A note on a Variational Bayes derivation of full Bayesian Latent Dirichlet Allocation

Daichi Mochihashi
ATR Spoken Language Translation Research Laboratories,
Kyoto, Japan
daichi.mochihashi@atr.jp

2004.10.20

Abstract

This note provides derivations and formulae for a full Bayesian treatment of Latent Dirichlet Allocation [1], which are mentioned but omitted in its full paper [2]. Further, we extend it a little to accommodate a non-uniform lexical prior. By using this treatment, we can get appropriately smoothed estimates of class unigrams $\boldsymbol{\beta} = p(w_n|z_k)$.

First, we assume that a corpus $\mathbf{w}$ consists of:

$$\mathbf{w} = \mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_D \tag{1}$$

$$\mathbf{w}_d = w_1, w_2, \ldots, w_{N_d}. \tag{2}$$

Then, we write a log likelihood for $\mathbf{w}$ given α, η, and approximate it as shown below.

$$\begin{aligned}
\log p(\mathbf{w}|\alpha,\eta) &= \log \int p(\mathbf{w},\beta|\alpha,\eta)d\beta & (3)\\
&= \log \int \frac{p(\mathbf{w},\beta|\alpha,\eta)}{q(\beta)} q(\beta)d\beta & (4)\\
&\geq \int q(\beta|\lambda) \log \frac{p(\mathbf{w},\beta|\alpha,\eta)}{q(\beta|\lambda)} d\beta & (5)\\
&= -\int q(\beta|\lambda)\log q(\beta|\lambda)d\beta + \int q(\beta|\lambda)\log p(\mathbf{w}|\alpha,\beta)p(\beta|\eta)d\beta & (6)\\
&= -\int q(\beta|\lambda)\log q(\beta|\lambda)d\beta + \int q(\beta|\lambda)\log p(\mathbf{w}|\alpha,\beta)d\beta + \int q(\beta|\lambda)\log p(\beta|\eta)d\beta & (7)\\
&\equiv L. & (8)
\end{aligned}$$

Here, $p(\mathbf{w}|\alpha,\beta)$ is a standard model of latent Dirichlet allocation, and decomposed as follows.

$$\begin{aligned}
\log p(\mathbf{w}|\alpha,\beta) &= \sum_{d=1}^{D} \log \int \sum_z p(\mathbf{w}_d, z, \theta|\alpha,\beta)d\theta & (9)\\
&= \sum_{d=1}^{D} \log \int \sum_z \frac{p(\mathbf{w}_d, z, \theta|\alpha,\beta)}{q(z,\theta)} q(z,\theta)d\theta & (10)\\
&\geq \sum_{d=1}^{D} \int \sum_z q(z,\theta) \log \frac{p(\mathbf{w}_d, z, \theta|\alpha,\beta)}{q(z,\theta)} d\theta & (11)\\
&= \sum_{d=1}^{D} \int \sum_z q(\theta)q(z) \left[\log p(\theta|\alpha) + \sum_n \log p(z_n|\theta) + \sum_n \log p(w_n|z_n,\beta) \right] d\theta \\
&\quad - \int \sum_z q(\theta)q(z) \log q(\theta)q(z) d\theta. & (12)
\end{aligned}$$

1 Solution for $q(\boldsymbol{\beta}|\boldsymbol{\lambda})$

Therefore, we can collect terms that contain $q(\boldsymbol{\beta}|\boldsymbol{\lambda})$ from L, and apply Lagrangians:

$$\begin{aligned} J(\beta) = &- \int \prod_k q(\beta_k|\lambda_k) \sum_k \log q(\beta_k|\lambda_k) d\boldsymbol{\beta} \\ &+ \sum_d \int \prod_k q(\beta_k|\lambda_k) \left[\sum_z q(z) \sum_n \log p(w_n|z_n, \beta) \right] d\boldsymbol{\beta} \\ &+ \int \prod_k q(\beta_k|\lambda_k) \sum_k \log p(\beta_k|\eta) d\boldsymbol{\beta} \\ &+ \sum_k \mu_k \left(\int q(\beta_k|\lambda_k) d\beta_k - 1 \right) \end{aligned} \tag{13}$$

$$\begin{aligned} \therefore \quad \frac{\partial J(\beta)}{\partial \beta_k} &= - \int \frac{\partial}{\partial \beta_k} \prod_k q(\beta_k|\lambda_k) \log q(\beta_k|\lambda_k) d\beta \\ &\quad + \mu_k \\ &\quad + \log p(\beta_k|\eta) \\ &\quad + \sum_d \sum_z q(z) \sum_n \log p(w_n|z_n, \beta) && (14) \\ &= - \log q(\beta_k|\lambda_k) + \mu_k + \log p(\beta_k|\eta) \\ &\quad + \sum_{d=1}^{D} \sum_{n=1}^{N} \sum_{v=1}^{V} \phi_{dnk} w_{dn}^v \log \beta_{kv} && (15) \\ &= 0. && (16) \end{aligned}$$

