Equation derivations for sparse additive generative models*

Tomonari MASADA @ Nagasaki University

September 12, 2012

The full joint distribution can be written as follows.

$$
\begin{aligned}
&p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\eta}, \boldsymbol{\tau} | \boldsymbol{m}, \theta, \gamma) = p(\boldsymbol{\tau}|\gamma) p(\boldsymbol{\eta}|\boldsymbol{\tau}) p(\boldsymbol{z}|\theta) p(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{\eta}, \boldsymbol{m}) \\
&= \prod_k \prod_w p(\tau_{kw}|\gamma) p(\eta_{kw}|\tau_{kw}) \cdot \prod_d p(z_d|\theta) \prod_i p(x_{di}|\eta_{z_d w}, \boldsymbol{m}) \\
&= \prod_k \prod_w \left\{ \gamma \exp(-\gamma \tau_{kw}) \cdot \frac{1}{\sqrt{2\pi\tau_{kw}}} \exp\Big(-\frac{\eta_{kw}^2}{2\tau_{kw}} \Big) \right\} \cdot \prod_k \theta_k^{m_k} \cdot \prod_k \prod_w \left\{ \frac{\exp(\eta_{kw} + m_w)}{\sum_w \exp(\eta_{kw} + m_w)} \right\}^{n_{kw}} . \quad (1)
\end{aligned}
$$

By marginalizing latent variables and τs out, we obtain the following marginal distribution.

$$
\begin{aligned}
&\ln p(\boldsymbol{x}, \boldsymbol{\eta} | \boldsymbol{m}, \boldsymbol{\theta}, \gamma) = \ln \int \sum_{\boldsymbol{z}} p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\eta}, \boldsymbol{\tau} | \boldsymbol{m}, \boldsymbol{\theta}, \gamma) d\boldsymbol{\tau} \\
&= \ln \int \sum_{\boldsymbol{z}} q(\boldsymbol{z}, \boldsymbol{\tau}) \frac{p(\boldsymbol{\eta}|\boldsymbol{\tau}) p(\boldsymbol{\tau}|\gamma) p(\boldsymbol{z}|\theta) p(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{\eta}, \boldsymbol{m})}{q(\boldsymbol{z}, \boldsymbol{\tau})} d\boldsymbol{\tau} \\
&\geq \int \sum_{\boldsymbol{z}} q(\boldsymbol{z}, \boldsymbol{\tau}) \ln \frac{p(\boldsymbol{\eta}|\boldsymbol{\tau}) p(\boldsymbol{\tau}|\gamma) p(\boldsymbol{z}|\theta) p(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{\eta}, \boldsymbol{m})}{q(\boldsymbol{z}, \boldsymbol{\tau})} d\boldsymbol{\tau} \\
&= \int \sum_{\boldsymbol{z}} q(\boldsymbol{z}) q(\boldsymbol{\tau}) \ln \frac{p(\boldsymbol{\eta}|\boldsymbol{\tau}) p(\boldsymbol{\tau}|\gamma) p(\boldsymbol{z}|\theta) p(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{\eta}, \boldsymbol{m})}{q(\boldsymbol{z}) q(\boldsymbol{\tau})} d\boldsymbol{\tau} \\
&= \int q(\boldsymbol{\tau}) \ln p(\boldsymbol{\eta}|\boldsymbol{\tau}) d\boldsymbol{\tau} + \int q(\boldsymbol{\tau}) \ln p(\boldsymbol{\tau}|\gamma) d\boldsymbol{\tau} + \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \ln p(\boldsymbol{z}|\theta) + \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \ln p(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{\eta}, \boldsymbol{m}) \\
&\quad - \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \ln q(\boldsymbol{z}) - \int q(\boldsymbol{\tau}) \ln q(\boldsymbol{\tau}) d\boldsymbol{\tau} , \quad (2)
\end{aligned}
$$

where we assume that the variational posterior $q(\boldsymbol{z}, \boldsymbol{\tau})$ can be factorized.

Assume $q(\boldsymbol{z}) = \prod_d \prod_k q(z_d = k)^{z_{dk}} = \prod_d \prod_k \lambda_{dk}^{z_{dk}}$.

We would like to estimate the probabilities of latent variable values.

