# Geographic Topic Model: Appendix

**Jacob Eisenstein**
Machine Learning Department
Carnegie Mellon University
jacobeis@cs.cmu.edu

## Abstract

Faceted topic models combine topical content with extraneous facets, such as ideology or dialect. In this model, the "pure" topics are corrupted by the facets, using a hierarchical generative model in which the pure topics act as priors on the faceted topics. This is most easily modeled using the logistic-normal distribution, which admits a normal prior on the mean.

## 1 Model

We build on latent Dirichlet allocation:

- For each document $d$, draw $\boldsymbol{\theta}_d \sim \text{Dirichlet}(\alpha_\theta)$
- For each token $n < N_d$
  - Draw a topic index $z_n \sim \boldsymbol{\theta}_d$
  - Draw a word token from the associated topic $w_n \sim \boldsymbol{\beta}_{z_n}$

We augment each document with an additional discrete facet variable, $v_d \sim \boldsymbol{\vartheta}$, which selects the appropriate faceted topic ($\boldsymbol{\vartheta} \sim \text{Dir}(\alpha_\vartheta)$): $w_n \sim \boldsymbol{\beta}_{z_n}^{(v_d)}$. Here $z_n$ indexes the topic and $v_d$ indicates the facet of the topic: for example, $v_d$ may select the "Detroit" version of the "electronic music" topic. We may also introduce metadata $\boldsymbol{y}_d$, such that $\boldsymbol{y}_d \sim f(\boldsymbol{y}; \boldsymbol{\rho}_{v_d})$, indicating an arbitrary probability distribution with parameters $\boldsymbol{\rho}_{v_d}$. In the geographical topic model, $f(\boldsymbol{y}; \boldsymbol{\rho}_{v_d})$ takes the form of a bivariate Gaussian; variational inference in this setting is deferred to [4].

The faceted topics take logistic-Normal priors, such that $\boldsymbol{\beta}_k^{(j)} \sim \mathcal{LN}(\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2)$.[1] It will be convenient to introduce the auxiliary variable $\boldsymbol{\eta}$, such that $\boldsymbol{\beta} = \exp \boldsymbol{\eta} / \sum_i \exp \eta[i]$ and $\boldsymbol{\eta}_k^{(j)} \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2)$. Throughout, $i$ will index the word, $j$ will index the facet, and $k$ will index the topic. We draw the pure topics from Normal priors, $\boldsymbol{\mu}_k \sim \mathcal{N}(\boldsymbol{a}, b)$), and the topic-variances from Gamma priors, $\sigma_k^2[i] \sim \mathcal{G}(c, d)$.

## 2 Variational Approximation

We'll make a fully-factorized approximation:

$$Q(\boldsymbol{z}, \boldsymbol{v}, \boldsymbol{\theta}, \boldsymbol{\vartheta}, \boldsymbol{\eta}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) = \left( \prod_k^K q(\boldsymbol{\mu}_k) q(\boldsymbol{\sigma}_k^2) \prod_j^J q(\boldsymbol{\eta}_k^{(j)}) \right) \prod_d^D q(v_d) q(\boldsymbol{\theta}_d) \prod_n^{N_d} q(z_{dn}). \quad (1)$$

[1] All covariance matrices are diagonal in this work.

We want to maximize the following bound (with $\langle x \rangle$ denoting the expectation of $x$ under the distribution $q$).

