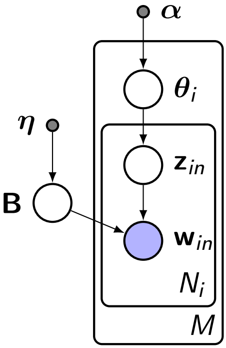
- Empirical words counts for document i: $\mathbf{x}_i \in \mathbb{R}^d$

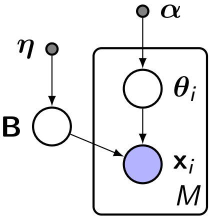
$$\mathbf{x}_i \sim \mathcal{M}(N_i, \mathbf{B}\boldsymbol{\theta}_i)$$

LDA with smoothing of the dictionary

Issue with *new words*: they will have probability 0 if **B** is optimized over the training data.

$\rightarrow$ Need to smooth **B** e.g. via Laplacian smoothing.





Learning with LDA

How do we learn with LDA?

- How do we learn for each **topic** its **word distribution** $\mathbf{b}_k$?
- How do we learn for each **document** its topic **composition** θ_i?
- How do we assign to each **word** of a document its **topic** z_{in} ?

How do we learn with LDA?

- How do we learn for each **topic** its **word distribution** $\mathbf{b}_k$?
- How do we learn for each **document** its topic **composition** $\boldsymbol{\theta}_i$?
- How do we assign to each **word** of a document its **topic** z_{in} ?

These quantities are treated in a Bayesian fashion, so the natural Bayesian answer are

$$p(\mathbf{B}|\mathbf{W}) \qquad p(\boldsymbol{\theta}_i|\mathbf{W}) \qquad p(\mathbf{z}_{in}|\mathbf{W})$$

How do we learn with LDA?

- How do we learn for each **topic** its **word distribution** $\mathbf{b}_k$?
- How do we learn for each **document** its topic **composition** θ_i?
- How do we assign to each **word** of a document its **topic** z_{in} ?

These quantities are treated in a Bayesian fashion, so the natural Bayesian answer are

$$p(\mathbf{B}|\mathbf{W}) \qquad p(\boldsymbol{\theta}_i|\mathbf{W}) \qquad p(\mathbf{z}_{in}|\mathbf{W})$$

or

$$\mathbb{E}(\mathbf{B}|\mathbf{W}) \qquad \mathbb{E}(\boldsymbol{\theta}_i|\mathbf{W}) \qquad \mathbb{E}(\mathbf{z}_{in}|\mathbf{W})$$

if *point-estimates* are needed.

Monte Carlo

Principle of Monte Carlo integration

Let Z be a random variable, to compute $\mathbb{E}[f(Z)]$ we can sample

$$Z^{(1)}, \dots, Z^{(B)} \overset{\text{i.i.d.}}{\sim} Z$$

and do the approximation

$$\mathbb{E}[f(X)] \approx \frac{1}{B} \sum_{b=1}^{B} f(Z^{(b)})$$

Monte Carlo

Principle of Monte Carlo integration

Let Z be a random variable, to compute $\mathbb{E}[f(Z)]$ we can sample

$$Z^{(1)}, \dots, Z^{(B)} \overset{\text{i.i.d.}}{\sim} Z$$

and do the approximation

$$\mathbb{E}[f(X)] \approx \frac{1}{B} \sum_{b=1}^{B} f(Z^{(b)})$$

Problem: In most situations sampling exactly from the distribution of Z is too hard, so this direct approach is impossible.

Markov Chain Monte Carlo (MCMC)

If we can cannot sample exact from the distribution of Z, i.e. from some $q(z) = \mathbb{P}(Z = z)$ or $q(z)$ is the density of r.v. Z, then we can create a sequence of random variables that approach the correct distribution.

Principle of MCMC

Construct a chain of random variables

$$Z^{(b,1)}, \dots, Z^{(b,T)} \qquad \text{with} \qquad Z^{(b,t)} \sim p_t(z^{(b,t)} \mid Z^{(b,t-1)} = z^{(b,t-1)})$$

such that

$$Z^{(b,T)} \xrightarrow[T\to\infty]{\mathcal{D}} Z$$

We can then approximate:

$$\mathbb{E}[f(Z)] \approx \frac{1}{B} \sum_{b=1}^{B} f\left(Z^{(b,T)}\right)$$

MCMC in practice

Run a single chain:

$$\mathbb{E}[f(Z)] \approx \frac{1}{T}\sum_{t=1}^{T} f\left(Z^{(T_0+k\cdot t)}\right)$$

MCMC in practice

Run a single chain:

$$\mathbb{E}[f(Z)] \approx \frac{1}{T} \sum_{t=1}^{T} f\left(Z^{(T_0+k\cdot t)}\right)$$

- T_0 is the burn-in time

MCMC in practice

Run a single chain:

$$\mathbb{E}[f(Z)] \approx \frac{1}{T}\sum_{t=1}^{T} f\left(Z^{(T_0+k\cdot t)}\right)$$

- T_0 is the burn-in time
- k is the thinning factor

MCMC in practice

Run a single chain:

$$\mathbb{E}[f(Z)] \approx \frac{1}{T}\sum_{t=1}^{T} f\left(Z^{(T_0+k\cdot t)}\right)$$

- T_0 is the burn-in time
- k is the thinning factor
  - $\rightarrow$ Useful to take $k > 1$ only if almost i.i.d. samples are required.
  - $\rightarrow$ To compute an expectation in which the correlation between $Z^{(t)}$ and $Z^{(t-1)}$ would not interfere take $k = 1$

MCMC in practice

Run a single chain:

$$\mathbb{E}[f(Z)] \approx \frac{1}{T}\sum_{t=1}^{T} f\left(Z^{(T_0+k\cdot t)}\right)$$

- T_0 is the burn-in time
- k is the thinning factor
  - $\rightarrow$ Useful to take $k > 1$ only if almost i.i.d. samples are required.
  - $\rightarrow$ To compute an expectation in which the correlation between $Z^{(t)}$ and $Z^{(t-1)}$ would not interfere take $k = 1$

Main difficulties:

- the **mixing time** of the chain can be very large

MCMC in practice

Run a single chain:

$$\mathbb{E}[f(Z)] \approx \frac{1}{T} \sum_{t=1}^{T} f\left(Z^{(T_0+k \cdot t)}\right)$$

- T_0 is the burn-in time
- k is the thinning factor
  - $\rightarrow$ Useful to take $k > 1$ only if almost i.i.d. samples are required.
  - $\rightarrow$ To compute an expectation in which the correlation between $Z^{(t)}$ and $Z^{(t-1)}$ would not interfere take $k = 1$

Main difficulties:

- the **mixing time** of the chain can be very large
- Assessing whether the chain has mixed or not is a hard problem

MCMC in practice

Run a single chain:

$$\mathbb{E}[f(Z)] \approx \frac{1}{T}\sum_{t=1}^{T} f\left(Z^{(T_0+k\cdot t)}\right)$$

- T_0 is the burn-in time
- k is the thinning factor
  - $\rightarrow$ Useful to take $k > 1$ only if almost i.i.d. samples are required.
  - $\rightarrow$ To compute an expectation in which the correlation between $Z^{(t)}$ and $Z^{(t-1)}$ would not interfere take $k = 1$

Main difficulties:

- the **mixing time** of the chain can be very large
- Assessing whether the chain has mixed or not is a hard problem

$\Rightarrow$ proper approximation only with T very large.

MCMC in practice

Run a single chain:

$$\mathbb{E}[f(Z)] \approx \frac{1}{T}\sum_{t=1}^{T} f\left(Z^{(T_0+k\cdot t)}\right)$$

- T_0 is the burn-in time
- k is the thinning factor
  - $\rightarrow$ Useful to take $k > 1$ only if almost i.i.d. samples are required.
  - $\rightarrow$ To compute an expectation in which the correlation between $Z^{(t)}$ and $Z^{(t-1)}$ would not interfere take $k = 1$

Main difficulties:

- the **mixing time** of the chain can be very large
- Assessing whether the chain has mixed or not is a hard problem

$\Rightarrow$ proper approximation only with T very large.

$\Rightarrow$ MCMC can be quite slow or just never converge and you will not necessarily know it.