Let λ_{dk} be the probability that document d belongs to cluster k.

$$
\begin{aligned}
l(\boldsymbol{z}) &= \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \ln p(\boldsymbol{z}|\theta) + \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \ln p(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{\eta}, \boldsymbol{m}) - \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \ln q(\boldsymbol{z}) \\
&= \sum_{\boldsymbol{z}} \prod_d \prod_k \lambda_{dk}^{z_{dk}} \ln \prod_d \prod_k \theta_k^{z_{dk}} + \sum_{\boldsymbol{z}} \prod_d \prod_k \lambda_{dk}^{z_{dk}} \ln \prod_d \prod_k \prod_w \left\{ \frac{\exp(\eta_{kw} + m_w)}{\sum_w \exp(\eta_{kw} + m_w)} \right\}^{n_{dw} z_{dk}} \\
&\quad - \sum_{\boldsymbol{z}} \prod_d \prod_k \lambda_{dk}^{z_{dk}} \ln \prod_d \prod_k \lambda_{dk}^{z_{dk}} \\
&= \sum_d \sum_k \lambda_{dk} \ln \theta_k + \sum_d \sum_k \lambda_{dk} \sum_w n_{dw} \ln \left\{ \frac{\exp(\eta_{kw} + m_w)}{\sum_w \exp(\eta_{kw} + m_w)} \right\} - \sum_d \sum_k \lambda_{dk} \ln \lambda_{dk} . \quad (3)
\end{aligned}
$$

By introducing Lagrange multipliers, we obtain the following function to be maximized.

$$
\begin{aligned}
l(\boldsymbol{z}) &= \sum_d \sum_k \lambda_{dk} \ln \theta_k + \sum_d \sum_k \lambda_{dk} \sum_w n_{dw} \ln \left\{ \frac{\exp(\eta_{kw} + m_w)}{\sum_w \exp(\eta_{kw} + m_w)} \right\} - \sum_d \sum_k \lambda_{dk} \ln \lambda_{dk} \\
&\quad + \sum_d s_d \Big(1 - \sum_k \lambda_{dk} \Big) \quad (4)
\end{aligned}
$$

*Eisenstein, Ahmed and Xing. Sparse Additive Generative Models of Text. Proceedings of ICML 2011.

By taking derivatives, we obtain the following result.

$$\frac{\partial l(\boldsymbol{z})}{\partial \lambda_{dk}} = \ln \theta_k + \sum_w n_{dw} \ln \left\{ \frac{\exp(\eta_{kw} + m_w)}{\sum_w \exp(\eta_{kw} + m_w)} \right\} - \ln \lambda_{dk} - 1 + s_d \ . \tag{5}$$

Solve $\frac{\partial l(\boldsymbol{z})}{\partial \lambda_{dk}} = 0$ and obtain the following result.

$$\lambda_{dk} \propto \theta_k \prod_w \left\{ \frac{\exp(\eta_{kw} + m_w)}{\sum_w \exp(\eta_{kw} + m_w)} \right\}^{n_{dw}} \ . \tag{6}$$

In the original paper, θ_k is set to $\frac{1}{K}$.

Next, we would like to estimate word probabilities for each document cluster.

Assume $q(\boldsymbol{\tau}) = \prod_k \prod_w \tau_{kw}^{a-1} \frac{\exp(-\tau_{kw}/b)}{\Gamma(a) b^a}$.

$$\begin{aligned} l(\boldsymbol{\eta}) &= \int q(\boldsymbol{\tau}) \ln p(\boldsymbol{\eta}|\boldsymbol{\tau}) d\boldsymbol{\tau} + \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \ln p(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{\eta}, \boldsymbol{m}) \\ &= \sum_k \sum_w \int \tau_{kw}^{a-1} \frac{\exp(-\tau_{kw}/b)}{\Gamma(a) b^a} \ln \left\{ \frac{1}{\sqrt{2\pi\tau_{kw}}} \exp\Big(- \frac{\eta_{kw}^2}{2\tau_{kw}} \Big) \right\} d\tau_{kw} \\ &\quad + \sum_d \sum_k \lambda_{dk} \sum_w n_{dw} \ln \left\{ \frac{\exp(\eta_{kw} + m_w)}{\sum_w \exp(\eta_{kw} + m_w)} \right\} \end{aligned} \tag{7}$$

The integral inside of the summation of the first term can be rewritten as follows.

