Chapter 12. Latent linear models

12.1.
The expected log likelihood for mixture of factor analysis is

$$Q = \mathbb{E}[\log \sum_i \sum_c ((2\pi)^{-\frac{D}{2}} |\mathbf{\Psi}|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_c - \mathbf{W}_c\mathbf{z})^T \mathbf{\Psi}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_c - \mathbf{W}_c\mathbf{z})})^{\pi_c}]$$

To jointly estimate the mean $\boldsymbol{\mu}_c$ and the factor loadings $\mathbf{W}_c$, it is useful to define an augmented column vector of factors $\tilde{\mathbf{z}} = [\mathbf{z}, 1]$ and an augmented factor loading matrix $\tilde{\mathbf{W}}_c = [\mathbf{W}_c; \boldsymbol{\mu}_c]$. The expected log likelihood is then

$$Q = C - \frac{n}{2}\log|\mathbf{\Psi}| - \sum_i \sum_c [\frac{1}{2} r_{ic} \mathbf{x}_i^T \mathbf{\Psi}^{-1} \mathbf{x}_i - r_{ic} \mathbf{x}_i^T \mathbf{\Psi}^{-1} \tilde{\mathbf{W}}_c \mathbb{E}[\tilde{\mathbf{z}} | \mathbf{x}_i, q_i = c]$$

$$+ \frac{1}{2} r_{ic} \mathrm{tr}(\tilde{\mathbf{W}}_c^T \mathbf{\Psi}^{-1} \tilde{\mathbf{W}}_c \mathbb{E}[\tilde{\mathbf{z}}\tilde{\mathbf{z}}^T | \mathbf{x}_i, q_i = c])]$$

where C is a constant. To estimate $\tilde{\mathbf{W}}_c$, we set

$$\frac{\partial Q}{\partial \tilde{\mathbf{W}}_c} = \sum_i (-r_{ic} \mathbf{\Psi}^{-1} \mathbf{x}_i \mathbf{b}_{ic}^T + r_{ic} \mathbf{\Psi}^{-1} \tilde{\mathbf{W}}_c \mathbf{C}_{ic}) = 0.$$

Solving this leads into

$$\hat{\tilde{\mathbf{W}}}_c = [\sum_i r_{ic} \mathbf{x}_i \mathbf{b}_{ic}^T][\sum_i r_{ic} \mathbf{C}_{ic}]^{-1}.$$

12.2.
The conjugate prior for $\mathbf{\Psi}$ would be a inverse Wishart distribution

$$p(\mathbf{\Psi}) \sim \mathrm{IW}(\mathbf{\Psi} | \mathbf{S}_0, \nu_0) \propto |\mathbf{\Psi}|^{-\frac{\nu_0 + D + 1}{2}} e^{-\frac{1}{2}\mathrm{tr}(\mathbf{S}_0^{-1} \mathbf{\Psi}^{-1})}$$

Since we have the fact that $\mathbf{\Psi}$ is diagonal, the prior $p(\mathbf{\Psi})$ is indeed a product of Gamma distributions.
The conjugate prior for $\tilde{\mathbf{W}}_k$ conditioned on $\mathbf{\Psi}$ is

$$p(\tilde{\mathbf{W}}_k|\mathbf{\Psi}) \propto |\mathbf{\Psi}|^{-\frac{L}{2}} e^{-\frac{1}{2}\mathrm{tr}(\mathbf{\Psi}^{-1}(\tilde{\mathbf{W}}_k-\tilde{\mathbf{W}}_{k0})\Delta_k(\tilde{\mathbf{W}}_k-\tilde{\mathbf{W}}_{k0})^T)}$$

with $\mathbf{S}^{-1}$ a diagonal matrix and $\Delta_k > 0$. This factors into the conditional prior for each row $\mathbf{w}_{kl}$ of $\tilde{\mathbf{W}}_k$ which has the form of Gaussian with shared covariance across the priors of rows.
Combining these, the joint full posterior of the parameters becomes

$$p(\tilde{\mathbf{W}}_k, \mathbf{\Psi}, \tilde{\mathbf{z}}_i, q_i = k|\mathbf{x}_i) \propto |\mathbf{\Psi}|^{-\frac{D+L+\nu_0}{2}}$$

$$\times e^{-\frac{1}{2}\mathrm{tr}[\mathbf{\Psi}^{-1}((\mathbf{x}_i-\tilde{\mathbf{W}}_k\tilde{\mathbf{z}}_i)^T(\mathbf{x}_i-\tilde{\mathbf{W}}_k\tilde{\mathbf{z}}_i)+\mathbf{S}^{-1}+(\tilde{\mathbf{W}}_k-\tilde{\mathbf{W}}_{k0})\Delta_k(\tilde{\mathbf{W}}_k-\tilde{\mathbf{W}}_{k0})^T)]}.$$

