Maximum A Posteriori Expectation-Maximization Clustering Method

Yingying Gu

June 11, 2015

1 A Short Review of EM Algorithms

In brief review of EM algorithm[1, 2, 3], we derived it with Gaussian Mixture Model. And it also can be extended into another statistical model, for example, the Bernoulli Model. Bernoulli Model is defined for the binary variables, which is useful for language/text processing.

1.1 Complete Data and Incomplete Data

First, we need to define the complete and incomplete data. For the complete data, we use x and $x = \{y, z\}$, where y is the observation data, and z is the indicator that which class the y belongs to. For the incomplete data, when the z is unknown, and only y is remained.
For Gaussian Mixture Model, each x_i came from Gaussian distribution $N(\mu_j, \Sigma_j)$ with weight/proportion of π_j, so

$$p(x_i|\theta) = \sum_{j=1}^{K} \pi_j p\left(x_i|\mu_j, \Sigma_j\right) \tag{1}$$

, where $\theta = \{\mu_1, \mu_2, ..., \mu_K; \Sigma_1, \Sigma_2, ..., \Sigma_K; \pi_1, \pi_2, ..., \pi_K\}$ and $\sum \pi_j = 1$, $\pi_j > 0$. The probability density of the multivariate normal distribution in d dimension is

$$p(x|\mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T(\Sigma)^{-1}(x-\mu)\right\} \tag{2}$$

, where $x \in \mathbb{R}^n$, $\mu \in \mathbb{R}^n$, $\Sigma \in \mathbb{R}^{n\times n}$.

1.2 Estimation (E) Step

In E step, we calculate the conditional expectation of the **log-MLE** (Maximum Likelihood Estimation) of the complete data.

$$
\begin{aligned}
Q(\theta|\theta^{(t)}) &= E[\log p(x|\theta)|y, \theta^{(t)}] \\
&= \sum_i \sum_j E[z_{i,j}|y_i, \theta^{(t)}] \log p(y_i|z_j, \theta^{(t)}) + \sum_i \sum_j E[z_{i,j}|y_i, \theta^{(t)}] \log \pi_j \\
&= \sum_i \sum_j E[z_{i,j}|y_i, \theta^{(t)}] \log \left(\frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} \exp\left\{ -\frac{1}{2}(y_i - \mu_j)^T (\Sigma_j)^{-1} (y_i - \mu_j) \right\} \right) \\
&+ \sum_i \sum_j E[z_{i,j}|y_i, \theta^{(t)}] \log \pi_j
\end{aligned}
\tag{3}
$$

$$
\begin{aligned}
E[z_{i,k}|y_i, \theta^{(t)}] &= p(z_{i,k}|y_i, \theta^{(t)}) \\
&= p(z_{i,k} = 1|y_i, \theta^{(t)}) \\
&= \frac{p(y_i|z_{i,k} = 1, \theta^{(t)}) p(z_{i,k} = 1|\theta^{(t)})}{\sum_{j=1}^{K} p(y_i|z_{i,j} = 1, \theta^{(t)}) p(z_{i,j} = 1|\theta^{(t)})}
\end{aligned}
\tag{4}
$$

1.3 Maximization (M) Step

In M step, we re-estimate the parameter $\theta^{(t+1)} = \arg\max_{\theta} \left\{ Q(\theta|\theta^{(t)}) \right\}$ by

$$
\frac{\partial Q(\theta|\theta^{(t)})}{\partial \theta} = 0
$$

For simplicity, the derivations are in APPENDIX A.
After taking the partial derivative of the **log-MLE** function with respect to $\{\mu, \Sigma, \pi\}$, each parameter is re-estimated as:

$$
\hat{\mu}_j^{(t+1)} = \frac{\sum_{i=1}^{n} E[z_{i,j}|y_i, \theta^{(t)}] y_i}{\sum_{i=1}^{n} E[z_{i,j}|y_i, \theta^{(t)}]} \tag{5}
$$

$$
\hat{\Sigma}_j^{(t+1)} = \frac{\sum_{i=1}^{n} E[z_{i,j}|y_i, \theta^{(t)}] (y_i - \hat{\mu}_j^{(t+1)})^T (y_i - \hat{\mu}_j^{(t+1)})}{\sum_{i=1}^{n} E[z_{i,j}|y_i, \theta^{(t)}]} \tag{6}
$$

$$
\hat{\pi}_j^{(t+1)} = \frac{\sum_{i=1}^{n} E[z_{i,j}|y_i, \theta^{(t)}]}{n} \tag{7}
$$

2 MAP EM Algorithms

In Bayesian theory, if the posterior distribution is in the same distribution family as the prior probability distribution, they are defined as conjugate distributions. And the prior is called the conjugate prior.