$$\begin{aligned} &\int \tau_{kw}^{a-1} \frac{\exp(-\tau_{kw}/b)}{\Gamma(a) b^a} \ln \left\{ \frac{1}{\sqrt{2\pi\tau_{kw}}} \exp\Big(- \frac{\eta_{kw}^2}{2\tau_{kw}} \Big) \right\} d\tau_{kw} \\ &= -\frac{1}{2} \int \tau_{kw}^{a-1} \frac{\exp(-\tau_{kw}/b)}{\Gamma(a) b^a} \ln \tau_{kw} d\tau_{kw} - \frac{\eta_{kw}^2}{2} \int \tau_{kw}^{a-1} \frac{\exp(-\tau_{kw}/b)}{\Gamma(a) b^a} \tau_{kw}^{-1} d\tau_{kw} + const. \\ &= -\frac{1}{2} \{ \Psi(a) + \ln b \} - \frac{\Gamma(a-1) b^{a-1}}{\Gamma(a) b^a} \frac{\eta_{kw}^2}{2} + const. \\ &= -\frac{\eta_{kw}^2}{2(a-1)b} + const. \end{aligned} \tag{8}$$

Note that $\langle \tau_{kw}^{-1} \rangle = \frac{1}{2(a-1)b}$. (See p.688 of C.M. Bishop's PRML.)

Therefore, we obtain the following derivative.

$$\begin{aligned} \frac{\partial l(\boldsymbol{\eta})}{\partial \eta_{kw}} &= \sum_d \lambda_{dk} n_{dw} - \frac{\exp(\eta_{kw} + m_w) \sum_d \lambda_{dk} \sum_w n_{dw}}{\sum_w \exp(\eta_{kw} + m_w)} - \frac{\eta_{kw}}{(a-1)b} \\ &= n_{kw} - n_k \frac{\exp(\eta_{kw} + m_w)}{\sum_w \exp(\eta_{kw} + m_w)} - \frac{\eta_{kw}}{(a-1)b} \end{aligned} \tag{9}$$

Let $\beta_{kw} \equiv \frac{\exp(\eta_{kw} + m_w)}{\sum_w \exp(\eta_{kw} + m_w)}$. We obtain second order derivatives as follows.

$$\begin{aligned} \frac{\partial^2 l(\boldsymbol{\eta})}{\partial \eta_{kw}^2} &= \frac{\partial}{\partial \eta_{kw}} \left\{ n_{kw} - n_k \frac{\exp(\eta_{kw} + m_w)}{\sum_w \exp(\eta_{kw} + m_w)} - \frac{\eta_{kw}}{(a-1)b} \right\} \\ &= -n_k \frac{\exp(\eta_{kw} + m_w)}{\sum_w \exp(\eta_{kw} + m_w)} + n_k \frac{\exp(\eta_{kw} + m_w)^2}{\{\sum_w \exp(\eta_{kw} + m_w)\}^2} - \frac{1}{(a-1)b} \\ \frac{\partial^2 l(\boldsymbol{\eta})}{\partial \eta_{kw} \partial \eta_{kw'}} &= \frac{\partial}{\partial \eta_{kw'}} \left\{ n_{kw} - n_k \frac{\exp(\eta_{kw} + m_w)}{\sum_w \exp(\eta_{kw} + m_w)} - \frac{\eta_{kw}}{(a-1)b} \right\} \\ &= n_k \frac{\exp(\eta_{kw} + m_w) \exp(\eta_{kw'} + m_{w'})}{\{\sum_w \exp(\eta_{kw} + m_w)\}^2} \end{aligned} \tag{10}$$

The original paper shows that the Hessian can be inverted by using Sherman-Morrison formula.

Finally, we would like to estimate variances. Recall that the variational posterior distribution of τ_{kw} is assumed to be $\tau_{kw}^{a-1}\frac{\exp(-\tau_{kw}/b)}{\Gamma(a)b^a}$, i.e., a Gamma distribution.