The expected log likelihood is then

$$Q = C-\frac{n+D+L+\nu_0}{2}\log|\mathbf{\Psi}|-\sum_i\sum_c[\frac{1}{2}r_{ic}\mathbf{x}_i^T\mathbf{\Psi}^{-1}\mathbf{x}_i-r_{ic}\mathbf{x}_i^T\mathbf{\Psi}^{-1}\tilde{\mathbf{W}}_c\mathbb{E}[\tilde{\mathbf{z}}|\mathbf{x}_i, q_i = c]$$

$$+\frac{1}{2}r_{ic}\mathrm{tr}(\tilde{\mathbf{W}}_c^T\mathbf{\Psi}^{-1}\tilde{\mathbf{W}}_c\mathbb{E}[\tilde{\mathbf{z}}\tilde{\mathbf{z}}^T|\mathbf{x}_i, q_i = c])]-\sum_c\frac{1}{2}\Delta_c(\tilde{\mathbf{W}}_c-\tilde{\mathbf{W}}_{c0})^T\mathbf{\Psi}^{-1}(\tilde{\mathbf{W}}_c-\tilde{\mathbf{W}}_{c0})$$

where C is a constant. To estimate $\tilde{\mathbf{W}}_c$, we set

$$\frac{\partial Q}{\partial\tilde{\mathbf{W}}_c} = \sum_i(-r_{ic}\mathbf{\Psi}^{-1}\mathbf{x}_i\mathbf{b}_{ic}^T + r_{ic}\mathbf{\Psi}^{-1}\tilde{\mathbf{W}}_c\mathbf{C}_{ic}) + \mathbf{\Psi}^{-1}\Delta_c(\tilde{\mathbf{W}}_c - \tilde{\mathbf{W}}_{c0}) = 0.$$

Solving this leads into

$$\hat{\tilde{\mathbf{W}}}_c = [\Delta_c\tilde{\mathbf{W}}_{c0} + \sum_i r_{ic}\mathbf{x}_i\mathbf{b}_{ic}^T][\Delta_c + \sum_i r_{ic}\mathbf{C}_{ic}]^{-1}.$$

We estimate $\mathbf{\Psi}$ through its inverse, setting

$$\frac{\partial Q}{\partial\mathbf{\Psi}^{-1}} = \frac{n+D+L+\nu_0}{2}\mathbf{\Psi}-\sum_i\sum_c[\frac{1}{2}r_{ic}\mathbf{x}_i\mathbf{x}_i^T-r_{ic}\hat{\tilde{\mathbf{W}}}_c\mathbf{b}_{ic}\mathbf{x}_i^T+\frac{1}{2}r_{ic}\hat{\tilde{\mathbf{W}}}_c\mathbf{C}_{ic}\hat{\tilde{\mathbf{W}}}_c^T]$$

$$-\frac{\Delta_c}{2}(\hat{\tilde{\mathbf{W}}}_c - \tilde{\mathbf{W}}_{c0})(\hat{\tilde{\mathbf{W}}}_c - \tilde{\mathbf{W}}_{c0})^T = 0.$$

Using the estimated value of $\hat{\tilde{\mathbf{W}}}_c$ and using the diagonal constraint on $\mathbf{\Psi}$ we obtain,

$$\hat{\mathbf{\Psi}} = \frac{1}{n + D + L + \nu_0}\text{diag}(\sum_i \sum_c r_{ic}(\mathbf{x}_i - \hat{\tilde{\mathbf{W}}}_c \mathbf{b}_{ic})\mathbf{x}_i^T + \Delta_c(\hat{\tilde{\mathbf{W}}}_c - \tilde{\mathbf{W}}_{c0})(\hat{\tilde{\mathbf{W}}}_c - \tilde{\mathbf{W}}_{c0})^T).$$

