Variational EM Algorithms for Correlated Topic Models

Mohammad Emtiyaz Khan and Guillaume Bouchard

September 14, 2009

Abstract

In this note, we derive a variational EM algorithm for correlated topic models. This algorithm was proposed in Blei and Lafferty's original paper [BL06] and is based on a simple bound on logarithm. Because of the form of this bound, E-step update are not available in closed form and need to be solved with a coordinate ascent algorithm.

1 Correlated Topic Model

Consider D number of documents with W words each. These words belong to a fixed vocabulary of size V. Let us say that there are T topics. The correlated topic model is a generative model for documents and is given as follows,

$$p(\boldsymbol{\eta}_d|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{\eta}_d|\boldsymbol{\mu},\boldsymbol{\Sigma}) \tag{1}$$

$$p(z_{n,d}|\boldsymbol{\eta}_d) = \text{Mult}(f(\boldsymbol{\eta}_d)) \tag{2}$$

$$p(w_{n,d}|z_{n,d},\boldsymbol{\beta}_{1:T}) = \text{Mult}(\boldsymbol{\beta}_{z_{n,d}}) \tag{3}$$

where $f(\mathbf{a}) = e^{\mathbf{a}}/\sum_j e^{a_j}$. Basically we sample probability vector for each topic using a logistic-normal distribution. Next using this probability vector we sample a topic for each word. Depending on the topic, words are then generated from a fixed probability distribution. We are interested in finding similarity between the topics and a clustering of words based on the topics. We use the following notation in the following: we denote vectors with small bold letters (e.g. $\mathbf{a}$) and matrices with capital bold letters (e.g. $\mathbf{A}$). For scalars we use both small/capital plain faced letters. We use $t = 1,\ldots,T$ as an index over topics, $v = 1,\ldots,V$ as an index over words in the vocabulary, $d = 1,\ldots,D$ as an index over documents, and $n = 1,\ldots,W_d$ as an index over words in d^{th} document.

The joint-distribution is the following,

$$\prod_{d=1}^{D} p(\mathbf{w}_d,\mathbf{z}_d,\boldsymbol{\eta}_d|\boldsymbol{\mu},\boldsymbol{\Sigma},\mathbf{B}) = \prod_{d=1}^{D}\left[\prod_{n=1}^{W_d} p(w_{n,d}|z_{n,d},\mathbf{B})p(z_{n,d}|\boldsymbol{\eta}_d)\right] p(\boldsymbol{\eta}_d|\boldsymbol{\mu},\boldsymbol{\Sigma}) \tag{4}$$

where $\mathbf{B} = [\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\ldots,\boldsymbol{\beta}_T]$. Our goal is to infer the posterior distribution over $\boldsymbol{\eta}_{1:D}$ given the data. Also we wish to estimate the parameters $\boldsymbol{\Theta} = \{\boldsymbol{\mu},\boldsymbol{\Sigma},\mathbf{B}\}$. We will take an empirical Bayes approach to estimate the parameters i.e. we maximize the marginal likelihood with respect to the parameters. The marginal likelihood of the data given parameters $\boldsymbol{\Theta}$ can be found as follows,

$$p(\mathbf{w}_{1:D}|\boldsymbol{\Theta}) = \prod_d \int_{\boldsymbol{\eta}_d} p(\mathbf{w}_d|\boldsymbol{\eta}_d,\mathbf{B})p(\boldsymbol{\eta}_d|\boldsymbol{\mu},\boldsymbol{\Sigma})d\boldsymbol{\eta}_d \tag{5}$$

Unfortunately this integral is intractable and hence we will resort to the variational methods for optimization. We first find a lower bound for which the integral is tractable, and then maximize the lower bound with respect to the parameters.