$$
\begin{aligned}
&l(\boldsymbol{\tau}) \\
&= \int q(\boldsymbol{\tau}) \ln p(\boldsymbol{\eta}|\boldsymbol{\tau}) d\boldsymbol{\tau} + \int q(\boldsymbol{\tau}) \ln p(\boldsymbol{\tau}|\gamma) d\boldsymbol{\tau} - \int q(\boldsymbol{\tau}) \ln q(\boldsymbol{\tau}) d\boldsymbol{\tau} \\
&= \sum_k \sum_w \int \tau_{kw}^{a-1} \frac{\exp(-\tau_{kw}/b)}{\Gamma(a)b^a} \ln \left\{ \frac{1}{\sqrt{2\pi\tau_{kw}}} \exp\Big(-\frac{\eta_{kw}^2}{2\tau_{kw}} \Big) \right\} d\tau_{kw} \\
&\quad + \sum_k \sum_w \int \tau_{kw}^{a-1} \frac{\exp(-\tau_{kw}/b)}{\Gamma(a)b^a} \ln \left\{ \gamma \exp(-\gamma\tau_{kw}) \right\} d\tau_{kw} \\
&\quad - \sum_k \sum_w \int \tau_{kw}^{a-1} \frac{\exp(-\tau_{kw}/b)}{\Gamma(a)b^a} \ln \left\{ \tau_{kw}^{a-1} \frac{\exp(-\tau_{kw}/b)}{\Gamma(a)b^a} \right\} d\tau_{kw} \\
&= \sum_k \sum_w \Bigg[-\frac{1}{2}\{\Psi(a) + \ln b\} - \frac{\eta_{kw}^2}{2(a-1)b} - \gamma ab - (a-1)\{\Psi(a) + \ln b\} + \frac{ab}{b} + \ln\Gamma(a) + a \ln b \Bigg] + const. \\
&= \sum_k \sum_w \Bigg[-\frac{\eta_{kw}^2}{2(a-1)b} - \gamma ab - \big(a - \frac{1}{2}\big)\{\Psi(a) + \ln b\} + a + \ln\Gamma(a) + a \ln b \Bigg] + const. \qquad (11)
\end{aligned}
$$

Assume that we have different as and bs for each τ_{kw}.

By taking derivatives with respect to a, we obtain the following results.

$$
\begin{aligned}
\frac{\partial l(\boldsymbol{\tau})}{\partial a} &= \frac{\eta_{kw}^2}{2(a-1)^2 b} - \gamma b - \{\Psi(a) + \ln b\} - \big(a - \frac{1}{2}\big)\Psi'(a) + 1 + \Psi(a) + \ln b \\
&= \frac{\eta_{kw}^2}{2(a-1)^2 b} - \gamma b - \big(a - \frac{1}{2}\big)\Psi'(a) + 1 \\
\frac{\partial^2 l(\boldsymbol{\tau})}{\partial a^2} &= -\frac{\eta_{kw}^2}{(a-1)^3 b} - \Psi'(a) - \big(a - \frac{1}{2}\big)\Psi''(a) \qquad (12)
\end{aligned}
$$

Further, we obtain a first order derivative with respect to b as follows.

$$
\frac{\partial l(\boldsymbol{\tau})}{\partial b} = \frac{\eta_{kw}^2}{2(a-1)b^2} - \gamma a - \big(a - \frac{1}{2}\big)\frac{1}{b} + \frac{a}{b} \ . \qquad (13)
$$

We can solve $\frac{\partial l(\boldsymbol{\tau})}{\partial b} = 0$ in a closed form.

When $p(\tau) \propto \frac{1}{\tau}$, we obtain the following result.

$$
\begin{aligned}
l(\boldsymbol{\tau}) &= \sum_k \sum_w \int \tau_{kw}^{a-1} \frac{\exp(-\tau_{kw}/b)}{\Gamma(a)b^a} \ln \left\{ \frac{1}{\sqrt{2\pi\tau_{kw}}} \exp\Big(-\frac{\eta_{kw}^2}{2\tau_{kw}} \Big) \right\} d\tau_{kw} \\
&\quad + \sum_k \sum_w \int \tau_{kw}^{a-1} \frac{\exp(-\tau_{kw}/b)}{\Gamma(a)b^a} \ln \frac{1}{\tau_{kw}} d\tau_{kw} \\
&\quad - \sum_k \sum_w \int \tau_{kw}^{a-1} \frac{\exp(-\tau_{kw}/b)}{\Gamma(a)b^a} \ln \left\{ \tau_{kw}^{a-1} \frac{\exp(-\tau_{kw}/b)}{\Gamma(a)b^a} \right\} d\tau_{kw} + const. \\
&= \sum_k \sum_w \Bigg[-\frac{1}{2}\{\Psi(a) + \ln b\} - \frac{\eta_{kw}^2}{2(a-1)b} - \{\Psi(a) + \ln b\} - (a-1)\{\Psi(a) + \ln b\} + \frac{ab}{b} + \ln\Gamma(a) + a \ln b \Bigg] \\
&\quad + const. \\
&= \sum_k \sum_w \Bigg[-\frac{\eta_{kw}^2}{2(a-1)b} - \big(a + \frac{1}{2}\big)\{\Psi(a) + \ln b\} + a + \ln\Gamma(a) + a \ln b \Bigg] + const. \qquad (14)
\end{aligned}
$$