$$
\begin{aligned}
L(\boldsymbol{w}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v}, \boldsymbol{\theta}, \boldsymbol{\vartheta}, \boldsymbol{\eta}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) =& \langle \log p(\boldsymbol{\vartheta} | \alpha_\vartheta) \rangle - \langle \log q(\boldsymbol{\vartheta}) \rangle \\
&+ \sum_k \langle \log p(\boldsymbol{\mu}_k | \boldsymbol{a}_k, b) \rangle - \langle \log q(\boldsymbol{\mu}_k) \rangle \\
&+ \langle \log p(\boldsymbol{\sigma}_k^2 | c, d) \rangle - \langle \log q(\boldsymbol{\sigma}_k^2) \rangle \\
&+ \sum_j \langle \log p(\boldsymbol{\eta}_k^{(j)} | \boldsymbol{\mu}_k, \sigma_k^2) \rangle - \langle \log q(\boldsymbol{\eta}_k^{(j)}) \rangle \\
&+ \langle \log p(\boldsymbol{\theta}_d | \alpha_\theta) \rangle - \langle \log q(\boldsymbol{\theta}_d) \rangle \\
&+ \langle \log p(v_d | \boldsymbol{\vartheta}) \rangle - \langle \log q(v_d) \rangle \\
&+ \sum_n^{N_d} \langle \log p(z_n | \boldsymbol{\theta}) \rangle - \langle \log q(z_n) \rangle \\
&+ \langle \log p(w_n | z_n, v, \boldsymbol{\beta}) \rangle
\end{aligned}
$$

The variational parameters on the normal distributions can be expressed in terms of expectations (e.g. $\langle \boldsymbol{\mu} \rangle$) and variances (e.g. $\mathcal{V}(\boldsymbol{\mu})$), so I won't introduce symbols for those. I'll write $\boldsymbol{\phi}$ for the multinomial variational parameter on $z_n$, where $\phi_{nk} = \langle \delta_{z_n,k} \rangle$, the expectation of a delta function that returns 1 when $z_n = k$ and zero otherwise. The parameter $\boldsymbol{\omega}_d$ plays an analagous role for $v_d$.

### 2.1 Taylor Approximation

The term $\langle \log p(w|z, v, \beta) \rangle$ will involve computing $\langle \log \beta_k^{(j)}[w_n] \rangle$, the expected log of the topic probability for word $w_n$. Approximation is required:

$$
\langle \log \beta_k^{(j)}[w_n] \rangle = \langle \eta_k^{(j)} \rangle - \langle \log \sum_i \exp \eta_k^{(j)}[i] \rangle \tag{2}
$$

$$
\langle \log \sum_i \exp \eta_k^{(j)}[i] \rangle \approx \log \zeta_k^{(j)} - 1 + \left( \sum_i \langle \exp \eta_k^{(j)}[i] \rangle \right) / \zeta_k^{(j)},
$$

making a first-order Taylor approximation of the intractable expectation of the normalizer. This approximation introduces the additional variational parameter $\zeta$. See [1] for details.

## 3 Variational updates

### 3.1 Updates: $\eta$

#### 3.1.1 Mean parameter $\langle \eta \rangle$

As shown above, the expected language models $\langle \log \beta \rangle$ can be computed deterministically from the parameters $\langle \eta \rangle$, $\langle \exp \eta \rangle$, and $\zeta$.

Because $q(\eta)$ is Gaussian, $\langle \exp \eta \rangle$ has a simple closed form; $\zeta$ can also be computed directly [1]:

$$
\langle \exp \eta_k^{(j)}[i] \rangle = \exp \left( \langle \eta_k^{(j)}[i] \rangle + \frac{1}{2} \mathcal{V}(\eta_k^{(j)}[i]) \right) \tag{3}
$$

$$
\zeta_k^{(j)} = \sum_i \langle \exp \eta_k^{(j)}[i] \rangle. \tag{4}
$$

This leaves us with the task of computing $\langle\eta\rangle$ and $\mathcal{V}(\eta)$. Unfortunately, these are not available in closed form. We derive gradient-based updates from the objective $L$,

$$
\begin{aligned}
L_{[\langle\eta_k^{(j)}\rangle[i]]} =& \langle\log p(w|z,v,\eta)\rangle + \langle\log p(\eta_k^{(j)}[i]|\mu_k[i],\sigma_k^2[i])\rangle \\
=& C + \sum_d \omega_{dj} \sum_n^{N_d} \phi_{nk}\left(\delta_{w_n=i}\langle\eta_k^{(j)}[i]\rangle - \langle\exp\eta_k^{(j)}[i]\rangle/\zeta_k^{(j)}\right) - \frac{1}{2}\langle\sigma_k^{-2}[i]\rangle\langle(\eta_k^{(j)}[i]-\mu_k[i])^2\rangle,
\end{aligned}
\tag{5}
$$

