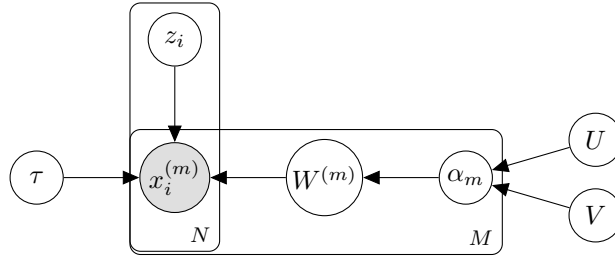


Advanced Machine Learning, Final Project

Diego A. Mancevo (890525-T650) damdc@kth.se

January 7, 2016

We begin by presenting the graphical model corresponding to group factor analysis:



Where:

$X \in \mathbb{R}^{N \times D}$ such that:

$$X = [X^{(1)}, \dots, X^{(M)}]$$

$$X^{(m)\top} = [x_1^{(m)}, \dots, x_N^{(m)}]$$

$$p(X|W, Z, \tau) = \prod_i \prod_m \mathcal{N}(x_i^{(m)} | W^{(m)\top} z_i, \tau_m^{-1} \mathbf{I})$$

$\tau \in \mathbb{R}^{1 \times M}$ such that:

$$\tau = [\tau_1, \dots, \tau_M]$$

$$p(\tau) = \prod_m \mathcal{G}(\tau_m | a^\tau = 10^{-14}, b^\tau = 10^{-14})$$

$Z \in \mathbb{R}^{N \times K}$ such that:

$$Z^\top = [z_1, \dots, z_N]$$

$$p(Z) = \prod_i \mathcal{N}(z_i | \mathbf{0}, \mathbf{I})$$

$W \in \mathbb{R}^{K \times D}$ such that:

$$W = [W^{(1)}, \dots, W^{(M)}]$$

$$W^{(m)\top} = [w_1^{(m)}, \dots, w_K^{(m)}]$$

$$\begin{aligned}
w_k^{(m)} &\in \mathbb{R}^{D_m} \\
\sum_{m=1}^M D_m &= D \\
p(W|\alpha) &= \prod_{m=1}^M \prod_{k=1}^K \prod_{d=1}^{D_m} \mathcal{N}(w_{k,d}^{(m)} | 0, \alpha_{m,k}^{-1})
\end{aligned}$$

$\alpha \in \mathbb{R}^{M \times K}$ such that:

$$\log(\alpha) = UV^\top + \mu_u \mathbf{1}^\top + \mathbf{1} \mu_v^\top.$$

$$U \in \mathbb{R}^{M \times R}$$

$$p(U) = \prod_m^M \prod_r^R \mathcal{N}(u_{m,r} | 0, (\lambda = 0.1)^{-1})$$

$$V \in \mathbb{R}^{K \times R}$$

$$p(V) = \prod_k^K \prod_r^R \mathcal{N}(v_{k,r} | 0, (\lambda = 0.1)^{-1})$$

Then the model's full joint probability can be written as:

$$p(\Theta, X) = p(Z, W, \tau, U, V, X) = p(Z)p(W|\alpha)p(\tau)p(U)p(V)p(X|W, Z, \tau)$$

Where in order to minimize the Kullback-Leibler divergence:

$$D_{KL}(q||p) = \int_{\Theta} q(\Theta) \log\left(\frac{q(\Theta)}{p(\Theta|X)}\right) d\Theta$$

or equivalently to maximize:

$$\mathcal{L}(\Theta) = \int_{\Theta} q(\Theta) \log\left(\frac{p(\Theta, X)}{q(\Theta)}\right) d\Theta$$

We assume:

$$q(\Theta) = q(Z)q(W)q(\tau)q(U)q(V)$$

In which case and by means of variational calculus we must have that $q(\theta_i)$ must have the form:

$$\begin{aligned}
q(\theta_i) &= \frac{e^{E_{i \neq j}[\log(p(\Theta, X))]} }{\int e^{E_{i \neq j}[\log(p(\Theta, X))]} d\theta_i} \\
\implies \log(q(\theta_i)) &= E_{i \neq j}[\log(p(\Theta, X))] + \text{constant}
\end{aligned}$$