Gibbs sampling

A nice special case of MCMC:

Gibbs sampling

A nice special case of MCMC:

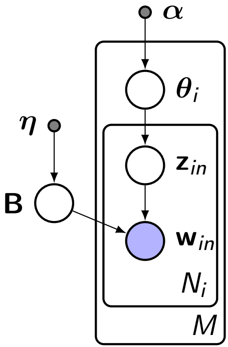
Principle of Gibbs sampling

For each node i in turn, sample the node conditionally on the other nodes, i.e.

$$\text{Sample} \qquad Z_i^{(t)} \sim p\Big(z_i \mid Z_{-i} = z_{-i}^{(t-1)}\Big)$$

Gibbs sampling

A nice special case of MCMC:



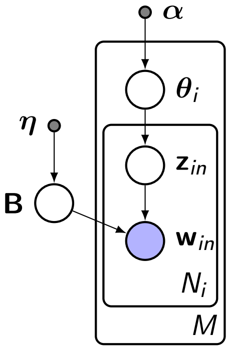
Principle of Gibbs sampling

For each node i in turn, sample the node conditionally on the other nodes, i.e.

$$\text{Sample} \qquad Z_i^{(t)} \sim p\left(z_i \mid Z_{-i} = z_{-i}^{(t-1)}\right)$$

Gibbs sampling

A nice special case of MCMC:



Principle of Gibbs sampling

For each node i in turn, sample the node conditionally on the other nodes, i.e.

$$\text{Sample} \qquad Z_i^{(t)} \sim p\Big(z_i \mid Z_{-i} = z_{-i}^{(t-1)}\Big)$$

Markov Blanket

Definition: Let V be the set of nodes of the graph. The Markov blanket of node i is the minimal set of nodes S (not containing i) such that

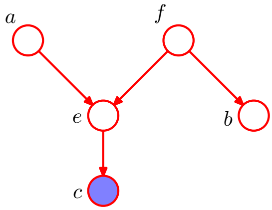
$$p(Z_i \mid Z_S) = p(Z_i \mid Z_{-i}) \quad \text{or equivalently} \quad Z_i \perp\!\!\!\perp Z_{V\backslash(S\cup\{i\})} \mid Z_S$$

d-separation

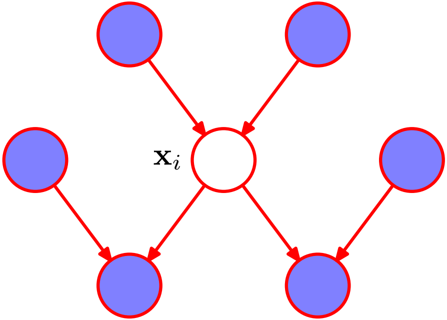
Theorem

Let A, B and C three disjoint sets of nodes. The property $X_A \perp\!\!\!\perp X_B | X_C$ holds if and only if all paths connecting A to B are *blocked*, which means that they contain at least one blocking node. Node j is a *blocking node*

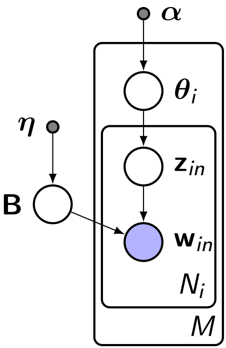
- if there is no "v-structure" in j and j is in C or
- if there is a "v-structure" in j and if neither j nor any of its descendants in the graph is in C.



Markov Blanket in a Directed Graphical model

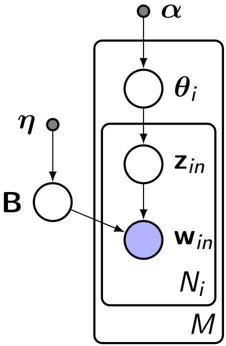


Markov Blankets in LDA?



Markov blankets for

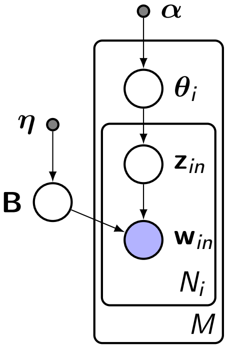
Markov Blankets in LDA?



Markov blankets for

- $\theta_i \rightarrow$

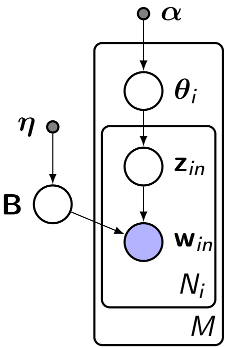
Markov Blankets in LDA?



Markov blankets for

- $\theta_i \to (\mathbf{z}_{in})_{n=1\dots N_i}$

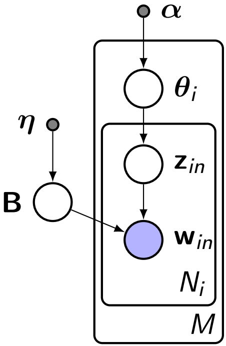
Markov Blankets in LDA?



Markov blankets for

- $\boldsymbol{\theta}_i \rightarrow (\mathbf{z}_{in})_{n=1\ldots N_i}$
- $\mathbf{z}_{in} \rightarrow$

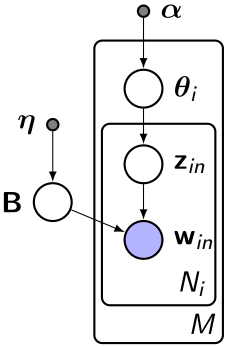
Markov Blankets in LDA?



Markov blankets for

- $\boldsymbol{\theta}_i \rightarrow (\mathbf{z}_{in})_{n=1\ldots N_i}$
- $\mathbf{z}_{in} \rightarrow \mathbf{w}_{in}$, $\boldsymbol{\theta}_i$ and $\mathbf{B}$

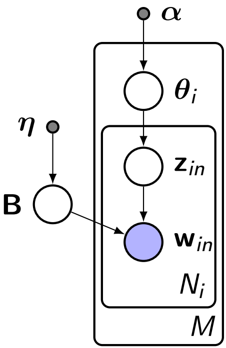
Markov Blankets in LDA?



Markov blankets for

- $\boldsymbol{\theta}_i \;\rightarrow (\mathbf{z}_{in})_{n=1\ldots N_i}$
- $\mathbf{z}_{in} \;\rightarrow \mathbf{w}_{in},\ \boldsymbol{\theta}_i$ and $\mathbf{B}$
- $\mathbf{B} \;\rightarrow$

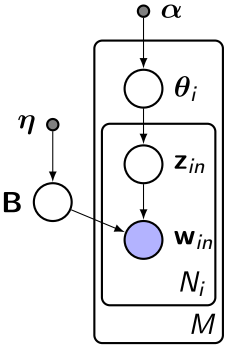
Markov Blankets in LDA?



Markov blankets for

- $\boldsymbol{\theta}_i \;\rightarrow (\mathbf{z}_{in})_{n=1\ldots N_i}$
- $\mathbf{z}_{in} \;\rightarrow \mathbf{w}_{in},\; \boldsymbol{\theta}_i$ and $\mathbf{B}$
- $\mathbf{B} \;\rightarrow (\mathbf{w}_{in}, \mathbf{z}_{in})_{n=1\ldots N_i, i=1\ldots,M}$

Markov Blankets in LDA?