2.1 Normal-Inverse-Wishart Prior

The conjugate prior of multivariate normal(Gaussian) distribution is the Normal-inverse-Wishart distribution. The notation is $(\mu, \Sigma) \sim \text{NIW}(m_0, \lambda, S_0, \nu_0)$. The probability density function can be written as:

$$p(\mu, \Sigma | m_0, \beta_0, S_0, \nu_0) = N\left(\mu | m_0, \frac{1}{\beta_0}\Sigma\right) W^{-1}(\Sigma | S_0, \nu_0) \quad (8)$$

The first term in Eq. 8 is the multivariate normal distribution in d dimension as:

$$N\left(\mu | m_0, \frac{1}{\beta_0}\Sigma\right) = \frac{1}{(2\pi)^{d/2}\left|\frac{1}{\beta_0}\Sigma\right|^{1/2}} \exp\left\{-\frac{1}{2}(\mu - m_0)^T \left(\frac{1}{\beta_0}\Sigma\right)^{-1} (\mu - m_0)\right\} \quad (9)$$

The second term in Eq. 8 is the probability density function of the Inverse-Wishart, which is written as:

$$W^{-1}(\Sigma | S_0, \nu_0) = \frac{|S_0|^{\nu_0/2}}{2^{\frac{\nu_0 d}{2}}\Gamma_d(\frac{\nu_0}{2})} |\Sigma_j|^{-(\nu_0+d+1)/2} \exp\left\{-\frac{1}{2}Tr(S_0 {\Sigma_j}^{-1})\right\} \quad (10)$$

, where Σ and S_0 are positive definite matrices, v_0 is degrees of freedom, Γ_d is the multivariate Gamma function, and $Tr(\cdot)$ is the trace of the matrix.
The parameters m_0, β_0, S_0, ν_0 have following interpretations: m_0 is the prior means of μ; β_0 represents how the prior m_0 is closed to the true means of μ; S_0 is the prior mean of Σ; ν_0 shows how the prior S_0 is closed to the true means of Σ.

2.2 Dirichlet Prior

Dirichlet prior is used on the proportions/weights (π) of the mixture, denoted as $\pi \sim Dir(\alpha)$. The probability function is written as:

$$Dir(\pi | \alpha) = \frac{1}{B(\alpha)} \sum_{i=1}^{K} {\pi_i}^{\alpha_i - 1} \quad (11)$$

The beta function is written as

$$B(\alpha) = \frac{\prod_{i=1}^{K}\Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K}\alpha_i)} \quad (12)$$

, where $\alpha = (\alpha_1, \alpha_2, ..., \alpha_K)$.

2.3 MAP Estimation

Now, we will introduce the MAP EM [4, 5] method, which has the same E step as the general EM algorithm, but the M step needs to be changed. After adding the priors on $\theta = \{\mu, \Sigma\}$ and π, the conditional expectation of the **log-MLE** (Maximum Likelihood Estimation) of the complete data can be re-written as:

$$Q'(\theta|\theta^{(t)}) = \sum_i \sum_j E[z_{i,j}|y_i, \theta^{(t)}] \log p(y_i|z_j, \theta^{(t)}) + \sum_i \sum_j E[z_{i,j}|y_i, \theta^{(t)}] \log \pi_j + \sum_j p(\pi_j|\alpha) + \sum_j \log p(\theta^{(t)}|m_0, \beta_0, S_0, \nu_0) \quad (13)$$