And so we proceed by taking the corresponding expectations with respect to the log of the model's full joint probability:

$$\log(q(Z)) = E_{W, \tau}[\log(p(\Theta, X))] = E_{W, \tau}[\log(p(Z))] + E_{W, \tau}[\log(p(X|W, Z, \tau))] + C_1$$

$$\begin{aligned}
&= E_{W,\tau} \left[\sum_i^N \log(\mathcal{N}(z_i | \mathbf{0}, \mathbf{I})) \right] + E_{W,\tau} \left[\sum_i^N \sum_m^M \log(\mathcal{N}(x_i^{(m)} | W^{(m)\top} z_i, \tau_m^{-1} \mathbf{I})) \right] + C_1 \\
&= -\frac{1}{2} \sum_i^N z_i^\top z_i - \frac{1}{2} \sum_m^M E_{W,\tau} \left[\tau_m (x_i^{(m)} - W^{(m)\top} z_i)^\top (x_i^{(m)} - W^{(m)\top} z_i) \right] + C_2 \\
&= -\frac{1}{2} \sum_i^N z_i^\top \mathbf{I}_k z_i + \sum_m^M \langle \tau_m \rangle (z_i^\top \langle W^{(m)} \rangle x_i^{(m)} - \frac{1}{2} z_i^\top \langle W^{(m)} W^{(m)\top} \rangle z_i) + C_3 \\
&= \sum_i^N \sum_m^M z_i^\top \langle W^{(m)} \rangle \langle \tau_m \rangle x_i^{(m)} - \frac{1}{2} z_i^\top (\mathbf{I}_k + \sum_m^M \langle \tau_m \rangle \langle W^{(m)} W^{(m)\top} \rangle) z_i + C_3
\end{aligned}$$

Note that above we denote the first moment by $E_{\theta_i}[\theta_i] = \langle \theta_i \rangle$ and the second moment by $E_{\theta_i}[\theta_i \theta_i^\top] = \langle \theta_i \theta_i^\top \rangle$. Note as well that we collect all constant factors with respect to z_i into C_1 , C_2 and C_3 respectively. Then recalling:

$$\mathcal{N}(x | \mu, \Sigma) \propto x^\top \Sigma^{-1} \mu - \frac{1}{2} x^\top \Sigma^{-1} x$$

We must have:

$$q(Z) = \prod_i^N \mathcal{N}(m_i^{(z)}, \Sigma^{(z)})$$

with:

$$\begin{aligned}
\Sigma^{(z)} &= \left(\mathbf{I}_k + \sum_m^M \langle \tau_m \rangle \langle W^{(m)} W^{(m)\top} \rangle \right)^{-1} \\
m_i^{(z)} &= \Sigma^{(z)} \langle W^{(m)} \rangle \langle \tau_m \rangle x_i^{(m)}
\end{aligned}$$

Similarly we proceed with $q(W)$ in which case we have:

$$\log(q(W)) = E_{\alpha,Z,\tau}[\log(p(\Theta, X))] = E_{\alpha,Z,\tau}[\log(p(W|\alpha))] + E_{\alpha,Z,\tau}[\log(p(X|W, Z, \tau))] + C_1$$

$$= E_{\alpha,Z,\tau} \left[\sum_m^M \sum_k^K \sum_d^{D_m} \log(\mathcal{N}(w_{k,d}^{(m)} | 0, \alpha_{m,k}^{-1})) \right] + E_{\alpha,Z,\tau} \left[\sum_i^N \sum_m^M \log(\mathcal{N}(x_i^{(m)} | W^{(m)\top} z_i, \tau_m^{-1} \mathbf{I})) \right] + C_1$$