Markov blankets for

- $\boldsymbol{\theta}_i \rightarrow (\mathbf{z}_{in})_{n=1\ldots N_i}$
- $\mathbf{z}_{in} \rightarrow \mathbf{w}_{in}$, $\boldsymbol{\theta}_i$ and $\mathbf{B}$
- $\mathbf{B} \rightarrow (\mathbf{w}_{in}, \mathbf{z}_{in})_{n=1\ldots N_i, i=1\ldots,M}$

Gibbs sampling for LDA with a single document

$$p(\mathbf{w}, \mathbf{z}, \boldsymbol{\theta}, \mathbf{B}; \boldsymbol{\alpha}, \boldsymbol{\eta}) =$$

Gibbs sampling for LDA with a single document

$$p(\mathbf{w}, \mathbf{z}, \boldsymbol{\theta}, \mathbf{B}; \boldsymbol{\alpha}, \boldsymbol{\eta}) \quad = \quad \left[\prod_{n=1}^{N} p(\mathbf{w}_n | \mathbf{z}_n, \mathbf{B}) \, p(\mathbf{z}_n | \boldsymbol{\theta}) \right] p(\boldsymbol{\theta} | \boldsymbol{\alpha}) \prod_k p(\mathbf{b}_k | \boldsymbol{\eta})$$

Gibbs sampling for LDA with a single document

$$\begin{aligned} p(\mathbf{w}, \mathbf{z}, \boldsymbol{\theta}, \mathbf{B}; \boldsymbol{\alpha}, \boldsymbol{\eta}) &= \left[\prod_{n=1}^{N} p(\mathbf{w}_n | \mathbf{z}_n, \mathbf{B})\, p(\mathbf{z}_n | \boldsymbol{\theta}) \right] p(\boldsymbol{\theta} | \boldsymbol{\alpha}) \prod_k p(\mathbf{b}_k | \boldsymbol{\eta}) \\ &\propto \left[\prod_{n=1}^{N} \prod_{j,k} (b_{jk} \theta_k)^{w_{nj}\, z_{nk}} \right] \prod_k \theta_k^{\alpha_k - 1} \prod_{j,k} b_{jk}^{\eta_j - 1} \end{aligned}$$

Gibbs sampling for LDA with a single document

$$
\begin{aligned}
p(\mathbf{w}, \mathbf{z}, \boldsymbol{\theta}, \mathbf{B}; \boldsymbol{\alpha}, \boldsymbol{\eta}) &= \left[\prod_{n=1}^{N} p(\mathbf{w}_n | \mathbf{z}_n, \mathbf{B})\, p(\mathbf{z}_n | \boldsymbol{\theta})\right] p(\boldsymbol{\theta} | \boldsymbol{\alpha}) \prod_k p(\mathbf{b}_k | \boldsymbol{\eta}) \\
&\propto \left[\prod_{n=1}^{N} \prod_{j,k} (b_{jk}\theta_k)^{w_{nj}\, z_{nk}}\right] \quad \prod_k \theta_k^{\alpha_k - 1} \quad \prod_{j,k} b_{jk}^{\eta_j - 1}
\end{aligned}
$$

We thus have:

- $(\mathbf{z}_n \mid \mathbf{w}_n, \boldsymbol{\theta}) \sim$

Gibbs sampling for LDA with a single document

$$
\begin{aligned}
p(\mathbf{w}, \mathbf{z}, \boldsymbol{\theta}, \mathbf{B}; \boldsymbol{\alpha}, \boldsymbol{\eta}) &= \left[\prod_{n=1}^{N} p(\mathbf{w}_n | \mathbf{z}_n, \mathbf{B})\, p(\mathbf{z}_n | \boldsymbol{\theta})\right] p(\boldsymbol{\theta} | \boldsymbol{\alpha}) \prod_k p(\mathbf{b}_k | \boldsymbol{\eta}) \\
&\propto \left[\prod_{n=1}^{N} \prod_{j,k} (b_{jk}\theta_k)^{w_{nj}\, z_{nk}}\right] \quad \prod_k \theta_k^{\alpha_k - 1} \quad \prod_{j,k} b_{jk}^{\eta_j - 1}
\end{aligned}
$$

We thus have:

- $(\mathbf{z}_n \mid \mathbf{w}_n, \boldsymbol{\theta}) \sim \mathcal{M}(1, \tilde{\mathbf{p}}_m)$ with $\tilde{p}_{nk} = \dfrac{b_{j(n),k}\, \theta_k}{\sum_{k'} b_{j(n),k'}\, \theta_{k'}}$.

Gibbs sampling for LDA with a single document

$$
\begin{aligned}
p(\mathbf{w}, \mathbf{z}, \boldsymbol{\theta}, \mathbf{B}; \boldsymbol{\alpha}, \boldsymbol{\eta}) &= \left[\prod_{n=1}^{N} p(\mathbf{w}_n | \mathbf{z}_n, \mathbf{B})\, p(\mathbf{z}_n | \boldsymbol{\theta}) \right] p(\boldsymbol{\theta} | \boldsymbol{\alpha}) \prod_k p(\mathbf{b}_k | \boldsymbol{\eta}) \\
&\propto \left[\prod_{n=1}^{N} \prod_{j,k} (b_{jk} \theta_k)^{w_{nj}\, z_{nk}} \right] \quad \prod_k \theta_k^{\alpha_k - 1} \quad \prod_{j,k} b_{jk}^{\eta_j - 1}
\end{aligned}
$$

We thus have:

- $(\mathbf{z}_n \mid \mathbf{w}_n, \boldsymbol{\theta}) \sim \mathcal{M}(1, \tilde{\mathbf{p}}_m)$ with $\tilde{p}_{nk} = \dfrac{b_{j(n),k}\, \theta_k}{\sum_{k'} b_{j(n),k'}\, \theta_{k'}}$.
- $(\boldsymbol{\theta} \mid (\mathbf{z}_n, \mathbf{w}_n)_n, \boldsymbol{\alpha}) \sim$

Gibbs sampling for LDA with a single document

$$
\begin{aligned}
p(\mathbf{w}, \mathbf{z}, \boldsymbol{\theta}, \mathbf{B}; \boldsymbol{\alpha}, \boldsymbol{\eta}) &= \left[\prod_{n=1}^{N} p(\mathbf{w}_n | \mathbf{z}_n, \mathbf{B})\, p(\mathbf{z}_n | \boldsymbol{\theta})\right] p(\boldsymbol{\theta} | \boldsymbol{\alpha}) \prod_{k} p(\mathbf{b}_k | \boldsymbol{\eta}) \\
&\propto \left[\prod_{n=1}^{N} \prod_{j,k} (b_{jk}\theta_k)^{w_{nj}\, z_{nk}}\right] \quad \prod_{k} \theta_k^{\alpha_k - 1} \quad \prod_{j,k} b_{jk}^{\eta_j - 1}
\end{aligned}
$$

We thus have:

- $(\mathbf{z}_n \mid \mathbf{w}_n, \boldsymbol{\theta}) \sim \mathcal{M}(1, \tilde{\mathbf{p}}_m)$ with $\tilde{p}_{nk} = \frac{b_{j(n),k}\, \theta_k}{\sum_{k'} b_{j(n),k'}\, \theta_{k'}}$.
- $(\boldsymbol{\theta} \mid (\mathbf{z}_n, \mathbf{w}_n)_n, \boldsymbol{\alpha}) \sim \mathrm{Dir}(\tilde{\boldsymbol{\alpha}})$ with $\tilde{\alpha}_k = \alpha_k + N_k,\quad N_k = \sum_{n=1}^{N} z_{nk}$.

Gibbs sampling for LDA with a single document

$$p(\mathbf{w}, \mathbf{z}, \boldsymbol{\theta}, \mathbf{B}; \boldsymbol{\alpha}, \boldsymbol{\eta}) = \Big[\prod_{n=1}^{N} p(\mathbf{w}_n|\mathbf{z}_n, \mathbf{B})\, p(\mathbf{z}_n|\boldsymbol{\theta})\Big]\, p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \prod_k p(\mathbf{b}_k|\boldsymbol{\eta})$$

$$\propto \Big[\prod_{n=1}^{N} \prod_{j,k} (b_{jk}\theta_k)^{w_{nj}\, z_{nk}}\Big] \quad \prod_k \theta_k^{\alpha_k - 1} \quad \prod_{j,k} b_{jk}^{\eta_j - 1}$$

We thus have:

- $(\mathbf{z}_n \mid \mathbf{w}_n, \boldsymbol{\theta}) \sim \mathcal{M}(1, \tilde{\mathbf{p}}_m)$ with $\tilde{p}_{nk} = \dfrac{b_{j(n),k}\,\theta_k}{\sum_{k'} b_{j(n),k'}\,\theta_{k'}}$.
- $(\boldsymbol{\theta} \mid (\mathbf{z}_n, \mathbf{w}_n)_n, \boldsymbol{\alpha}) \sim \mathrm{Dir}(\tilde{\boldsymbol{\alpha}})$ with $\tilde{\alpha}_k = \alpha_k + N_k$, $N_k = \sum_{n=1}^{N} z_{nk}$.
- $(\mathbf{b}_k \mid (\mathbf{z}_n, \mathbf{w}_n)_n, \boldsymbol{\eta}) \sim$