2 A lower bound for the marginal likelihood

To make the integral tractable, we introduce an auxiliary distribution $q(\boldsymbol{\eta}_{1:D}) = \prod_d q_d(\boldsymbol{\eta}_d)$ and use Jensen's inequality to find a lower bound. We fix q_d to be a normal distribution with mean $\mathbf{m}_d$ and covariance $\mathbf{V}_d$. Taking log of the marginal likelihood,

$$\log p(\mathbf{w}_{1:D}|\boldsymbol{\Theta}) = \sum_d \log \int_{\boldsymbol{\eta}_d} p(\mathbf{w}_d|\boldsymbol{\eta}_d, \mathbf{B}) p(\boldsymbol{\eta}_d|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\boldsymbol{\eta}_d \tag{6}$$

$$= \sum_d \log \int_{\boldsymbol{\eta}_d} \frac{p(\mathbf{w}_d|\boldsymbol{\eta}_d, \mathbf{B}) p(\boldsymbol{\eta}_d|\boldsymbol{\mu}, \boldsymbol{\Sigma})}{q(\boldsymbol{\eta}_d|\mathbf{m}_d, \mathbf{V}_d)} q(\boldsymbol{\eta}_d|\mathbf{m}_d, \mathbf{v}_d) d\boldsymbol{\eta}_d \tag{7}$$

$$\geq \sum_d \langle \log p(\mathbf{w}_d|\boldsymbol{\eta}_d, \mathbf{B}) \rangle_{q_d} + \langle \log p(\boldsymbol{\eta}_d|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \rangle_{q_d} + \mathbb{H}(q_d) \tag{8}$$

$$= \sum_d \langle \log p(\mathbf{w}_d|\boldsymbol{\eta}_d, \mathbf{B}) \rangle_{q_d} - KL\left[q_d(\boldsymbol{\eta}_d|\mathbf{m}_d, \mathbf{V}_d)||p(\boldsymbol{\eta}_d|\boldsymbol{\mu}, \boldsymbol{\Sigma})\right] \tag{9}$$

Here $\langle \cdot \rangle_q$ denotes the expectation with respect to the distribution q, $\mathbb{H}(\cdot)$ denotes the entropy of a distribution, and $KL(\cdot)$ denotes the KL divergence. Last term is given as follows (see Wikipedia),

$$-KL\left(q_d||p\right) = -\frac{1}{2}\log|\boldsymbol{\Sigma}| + \frac{1}{2}\log|\mathbf{V}_d| - \frac{1}{2}\left\{\mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{V}_d\right) + (\mathbf{m}_d - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{m}_d - \boldsymbol{\mu})\right\} + \text{cnst} \tag{10}$$

It is not possible to get a closed form expression for the first term, but we can find another lower bound to it.

2.1 Lower bound for the first term

To make notation simpler, we drop the subscript d. We can integrate out $\mathbf{z}$ to get the following (see Appendix A),

$$p(\mathbf{w}|\boldsymbol{\eta}, \mathbf{B}) = \frac{\prod_{v=1}^{V}\left(\sum_{t=1}^{T}\beta_{v,t}e^{\eta_t}\right)^{c_v}}{\left(\sum_{t=1}^{T}e^{\eta_t}\right)^{W}} \tag{11}$$

Here c_v is the count of v^{th} word. Taking log,

$$\log p(\mathbf{w}|\boldsymbol{\eta}, \mathbf{B}) = \sum_v c_v \log \sum_t \beta_{v,t}e^{\eta_t} - W \log \sum_t e^{\eta_t} \tag{12}$$

Expectation of these terms is hard to compute, so we find a tractable lower bound. The first term can be lower bounded using Jensen's inequality by introducing an auxiliary distribution $\mathbf{s}_v$:

$$\log \sum_t \beta_{v,t}e^{\eta_t} \geq \sum_t s_{v,t}(\eta_t + \log\beta_{v,t}) - \sum_t s_{v,t}\log s_{v,t} \tag{13}$$

For the second term we note that,

$$\log x \leq \xi^{-1}x + \log\xi - 1 \tag{14}$$

This can be derived using the concave-conjugate of log function (see [BV04], Chapter 3). Using this we get the following lower bound for the second term,

$$-\log \sum_t e^{\eta_t} \geq -\xi^{-1}\sum_t e^{\eta_t} - \log\xi + 1 \tag{15}$$

Using these two lower bounds we get the following expression for the expectation,

$$\langle \log p(\mathbf{w}|\boldsymbol{\eta}, \mathbf{B}) \rangle \geq \sum_{v,t} c_v s_{v,t}\left(\log\beta_{v,t} + \langle\eta_t\rangle - \log s_{v,t}\right) - W\xi^{-1}\sum_t \langle e^{\eta_t}\rangle_q - W\log\xi + W \tag{16}$$