Finally, to reestimate the mixing proportions, we use the definition

$$\pi_c = p(q = c) = \int p(q = c|\mathbf{x})p(\mathbf{x})d\mathbf{x}.$$

Since $r_{ic} = p(q_i = c|\mathbf{x}_i)$, using the empirical distribution as an estimate, we get

$$\hat{\pi}_c = \frac{1}{n}\sum_i r_{ic}.$$

12.3.
The mean of eigenvalues is a fraction of the total variance of the dataset over the size of the dataset. As all eigenvalues are nonnegative (because we're dealing with the covariance matrix), all eigenvalues are lower or equal to the total variance. Hence the maximum variance among eigenvalues is reached when one of eigenvalues is equal to the total variance and all other eigenvalues are zero. This is the optimal case because it means that all features are reduced to one dimensional subspace. Since the sum of square function is convex, the higher the value of the variance of the eigenvalues, the more useful PCA is.

12.4.
a. Since $\mathbf{v}_1^T\mathbf{v}_1 = \mathbf{v}_2^T\mathbf{v}_2 = 1$ and $\mathbf{v}_1^T\mathbf{v}_2 = 0$,

$$J(\mathbf{v}_2, \mathbf{z}_2) = \frac{1}{N}\sum_i (\mathbf{x}_i^T\mathbf{x}_i + z_{i1}^2 + z_{i2}^2 - 2z_{i1}\mathbf{v}_1^T\mathbf{x}_i - 2z_{i2}\mathbf{v}_2^T\mathbf{x}_i).$$

$$\frac{\partial J}{\partial \mathbf{z}_2} = 0 \Rightarrow \frac{\partial J}{\partial \mathbf{z}_{i2}} = \frac{1}{N}(2z_{i2} - 2\mathbf{v}_2^T\mathbf{x}_i) = 0 \Rightarrow z_{i2} = \mathbf{v}_2^T\mathbf{x}_i.$$

b.

$$\frac{\partial \tilde{J}}{\partial \mathbf{v}_2} = -2\mathbf{C}\mathbf{v}_2 + 2\lambda_2\mathbf{v}_2 + \lambda_{12}\mathbf{v}_1 = 0.$$

Multiplying by $\mathbf{v}_1^T$ gives

$$-2\mathbf{v}_1^T\mathbf{C}\mathbf{v}_2 + 2\lambda_2\mathbf{v}_1^T\mathbf{v}_2 + \lambda_{12}\mathbf{v}_1^T\mathbf{v}_1.$$

Since $\mathbf{C}\mathbf{v}_1 = \lambda_1\mathbf{v}_1$, $\mathbf{C} = \mathbf{C}^T$, $\mathbf{v}_1^T\mathbf{v}_2 = 0$, $\mathbf{v}_1^T\mathbf{v}_1 = 1$, we have

$$-2\lambda_1\mathbf{v}_1^T\mathbf{v}_2 + 2\lambda_2\mathbf{v}_1^T\mathbf{v}_2 + \lambda_{12} = 0 \Rightarrow \lambda_{12} = 0.$$

Plugging this into the original equation gives $-2\mathbf{C}\mathbf{v}_2 + 2\lambda_2\mathbf{v}_2 = 0$, therefore $\mathbf{v}_2$ is an eigenvector with the second largest eigenvalue.

12.5.

a.

$$\|\mathbf{x}_i - \sum_j z_{ij}\mathbf{v}_j\|^2 = (\mathbf{x}_i - \sum_j z_{ij}\mathbf{v}_j)^T(\mathbf{x}_i - \sum_j z_{ij}\mathbf{v}_j)$$

$$= \mathbf{x}_i^T\mathbf{x}_i - 2\sum_j z_{ij}\mathbf{x}_i^T\mathbf{v}_j + \sum_j z_{ij}^2\mathbf{v}_j^T\mathbf{v}_j + \sum_j\sum_k z_{ij}z_{ik}\mathbf{v}_j^T\mathbf{v}_k$$

$$= \mathbf{x}_i^T\mathbf{x}_i - 2\sum_j \mathbf{v}_j^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{v}_j + \sum_j z_{ij}^2 = \mathbf{x}_i^T\mathbf{x}_i - \sum_j \mathbf{v}_j^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{v}_j.$$

b.