Gibbs sampling for LDA with a single document

$$
\begin{aligned}
p(\mathbf{w},\mathbf{z},\boldsymbol{\theta},\mathbf{B};\boldsymbol{\alpha},\boldsymbol{\eta}) &= \left[\prod_{n=1}^{N} p(\mathbf{w}_n|\mathbf{z}_n,\mathbf{B})\, p(\mathbf{z}_n|\boldsymbol{\theta})\right] p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \prod_k p(\mathbf{b}_k|\boldsymbol{\eta}) \\
&\propto \left[\prod_{n=1}^{N}\prod_{j,k}(b_{jk}\theta_k)^{w_{nj}\, z_{nk}}\right] \quad \prod_k \theta_k^{\alpha_k-1} \quad \prod_{j,k} b_{jk}^{\eta_j-1}
\end{aligned}
$$

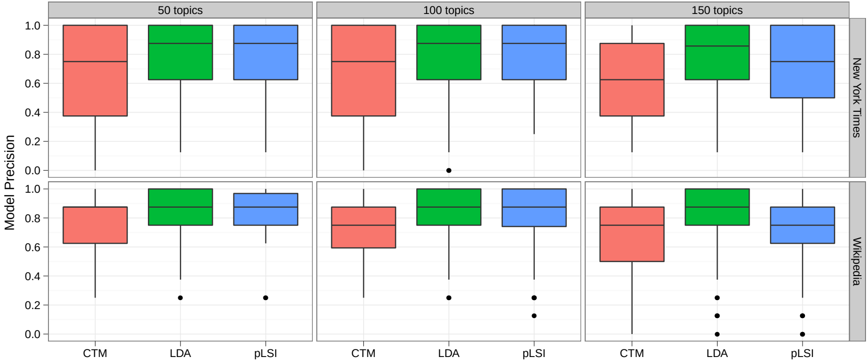
We thus have:

- $(\mathbf{z}_n \mid \mathbf{w}_n, \boldsymbol{\theta}) \sim \mathcal{M}(1, \tilde{\mathbf{p}}_m)$ with $\tilde{p}_{nk} = \dfrac{b_{j(n),k}\,\theta_k}{\sum_{k'} b_{j(n),k'}\,\theta_{k'}}$.
- $(\boldsymbol{\theta} \mid (\mathbf{z}_n, \mathbf{w}_n)_n, \boldsymbol{\alpha}) \sim \mathrm{Dir}(\tilde{\boldsymbol{\alpha}})$ with $\tilde{\alpha}_k = \alpha_k + N_k, \quad N_k = \displaystyle\sum_{n=1}^{N} z_{nk}$.
- $(\mathbf{b}_k \mid (\mathbf{z}_n, \mathbf{w}_n)_n, \boldsymbol{\eta}) \sim \mathrm{Dir}(\tilde{\boldsymbol{\eta}})$ with $\tilde{\eta}_j = \eta_j + \sum_{n=1}^{N} w_{nj} z_{nk}$.

LDA Results (Blei et al., 2003)

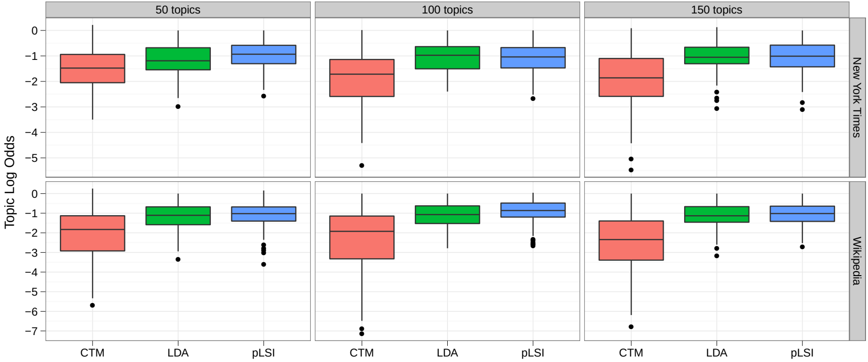
"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

Reading Tea leaves: word precision (Boyd-Graber et al., 2009)



Precision of the identification of **word** outliers,
by humans and for different models.

Reading Tea leaves: topic precision (Boyd-Graber et al., 2009)



Precision of the identification of **topic** outliers,
by humans and for different models.

Reading Tea leaves: log-likelihood on held out data

(Boyd-Graber et al., 2009)

Corpus	Topics	LDA	CTM	pLSI
New York Times	50	**-7.3214 / 784.38**	-7.3335 / 788.58	-7.3384 / 796.43
	100	-7.2761 / 778.24	**-7.2647 / 762.16**	-7.2834 / 785.05
	150	-7.2477 / 777.32	-7.2467 / **755.55**	**-7.2382** / 770.36
Wikipedia	50	**-7.5257** / 961.86	-7.5332 / **936.58**	-7.5378 / 975.88
	100	-7.4629 / 935.53	**-7.4385 / 880.30**	-7.4748 / 951.78
	150	-7.4266 / 929.76	**-7.3872 / 852.46**	-7.4355 / 945.29

Log-likelihoods of several models including LDA, pLSI and CTM (CTM=correlated topic model)

Variational inference for LDA

Principle of Variational Inference

Problem: it is hard to compute:

$$p(\mathbf{B}, \boldsymbol{\theta}_i, z_{in}|\mathbf{W}), \quad \mathbb{E}(\mathbf{B}|\mathbf{W}), \quad \mathbb{E}(\boldsymbol{\theta}_i|\mathbf{W}), \quad \mathbb{E}(\mathbf{z}_{in}|\mathbf{W}).$$

Idea of Variational Inference:

Find a distribution q which is

- as close as possible to $p(\cdot|\mathbf{W})$
- for which it is not too hard to compute $\mathbb{E}_q(\mathbf{B})$, $\mathbb{E}_q(\boldsymbol{\theta}_i)$, $\mathbb{E}_q(\mathbf{z}_{in})$.

Principle of Variational Inference

Problem: it is hard to compute:

$$p(\mathbf{B}, \boldsymbol{\theta}_i, z_{in} | \mathbf{W}), \quad \mathbb{E}(\mathbf{B}|\mathbf{W}), \quad \mathbb{E}(\boldsymbol{\theta}_i|\mathbf{W}), \quad \mathbb{E}(\mathbf{z}_{in}|\mathbf{W}).$$

Idea of Variational Inference:

Find a distribution q which is

- as close as possible to $p(\cdot|\mathbf{W})$
- for which it is not too hard to compute $\mathbb{E}_q(\mathbf{B})$, $\mathbb{E}_q(\boldsymbol{\theta}_i)$, $\mathbb{E}_q(\mathbf{z}_{in})$.

Usual approach:

1. Choose a simple parametric family $\mathcal{Q}$ for q.

Principle of Variational Inference

Problem: it is hard to compute:

$$p(\mathbf{B}, \boldsymbol{\theta}_i, z_{in}|\mathbf{W}), \quad \mathbb{E}(\mathbf{B}|\mathbf{W}), \quad \mathbb{E}(\boldsymbol{\theta}_i|\mathbf{W}), \quad \mathbb{E}(\mathbf{z}_{in}|\mathbf{W}).$$

Idea of Variational Inference:

Find a distribution q which is

- as close as possible to $p(\cdot|\mathbf{W})$
- for which it is not too hard to compute $\mathbb{E}_q(\mathbf{B})$, $\mathbb{E}_q(\boldsymbol{\theta}_i)$, $\mathbb{E}_q(\mathbf{z}_{in})$.

Usual approach:

1. Choose a simple parametric family $\mathcal{Q}$ for q.
2. Solve the *variational formulation* $\quad \min_{q \in \mathcal{Q}} KL\big(q \,\|\, p(\cdot|\mathbf{W})\big)$

Principle of Variational Inference

Problem: it is hard to compute:

$$p(\mathbf{B}, \boldsymbol{\theta}_i, z_{in}|\mathbf{W}), \quad \mathbb{E}(\mathbf{B}|\mathbf{W}), \quad \mathbb{E}(\boldsymbol{\theta}_i|\mathbf{W}), \quad \mathbb{E}(\mathbf{z}_{in}|\mathbf{W}).$$

Idea of Variational Inference:

Find a distribution q which is

- as close as possible to $p(\cdot|\mathbf{W})$
- for which it is not too hard to compute $\mathbb{E}_q(\mathbf{B})$, $\mathbb{E}_q(\boldsymbol{\theta}_i)$, $\mathbb{E}_q(\mathbf{z}_{in})$.