where $C$ is constant in $\langle\eta\rangle$. The final term will be relevant for several of the updates – it expands to

$$
\langle(\eta^{(j)}-\mu)^2\rangle = \langle\eta^{(j)}\rangle^2 + \langle\mu\rangle^2 + \mathcal{V}(\eta^{(j)}) + \mathcal{V}(\mu) - 2\langle\mu\rangle\langle\eta^{(j)}\rangle \tag{6}
$$

To take the gradient $\partial L/\partial\langle\eta\rangle$, we first observe that $\frac{\partial}{\partial\langle\eta\rangle}\langle\exp\eta\rangle = \langle\exp\eta\rangle$, because the quantity inside the exponent (equation 3) is linear in $\langle\eta\rangle$. Applying equation 6, the gradient of the objective is,

$$
\frac{\partial L}{\partial\langle\eta_k^{(j)}\rangle[i]} = N(i,j,k) - N(j,k)\langle\exp\eta_k^{(j)}[i]\rangle/\zeta_k^{(j)} - (\langle\eta_k^{(j)}[i]\rangle - \langle\mu_k[i]\rangle)\langle\sigma_k^{-2}[i]\rangle, \tag{7}
$$

where $N(i,j,k)$ is the expected count of term $i$, facet $j$, and topic $k$ (summing over $\phi$ and $\omega$). Because $\zeta_k^{(j)} = \sum_{i'}\langle\exp\eta_k^{(j)}[i']\rangle$, we have $\langle\exp\eta_k^{(j)}[i]\rangle/\zeta_k^{(j)} = \langle\beta_k^i\rangle$. This yields the intuitively appealing alternative form for the gradient,

$$
\frac{\partial L}{\partial\langle\eta_k^{(j)}\rangle[i]} = N(i,j,k) - N(j,k)\beta_k^{(j)}[i] - (\langle\eta_k^{(j)}[i]\rangle - \langle\mu_k[i]\rangle)\langle\sigma_k^{-2}[i]\rangle,
$$

in which the first two terms compare expected counts using observed words with the expected counts under the topic $\langle\beta\rangle$, and the third term penalizes divergence from the prior.

#### 3.1.2 Variance $\mathcal{V}(\eta)$

We now compute the posterior variance $\mathcal{V}(\eta)$. Again, we begin by identifying the terms of the objective that relate to this variational parameter:

$$
\begin{aligned}
L_{[\mathcal{V}(\eta_k^{(j)}[i])]} =& \langle\log p(w|z,v,\eta)\rangle + \langle\log p(\eta_k^{(j)}[i]|\mu_k[i],\sigma_k^2[i])\rangle - \langle\log q(\eta_k^{(j)})\rangle \\
=& C - \sum_d \omega_{dj} \sum_n^{N_d} \phi_{nk}\langle\exp\eta_k^{(j)}[i]\rangle/\zeta_k^{(j)} - \frac{1}{2}\langle\sigma_k^{-2}[i]\rangle\mathcal{V}(\eta_k^{(j)}[i]) + \frac{1}{2}\log\mathcal{V}(\eta_k^{(j)}[i]),
\end{aligned}
\tag{8}
$$

The first term is from Equation 3, which shows that $\langle\exp\eta\rangle$ includes the variance; the second term is from Equation 6; the third term is obtained from the entropy of the normal distribution. Taking the derivative with respect to the parameter of interest,

$$
\begin{aligned}
\frac{\partial L}{\partial\mathcal{V}(\eta_k^{(j)}[i])} =& -\frac{N(j,k)\langle\exp\eta_k^{(j)}[i]\rangle}{2\zeta_k^{(j)}} - \langle\sigma_k^{-2}[i]\rangle/2 + \frac{1}{2\mathcal{V}(\eta_k^{(j)}[i])} \\
=& -\frac{1}{2}\left(N(j,k)\langle\beta_k^{(j)}[i]\rangle + \langle\sigma_k^{-2}[i]\rangle - \mathcal{V}(\eta_k^{(j)}[i])^{-1}\right) \\
\mathcal{V}(\eta_k^{(j)}[i]) =& \left(\sum_d \omega_{dj}\sum_n \phi_{nk}\langle\beta_k^{(j)}[i]\rangle + \langle\sigma_k^{-2}[i]\rangle\right)^{-1}.
\end{aligned}
\tag{9}
$$