$$\frac{1}{n}\sum_i(\mathbf{x}_i^T\mathbf{x}_i - \sum_j \mathbf{v}_j^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{v}_j) = \frac{1}{n}\sum_i \mathbf{x}_i^T\mathbf{x}_i - \frac{1}{n}\sum_i\sum_j \mathbf{v}_j^T\mathbf{C}\mathbf{v}_j$$

$$= \frac{1}{n}\sum_i \mathbf{x}_i^T\mathbf{x}_i - \frac{1}{n}\sum_i\sum_j \mathbf{v}_j^T\lambda_j\mathbf{v}_j = \frac{1}{n}\sum_i \mathbf{x}_i^T\mathbf{x}_i - \frac{1}{n}\sum_i\sum_j \lambda_j$$

$$\frac{1}{n}\sum_i \mathbf{x}_i^T\mathbf{x}_i - \sum_j \lambda_j.$$

c. Since $J_d = 0$, we have

$$\frac{1}{n}\sum_i \mathbf{x}_i^T\mathbf{x}_i = \sum_{j=1}^{d}\lambda_j.$$

$$\Rightarrow J_k = \sum_{j=1}^{d}\lambda_j - \sum_{j=1}^{K}\lambda_j = \sum_{j=K+1}^{d}\lambda_j.$$

12.6.
Since $\mathbf{S}_B$ and $\mathbf{S}_W$ are symmetric, we have

$$\frac{\partial J}{\partial \mathbf{w}} = \frac{2\mathbf{S}_B\mathbf{w}\mathbf{w}^T\mathbf{S}_W\mathbf{w} - 2\mathbf{S}_W\mathbf{w}\mathbf{w}^T\mathbf{S}_B\mathbf{w}}{(\mathbf{w}^T\mathbf{S}_W\mathbf{w})^2}.$$

Equating this to 0 gives

$$\mathbf{S}_B\mathbf{w} = \mathbf{S}_W\mathbf{w}\frac{\mathbf{w}^T\mathbf{S}_B\mathbf{w}}{\mathbf{w}^T\mathbf{S}_W\mathbf{w}} = \lambda\mathbf{S}_W\mathbf{w}.$$

12.7.
a.

$$\tilde{\mathbf{C}} = \frac{1}{n}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T = \frac{1}{n}(\mathbf{I}-\mathbf{v}_1\mathbf{v}_1^T)\mathbf{X}\mathbf{X}^T(\mathbf{I}-\mathbf{v}_1\mathbf{v}_1^T)^T = \frac{1}{n}(\mathbf{X}\mathbf{X}^T-\mathbf{v}_1\mathbf{v}_1^T\mathbf{X}\mathbf{X}^T)(\mathbf{I}-\mathbf{v}_1\mathbf{v}_1^T)$$

$$= \frac{1}{n}(\mathbf{X}\mathbf{X}^T-2\mathbf{v}_1\mathbf{v}_1^T\mathbf{X}\mathbf{X}^T+\mathbf{v}_1\mathbf{v}_1^T\mathbf{X}\mathbf{X}^T\mathbf{v}_1\mathbf{v}_1^T) = \frac{1}{n}(\mathbf{X}\mathbf{X}^T-2n\lambda_1\mathbf{v}_1\mathbf{v}_1^T+n\lambda_1\mathbf{v}_1\mathbf{v}_1^T\mathbf{v}_1\mathbf{v}_1^T)$$

$$= \frac{1}{n}\mathbf{X}\mathbf{X}^T - \lambda_1\mathbf{v}_1\mathbf{v}_1^T.$$

b. Let $\mathbf{v}_j$ be an eigenvector of $\mathbf{X}\mathbf{X}^T$ with $j \neq 1$. Then Since

$$\tilde{\mathbf{C}}\mathbf{v}_j = \frac{1}{n}\mathbf{X}\mathbf{X}^T\mathbf{v}_j - \lambda_1\mathbf{v}_1\mathbf{v}_1^T\mathbf{v}_j = \lambda = \lambda_j\mathbf{v}_j,$$

we have that $\mathbf{v}_j$ is an eigenvector of $\tilde{\mathbf{C}}$. Since $\mathbf{v}_1$ is in the null space of $\tilde{\mathbf{C}}$, we can conclude that the set of $\mathbf{v}_j$ with $j \neq 1$ is the set of whole eigenvectors of $\tilde{\mathbf{C}}$, therefore the principal eigenvector of $\tilde{\mathbf{C}}$ is equivalent to $\mathbf{v}_2$.
c.