Then,

$$\log q(\beta_k|\lambda_k) = \mu_k + \log p(\beta_k|\eta) + \sum_{d=1}^{D} \sum_{n=1}^{N} \sum_{v=1}^{V} \phi_{dnk} w_{dn}^v \log \beta_{kv} \tag{17}$$

$$\begin{aligned} \therefore \quad q(\beta_k|\lambda_k) &\propto p(\beta_k|\eta) \exp(\sum_{d=1}^{D} \sum_{n=1}^{N} \sum_{v=1}^{V} \phi_{dnk} w_{dn}^v \log \beta_{kv}) && (18) \\ &= p(\beta_k|\eta) \cdot \beta_{kv}^{\sum_{d=1}^{D} \sum_{n=1}^{N} \sum_{v=1}^{V} \phi_{dnk} w_{dn}^v} && (19) \\ &\propto \mathrm{Dir}(\beta_k|\eta + \sum_{d=1}^{D} \sum_{n=1}^{N} \sum_{v=1}^{V} \phi_{dnk} w_{dn}^v) && (20) \end{aligned}$$

$$\iff \lambda_k = \eta + \sum_{d=1}^{D} \sum_{n=1}^{N} \sum_{v=1}^{V} \phi_{dnk} w_{dn}^v. \quad \blacksquare \tag{21}$$

2 Newton-Raphson iteration for η

Here, we derive a Newton-Raphson iteration for η, a hyperparameter that works as a smoothing term for $\boldsymbol{\beta}$.

First, we extract a term that contains η, from L:

$$L_\eta = \int q(\beta|\lambda) \log p(\beta|\eta) d\beta \tag{22}$$

$$= \int \prod_k q(\beta_k|\lambda_k) \sum_k \log p(\beta_k|\eta) d\beta \tag{23}$$

$$= \sum_k \int q(\beta_k|\lambda_k) \log \left(\frac{\Gamma(V\eta)}{\Gamma(\eta)^V} \prod_v \beta_{kv}^{\eta-1} \right) d\beta_k \tag{24}$$

$$= \sum_k \left[\log \Gamma(V\eta) - V \log \Gamma(\eta) + \int \mathrm{Dir}(\beta_k|\lambda_k) \sum_v (\eta - 1) \log \beta_{kv} d\beta_k \right] \tag{25}$$

$$= K(\log \Gamma(V\eta) - V \log \Gamma(\eta)) + \sum_k \sum_v (\eta - 1) \int \mathrm{Dir}(\beta_k|\lambda_k) \log \beta_{kv} d\beta_k \tag{26}$$

$$= K(\log \Gamma(V\eta) - V \log \Gamma(\eta)) + (\eta - 1) \sum_k \sum_v \{\Psi(\lambda_{kv}) - \Psi(\sum_v \lambda_{kv})\} \tag{27}$$

We denote $\sum_k \sum_v \Psi(\lambda_{kv}) - \Psi(\sum_v \lambda_{kv})$ as P. Then,

$$\frac{\partial L_\eta}{\partial \eta} = K(V\Psi(V\eta) - V\Psi(\eta)) + P. \tag{28}$$

Here, we can derive a Newton-Raphson update for scalar hyperparameter η.

$$\therefore\ \eta^{\text{new}} = \eta - \frac{K(\log \Gamma(V\eta) - V \log \Gamma(\eta)) + (\eta - 1) \cdot P}{K(V\Psi(V\eta) - V\Psi(\eta)) + P} \tag{29}$$

$$= \eta - \frac{\log \Gamma(V\eta) - V \log \Gamma(\eta) + (\eta - 1) \cdot P/K}{V\Psi(V\eta) - V\Psi(\eta) + P/K} \tag{30}$$

$$= \eta - \frac{\log \Gamma(V\eta)/V - \log \Gamma(\eta) + (\eta - 1) \cdot P/(KV)}{\Psi(V\eta) - \Psi(\eta) + P/(KV)}. \quad \blacksquare \tag{31}$$

3 Newton-Raphson iteration for $\boldsymbol{\eta}$ (extended)

Equation (31) is the update formula for scalar η, that is mentioned but omitted in [2].

