EM for factor analysis

I. Memming Park

August 6, 2014

Factor Analysis (FA) is usually not fit by expectation-maximization (EM), but we do it anyways. The generative model for FA is given by [2],

$$\begin{aligned}\mathbf{x}_i &\sim \mathcal{N}(0, \mathbf{I}) \\ \mathbf{y}_i &\sim \mathcal{N}(\mathbf{C}\mathbf{x}_i, \mathbf{R}) \\ \mathbf{R} &= \operatorname{diag}(\sigma_1^2, \cdots, \sigma_p^2)\end{aligned} \tag{1}$$

where $\mathbf{x}$ is a q-dimensional latent variable, $\mathbf{y}$ is a p-dimensional observation, and $q < p$. $\mathbf{C}$ is a $(p \times q)$ factor loadings matrix, and $\mathbf{R}$ is the covariance representing independent noise in each observed dimension.

Note that $\mathbf{x}$ and $\mathbf{y}$ are jointly normal, in fact [1],

$$\begin{bmatrix}\mathbf{x}_i \\ \mathbf{y}_i\end{bmatrix} = \mathcal{N}\left(0, \begin{bmatrix}\mathbf{I} & \mathbf{C}^\top \\ \mathbf{C} & \mathbf{R} + \mathbf{C}\mathbf{C}^\top\end{bmatrix}\right). \tag{2}$$

Also, the posterior over the latents is given by,

$$\mathbf{x}_i | \mathbf{y}_i \sim \mathcal{N}\left(\mu, \Lambda\right), \tag{3}$$

where

$$\begin{aligned}\Lambda &= \left(\mathbf{I} + \mathbf{C}^\top \mathbf{R}^{-1} \mathbf{C}\right)^{-1}, \\ \mu_i &= \Lambda \mathbf{C}^\top \mathbf{R}^{-1} \mathbf{y}_i.\end{aligned} \tag{4}$$

Note that the covariance of the latent doesn't depend on the sample. Our goal is to find the parameters $\theta = \{\mathbf{C}, \mathbf{R}\}$ that maximizes the likelihood $p(\mathbf{y}|\theta)$.

In the E-step, we minimize the KL divergence between $Q(x)$ and $p(\mathbf{x}|\mathbf{y}, \theta)$, which can be achieved by setting $Q(x) = p(\mathbf{x}|\mathbf{y}, \theta)$. Hence, (4) corresponds to the computation required for the E-step.

In the M-step, we maximize the conditional expectation of the total data log-likelihood with respect to θ, i.e.,

$$\theta_{\text{new}} = \operatorname*{argmax}_{\theta} \mathop{\mathrm{E}}_{\mathbf{x}\sim Q}\left[\sum_i \log p(\mathbf{y}_i, \mathbf{x}_i | \theta)\right] \tag{5}$$

To compute the expectation, it's convenient to define the following quantities [3]:

$$\delta = \Lambda \mathbf{C} \mathbf{R}^{-1} \tag{6}$$

$$\Sigma_{yy} = \frac{1}{N} \sum_i \mathbf{y}_i {\mathbf{y}_i}^\top \tag{7}$$

$$\Sigma_{yx} = \frac{1}{N} \mathop{\mathrm{E}}_{\mathbf{x}\sim Q}\left[\sum_i \mathbf{y}_i {\mathbf{x}_i}^\top\right] = \frac{1}{N}\sum_i \mathbf{y}_i {\mu_i}^\top = \frac{1}{N}\sum_i \mathbf{y}_i {\mathbf{y}_i}^\top \delta^\top = \Sigma_{yy}\delta^\top \tag{8}$$

$$\Sigma_{xx} = \frac{1}{N} \mathop{\mathrm{E}}_{\mathbf{x}\sim Q}\left[\sum_i \mathbf{x}_i {\mathbf{x}_i}^\top\right] = \frac{1}{N}\sum_i \left(\mathop{\mathrm{cov}}_{\mathbf{x}\sim Q}(\mathbf{x}_i) + \mu_i {\mu_i}^\top\right) = \Lambda + \delta \Sigma_{yy} \delta^\top \tag{9}$$

Note that the total data log-likelihood can be written as,

$$\sum_i \log p(\mathbf{y}_i, \mathbf{x}_i|\theta) = \sum_i \log p(\mathbf{y}_i|\mathbf{x}_i\theta) + \sum_i \log p(\mathbf{x}_i), \tag{10}$$

and the second term does not depend on the parameters θ, hence our objective in (5) can simply be written as,

$$\theta_{\text{new}} = \operatorname*{argmax}_\theta \operatorname*{E}_{\mathbf{x}\sim Q}\left[\sum_i \log p(\mathbf{y}_i, \mathbf{x}_i|\theta)\right] = \frac{1}{N}\operatorname*{E}_{\mathbf{x}\sim Q}\left[\sum_i \log p(\mathbf{y}_i|\mathbf{x}_i, \theta)\right] \tag{11}$$

where N is the number of i.i.d. samples.

