Written by Peter Sarvari

Based on the exercise provided by Dr. Stefanos Zafeiriou as part of CO-495, Imperial College London

Factor analysis is a generalization of PPCA, where we have a different noise variance per dimension.

$p(x_i|y_i) = N(x_i|W * y_i + \mu, \Psi)$, where Ψ is a diagonal matrix with diagonal elements $\sigma_1^2, \sigma_2^2, \dots, \sigma_F^2$.

$p(y_i) = N(y_i|0, I)$, as in standard PPCA.

Note that x_i has dimension $F * 1$ and y_i has dimension $d * 1, d < F$. Accordingly, W has dimension $F * d$ and Ψ has dimension $F * F$.

1. Finding the marginal, $p(x_i)$ and the posterior, p($y_i|x_i$)
$$p(x_i|y_i) \sim \left(x_i - (W * y_i + \mu)\right)' * \Psi^{-1} * \left(x_i - (W * y_i + \mu)\right)$$
$$p(y_i) \sim y_i' * y_i$$
Since $p(x_i) = \int p(x_i|y_i) * p(y_i) * dy_i$, we need to use "completing the squares" technique, let's denote $\tilde{x} = x_i - \mu$
   a. Inside the integral, in the exponential we have:
   $$(\tilde{x} - W * y_i)' * \Psi^{-1} * (\tilde{x} - W * y_i) + y_i' * y_i \quad (1)$$
   b. Considering only terms containing the variable that we are marginalizing out (y_i)
   $$y_i' * (I + W' * \Psi^{-1} * W) * y_i - 2 * \tilde{x}' * \Psi^{-1} * W * y_i$$
   c. For any arbitrary, but symmetric matrix B, we can write
   $$= y_i' * (I + W' * \Psi^{-1} * W) * y_i - 2 * (B^{-1} * W' * \Psi^{-1} * \tilde{x})' * B * y_i \quad (2)$$
   d. Matching terms: a general quadratic expression in y_i: $(y_i - m)' * S * (y - m) = {y_i}' * S * y_i - 2 * m' * S * y_i + m' * S * m \quad (3)$, for symmetric S
   Comparison of (2) with (3) yields:
   $$S = I + W' * \Psi^{-1} * W$$
   $$m' * S = (B^{-1} * W' * \Psi^{-1} * \tilde{x})' * B$$
   Solution:
   $$B = S = I + W' * \Psi^{-1} * W$$
   $$m = B^{-1} * W' * \Psi^{-1} * \tilde{x}$$
   If we add the following term to (2):
   $$m' * S * m = (B^{-1} * W' * \Psi^{-1} * \tilde{x})' * B * (B^{-1} * W' * \Psi^{-1} * \tilde{x})$$
   $$= \tilde{x}' * \Psi^{-1} * W * B^{-1} * W' * \Psi^{-1} * \tilde{x}$$
   Then we can complete the square and it yields the quadratic term in (3) that is in the exponential of the following Gaussian (after substituting m and S):
   $$N(y_i|B^{-1} * W' * \Psi^{-1} * \tilde{x}, B^{-1})$$
   Since this is a Gaussian for y_i, but depends on x_i, this must be $p(y_i|x_i)$. Hence
   $$p(y_i|x_i) = N(y_i|B^{-1} * W' * \Psi^{-1} * (x_i - \mu), B^{-1}) \quad (4)$$
   e. We had a term in (1) that we did not consider, which is $\tilde{x}' * \Psi^{-1} * \tilde{x}$, plus we have to subtract what we added, which was $\tilde{x}' * \Psi^{-1} * W * B^{-1} * W' * \Psi^{-1} * \tilde{x}$. So the remaining terms (that obviously go outside the integral) are
   $$\tilde{x}' * \Psi^{-1} * \tilde{x} - \tilde{x}' * \Psi^{-1} * W * B^{-1} * W' * \Psi^{-1} * \tilde{x}$$
   $$= \tilde{x}'(\Psi^{-1} - \Psi^{-1} * W * B^{-1} * W' * \Psi^{-1}) * \tilde{x}$$
   Since whatever was inside the integral integrates to one (since the integral of a Gaussian is one), we have that
   $$p(x_i) = N(\tilde{x}|0, (\Psi^{-1} - \Psi^{-1} * W * B^{-1} * W' * \Psi^{-1})^{-1})$$
   $$= N(x_i|\mu, (\Psi^{-1} - \Psi^{-1} * W * B^{-1} * W' * \Psi^{-1})^{-1}) \quad (5)$$