**Note that the last line is not a closed-form solution**, because $\langle\beta_k^{(j)}\rangle$ implicitly includes $\mathcal{V}(\eta_k^{(j)})$; the goal is to give intuition. If the expected word counts are zero, then the posterior variance $\mathcal{V}(\eta)$ will be identical to the prior $\langle\sigma^2\rangle$; as the expected word counts increase, the posterior variance goes to zero.

### 3.2 Updates: $\boldsymbol{\mu}$

The update for the expectation $\langle \boldsymbol{\mu}_k \rangle$ follows directly from the normal equations:

$$\langle \mu_k[i] \rangle = \frac{b^2 \sum_j^J \langle \eta_k^{(j)}[i] \rangle + \langle \sigma_k^2[i] \rangle a[i]}{b^2 J + \langle \sigma_k^2[i] \rangle}. \tag{10}$$

We'll also need the variance of the variational distribution, $\mathcal{V}(\boldsymbol{\mu}_k)$. For simplicity we'll elide the topic index $k$ and the word index $i$, but keep in mind that we need to compute $\mathcal{V}(\mu_k[i])$ for all $k$ and $i$. The relevant terms for the variational bound are:

$$\begin{aligned}
L_{[\mathcal{V}(\mu)]} =& -\langle \log q(\mu) \rangle + \langle \log p(\mu | a, b^2) \rangle + \sum_j^J \langle \log p(\eta^{(j)} | \mu, \sigma^2) \rangle \\
=& \frac{1}{2} \log \mathcal{V}(\mu) - \frac{1}{2b^2} \langle (\mu - a)^2 \rangle - \sum_j^J \frac{1}{2} \langle (\eta^{(j)} - \mu)^2 \rangle \langle \sigma^{-2} \rangle \\
=& \frac{1}{2} \left( \log \mathcal{V}(\mu) - \mathcal{V}(\mu) b^{-2} - J \mathcal{V}(\mu) \langle \sigma^{-2} \rangle \right) \\
\partial L / \partial \mathcal{V}(\mu) =& \mathcal{V}(\mu)^{-1} - \left( b^{-2} + J \langle \sigma^{-2} \rangle \right) \\
\mathcal{V}(\mu) =& \left( b^{-2} + J \langle \sigma^{-2} \rangle \right)^{-1}.
\end{aligned} \tag{11}$$

Again, the solution is available in closed form, with a straightforward interpretation.

### 3.3 Updates: $\sigma^2$

Eliding the word index $i$ and the topic index $k$, the relevant parts of the variational bound for the expected variation $\langle \sigma_{ki}^2 \rangle$ are:

$$L_{[\langle \sigma^2 \rangle]} = \sum_j^J \langle \log p(\eta^{(j)} | \mu, \sigma^2) \rangle + \langle \log p(\sigma^2 | c, d) \rangle - \langle \log q(\sigma^2) \rangle. \tag{12}$$

The relevant elements of the first term of equation 12 are:

$$\begin{aligned}
L_{[\eta | \sigma^2]} =& \sum_j -\frac{1}{2} \langle \log 2\pi\sigma^2 \rangle - \langle (\eta_j - \mu)^2 / 2\sigma^2 \rangle \\
=& C - \frac{1}{2} \sum_j \left( \langle \log \sigma^2 \rangle + \langle (\eta_j - \mu)^2 \rangle \langle \sigma^{-2} \rangle \right) \\
=& C - \frac{J}{2} \langle \log \sigma^2 \rangle - S \langle \sigma^{-2} \rangle,
\end{aligned}$$

where $S = \sum_j \langle (\eta_j - \mu)^2 \rangle / 2$.