The second expectation is given as follows (see derivation in Appendix B),

$$\langle e^{\eta_t}\rangle_q = e^{\frac{1}{2}V_{tt} + m_t} \tag{17}$$

This gives us a lower bound for the first term,

$$\langle \log p(\mathbf{w}|\boldsymbol{\eta}, \mathbf{B}) \rangle \geq \sum_{v,t} c_v s_{v,t} \left(\log \beta_{v,t} + m_t - \log s_{v,t} \right) - W \xi^{-1} \sum_t e^{\frac{1}{2} V_{tt} + m_t} - W \log \xi + W \tag{18}$$

Including the document subscripts,

$$\langle \log p(\mathbf{w}_d|\boldsymbol{\eta}_d, \mathbf{B}) \rangle \geq \sum_{v,t,d} c_{v,d} s_{v,t,d} \left(\log \beta_{v,t} + m_{t,d} - \log s_{v,t,d} \right) - W_d \xi_d^{-1} \sum_t e^{\frac{1}{2} V_{tt,d} + m_{t,d}} - W \log \xi_d + W \tag{19}$$

2.2 Final lower bound

We get the following lower bound to the marginal likelihood,

$$\begin{aligned} \mathcal{L}_q(\boldsymbol{\Theta}) = & \sum_{v,t,d} c_{v,d} s_{v,t,d} \left(\log \beta_{v,t} + m_{t,d} - \log s_{v,t,d} \right) - \sum_d W_d \xi_d^{-1} \sum_t e^{\frac{1}{2} V_{tt,d} + m_{t,d}} - \sum_d W_d \log \xi_d + \sum_d W_d \\ & - \frac{D}{2} \log |\boldsymbol{\Sigma}| + \frac{1}{2} \sum_d \log |\mathbf{V}_d| - \frac{1}{2} \sum_d \left\{ \mathrm{Tr} \left(\boldsymbol{\Sigma}^{-1} \mathbf{V}_d \right) + (\mathbf{m}_d - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{m}_d - \boldsymbol{\mu}) \right\} \end{aligned} \tag{20}$$

subject to the constraints $\sum_v \beta_{v,t} = 1, \forall t$ and $\sum_t s_{v,t,d} = 1$ for all t and d.

3 A variational EM algorithm

We now use EM algorithm to maximize the marginal likelihood. In E-step we maximize with respect to the parameters of q_d, and in M-step we maximize with respect to $\boldsymbol{\Theta}$.

3.1 Optimizing with respect to the parameters (M-step)

Differentiating the lower bound with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}^{-1}$,

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}} = - \sum_d \boldsymbol{\Sigma}^{-1} (\mathbf{m}_d - \boldsymbol{\mu}) \tag{21}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\Sigma}^{-1}} = \frac{D}{2} \boldsymbol{\Sigma} - \frac{1}{2} \sum_d \mathbf{V}_d - \frac{1}{2} \sum_d (\mathbf{m}_d - \boldsymbol{\mu})(\mathbf{m}_d - \boldsymbol{\mu})^T \tag{22}$$

and setting it to zero we get the updates,

$$\boldsymbol{\mu} = \frac{1}{D} \sum_d \mathbf{m}_d \tag{23}$$

$$\boldsymbol{\Sigma} = \frac{1}{D} \sum_d \mathbf{V}_d + (\mathbf{m}_d - \boldsymbol{\mu})(\mathbf{m}_d - \boldsymbol{\mu})^T \tag{24}$$