We will write the completed form by inserting the Eq. 8 and Eq. 11.
For a clean form, we let $r_{i,j} = E[z_{i,j}|y_i, \theta^{(t)}]$. So,

$$\begin{aligned} Q'(\theta|\theta^{(t)}) = & \sum_i \sum_j r_{i,j} \log \left(\frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} \exp \left\{ -\frac{1}{2}(y_i - \mu_j)^T (\Sigma_j)^{-1} (y_i - \mu_j) \right\} \right) \\ & + \sum_i \sum_j r_{i,j} \log \pi_j + \sum_j \log \left(\frac{1}{B(\alpha)} \right) + \sum_j (\alpha_i - 1) \log \pi_j \\ & + \sum_j \log \left(\frac{1}{(2\pi)^{d/2} \left| \frac{1}{\beta_0}\Sigma_j \right|^{1/2}} \exp \left\{ -\frac{1}{2}(\mu_j - m_0)^T \left(\frac{1}{\beta_0}\Sigma_j \right)^{-1} (\mu_j - m_0) \right\} \right) \\ & + \sum_j \log \left(\frac{|S_0|^{\nu_0/2}}{2^{\frac{\nu_0 d}{2}} \Gamma_d(\frac{\nu_0}{2})} |\Sigma_j|^{-(\nu_0+d+1)/2} \exp \left\{ -\frac{1}{2} Tr(S_0 {\Sigma_j}^{-1}) \right\} \right) \end{aligned} \tag{14}$$

Then, we re-estimate the parameters $\theta^{(t+1)} = \arg\max_\theta \left\{ Q'(\theta|\theta^{(t)}) \right\}$ by

$$\frac{\partial Q'(\theta|\theta^{(t)})}{\partial \theta} = 0$$

For simplicity, the derivations are in APPENDIX B. So each parameter is re-estimated as:

$$\hat{\mu}_j^{(t+1)} = \frac{\sum_{i=1}^n r_{i,j} y_i + \beta_0 m_0}{\sum_{i=1}^n r_{i,j} + \beta_0} \tag{15}$$

$$\hat{\Sigma}_j^{(t+1)} = \frac{S_0 + \sum_{i=1}^n r_{i,j} \left(y_i - \hat{\mu}_j^{(t+1)} \right)^T \left(y_i - \hat{\mu}_j^{(t+1)} \right) + \beta_0 \left(\hat{\mu}_j^{(t+1)} - m_0 \right)^T \left(\hat{\mu}_j^{(t+1)} - m_0 \right)}{\sum_{i=1}^n r_{i,j} + \beta_0 + \nu_0 + D + 2} \tag{16}$$

$$\hat{\pi}_j^{(t+1)} = \frac{\sum_{i=1}^n r_{i,j} + \alpha_j - 1}{n + \sum_{j=1}^K \alpha_j - K} \tag{17}$$

Compared with the parameters from General EM in Eq.5, Eq.6 Eq.7, the parameters from MAP EM in Eq.15, Eq.16 Eq.17 are different by including the priors' parameters of $\alpha, m_0, \beta_0, S_0, \nu_0$. These priors' parameters in MAP EM algorithm can be used to avoid/improve the local maximum, which happens in General EM algorithm.

2.4 Convergence

We have two convergence criteria to measure if the iteration of MAP EM algorithm is need to stop. The first stopping criterion is if the maximum number of iteration (for example $max\_iter = 5000$) is reached. The second stopping criterion is if the error between successive iterations is smaller than a certain value ϵ (for example $\epsilon = 10^{-10}$).
The error can be calculated as:

$$Err \left(\theta^{(t+1)}, \theta^{(t)} \right) = \frac{|Q(\theta^{(t+1)}) - Q(\theta^{(t)})|}{|Q(\theta^{(t)})|} \tag{18}$$

We can use the gradient of the Error function to evaluate the convergence conditions.

$$\nabla_\theta Err \left(\theta^{(t+1)}, \theta^{(t)} \right) = const * (|\theta^{(t+1)} - \theta^{(t)}|) \tag{19}$$

Then, the convergence condition can be simplified as:

$$\frac{|\theta^{(t+1)} - \theta^{(t)}|}{|\theta^{(t)}|} < \epsilon \tag{20}$$

References

[1] A. P. Dempster, N. M. Laird and D.B. Rubin, Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, Series B,39(1), pp. 1-38, 1977.

[2] Z. Ghahramani, and M. I. Jordan, Supervised learning from incomplete data via an EM approach. *Advances in Neural information Processing Systems*, 6, pp. 120-127, 1994.

[3] L. Xu and M. I. Jordan, On Convergence Properties of the EM Algorithm for Gaussian Mixtures. *Neural Computation*, 8, pp. 129-151, 1995.