#### 3.3.1 Exponential version

There are several ways we can define the variational distribution $q(\sigma_{ki}^{-2})$, but the simplest is to use the exponential distribution, $q(\sigma_{ki}^{-2}) = \text{Exp}(\sigma_{ki}^{-2}; \gamma_{ki})$, where $\gamma_{ki}$ is a variational parameter. Then we have, $\langle \sigma_{ki}^{-2} \rangle = 1/\gamma_{ki}$ and $\langle \log \sigma_{ki}^2 \rangle = -\langle \log \sigma_{ki}^{-2} \rangle = -\psi(1) + \log \gamma_{ki}$.

Recall that the prior on $\sigma^2$ is a Gamma distribution, $\sigma^2 \sim \mathcal{G}(c, d)$. Then the relevant elements for the second two terms of equation 12 are

$$\begin{aligned}
L =& C - (c-1) \langle \log \sigma^{-2} \rangle - d \langle \sigma^{-2} \rangle - \log \gamma - 1 \\
=& C - c \log \gamma - d / \gamma,
\end{aligned}$$

so the overall likelihood bound is:

$$\begin{aligned}
L =& C - \frac{J}{2}\langle \log \sigma^2 \rangle - S\langle \sigma^{-2} \rangle - c \log \gamma - d/\gamma \\
=& C - (\frac{J}{2} + c) \log \gamma - (S + d)/\gamma \\
dL/d\gamma =& -(\frac{J}{2} + c)\gamma^{-1} + (S + d)\gamma^{-2} \\
=& -(\frac{J}{2} + c)\gamma + (S + d) \\
\gamma =& \frac{S + d}{\frac{J}{2} + c}
\end{aligned}$$

Thus we can update $\gamma$ in closed form. Remember that we are eliding the topic index $k$ and the term index $i$, so we need to compute each $\gamma_{ki}$.

### 3.3.2 Gamma version

In the Geographical Topic Model paper [2], the variational distribution over $\sigma^2_{ki}$ is a Gamma distribution, $q(\sigma^2_{ki}) = \mathcal{G}(\sigma^2_{ki}; \tilde{c}_{ki}, \tilde{d}_{ki})$, with variational parameters $\tilde{c}_{ki}$ and $\tilde{d}_{ki}$. Those updates cannot be computed in closed form, but the derivatives are given here. Eliding $k$ and $i$, we note the following properties of the Gamma distribution:

$$\begin{aligned}
\langle \sigma^2 \rangle &= \tilde{c}/\tilde{d} \\
\langle \sigma^{-2} \rangle &= \tilde{d}/(\tilde{c} - 1) \\
\langle \log \sigma^2 \rangle &= \Psi(\tilde{c}) - \log \tilde{d}.
\end{aligned}$$

The relevant elements from the remaining two terms of equation 12 are found in the PDF of the Gamma distribution:

$$\begin{aligned}
L_{[\sigma^2|\tilde{c},\tilde{d}]} &= c \log d - \log \Gamma(c) + (c-1)(\Psi(\tilde{c}) - \log \tilde{d}) - \tilde{c}d/\tilde{d} \\
&\quad - \tilde{c} \log \tilde{d} + \log \Gamma(\tilde{c}) - (\tilde{c} - 1)(\Psi(\tilde{c}) - \log \tilde{d}) + \tilde{c} && (13) \\
&= (c - \tilde{c})(\Psi(\tilde{c}) - \log \tilde{d}) + \tilde{c}(1 - d/\tilde{d} - \log \tilde{d}) + \log \Gamma(\tilde{c}) && (14) \\
&= (c - \tilde{c})\langle \log \sigma^2 \rangle + \tilde{c}(1 - d/\tilde{d} - \log \tilde{d}) + \log \Gamma(\tilde{c}) && (15)
\end{aligned}$$