$\mathbf{C}' = \mathbf{C}$;
for *j = 1 to K* **do**
  $[\lambda_j, \mathbf{v}_j] = f(\mathbf{C}')$;
  $\mathbf{C}' = \frac{1}{n}\mathbf{C}' - \lambda_j\mathbf{v}_j\mathbf{v}_j^T$;
end

Algorithm 1: Finding the first K principal vectors($\mathbf{C}$, K)

12.9.

$p(\mathbf{x}_h|\mathbf{z}_h, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{x}_h|\mathbf{W}\mathbf{z}_h + \boldsymbol{\mu}_h, \boldsymbol{\Psi}_h)$, $p(\mathbf{x}_v|\mathbf{z}_v, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{x}_h|\mathbf{W}\mathbf{z}_v + \boldsymbol{\mu}_v, \boldsymbol{\Psi}_v)$
Since $\boldsymbol{\Psi}$ is diagonal, we have

$$p(\mathbf{x}_h|\mathbf{x}_v, \mathbf{z}_h, \mathbf{z}_v, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{x}_h|\mathbf{W}\mathbf{z}_h + \boldsymbol{\mu}_h, \boldsymbol{\Psi}_h)$$

If we let the prior for $\mathbf{z}$ to be $\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$,

$$p(\mathbf{x}_h|\mathbf{x}_v, \boldsymbol{\theta}) = \int \mathcal{N}(\mathbf{x}_h|\mathbf{W}\mathbf{z}_h + \boldsymbol{\mu}_h, \boldsymbol{\Psi}_h)\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)d\mathbf{z}_h$$

$$= \mathcal{N}(\mathbf{x}_h|\mathbf{W}\boldsymbol{\mu}_0 + \boldsymbol{\mu}_h, \boldsymbol{\Psi}_h + \mathbf{W}\boldsymbol{\Sigma}_0\mathbf{W}^T).$$

12.10.

$$\log p(\mathbf{X}|\mathbf{W}, \sigma^2) = -\frac{N}{2}\ln|\mathbf{C}| - \frac{1}{2}\sum_i \mathbf{x}_i^T\mathbf{C}^{-1}\mathbf{x}_i$$

By the matrix inversion lemma,

$$\hat{\mathbf{C}}^{-1} = \frac{1}{\hat{\sigma}^2}(\hat{\sigma}^2\hat{\mathbf{W}}\hat{\mathbf{W}}^T + \mathbf{I})^{-1} = \frac{1}{\hat{\sigma}^2}(\mathbf{I} - \hat{\sigma}^2\hat{\mathbf{W}}(\mathbf{I} + \hat{\sigma}^2\hat{\mathbf{W}}^T\hat{\mathbf{W}})^{-1}\hat{\mathbf{W}}^T)$$

Substituting $\hat{\mathbf{W}} = \mathbf{V}(\boldsymbol{\Lambda} - \hat{\sigma}^2\mathbf{I})^{\frac{1}{2}}$ into the original log-likelihood gives

$$\log p(\mathbf{X}|\hat{\mathbf{W}}, \sigma^2) = -\frac{N}{2}(D\ln(2\pi) + \sum_{j=1}^{L}\ln(\lambda_j) + \frac{1}{\sigma^2}\sum_{j=L+1}^{D}\lambda_j + (D-L)\ln(\sigma^2) + L)$$

Substituting $\hat{\sigma}^2 = \frac{1}{D-L}\sum_{j=L+1}^{D}\lambda_j$ gives

$$\log p(\mathbf{X}|\hat{\mathbf{W}}, \hat{\sigma}^2) = -\frac{N}{2}(D\ln(2\pi) + D + \sum_{j=1}^{L}\ln(\lambda_j) + (D-L)\ln(\frac{1}{D-L}\sum_{j=L+1}^{D}\lambda_j)).$$