However, this means we are doing a generalized Laplace smoothing (Lidstone's law) [3]; that is, it gives a *uniform* smoothing term to class unigrams β_{kv}, no matter what word v is.

Apparently, this is not an adequate approach to smoothing. Instead, when we introduce a vector hyperparameter $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_V)$ and assume a prior distribution $p(\boldsymbol{\beta}|\boldsymbol{\eta}) = \mathrm{Dir}(\boldsymbol{\beta}|\boldsymbol{\eta})$, we get an another Bayesian estimate of $\boldsymbol{\beta}$, appropriately smoothed word by word.

Fortunately, inferring $\boldsymbol{\eta}$ can be done by a linear-time Newton-Raphson iteration, as shown below.

$$
\begin{aligned}
L_\eta &= \int q(\beta|\lambda) \log p(\beta|\boldsymbol{\eta}) d\beta && (32)\\
&= \int \prod_k q(\beta_k|\lambda_k) \sum_k \log p(\beta_k|\boldsymbol{\eta}) d\beta && (33)\\
&= \sum_k \int q(\beta_k|\lambda_k) \log \left(\frac{\Gamma(\sum_v \eta_v)}{\prod_v \Gamma(\eta_v)} \prod_v \beta_{kv}^{\eta_v - 1} \right) d\beta_k && (34)\\
&= \sum_k \left[\log \Gamma(\sum_v \eta_v) - \sum_v \log \Gamma(\eta_v) + \int \mathrm{Dir}(\beta_k|\lambda_k) \sum_v (\eta_v - 1) \log \beta_{kv} d\beta_k \right] && (35)\\
&= K \left(\log \Gamma(\sum_v \eta_v) - \sum_v \log \Gamma(\eta_v) \right) + \sum_k \sum_v (\eta_v - 1) \left(\Psi(\lambda_{kv}) - \Psi(\sum_v \lambda_{kv}) \right) && (36)\\
&= K \left(\log \Gamma(\sum_v \eta_v) - \sum_v \log \Gamma(\eta_v) \right) + \sum_v (\eta_v - 1) \sum_k \left(\Psi(\lambda_{kv}) - \Psi(\sum_v \lambda_{kv}) \right) && (37)
\end{aligned}
$$

We denote $\sum_k \Psi(\lambda_{kv}) - \Psi(\sum_v \lambda_{kv})$ as P_v.

Then,

$$
\frac{\partial L_\eta}{\partial \eta_i} = K \left(\Psi(\sum_i \eta_i) - \Psi(\eta_i) \right) + P_i \quad \equiv g(\eta_i) \tag{38}
$$

$$
\frac{\partial^2 L_\eta}{\partial \eta_i \partial \eta_j} = \begin{cases} K\Psi'(\sum_i \eta_i) - K\Psi'(\eta_i) & \text{if } i = j \\ K\Psi'(\sum_i \eta_i) & \text{otherwise} \end{cases} \tag{39}
$$

Therefore, the Hessian is of the form

$$
H = K \cdot \left(\mathrm{diag}(\mathbf{h}) + \mathbf{1} z \mathbf{1}^T \right) \tag{40}
$$

where

$$
h_i = -\Psi'(\eta_i) \tag{41}
$$

$$
z = \Psi'(\sum_i \eta_i). \tag{42}
$$

So we can derive a linear-time Newton-Raphson iteration as outlined in [2], as follows.

$$
\boldsymbol{\eta}^{\text{new}} = \boldsymbol{\eta} - H(\boldsymbol{\eta})^{-1} g(\boldsymbol{\eta}) \tag{43}
$$

$$
\left(H^{-1} g\right)_v = \frac{1}{K} \cdot \frac{c - P_v + K\left(\Psi(\eta_v) - \Psi(\sum_v \eta_v)\right)}{\Psi'(\eta_v)} \tag{44}
$$

$$
c = \frac{\sum_v \left(K\left(\Psi(\eta_v) - \Psi(\sum_v \eta_v)\right) - P_v \right) / \Psi'(\eta_v)}{\Psi'(\sum_v \eta_v)^{-1} - \sum_v \Psi'(\eta_v)^{-1}} \ . \quad \blacksquare \tag{45}
$$

References

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. In *Neural Information Processing Systems 14*, 2001.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.