We can now take derivatives. Note that $\Psi'$ equals the trigamma function:

$$\begin{aligned}
L_{[\tilde{c}]} =& (c - \tilde{c} - J/2)\langle \log \sigma^2 \rangle - S\langle \sigma^{-2} \rangle + (1 - d/\tilde{d} - \log \tilde{d})\tilde{c} + \log \Gamma(\tilde{c}) \\
dL/d\tilde{c} =& -\langle \log \sigma^2 \rangle + (c - \tilde{c} - J/2)\Psi'(\tilde{c}) + S\frac{\tilde{d}}{(1 - \tilde{c})^2} + 1 - d/\tilde{d} - \log \tilde{d} + \Psi(\tilde{c}) \\
=& (c - \tilde{c} - J/2)\Psi'(\tilde{c}) + S\frac{\tilde{d}}{(1 - \tilde{c})^2} + 1 - d/\tilde{d} && (16)
\end{aligned}$$

Now we'll do the other parameter, $\tilde{d}$:

$$\begin{aligned}
L_{[\tilde{d}]} =& -(c - 1) \log \tilde{d} - \tilde{c}d/\tilde{d} - \tilde{c} \log \tilde{d} + (\tilde{c} - 1) \log \tilde{d} + \frac{J}{2} \log \tilde{d} - \frac{S}{\tilde{c} - 1}\tilde{d} \\
=& (J/2 - c) \log \tilde{d} - \tilde{c}d/\tilde{d} - \frac{S}{\tilde{c} - 1}\tilde{d} \\
\frac{\partial L}{\partial \tilde{d}} =& (J/2 - c)/\tilde{d} + \tilde{c}d/\tilde{d}^2 - \frac{S}{\tilde{c} - 1} \\
0 =& (J/2 - c)\tilde{d} + \tilde{c}d - \frac{S}{\tilde{c} - 1}\tilde{d}^2
\end{aligned}$$

where we have multiplied through by $\tilde{d}^2$. We can now apply the quadratic formula to solve for $\tilde{d}$, obtaining

$$\tilde{d} = -\frac{\tilde{c}-1}{2S}\left(c - J/2 - \sqrt{(J/2-c)^2 + 4d\frac{\tilde{c}}{\tilde{c}-1}S}\right) \tag{17}$$

In practice, we iterate between gradient-based updates for $\tilde{c}$ and closed-form updates for $\tilde{d}$.

### 3.4 Update: $\theta$

We'll need the expectation $\langle \log \theta \rangle$. The update is identical to LDA:

$$\begin{aligned}\langle \log \theta_k \rangle &= \Psi(\alpha_\theta + \sum_n^N \phi_{nk}) - \Psi(K\alpha_\theta + \sum_n^N \sum_k^K \phi_{nk}) \\ &= \Psi(\alpha_\theta + \sum_n^N \phi_{nk}) - \Psi(K(N + \alpha_\theta))\end{aligned} \tag{18}$$

where $\Psi$ is the digamma function, $K$ is the number of topics, and the document index $d$ is implicit.

### 3.5 Update: $\vartheta$

The update for $\vartheta$ is similar, but we sum over all documents rather than indices in a single document:

$$\langle \log \vartheta_j \rangle = \Psi(\alpha_\vartheta + \sum_d^D \omega_{dj}) - \Psi(J(D + \alpha_\vartheta)), \tag{19}$$

where $J$ is the number of facets and $\omega_{dj}$ is a variational parameter that characterizes a discrete distribution over the set $1, 2, \dots J$ (defined above).

### 3.6 Updates: $\omega$ and $\phi$

The parameter $\phi$ parametrizes the topic indicator $z$; its update is identical to standard LDA, except that you have to marginalize over settings of $v$:

$$\phi_{nk} \propto \exp\{\langle \log \theta_{dk} \rangle + \sum_j^J \omega_{dj} \langle \log \beta_k^{(j)}[w_n] \rangle\}, \tag{20}$$

where $\langle \log \beta_k^{(j)}[i] \rangle$ is defined in Equation 2. The update for $\omega$ is similar:

$$\omega_{dj} \propto \exp\{\langle \log \vartheta_j \rangle + \sum_n^{N_d} \sum_k^K \phi_{nk} \langle \log \beta_k^{(j)}[w_n] \rangle\} p(y_d|\rho_j), \tag{21}$$

where we append a probability distribution over the metadata $y_d$ at the end.