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^N \log\left(p(\mathbf{y}_i|\mathbf{x}_i, \theta)\right) \\
&\quad= -\frac{1}{2}\log|\mathbf{R}| - \frac{1}{2}\frac{1}{N}\sum_i \left(\mathbf{C}\mathbf{x}_i - \mathbf{y}_i\right)^\top \mathbf{R}^{-1}\left(\mathbf{C}\mathbf{x}_i - \mathbf{y}_i\right) \\
&\quad= -\frac{1}{2}\log|\mathbf{R}| - \frac{1}{2}\frac{1}{N}\sum_i \left({\mathbf{x}_i}^\top \mathbf{C}^\top\mathbf{R}^{-1}\mathbf{C}\mathbf{x}_i + {\mathbf{y}_i}^\top\mathbf{R}^{-1}\mathbf{y}_i - 2{\mathbf{y}_i}^\top\mathbf{R}^{-1}\mathbf{C}\mathbf{x}_i\right) \\
&\quad= -\frac{1}{2}\log|\mathbf{R}| - \frac{1}{2}\left(\operatorname{tr}\left[\mathbf{C}^\top\mathbf{R}^{-1}\mathbf{C}\frac{1}{N}\sum_i \mathbf{x}_i{\mathbf{x}_i}^\top\right] + \operatorname{tr}\left[\mathbf{R}^{-1}\frac{1}{N}\sum_i \mathbf{y}_i{\mathbf{y}_i}^\top\right] - 2\operatorname{tr}\left[\mathbf{R}^{-1}\mathbf{C}\frac{1}{N}\sum_i \mathbf{x}_i{\mathbf{y}_i}^\top\right]\right)
\end{aligned}$$

Now, taking the expectation over $\mathbf{x} \sim Q$,

$$\begin{aligned}
&\frac{1}{N}\operatorname*{E}_{\mathbf{x}\sim Q}\left[\sum_{i=1}^N \log\left(p(\mathbf{y}_i|\mathbf{x}_i, \theta)\right)\right] \\
&\qquad= -\frac{1}{2}\log|\mathbf{R}| - \frac{1}{2}\left(\operatorname{tr}\left[\mathbf{C}^\top\mathbf{R}^{-1}\mathbf{C}\Sigma_{xx}\right] + \operatorname{tr}\left[\mathbf{R}^{-1}\Sigma_{yy}\right] - 2\operatorname{tr}\left[\mathbf{R}^{-1}\mathbf{C}\Sigma_{xy}\right]\right),
\end{aligned} \tag{12}$$

where constant terms are omitted. Keep in mind that Σ_{xx} and Σ_{xy}'s dependence on parameters are through Q, and hence fixed during the M-step update. We find the stationary points by inspecting the partial derivatives with respect to the parameters.

$$\frac{\partial}{\partial C} = \mathbf{R}^{-1}\hat{\mathbf{C}}\Sigma_{xx} - \mathbf{R}^{-1}{\Sigma_{xy}}^\top = 0 \tag{13}$$

$$\implies \hat{\mathbf{C}} = \Sigma_{yx}\Sigma_{xx}^{-1} = \Sigma_{yy}\delta^\top(\Lambda + \delta\Sigma_{yy}\delta^\top)^{-1} \tag{14}$$

$$\frac{\partial}{\partial R^{-1}} = \frac{1}{2}\mathbf{R} - \frac{1}{2}\mathbf{C}\Sigma_{xx}\mathbf{C}^\top \circ \mathbf{I} - \frac{1}{2}\Sigma_{yy} \circ \mathbf{I} + \Sigma_{yx}\mathbf{C}^\top \circ \mathbf{I} = 0 \tag{15}$$

$$\begin{aligned}
\implies \mathbf{R} &= \left(\mathbf{C}\Sigma_{xx}\mathbf{C}^\top + \Sigma_{yy} - 2\Sigma_{yx}\mathbf{C}^\top\right) \circ \mathbf{I} &(16)\\
&= \left(\Sigma_{yy}\delta^\top(\Lambda + \delta\Sigma_{yy}\delta^\top)^{-1}\delta\Sigma_{yy} + \Sigma_{yy} - 2\Sigma_{yy}\delta^\top(\Lambda + \delta\Sigma_{yy}\delta^\top)^{-1}\delta\Sigma_{yy}\right) \circ \mathbf{I} &(17)\\
&= \left(\Sigma_{yy} - \Sigma_{yy}\delta^\top(\Lambda + \delta\Sigma_{yy}\delta^\top)^{-1}\delta\Sigma_{yy}\right) \circ \mathbf{I} &(18)\\
&= \left(\Sigma_{yy}^{-1} + \delta^\top\Lambda\delta\right)^{-1} \circ \mathbf{I} &(19)
\end{aligned}$$

(18) is proposed by [3].

References

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1st ed. 2006. corr. 2nd printing 2011 edition, October 2007.

[2] Sam Roweis and Zoubin Ghahramani. A unifying review of linear gaussian models. *Neural Comput.*, 11(2):305–345, February 1999.

[3] DonaldB Rubin and DorothyT Thayer. EM algorithms for ML factor analysis. *Psychometrika*, 47(1):69–76, March 1982.