Next we optimize over $s_{v,t}, \beta_{v,t}$. To include the constraint we use the Lagrangian multiplier. The objective function is given as follows,

$$\sum_{v,t,d} c_{v,d} s_{v,t,d} \left(\log \beta_{v,t} + m_{t,d} - \log s_{v,t,d} \right) + \gamma_1 (1 - \sum_v \beta_{v,t}) + \gamma_2 (1 - \sum_v s_{v,t}) \tag{25}$$

where γ is the Lagrangian multiplier. Differentiating with respect to $\beta_{v,t}, \gamma_1, s_{v,t,d}, \gamma_2$ we get the following,

$$\beta_{v,t}^{-1} \sum_d c_{v,d} s_{v,t,d} - \gamma_1 = 0 \tag{26}$$

$$1 - \sum_v \beta_{v,t} = 0 \tag{27}$$

$$c_{v,d} \left(\log \beta_{v,t} + m_{t,d} \right) - c_{v,d} \log s_{v,t,d} - c_{v,d} - \gamma_2 = 0 \tag{28}$$

$$1 - \sum_v s_{v,t,d} = 0 \tag{29}$$

Solving these equations we get the following updates,

$$\beta_{v,t} \propto \sum_d c_{v,d} s_{v,t,d} \tag{30}$$

$$s_{v,t,d} \propto \beta_{v,t} e^{m_{t,d}} \tag{31}$$

with normalization to 1 over v for first quantity and over t for second quantity.

3.2 Optimizing with respect to the variational parameters (E-step)

First we optimize with respect to $\xi_{1:D}$. Derivative is given as follows,

$$\frac{\partial \mathcal{L}}{\partial \xi_d} = W_d \xi_d^{-2} \sum_t e^{\frac{1}{2} V_{tt,d} + m_{t,d}} - W_d \xi_d^{-1} \tag{32}$$

which gives us the following update for ξ_d,

$$\xi_d = \sum_t e^{\frac{1}{2} V_{tt,d} + m_{t,d}} \tag{33}$$

Substituting this we get the following objective function for $\mathbf{m}_d, \mathbf{V}_d$,

$$\mathcal{L}'(\boldsymbol{\mu}_d, \mathbf{V}_d) = \mathbf{c}_d^T \mathbf{S}_d \mathbf{m}_d - W_d \log \sum_t e^{\frac{1}{2} V_{tt,d} + m_{t,d}} + \frac{1}{2} \log |\mathbf{V}_d| - \frac{1}{2} \left\{ \mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1} \mathbf{V}_d\right) + (\mathbf{m}_d - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{m}_d - \boldsymbol{\mu}) \right\} \tag{34}$$

where $\mathbf{S}_d$ is $V \times T$ matrix containing $s_{v,t,d}$ for a fixed d. Closed form updates are not possible for $\mathbf{m}_d$ and $\mathbf{V}_d$ because of the presence of the exponential term. However we can still use a gradient based methods using the following gradients,

$$\frac{\partial \mathcal{L}}{\partial \mathbf{m}_d} = \mathbf{S}_d \mathbf{c}_d - \boldsymbol{\Sigma}^{-1} (\mathbf{m}_d - \boldsymbol{\mu}) - \frac{W_d}{\xi_d} e^{\mathrm{diag}\left(\frac{1}{2} V_{tt,d} + m_{t,d}\right)_{1:T}} \tag{35}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{V}_d} = -\frac{1}{2} \boldsymbol{\Sigma}^{-1} + \frac{1}{2} \mathbf{V}_d^{-1} - \frac{W_d}{2\xi_d} e^{\mathrm{diag}\left(\frac{1}{2} V_{tt,d} + m_{t,d}\right)_{1:T}} \tag{36}$$

Here $\mathrm{diag}(a_t)_{1:T}$ means a diagonal matrix with diagonal entries a_t. The pseudo-code is summarized in Algorithm 1 In the case of diagonal V_d, the above equations reduce to Eq. (16-17) in [BL06]. Note that we have to compute both V_d and inverse of V_d for the E-step. The case where we fix V_d to be diagonal will be much faster as we don't have to compute inverse of V_d at every iteration.