## 4 Algorithm

The basic idea here is variational expectation-maximization: iterate over the variational parameters until convergence, then update the remaining (non-variational) parameters (mostly priors, but also $\rho$).

Initialize all $\phi$ from Latent Dirichlet Allocation
Initialize all $\omega$ from a mixture model on $y$ with $J$ components
Initialize all $\langle \eta_k^{(j)} \rangle, \langle \mu_k \rangle, \langle \theta_d \rangle, \langle \vartheta \rangle$ empirically (by counting)
Initialize variance parameters sensibly
**while** not done with EM **do**
  **while** not done with E-step **do**
    **for** Each topic $k$ **do**

for Each facet $j$ do
Use LBFGS to update $\langle \eta_k^{(j)} \rangle$ from the Equations 5 (objective) and 7 (gradient).
Use LBFGS to update $\mathcal{V}(\eta_k^{(j)})$ from the Equation 8 (objective) and 9 (gradient).
Update $\zeta_k^{(j)}$ from Equation 4.
end for
Use the closed-form equations 10 and 11 to update $\langle \mu_k \rangle$ and $\mathcal{V}(\mu_k)$.
Use LBFGS to update $\tilde{\mathbf{c}}_k$ from the Equations 15 (objective) and 16 (gradient).
Use the closed-form equation 17 to update $\tilde{\mathbf{d}}_k$,
end for
for Each document $d$ do
while not converged do
for Each token $n < N_d$ do
Update $\phi_n$ from Equation 20
end for
Update $\omega_d$ from Equation 21
Update $\langle \theta_d \rangle$ from Equation 18
end while
end for
Update $\langle \vartheta \rangle$ from Equation 19
end while
Update the metadata parameters $\rho$.
Optionally: update the Dirichlet priors $\alpha_\theta$ and $\alpha_\vartheta$ (see [3]).
Optionally: update the Normal priors $\mathbf{a}$ and $b$ by taking point estimates of all the $\mu_j$ and fitting a Gaussian.
Optionally: update the Gamma priors $c$ and $d$ by taking point estimates of all the $\sigma_j$.[2]
end while

## 5 Simplified Approximation

We now consider a simplified variational approximation, in which $\langle \sigma_k^2[i] \rangle$ and $\mathcal{V}(\mu_k)$ are identical for all terms $i$. Thus, the variational gamma parameters $\tilde{c}$ and $\tilde{d}$ are now scalars rather than vectors. We simply modify the derivation in Section 3.3, starting with equation 12:

$$L_{[\tilde{c}_k, \tilde{d}_k]} = W \left( \langle \log p(\sigma_k^2 | c, d) \rangle - \langle \log q(\sigma_k^2) \rangle \right) + \sum_j^J \sum_i^W \langle \log p(\eta_k^{(j)}[i] | \mu_k[i], \sigma_k^2[i]) \rangle$$

Essentially, the gradient and objective work out to be the same as in Equations 15 and 16, except that we use the average $\frac{1}{W} \sum_j^J \sum_i^W \langle (\eta^{(j)}[i] - \mu[i])^2 \rangle$. Similarly, we obtain a closed-form for $\tilde{d}$ in which the $S$ term is computed in average over all terms.

## References


[1] D. M. Blei and J. Lafferty. Correlated topic models. In *NIPS*, 2006.

[2] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing. A latent variable model of geographic lexical variation. In *Proceedings of EMNLP*, 2010.

[3] T. P. Minka. Estimating a Dirichlet distribution. Technical report, Massachusetts Institute of Technology, 2003.

[4] W. D. Penny. Variational Bayes for $d$-dimensional Gaussian mixture models. Technical report, University College London, 2001.


[2] Unfortunately there's no closed-form for this. Consider playing sneaky games with these priors to control how much variance is in the facets.