[4] C. M. Bishop, Pattern Recognition and Machine Learning. *Springer*, 2007.

[5] K. P. Murphy, Machine Learning: A Probabilistic Perspective. *the MIT Press*, 2012.

3 APPENDIX A

$$\theta^{(t+1)} = \arg\max_{\theta} \left\{ Q(\theta|\theta^{(t)}) \right\} \tag{21}$$

$$\nabla_\theta \left\{ Q(\theta|\theta^{(t)}) \right\} = 0 \tag{22}$$

$$\nabla_\theta \left\{ \sum_i \sum_j E[z_{i,j}|y_i, \theta^{(t)}] \log p(y_i|z_j, \theta^{(t)}) + \sum_i \sum_j E[z_{i,j}|y_i, \theta^{(t)}] \log \pi_j \right\} = 0 \tag{23}$$

$$\nabla_\theta \left\{ \sum_i \sum_j E[z_{i,j}|y_i, \theta^{(t)}] \log \left(\frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} \exp\left\{ -\frac{1}{2}(y_i - \mu_j)^T (\Sigma_j)^{-1} (y_i - \mu_j) \right\} \right) \right. \\ \left. + \sum_i \sum_j E[z_{i,j}|y_i, \theta^{(t)}] \log \pi_j \right\} = 0 \tag{24}$$

Estimation of $\hat{\mu}$

$$\nabla_\mu \left\{ Q(\theta|\theta^{(t)}) \right\} = 0 \tag{25}$$

$$\nabla_{\mu_j} \left\{ \sum_i \sum_j E[z_{i,j}|y_i, \theta^{(t)}] \log \left(\frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} \exp\left\{ -\frac{1}{2}(y_i - \mu_j)^T (\Sigma_j)^{-1} (y_i - \mu_j) \right\} \right) \right\} = 0 \tag{26}$$

$$\sum_i E[z_{i,j}|y_i,\theta^{(t)}]\frac{y_i-\mu_j}{\Sigma_j}=0 \tag{27}$$

$$\hat{\mu}_j^{(t+1)}=\frac{\sum_{i=1}^n E[z_{i,j}|y_i,\theta^{(t)}]y_i}{\sum_{i=1}^n E[z_{i,j}|y_i,\theta^{(t)}]} \tag{28}$$

Estimation of $\hat{\Sigma}$

$$\nabla_\Sigma\left\{Q(\theta|\theta^{(t)})\right\}=0 \tag{29}$$

$$\nabla_{\Sigma_j}\left\{\sum_i\sum_j E[z_{i,j}|y_i,\theta^{(t)}]\log\left(\frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}}\exp\left\{-\frac{1}{2}(y_i-\mu_j)^T(\Sigma_j)^{-1}(y_i-\mu_j)\right\}\right)\right\}=0 \tag{30}$$

$$-\frac{1}{2}\sum_i E[z_{i,j}|y_i,\theta^{(t)}]\frac{1}{|\Sigma_j|}|\Sigma_j|(\Sigma_j)^{-1}+\frac{1}{2}\sum_i E[z_{i,j}|y_i,\theta^{(t)}](\Sigma_j)^{-1}(y_i-\mu_j)^T(y_i-\mu_j)(\Sigma_j)^{-1}=0 \tag{31}$$

$$\hat{\Sigma}_j^{(t+1)}=\frac{\sum_{i=1}^n E[z_{i,j}|y_i,\theta^{(t)}](y_i-\hat{\mu}_j^{(t+1)})^T(y_i-\hat{\mu}_j^{(t+1)})}{\sum_{i=1}^n E[z_{i,j}|y_i,\theta^{(t)}]} \tag{32}$$

Estimation of $\hat{\pi}$

$$\nabla_\pi\left\{Q(\theta|\theta^{(t)})\right\}=0 \tag{33}$$

$$\nabla_\pi\left\{\sum_i\sum_j E[z_{i,j}|y_i,\theta^{(t)}]\log\pi_j\right\}=0 \tag{34}$$