A Integrating out z

$$p(\mathbf{w}|\boldsymbol{\eta}, \mathbf{B}) = \sum_{\mathbf{z}} p(\mathbf{w}, \mathbf{z}|\boldsymbol{\eta}, \mathbf{B}) \tag{43}$$

$$= \sum_{\mathbf{z}} \prod_{n=1}^{W} p(w_n|z_n, \mathbf{B}) p(z_n|\boldsymbol{\eta}) \tag{44}$$

$$= \prod_{v=1}^{V} \left[\sum_{\mathbf{z}} p(w_n|z_n, \mathbf{B}) p(z_n|\boldsymbol{\eta}) \right]^{c_v} \tag{45}$$

$$= \prod_{v=1}^{V} \left[\sum_{t=1}^{T} \beta_{v,t} \frac{e^{\eta_t}}{\sum_{t'} e^{\eta_{t'}}} \right]^{c_v} \tag{46}$$

$$= \frac{\prod_{v=1}^{V} \left(\sum_{t=1}^{T} \beta_{v,t} e^{\eta_t} \right)^{c_v}}{\left(\sum_{t=1}^{T} e^{\eta_t} \right)^{W}} \tag{47}$$

Algorithm 1 (Variational EM, Blei & Lafferty)

1: Initialize $\boldsymbol{\mu}^{(0)} = \mathbf{m}_d = 0, \boldsymbol{\Sigma}^{(0)} = \mathbf{I}_T, \mathbf{V}_d = \mathbf{I}_T$.
2: Iterate between E and M step until convergence.
3: E-Step: For $d = 1, 2, \ldots, D$, solve the following equations,

$$\mathbf{S}_d \mathbf{c}_d - \boldsymbol{\Sigma}^{-1}(\mathbf{m}_d - \boldsymbol{\mu}) - \frac{W_d}{\xi_d} e^{\text{diag}\left(\frac{1}{2}V_{tt,d} + m_{t,d}\right)_{1:T}} = 0 \tag{37}$$

$$-\frac{1}{2}\boldsymbol{\Sigma}^{-1} + \frac{1}{2}\mathbf{V}_d^{-1} - \frac{W_d}{2\xi_d} e^{\text{diag}\left(\frac{1}{2}V_{tt,d} + m_{t,d}\right)_{1:T}} = 0 \tag{38}$$

where $\xi_d = \sum_t e^{\frac{1}{2}V_{tt,d} + m_{t,d}}$, $\mathbf{S}_d = [\mathbf{s}_1, \ldots, \mathbf{s}_V]$ and $\mathbf{s}_v$ is a vector of $s_{v,t} \log \beta_{v,t}$ for all t.
4: M-Step

$$\boldsymbol{\mu} = \frac{1}{D}\sum_d \mathbf{m}_d \tag{39}$$

$$\boldsymbol{\Sigma} = \frac{1}{D}\sum_d \mathbf{V}_d + (\mathbf{m}_d - \boldsymbol{\mu})(\mathbf{m}_d - \boldsymbol{\mu})^T \tag{40}$$

$$\beta_{v,t} \propto \sum_d c_{v,d} s_{v,t,d} \tag{41}$$

$$s_{v,t,d} \propto \beta_{v,t} e^{m_{t,d}} \tag{42}$$

with normalization to 1 over v for first quantity and over t for second quantity.

B Expression for $\left\langle e^{\eta_t} \right\rangle_q$

$$\langle e^{\eta_t} \rangle_q = \int e^{\eta_t} q(\eta_t | m_t, V_{tt}) d\eta_t \tag{48}$$

$$= \int \frac{1}{\sqrt{2\pi V_{tt}}} e^{-\frac{1}{2V_{tt}}(\eta_t - m_t)^2 + \eta_t} d\eta_t \tag{49}$$

We now complete squares,

$$(\eta_t - m_t)^2 - 2V_{tt}\eta_t \tag{50}$$

$$= \eta_t^2 + m_t^2 - 2\eta_t m_t - 2V_{tt}\eta_t \tag{51}$$

$$= \eta_t^2 + m_t^2 - 2\eta_t(m_t + V_{tt}) \tag{52}$$

$$= \eta_t^2 + (m_t + V_{tt})^2 - 2\eta_t(m_t + V_{tt}) + m_t^2 - (m_t + V_{tt})^2 \tag{53}$$

$$= (\eta_t - m_t - V_{tt})^2 - V_{tt}^2 - 2m_t V_{tt} \tag{54}$$

Using this we get the following expression for the expectation,

$$\langle e^{\eta_t} \rangle_q = e^{\frac{1}{2}V_{tt} + m_t} \tag{55}$$

References

[BL06] D. Blei and J. Lafferty. Correlated topic models. *Advances in neural information processing systems*, 18:147, 2006.

[BV04] S.P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.