We continue by looking at the group columns $w_{:,d}^{(m)}$ in W as opposed to the group rows $w_k^{(m)}$ such that $W^{(m)} = [w_{:,1}^{(m)}, \dots, w_{:,D_m}^{(m)}]$. Then note that the number of columns in X is equal to the number of columns in W and so we have:

$$= E_{\alpha,Z,\tau} \left[\sum_m^M \sum_d^{D_m} \log(\mathcal{N}(w_{:,d}^{(m)} | \mathbf{0}, \bar{\alpha}_m^{-1})) \right] + E_{\alpha,Z,\tau} \left[\sum_m^M \sum_d^{D_m} \sum_i^N \log(\mathcal{N}(x_{i,d}^{(m)} | w_{:,d}^{(m)\top} z_i, \tau_m^{-1})) \right] + C_1$$

Where $\bar{\alpha}_m$ is the m -th row of α transformed into a diagonal $K \times K$ matrix.

$$\begin{aligned} &= -E_{\alpha,Z,\tau} \left[\frac{1}{2} \sum_m^M \sum_d^{D_m} w_{:,d}^{(m)\top} \bar{\alpha}_m w_{:,d}^{(m)} \right] - E_{\alpha,Z,\tau} \left[\frac{1}{2} \sum_m^M \sum_d^{D_m} \sum_i^N \tau_m (x_{i,d}^{(m)} - w_{:,d}^{(m)\top} z_i)^2 \right] + C_2 \\ &= -\frac{1}{2} \sum_m^M \sum_d^{D_m} w_{:,d}^{(m)\top} \langle \bar{\alpha}_m \rangle w_{:,d}^{(m)} - \frac{1}{2} \sum_m^M \sum_d^{D_m} \sum_i^N \langle \tau_m \rangle (-2x_{i,d}^{(m)} w_{:,d}^{(m)\top} \langle z_i \rangle + w_{:,d}^{(m)\top} \langle z_i z_i^\top \rangle w_{:,d}^{(m)}) + C_3 \\ &= \sum_m^M \sum_d^{D_m} \langle \tau_m \rangle \sum_i^N w_{:,d}^{(m)\top} x_{i,d}^{(m)} \langle z_i \rangle - \frac{1}{2} \sum_m^M \sum_d^{D_m} w_{:,d}^{(m)\top} (\langle \tau_m \rangle \sum_i^N \langle z_i z_i^\top \rangle + \langle \bar{\alpha}_m \rangle) w_{:,d}^{(m)} + C_3 \end{aligned}$$

Then again recalling that $\mathcal{N}(x|\mu, \Sigma) \propto x^\top \Sigma^{-1} \mu - \frac{1}{2} x^\top \Sigma^{-1} x$, we must have that $q(W) = \prod_m^M \prod_d^{D_m} \mathcal{N}(w_{:,d}^{(m)} | m_{m,d}^{(w)}, \Sigma_m^{(w)})$ with:

$$\begin{aligned} \Sigma_m^{(w)} &= \left(\langle \tau_m \rangle \sum_i^N \langle z_i z_i^\top \rangle + \langle \bar{\alpha}_m \rangle \right)^{-1} \\ m_{m,d}^{(w)} &= \Sigma_m^{(w)} \langle \tau_m \rangle \sum_i^N x_{i,d}^{(m)} \langle z_i \rangle \end{aligned}$$

Moving on to $q(\tau)$ we have:

$$\begin{aligned} \log(q(\tau)) &= E_{W,Z}[\log(p(\Theta, X))] = E_{W,Z}[\log(p(\tau)) + E_{W,Z}[\log(p(X|W, Z, \tau))] + C_1 \\ &= E_{W,Z} \left[\sum_m^M \log(\mathcal{G}(\tau_m | a^\tau, b^\tau)) \right] + E_{W,Z} \left[\sum_i^N \sum_m^M \log(\mathcal{N}(x_i^{(m)} | W^{(m)\top} z_i, \tau_m^{-1} \mathbf{I})) \right] + C_1 \\ &= E_{W,Z} \left[\sum_m^M (a^\tau - 1) \log(\tau_m) - b^\tau \tau_m \right] + E_{W,Z} \left[-\frac{1}{2} \sum_m^M \sum_i^N \log(|\tau_m^{-1} \mathbf{I}|) - (x_i^{(m)} - W^{(m)\top} z_i)^2 \tau_m \right] + C_2 \end{aligned}$$