Usual approach:

1. Choose a simple parametric family $\mathcal{Q}$ for q.
2. Solve the *variational formulation* $\quad \min_{q \in \mathcal{Q}} KL\big(q \,\|\, p(\cdot|\mathbf{W})\big)$
3. Compute the desired expectations: $\mathbb{E}_q(\mathbf{B})$, $\mathbb{E}_q(\boldsymbol{\theta}_i)$, $\mathbb{E}_q(\mathbf{z}_{in})$.

Principle of Variational Inference

Problem: it is hard to compute:

$$p(\mathbf{B}, \boldsymbol{\theta}_i, z_{in}|\mathbf{W}), \quad \mathbb{E}(\mathbf{B}|\mathbf{W}), \quad \mathbb{E}(\boldsymbol{\theta}_i|\mathbf{W}), \quad \mathbb{E}(\mathbf{z}_{in}|\mathbf{W}).$$

Idea of Variational Inference:

Find a distribution q which is

- as close as possible to $p(\cdot|\mathbf{W})$
- for which it is not too hard to compute $\mathbb{E}_q(\mathbf{B})$, $\mathbb{E}_q(\boldsymbol{\theta}_i)$, $\mathbb{E}_q(\mathbf{z}_{in})$.

Usual approach:

1. Choose a simple parametric family $\mathcal{Q}$ for q.
2. Solve the *variational formulation* $\min_{q \in \mathcal{Q}} KL\big(q \,\|\, p(\cdot|\mathbf{W})\big)$
3. Compute the desired expectations: $\mathbb{E}_q(\mathbf{B})$, $\mathbb{E}_q(\boldsymbol{\theta}_i)$, $\mathbb{E}_q(\mathbf{z}_{in})$.

Principle of Variational Inference

Problem: it is hard to compute:

$$p(\mathbf{B}, \boldsymbol{\theta}_i, z_{in} | \mathbf{W}), \quad \mathbb{E}(\mathbf{B}|\mathbf{W}), \quad \mathbb{E}(\boldsymbol{\theta}_i|\mathbf{W}), \quad \mathbb{E}(\mathbf{z}_{in}|\mathbf{W}).$$

Idea of Variational Inference:

Find a distribution q which is

- as close as possible to $p(\cdot|\mathbf{W})$
- for which it is not too hard to compute $\mathbb{E}_q(\mathbf{B})$, $\mathbb{E}_q(\boldsymbol{\theta}_i)$, $\mathbb{E}_q(\mathbf{z}_{in})$.

Usual approach:

1. Choose a simple parametric family $\mathcal{Q}$ for q.
2. Solve the *variational formulation* $\quad \min_{q \in \mathcal{Q}} KL\big(q \,\|\, p(\cdot|\mathbf{W})\big)$
3. Compute the desired expectations: $\mathbb{E}_q(\mathbf{B})$, $\mathbb{E}_q(\boldsymbol{\theta}_i)$, $\mathbb{E}_q(\mathbf{z}_{in})$.

Variational Inference for LDA (Blei et al., 2003)

Assume $\mathbf{B}$ is a parameter, assume there is a single document, and focus on the inference on $\boldsymbol{\theta}$ and $(\mathbf{z}_n)_n$. Choose q in a factorized form (*mean field* approximation)

$$q(\boldsymbol{\theta}, (\mathbf{z}_n)_n) = q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \prod_{n=1}^{N} q_{\mathbf{z}_n}(\mathbf{z}_n)$$

Variational Inference for LDA (Blei et al., 2003)

Assume $\mathbf{B}$ is a parameter, assume there is a single document, and focus on the inference on $\boldsymbol{\theta}$ and $(\mathbf{z}_n)_n$. Choose q in a factorized form (*mean field* approximation)

$$q(\boldsymbol{\theta}, (\mathbf{z}_n)_n) = q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \prod_{n=1}^{N} q_{\mathbf{z}_n}(\mathbf{z}_n) \qquad \text{with}$$

$$q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \frac{\Gamma(\sum_k \gamma_k)}{\prod_k \Gamma(\gamma_k)} \prod_k \theta_k^{\gamma_k - 1}$$

Variational Inference for LDA (Blei et al., 2003)

Assume $\mathbf{B}$ is a parameter, assume there is a single document, and focus on the inference on $\boldsymbol{\theta}$ and $(\mathbf{z}_n)_n$. Choose q in a factorized form (*mean field* approximation)

$$q(\boldsymbol{\theta}, (\mathbf{z}_n)_n) = q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \prod_{n=1}^{N} q_{\mathbf{z}_n}(\mathbf{z}_n) \qquad \text{with}$$

$$q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \frac{\Gamma(\sum_k \gamma_k)}{\prod_k \Gamma(\gamma_k)} \prod_k \theta_k^{\gamma_k - 1} \qquad \text{and} \qquad q_{\mathbf{z}_n}(\mathbf{z}_n) = \prod_k \phi_{nk}^{z_{nk}}.$$

Variational Inference for LDA (Blei et al., 2003)

Assume **B** is a parameter, assume there is a single document, and focus on the inference on $\boldsymbol{\theta}$ and $(\mathbf{z}_n)_n$. Choose q in a factorized form (*mean field* approximation)

$$q(\boldsymbol{\theta}, (\mathbf{z}_n)_n) = q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \prod_{n=1}^{N} q_{\mathbf{z}_n}(\mathbf{z}_n) \qquad \text{with}$$

$$q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \frac{\Gamma(\sum_k \gamma_k)}{\prod_k \Gamma(\gamma_k)} \prod_k \theta_k^{\gamma_k - 1} \qquad \text{and} \qquad q_{\mathbf{z}_n}(\mathbf{z}_n) = \prod_k \phi_{nk}^{z_{nk}}.$$

$$KL\big(q \,\|\, p(\cdot|\mathbf{W})\big) = \mathbb{E}_q\Big[\log \frac{q(\boldsymbol{\theta}, (\mathbf{z}_n)_n)}{p(\boldsymbol{\theta}, (\mathbf{z}_n)_n \mid \mathbf{W})}\Big] = \mathbb{E}_q\Bigg[\log q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) + \sum_n \log q_{\mathbf{z}_n}(\mathbf{z}_n)$$

$$\ldots - \log p(\boldsymbol{\theta}|\boldsymbol{\alpha}) - \sum_n \big(\log p(\mathbf{z}_n|\boldsymbol{\theta}) + \log p(\mathbf{w}_n|\mathbf{z}_n, \mathbf{B})\big)\Bigg] - p\big((\mathbf{w}_n)_n\big)$$

Variational Inference for LDA (Blei et al., 2003)

Assume $\mathbf{B}$ is a parameter, assume there is a single document, and focus on the inference on $\boldsymbol{\theta}$ and $(\mathbf{z}_n)_n$. Choose q in a factorized form (*mean field* approximation)

$$q(\boldsymbol{\theta}, (\mathbf{z}_n)_n) = q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \prod_{n=1}^{N} q_{\mathbf{z}_n}(\mathbf{z}_n) \qquad \text{with}$$

$$q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \frac{\Gamma(\sum_k \gamma_k)}{\prod_k \Gamma(\gamma_k)} \prod_k \theta_k^{\gamma_k - 1} \qquad \text{and} \qquad q_{\mathbf{z}_n}(\mathbf{z}_n) = \prod_k \phi_{nk}^{z_{nk}}.$$

$$KL\big(q \,\|\, p(\cdot|\mathbf{W})\big) = \qquad \mathbb{E}_q\Big[\log q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) + \sum_n \log q_{\mathbf{z}_n}(\mathbf{z}_n)$$

$$\ldots - \log p(\boldsymbol{\theta}|\boldsymbol{\alpha}) - \sum_n \big(\log p(\mathbf{z}_n|\boldsymbol{\theta}) + \log p(\mathbf{w}_n|\mathbf{z}_n, \mathbf{B})\big)\Big] - p\big((\mathbf{w}_n)_n\big)$$

Variational Inference for LDA (Blei et al., 2003)