Given the equation of $\sum_{j=1}^K \pi_j=1$,

$$\Lambda=\sum_i\sum_j E[z_{i,j}|y_i,\theta^{(t)}]\log\pi_j+\lambda\left(\sum_j\pi_j-1\right) \tag{35}$$

$$\nabla_\pi\{\Lambda\}=0 \tag{36}$$

$$\begin{aligned}\sum_i E[z_{i,j}|y_i,\theta^{(t)}]*\frac{1}{\pi_j}+\lambda&=0\\ \sum_i E[z_{i,j}|y_i,\theta^{(t)}]*\frac{1}{\pi_j}&=-\lambda\pi_j\\ \sum_j\sum_i E[z_{i,j}|y_i,\theta^{(t)}]*\frac{1}{\pi_j}&=\sum_j(-\lambda\pi_j)\\ n&=-\lambda\end{aligned}$$

$$\hat{\pi}_j^{(t+1)}=\frac{\sum_{i=1}^n E[z_{i,j}|y_i,\theta^{(t)}]}{n} \tag{37}$$

4 APPENDIX B

$$
\begin{aligned}
Q'(\theta|\theta^{(t)}) = & \sum_i \sum_j r_{i,j} \log \left(\frac{1}{(2\pi)^{d/2} |\Sigma_j|^{1/2}} \exp \left\{ -\frac{1}{2} (y_i - \mu_j)^T (\Sigma_j)^{-1} (y_i - \mu_j) \right\} \right) \\
& + \sum_i \sum_j r_{i,j} \log \pi_j + \sum_j \log \left(\frac{1}{B(\alpha)} \right) + \sum_j (\alpha_i - 1) \log \pi_j \\
& + \sum_j \log \left(\frac{1}{(2\pi)^{d/2} \left| \frac{1}{\beta_0} \Sigma_j \right|^{1/2}} \exp \left\{ -\frac{1}{2} (\mu_j - m_0)^T \left(\frac{1}{\beta_0} \Sigma_j \right)^{-1} (\mu_j - m_0) \right\} \right) \\
& + \sum_j \log \left(\frac{|S_0|^{\nu_0/2}}{2^{\frac{\nu_0 d}{2}} \Gamma_d(\frac{\nu_0}{2})} |\Sigma_j|^{-(\nu_0+d+1)/2} \exp \left\{ -\frac{1}{2} Tr(S_0 {\Sigma_j}^{-1}) \right\} \right)
\end{aligned} \tag{38}
$$

$$
\theta^{(t+1)} = \arg\max_{\theta} \left\{ Q'(\theta|\theta^{(t)}) \right\} \tag{39}
$$

$$
\nabla_\theta \left\{ Q'(\theta|\theta^{(t)}) \right\} = 0 \tag{40}
$$

Estimation of $\hat{\mu}'$

$$
\nabla_\mu \left\{ Q'(\theta|\theta^{(t)}) \right\} = 0 \tag{41}
$$

$$
\begin{aligned}
\nabla_{\mu_j} \Bigg\{ & \sum_i \sum_j r_{i,j} \log \left(\frac{1}{(2\pi)^{d/2} |\Sigma_j|^{1/2}} \exp \left\{ -\frac{1}{2} (y_i - \mu_j)^T (\Sigma_j)^{-1} (y_i - \mu_j) \right\} \right) \\
& + \sum_j \log \left(\frac{1}{(2\pi)^{d/2} \left| \frac{1}{\beta_0} \Sigma_j \right|^{1/2}} \exp \left\{ -\frac{1}{2} (\mu_j - m_0)^T \left(\frac{1}{\beta_0} \Sigma_j \right)^{-1} (\mu_j - m_0) \right\} \right) \Bigg\} = 0
\end{aligned} \tag{42}
$$

$$
\sum_i r_{i,j} \frac{y_i - \mu_j}{\Sigma_j} + \frac{\beta_0 (\mu_j - m_0)}{\Sigma_j} = 0 \tag{43}
$$

$$
\hat{\mu}_j^{(t+1)} = \frac{\sum_{i=1}^n r_{i,j} y_i + \beta_0 m_0}{\sum_{i=1}^n r_{i,j} + \beta_0} \tag{44}
$$