Then notice that $\log(|\tau_m^{-1}\mathbf{I}|) = -D_m \log(\tau_m)$ and so we have:

$$= \sum_m^M \left(a^\tau + \frac{ND_m}{2} - 1 \right) \log(\tau_m) - \left(b^\tau + \sum_i^N \left\langle (x_i^{(m)} - W^{(m)\top} z_i)^2 \right\rangle \right) \tau_m + C_2$$

Which has the form of a new Gamma distribution and thus we must have that $q(\tau) = \prod_m^M \mathcal{G}(\tau_m | a_m^\tau, b_m^\tau)$ where:

$$a_m^\tau = a^\tau + \frac{ND_m}{2}$$

$$b_m^\tau = b^\tau + \sum_i^N \left\langle (x_i^{(m)} - W^{(m)\top} z_i)^2 \right\rangle$$

Finally we turn our attention to U and V at which point we recall that $\log(\alpha) = UV^\top + \mu_u + \mu_v$ but then notice that we can append both μ_u and μ_v to U and V respectively if we let:

$$U = \begin{bmatrix} U \\ \mu_u \end{bmatrix}$$

$$V^\top = \begin{bmatrix} V^\top \\ \mu_v \end{bmatrix}$$

Such that $\log(\alpha) = UV^\top$, so let us use the expanded version of U and V for what follows and then recall:

$$\begin{aligned} \mathcal{L}(\Theta) &= \int_{\Theta} q(\Theta) \log\left(\frac{p(\Theta, X)}{q(\Theta)}\right) d\Theta \\ &= \int_{\Theta} q(Z, W, \tau) q(U) q(V) \log\left(\frac{p(Z, \tau, X) p(U, V) p(W|\alpha)}{q(Z, W, \tau) q(U) q(V)}\right) dZ dW d\tau dU dV \end{aligned}$$

If we concentrate on U and V and regard the remaining variables as constant we then have:

$$\propto \int_{UV} q(U) q(V) \log\left(\frac{p(U, V) p(W|\alpha)}{q(U) q(V)}\right) dU dV$$

At this point we use fixed-form distributions for $q(U)$ and $q(V)$ such that $q(U) = \delta_U$ and $q(V) = \delta_V$ such that:

$$\begin{aligned} &\propto \int_{UV} \log(p(U, V)) + \log(p(W|U, V)) dU dV \\ &= \int_{UV} \log(p(U, V)) + \sum_m^M \sum_k^K \sum_d^{D_m} \log\left(\mathcal{N}(w_{k,d}^{(m)} | 0, \alpha_{m,k}^{-1})\right) dU dV \end{aligned}$$

$$= \int_{UV} \log(p(U, V)) + \sum_m^M \sum_k^K \sum_d^{D_m} \frac{1}{2} \log(\alpha_{m,k}) - \frac{1}{2} \alpha_{m,k} \langle w_{k,d}^{(m)2} \rangle dU dV$$

We then express $p(W|\alpha)$ in terms of U and V as opposed to α such that $\alpha_{m,k} = e^{u_m v_k^\top}$ and notice that the sum from d to D_m of the second moments $\langle w_{k,d}^{(m)2} \rangle$ is the entry in the k -th column and k -th row of the matrix $\langle W^{(m)} W^{(m)\top} \rangle$ and thus:

$$\propto \int_{UV} 2 \log(p(U, V)) + \sum_m^M \sum_k^K \left(D_m u_m v_k^\top - \langle W^{(m)} W^{(m)\top} \rangle_{k,k} e^{u_m v_k^\top} \right) dU dV$$