Assume $\mathbf{B}$ is a parameter, assume there is a single document, and focus on the inference on $\boldsymbol{\theta}$ and $(\mathbf{z}_n)_n$. Choose q in a factorized form (*mean field* approximation)

$$q(\boldsymbol{\theta}, (\mathbf{z}_n)_n) = q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \prod_{n=1}^{N} q_{\mathbf{z}_n}(\mathbf{z}_n) \qquad \text{with}$$

$$q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \frac{\Gamma(\sum_k \gamma_k)}{\prod_k \Gamma(\gamma_k)} \prod_k \theta_k^{\gamma_k - 1} \qquad \text{and} \qquad q_{\mathbf{z}_n}(\mathbf{z}_n) = \prod_k \phi_{nk}^{z_{nk}}.$$

$$KL\big(q \,\|\, p(\cdot|\mathbf{W})\big) = \mathbb{E}_q\Big[\log \frac{q(\boldsymbol{\theta}, (\mathbf{z}_n)_n)}{p(\boldsymbol{\theta}, (\mathbf{z}_n)_n \mid \mathbf{W})}\Big] = \mathbb{E}_q\Big[\log q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) + \sum_n \log q_{\mathbf{z}_n}(\mathbf{z}_n)$$

$$\ldots - \log p(\boldsymbol{\theta}|\boldsymbol{\alpha}) - \sum_n \big(\log p(\mathbf{z}_n|\boldsymbol{\theta}) + \log p(\mathbf{w}_n|\mathbf{z}_n, \mathbf{B})\big)\Big] - p\big((\mathbf{w}_n)_n\big)$$

Variational Inference for LDA II

$$\mathbb{E}\Big[\log q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}|\boldsymbol{\alpha}) + \sum_n \big(\log q_{\mathbf{z}_n}(\mathbf{z}_n) - \log p(\mathbf{z}_n|\boldsymbol{\theta}) - \log p(\mathbf{w}_n|\mathbf{z}_n, \mathbf{B})\big)\Big]$$

Variational Inference for LDA II

$$\mathbb{E}\Big[\log q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}|\boldsymbol{\alpha}) + \sum_n \big(\log q_{\mathbf{z}_n}(\mathbf{z}_n) - \log p(\mathbf{z}_n|\boldsymbol{\theta}) - \log p(\mathbf{w}_n|\mathbf{z}_n, \mathbf{B})\big)\Big]$$

$$\mathbb{E}_q\big[\log q_{\boldsymbol{\theta}}(\boldsymbol{\theta})\big] =$$

Variational Inference for LDA II

$$\mathbb{E}\Big[\log q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}|\boldsymbol{\alpha}) + \sum_n \big(\log q_{\mathbf{z}_n}(\mathbf{z}_n) - \log p(\mathbf{z}_n|\boldsymbol{\theta}) - \log p(\mathbf{w}_n|\mathbf{z}_n, \mathbf{B})\big)\Big]$$

$$\mathbb{E}_q\big[\log q_{\boldsymbol{\theta}}(\boldsymbol{\theta})\big] \;=\; \mathbb{E}_q\big[\log \Gamma(\textstyle\sum_k \gamma_k) - \sum_k \log \Gamma(\gamma_k) + \sum_k \big((\gamma_k - 1)\log(\theta_k)\big)\big]$$

Variational Inference for LDA II

$$\mathbb{E}\Big[\log q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}|\boldsymbol{\alpha}) + \sum_n \big(\log q_{\mathbf{z}_n}(\mathbf{z}_n) - \log p(\mathbf{z}_n|\boldsymbol{\theta}) - \log p(\mathbf{w}_n|\mathbf{z}_n, \mathbf{B})\big)\Big]$$

$$\begin{aligned}
\mathbb{E}_q\big[\log q_{\boldsymbol{\theta}}(\boldsymbol{\theta})\big] &= \mathbb{E}_q\big[\log \Gamma(\textstyle\sum_k \gamma_k) - \sum_k \log \Gamma(\gamma_k) + \sum_k \big((\gamma_k - 1) \log(\theta_k)\big)\big] \\
&= \log \Gamma(\textstyle\sum_k \gamma_k) - \sum_k \log \Gamma(\gamma_k) + \sum_k \big((\gamma_k - 1)\, \mathbb{E}_q[\log(\theta_k)]\big)
\end{aligned}$$

Variational Inference for LDA II

$$\mathbb{E}\Big[\log q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}|\boldsymbol{\alpha}) + \sum_n \big(\log q_{\mathbf{z}_n}(\mathbf{z}_n) - \log p(\mathbf{z}_n|\boldsymbol{\theta}) - \log p(\mathbf{w}_n|\mathbf{z}_n, \mathbf{B})\big)\Big]$$

$$\begin{aligned}
\mathbb{E}_q\big[\log q_{\boldsymbol{\theta}}(\boldsymbol{\theta})\big] &= \mathbb{E}_q\big[\log \Gamma(\textstyle\sum_k \gamma_k) - \sum_k \log \Gamma(\gamma_k) + \sum_k \big((\gamma_k - 1) \log(\theta_k)\big)\big] \\
&= \log \Gamma(\textstyle\sum_k \gamma_k) - \sum_k \log \Gamma(\gamma_k) + \sum_k \big((\gamma_k - 1)\, \mathbb{E}_q[\log(\theta_k)]\big)
\end{aligned}$$

$$\mathbb{E}_q[p(\boldsymbol{\theta}|\boldsymbol{\alpha})] \quad =$$

Variational Inference for LDA II

$$\mathbb{E}\Big[\log q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}|\boldsymbol{\alpha}) + \sum_n \big(\log q_{\mathbf{z}_n}(\mathbf{z}_n) - \log p(\mathbf{z}_n|\boldsymbol{\theta}) - \log p(\mathbf{w}_n|\mathbf{z}_n, \mathbf{B})\big)\Big]$$

$$\begin{aligned}\mathbb{E}_q\big[\log q_{\boldsymbol{\theta}}(\boldsymbol{\theta})\big] &= \mathbb{E}_q\big[\log \Gamma(\textstyle\sum_k \gamma_k) - \sum_k \log \Gamma(\gamma_k) + \sum_k \big((\gamma_k - 1) \log(\theta_k)\big)\big] \\ &= \log \Gamma(\textstyle\sum_k \gamma_k) - \sum_k \log \Gamma(\gamma_k) + \sum_k \big((\gamma_k - 1)\, \mathbb{E}_q[\log(\theta_k)]\big)\end{aligned}$$

$$\mathbb{E}_q[p(\boldsymbol{\theta}|\boldsymbol{\alpha})] = \mathbb{E}[(\alpha_k - 1)\log(\theta_k)] + cst = (\alpha_k - 1)\mathbb{E}_q[\log(\theta_k)] + cst$$

Variational Inference for LDA II

$$\mathbb{E}\Big[\log q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}|\boldsymbol{\alpha}) + \sum_n \big(\log q_{\mathbf{z}_n}(\mathbf{z}_n) - \log p(\mathbf{z}_n|\boldsymbol{\theta}) - \log p(\mathbf{w}_n|\mathbf{z}_n, \mathbf{B})\big)\Big]$$

$$\begin{aligned}\mathbb{E}_q\big[\log q_{\boldsymbol{\theta}}(\boldsymbol{\theta})\big] &= \mathbb{E}_q\big[\log \Gamma(\textstyle\sum_k \gamma_k) - \sum_k \log \Gamma(\gamma_k) + \sum_k \big((\gamma_k - 1) \log(\theta_k)\big)\big] \\ &= \log \Gamma(\textstyle\sum_k \gamma_k) - \sum_k \log \Gamma(\gamma_k) + \sum_k \big((\gamma_k - 1)\, \mathbb{E}_q[\log(\theta_k)]\big)\end{aligned}$$

$$\mathbb{E}_q[p(\boldsymbol{\theta}|\boldsymbol{\alpha})] = \mathbb{E}[(\alpha_k - 1)\log(\theta_k)] + cst = (\alpha_k - 1)\mathbb{E}_q[\log(\theta_k)] + cst$$

$$\mathbb{E}_q[\log q_{\mathbf{z}_n}(\mathbf{z}_n) - \log p(\mathbf{z}_n)] =$$

Variational Inference for LDA II

$$\mathbb{E}\Big[\log q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}|\boldsymbol{\alpha}) + \sum_n \big(\log q_{\mathbf{z}_n}(\mathbf{z}_n) - \log p(\mathbf{z}_n|\boldsymbol{\theta}) - \log p(\mathbf{w}_n|\mathbf{z}_n, \mathbf{B})\big)\Big]$$