Estimation of $\hat{\Sigma}'$

$$
\nabla_\Sigma \left\{ Q'(\theta|\theta^{(t)}) \right\} = 0 \tag{45}
$$

$$\nabla_{\Sigma_j}\left\{\sum_i\sum_j r_{i,j}\log\left(\frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}}\exp\left\{-\frac{1}{2}(y_i-\mu_j)^T(\Sigma_j)^{-1}(y_i-\mu_j)\right\}\right)\right.$$
$$+\sum_j\log\left(\frac{1}{(2\pi)^{d/2}\left|\frac{1}{\beta_0}\Sigma_j\right|^{1/2}}\exp\left\{-\frac{1}{2}(\mu_j-m_0)^T\left(\frac{1}{\beta_0}\Sigma_j\right)^{-1}(\mu_j-m_0)\right\}\right)$$
$$\left.+\sum_j\log\left(\frac{|S_0|^{\nu_0/2}}{2^{\frac{\nu_0 d}{2}}\Gamma_d(\frac{\nu_0}{2})}|\Sigma_j|^{-(\nu_0+d+1)/2}\exp\left\{-\frac{1}{2}Tr(S_0{\Sigma_j}^{-1})\right\}\right)\right\}=0 \quad (46)$$

$$-\frac{1}{2}\sum_i r_{i,j}\frac{1}{|\Sigma_j|}|\Sigma_j|(\Sigma_j)^{-1}+\frac{1}{2}\sum_i r_{i,j}(\Sigma_j)^{-1}(y_i-\mu_j)^T(y_i-\mu_j)(\Sigma_j)^{-1}$$
$$-\frac{1}{2}\frac{1}{\beta_0}\frac{1}{\left|\frac{1}{\beta_0}\Sigma_j\right|}\left|\frac{1}{\beta_0}\Sigma_j\right|\left(\frac{1}{\beta_0}\Sigma_j\right)^{-1}+\frac{1}{2}\frac{1}{\beta_0}\left(\frac{1}{\beta_0}\Sigma_j\right)^{-1}(\mu_j-m_0)^T(\mu_j-m_0)\left(\frac{1}{\beta_0}\Sigma_j\right)^{-1}$$
$$-\frac{\nu_0+d+1}{2}\frac{1}{|\Sigma_j|}|\Sigma_j|(\Sigma_j)^{-1}+\frac{1}{2}(\Sigma_j)^{-1}S_0(\Sigma_j)^{-1}=0 \quad (47)$$

After simplifying the above equation, we can get

$$\hat{\Sigma}_j^{(t+1)}=\frac{S_0+\sum_{i=1}^n r_{i,j}\left(y_i-\hat{\mu}_j^{(t+1)}\right)^T\left(y_i-\hat{\mu}_j^{(t+1)}\right)+\beta_0\left(\hat{\mu}_j^{(t+1)}-m_0\right)^T\left(\hat{\mu}_j^{(t+1)}-m_0\right)}{\sum_{i=1}^n r_{i,j}+\beta_0+\nu_0+D+2} \quad (48)$$

Estimation of $\hat{\pi}'$

$$\nabla_\pi\left\{Q'(\theta|\theta^{(t)})\right\}=0 \quad (49)$$

$$\nabla_\pi\left\{\sum_i\sum_j r_{i,j}\log\pi_j+\sum_j(\alpha_i-1)\log\pi_j\right\}=0 \quad (50)$$

Given the equation of $\sum_{j=1}^K\pi_j=1$,

$$\Lambda'=\sum_i\sum_j r_{i,j}\log\pi_j+\sum_j(\alpha_i-1)\log\pi_j+\lambda'\left(\sum_j\pi_j-1\right) \quad (51)$$

$$\nabla_\pi\left\{\Lambda'\right\}=0 \quad (52)$$

$$
\begin{aligned}
\sum_i r_{i,j} * \frac{1}{\pi_j} + \alpha_j * \frac{1}{\pi_j} - \frac{1}{\pi_j} + \lambda' = 0 \\
\sum_i r_{i,j} + (\alpha_j - 1) = -\lambda' \pi_j \\
\sum_j \sum_i r_{i,j} + \sum_j (\alpha_j - 1) = \sum_j (-\lambda' \pi_j) \\
n + \sum_j \alpha_j - K = -\lambda'
\end{aligned}
$$

$$
\hat{\pi}_j^{(t+1)} = \frac{\sum_{i=1}^n r_{i,j} + \alpha_j - 1}{n + \sum_{j=1}^K \alpha_j - K} \tag{53}
$$