$$\begin{aligned}
\mathbb{E}_q\big[\log q_{\boldsymbol{\theta}}(\boldsymbol{\theta})\big] &= \mathbb{E}_q\big[\log \Gamma(\textstyle\sum_k \gamma_k) - \sum_k \log \Gamma(\gamma_k) + \sum_k \big((\gamma_k - 1) \log(\theta_k)\big)\big] \\
&= \log \Gamma(\textstyle\sum_k \gamma_k) - \sum_k \log \Gamma(\gamma_k) + \sum_k \big((\gamma_k - 1)\, \mathbb{E}_q[\log(\theta_k)]\big)
\end{aligned}$$

$$\mathbb{E}_q[p(\boldsymbol{\theta}|\boldsymbol{\alpha})] = \mathbb{E}[(\alpha_k - 1)\log(\theta_k)] + cst = (\alpha_k - 1)\mathbb{E}_q[\log(\theta_k)] + cst$$

$$\begin{aligned}
\mathbb{E}_q[\log q_{\mathbf{z}_n}(\mathbf{z}_n) - \log p(\mathbf{z}_n)] &= \mathbb{E}_q\Big[\sum_k \big(z_{nk}\log(\phi_{nk}) - z_{nk}\log(\theta_k)\big)\Big] \\
&= \sum_k \mathbb{E}_q[z_{nk}]\big(\log(\phi_{nk}) - \mathbb{E}_q[\log(\theta_k)]\big)
\end{aligned}$$

Variational Inference for LDA II

$$\mathbb{E}\Big[\log q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}|\boldsymbol{\alpha}) + \sum_n \big(\log q_{\mathbf{z}_n}(\mathbf{z}_n) - \log p(\mathbf{z}_n|\boldsymbol{\theta}) - \log p(\mathbf{w}_n|\mathbf{z}_n, \mathbf{B})\big)\Big]$$

$$\begin{aligned}\mathbb{E}_q\big[\log q_{\boldsymbol{\theta}}(\boldsymbol{\theta})\big] &= \mathbb{E}_q\big[\log \Gamma(\textstyle\sum_k \gamma_k) - \sum_k \log \Gamma(\gamma_k) + \sum_k \big((\gamma_k - 1) \log(\theta_k)\big)\big] \\ &= \log \Gamma(\textstyle\sum_k \gamma_k) - \sum_k \log \Gamma(\gamma_k) + \sum_k \big((\gamma_k - 1)\, \mathbb{E}_q[\log(\theta_k)]\big)\end{aligned}$$

$$\mathbb{E}_q[p(\boldsymbol{\theta}|\boldsymbol{\alpha})] = \mathbb{E}[(\alpha_k - 1)\log(\theta_k)] + cst = (\alpha_k - 1)\mathbb{E}_q[\log(\theta_k)] + cst$$

$$\begin{aligned}\mathbb{E}_q[\log q_{\mathbf{z}_n}(\mathbf{z}_n) - \log p(\mathbf{z}_n)] &= \mathbb{E}_q\Big[\sum_k \big(z_{nk}\log(\phi_{nk}) - z_{nk}\log(\theta_k)\big)\Big] \\ &= \sum_k \mathbb{E}_q[z_{nk}]\big(\log(\phi_{nk}) - \mathbb{E}_q[\log(\theta_k)]\big)\end{aligned}$$

$$\mathbb{E}_q[\log p(\mathbf{w}_n|\mathbf{z}_n, \mathbf{B})] = \mathbb{E}_q\Big[\sum z_{nk} w_{nj} \log(b_{jk})\Big] =$$

Variational Inference for LDA II

$$\mathbb{E}\Big[\log q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}|\boldsymbol{\alpha}) + \sum_n \big(\log q_{\mathbf{z}_n}(\mathbf{z}_n) - \log p(\mathbf{z}_n|\boldsymbol{\theta}) - \log p(\mathbf{w}_n|\mathbf{z}_n, \mathbf{B})\big)\Big]$$

$$\begin{aligned}\mathbb{E}_q\big[\log q_{\boldsymbol{\theta}}(\boldsymbol{\theta})\big] &= \mathbb{E}_q\big[\log\Gamma(\textstyle\sum_k \gamma_k) - \sum_k \log\Gamma(\gamma_k) + \sum_k \big((\gamma_k - 1)\log(\theta_k)\big)\big] \\ &= \log\Gamma(\textstyle\sum_k \gamma_k) - \sum_k \log\Gamma(\gamma_k) + \sum_k \big((\gamma_k - 1)\,\mathbb{E}_q[\log(\theta_k)]\big)\end{aligned}$$

$$\mathbb{E}_q[p(\boldsymbol{\theta}|\boldsymbol{\alpha})] = \mathbb{E}[(\alpha_k - 1)\log(\theta_k)] + cst = (\alpha_k - 1)\mathbb{E}_q[\log(\theta_k)] + cst$$

$$\begin{aligned}\mathbb{E}_q[\log q_{\mathbf{z}_n}(\mathbf{z}_n) - \log p(\mathbf{z}_n)] &= \mathbb{E}_q\Big[\sum_k \big(z_{nk}\log(\phi_{nk}) - z_{nk}\log(\theta_k)\big)\Big] \\ &= \sum_k \mathbb{E}_q[z_{nk}]\big(\log(\phi_{nk}) - \mathbb{E}_q[\log(\theta_k)]\big)\end{aligned}$$

$$\mathbb{E}_q[\log p(\mathbf{w}_n|\mathbf{z}_n, \mathbf{B})] = \mathbb{E}_q\Big[\sum z_{nk} w_{nj}\log(b_{jk})\Big] = \sum \mathbb{E}_q[z_{nk}]\, w_{nj}\log(b_{jk})$$

VI for LDA: Computing the expectations

The expectation of the logarithm of a Dirichlet r.v. can be computed exactly with the digamma function Ψ:

$$\mathbb{E}_q[\log(\theta_k)] = \Psi(\gamma_k) - \Psi(\textstyle\sum_k \gamma_k), \quad \text{with} \quad \Psi(x) := \frac{\partial}{\partial x}\big(\log \Gamma(x)\big).$$

VI for LDA: Computing the expectations

The expectation of the logarithm of a Dirichlet r.v. can be computed exactly with the digamma function Ψ:

$$\mathbb{E}_q[\log(\theta_k)] = \Psi(\gamma_k) - \Psi(\textstyle\sum_k \gamma_k), \quad \text{with} \quad \Psi(x) := \frac{\partial}{\partial x}\big(\log \Gamma(x)\big).$$

We obviously have $\mathbb{E}_q[z_{nk}] = \phi_{nk}$.

VI for LDA: Computing the expectations

The expectation of the logarithm of a Dirichlet r.v. can be computed exactly with the digamma function Ψ:

$$\mathbb{E}_q[\log(\theta_k)] = \Psi(\gamma_k) - \Psi(\textstyle\sum_k \gamma_k), \quad \text{with} \quad \Psi(x) := \frac{\partial}{\partial x}\big(\log \Gamma(x)\big).$$

We obviously have $\mathbb{E}_q[z_{nk}] = \phi_{nk}$.

The problem $\min\limits_{q \in \mathcal{Q}} KL\big(q \,\|\, p(\cdot|\mathbf{W})\big)$ is therefore equivalent to

$$\min_{\gamma,(\phi_n)_n} D(\gamma, (\phi_n)_n) \qquad \text{with}$$

VI for LDA: Computing the expectations

The expectation of the logarithm of a Dirichlet r.v. can be computed exactly with the digamma function Ψ:

$$\mathbb{E}_q[\log(\theta_k)] = \Psi(\gamma_k) - \Psi(\textstyle\sum_k \gamma_k), \quad \text{with} \quad \Psi(x) := \frac{\partial}{\partial x}\big(\log\Gamma(x)\big).$$

We obviously have $\mathbb{E}_q[z_{nk}] = \phi_{nk}$.

The problem $\min\limits_{q\in\mathcal{Q}} KL\big(q \,\|\, p(\cdot|\mathbf{W})\big)$ is therefore equivalent to

$$\min_{\boldsymbol{\gamma},(\boldsymbol{\phi}_n)_n} D(\boldsymbol{\gamma},(\boldsymbol{\phi}_n)_n) \qquad \text{with}$$

$$D(\boldsymbol{\gamma},(\boldsymbol{\phi}_n)_n) = \log\Gamma(\sum_k \gamma_k) - \sum_k \log\Gamma(\gamma_k) + \sum_{n,k} \phi_{nk}\log(\phi_{nk})$$

$$-\sum_{n,k}\phi_{nk}\sum_j w_{nj}\log(b_{jk}) - \sum_k\big((\alpha_k + \textstyle\sum_n \phi_{nk} - \gamma_k)\big)\big(\Psi(\gamma_k) - \Psi(\textstyle\sum_k \gamma_k)\big)$$

VI for LDA: Solving for the variational updates

Introducing a Lagrangian to account for the constraints $\sum_{k=1}^{K} \phi_{nk} = 1$:

$$\mathcal{L}(\gamma, (\phi_n)_n) = D(\gamma, (\phi_n)_n) + \sum_{n=1}^{N} \lambda_n \left(1 - \sum_k \phi_{nk}\right)$$

VI for LDA: Solving for the variational updates

Introducing a Lagrangian to account for the constraints $\sum_{k=1}^{K} \phi_{nk} = 1$:

$$\mathcal{L}(\gamma, (\phi_n)_n) = D(\gamma, (\phi_n)_n) + \sum_{n=1}^{N} \lambda_n \left(1 - \sum_k \phi_{nk}\right)$$

Computing the gradient of the Lagrangian:

VI for LDA: Solving for the variational updates

Introducing a Lagrangian to account for the constraints $\sum_{k=1}^{K} \phi_{nk} = 1$:

$$\mathcal{L}(\gamma, (\phi_n)_n) = D(\gamma, (\phi_n)_n) + \sum_{n=1}^{N} \lambda_n \left(1 - \sum_k \phi_{nk}\right)$$

Computing the gradient of the Lagrangian:

$$\frac{\partial \mathcal{L}}{\partial \gamma_k} = -(\alpha_k + \sum_n \phi_{nk} - \gamma_k)(\Psi'(\gamma_k) - \Psi'(\sum_k \gamma_k))$$

$$\frac{\partial \mathcal{L}}{\partial \phi_{nk}} = \log(\phi_{nk}) + 1 - \sum_j w_{nj} \log(b_{jk}) - (\Psi(\gamma_k) - \Psi(\sum_k \gamma_k)) - \lambda_n$$

Partial minimizations in γ and ϕ_{nk} are therefore respectively solved by

$$\boxed{\gamma_k = \alpha_k + \sum_n \phi_{nk}} \qquad \text{and} \qquad \boxed{\phi_{nk} \propto b_{j(n),k} \exp(\Psi(\gamma_k) - \Psi(\sum_k \gamma_k)),}$$

where $j(n)$ is the one and only j such that $w_{nj} = 1$.

VI for LDA: Solving for the variational updates

Introducing a Lagrangian to account for the constraints $\sum_{k=1}^{K} \phi_{nk} = 1$:

$$\mathcal{L}(\gamma, (\phi_n)_n) = D(\gamma, (\phi_n)_n) + \sum_{n=1}^{N} \lambda_n \left(1 - \sum_k \phi_{nk}\right)$$

Computing the gradient of the Lagrangian:

$$\frac{\partial \mathcal{L}}{\partial \gamma_k} = -(\alpha_k + \sum_n \phi_{nk} - \gamma_k)(\Psi'(\gamma_k) - \Psi'(\sum_k \gamma_k))$$

$$\frac{\partial \mathcal{L}}{\partial \phi_{nk}} = \log(\phi_{nk}) + 1 - \sum_j w_{nj} \log(b_{jk}) - (\Psi(\gamma_k) - \Psi(\sum_k \gamma_k)) - \lambda_n$$

VI for LDA: Solving for the variational updates

Introducing a Lagrangian to account for the constraints $\sum_{k=1}^{K} \phi_{nk} = 1$:

$$\mathcal{L}(\gamma, (\phi_n)_n) = D(\gamma, (\phi_n)_n) + \sum_{n=1}^{N} \lambda_n \left(1 - \sum_k \phi_{nk}\right)$$

Computing the gradient of the Lagrangian:

$$\frac{\partial \mathcal{L}}{\partial \gamma_k} = -(\alpha_k + \sum_n \phi_{nk} - \gamma_k)(\Psi'(\gamma_k) - \Psi'(\sum_k \gamma_k))$$

$$\frac{\partial \mathcal{L}}{\partial \phi_{nk}} = \log(\phi_{nk}) + 1 - \sum_j w_{nj} \log(b_{jk}) - (\Psi(\gamma_k) - \Psi(\sum_k \gamma_k)) - \lambda_n$$

Partial minimizations in γ and ϕ_{nk} are therefore respectively solved by

$$\boxed{\gamma_k = \alpha_k + \sum_n \phi_{nk}} \qquad \text{and} \qquad \boxed{\phi_{nk} \propto b_{j(n),k} \exp(\Psi(\gamma_k) - \Psi(\sum_k \gamma_k)),}$$

where $j(n)$ is the one and only j such that $w_{nj} = 1$.

Variational Algorithm

Algorithm 1 Variational inference for LDA

Require: $\mathbf{W}, \alpha, \gamma_{\text{init}}, (\phi_{n,\text{init}})_n$

while Not converged **do**
  $\gamma_k \leftarrow \alpha_k + \sum_n \phi_{nk}$
  for n=1..N **do**
    for k=1..K **do**
      $\phi_{nk} \leftarrow b_{j(n),k} \exp(\Psi(\gamma_k) - \Psi(\sum_k \gamma_k))$
    end for
    $\phi_n \leftarrow \frac{1}{\sum_k \phi_{nk}} \phi_n$
  end for
end while
return $\gamma, (\phi_n)_n$

Variational Algorithm

Algorithm 2 Variational inference for LDA

Require: $\mathbf{W}, \alpha, \gamma_{\text{init}}, (\phi_{n,\text{init}})_n$

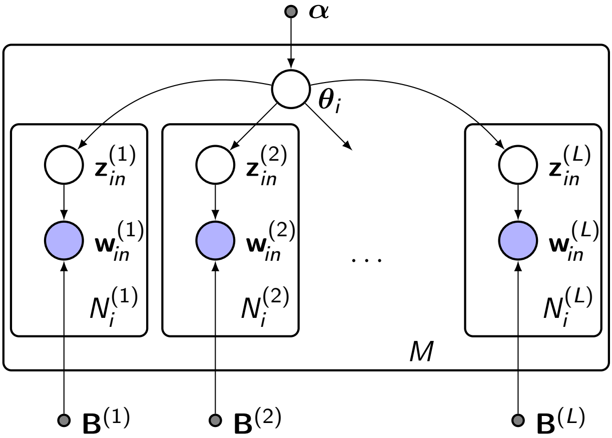
while Not converged **do**
    $\gamma_k \leftarrow \alpha_k + \sum_n \phi_{nk}$
    for n=1..N **do**
        for k=1..K **do**
            $\phi_{nk} \leftarrow b_{j(n),k} \exp(\Psi(\gamma_k) - \Psi(\sum_k \gamma_k))$
        end for
        $\phi_n \leftarrow \frac{1}{\sum_k \phi_{nk}} \phi_n$
    end for
end while
return $\gamma, (\phi_n)_n$

With the quantities computed we can approximate:

$$\boxed{\mathbb{E}[\theta_k|\mathbf{W}] \approx \frac{\gamma_k}{\sum_{k'} \gamma_{k'}}} \quad \text{and} \quad \boxed{\mathbb{E}[\mathbf{z}_n|\mathbf{W}] \approx \phi_n}$$

Polylingual Topic Model (Mimno et al., 2009)

Generalization of LDA to documents *available simultaneously in several languages* such as Wikipedia articles, which are not literal translations of one another but share the same topics.



References I

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

Boyd-Graber, J., Chang, J., Gerrish, S., Wang, C., and Blei, D. (2009). Reading tea leaves: How humans interpret topic models. In *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems*.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1):177–196.

Mimno, D., Wallach, H., Naradowsky, J., Smith, D., and McCallum, A. (2009). Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 880–889. Association for Computational Linguistics.