Using the Woodbury identity,

$$(A + U * C * V)^{-1} = A^{-1} - A^{-1} * U(C^{-1} + V * A^{-1} * U)^{-1} * V * A^{-1}$$

Choosing

$$A = \Psi$$
$$U = W$$
$$C = I$$
$$V = W'$$

We can easily see that $(\Psi + \mathrm{W} * \mathrm{W}')^{-1} = \Psi^{-1} - \Psi^{-1} * W * B^{-1} * W' * \Psi^{-1}$

Hence $(\Psi^{-1} - \Psi^{-1} * W * B^{-1} * W' * \Psi^{-1})^{-1} = \Psi + \mathrm{W} * \mathrm{W}'$

Substituting this into (5) yields:

$$p(x_i) = N(x_i|\mu, \Psi + \mathrm{W} * \mathrm{W}')$$

2. Devise an Expectation-Maximization algorithm for Factor Analysis

We aim to maximize the joint probability:

$$p(x_i,y_i) = p(x_i|y_i) * p(y_i) = N(x_i|W * y_i + \mu, \Psi) * N(y_i|0, I) \qquad (6)$$

Assuming independent samples:

$$p(X,Y) = \prod_{i=1}^{N} p(x_i,y_i) \qquad (7)$$

Substituting (6) into (7) and taking the natural logarithm yields:

$$\ln\big(p(X,Y)\big) = \frac{-1}{2} * \sum_{i=1}^{N} F * \ln(2\pi) + \ln(\Psi) + \big(x_i - (W * y_i + \mu)\big)' * \Psi^{-1} * \big(x_i - (W * y_i + \mu)\big) + d * \ln(2\pi) + * {y_i}' * y_i \qquad (8)$$

Taking the expectation of (8) on the posterior

$$L(W,\mu,\Psi) \overset{\text{def}}{=} E_{p(Y|X)}\{\ln\big(p(X,Y)\big)\}$$
$$= \frac{-1}{2} * [\sum_{i=1}^{N} +F * \ln(2\pi) + \ln(|\Psi|) + (x_i - \mu)' * \Psi^{-1} * (x_i - \mu) - 2 * E\{y_i\}' * W' * \Psi^{-1} * (x_i - \mu) + E\{(W * y_i)' * \Psi^{-1} * (W * y_i)\} + d * \ln(2\pi) + E\{{y_i}' * y_i\}] \qquad (9)$$

Note that $E\{(W * y_i)' * \Psi^{-1} * (W * y_i)\} = tr(E\{y_i * y_i'\} * W' * \Psi^{-1} * W)$, which – since Ψ^{-1} is a diagonal matrix – is $tr(\Psi^{-1} * E\{y_i * y_i'\} * W' * W)$

Substituting this to (9) yields:

$$L(W,\mu,\Psi) \overset{\text{def}}{=} E_{p(Y|X)}\{\ln\big(p(X,Y)\big)\}$$
$$= \frac{-1}{2} * [\sum_{i=1}^{N} +F * \ln(2\pi) + \ln(\Psi) + (x_i - \mu)' * \Psi^{-1} * (x_i - \mu) - 2 * E\{y_i\}' * W' * \Psi^{-1} * (x_i - \mu) + tr(\Psi^{-1} * E\{y_i * y_i'\} * W' * W) + d * \ln(2\pi) + E\{{y_i}' * y_i\}] \qquad (10)$$

We know $E\{y_i\}$ directly from (4):

$$E\{y_i\} = B^{-1} * W' * \Psi^{-1} * (x_i - \mu) \qquad (11)$$

To calculate $E\{{y_i}' * y_i\}$, we will use that

$$E\{{y_i}' * y_i\} = cov(y_i) + E\{y_i\} * E\{y_i\}'$$

Hence

$$E\{{y_i}' * y_i\} = B^{-1} + E\{y_i\} * E\{y_i\}' \qquad (12)$$

a) Finding the optimal W:

a. Examining the term: $-2 * E\{ y_i\}' * W' * \Psi^{-1} * (x_i - \mu)$

$$\frac{\delta}{\delta W}\left(-2 * E\{ y_i\}' * W' * \Psi^{-1} * (x_i - \mu)\right)$$
$$= \frac{\delta}{\delta W}(-2 * tr(\Psi^{-1} * (x_i - \mu) * E\{ y_i\}' * W')$$
$$= -2 * \Psi^{-1} * (x_i - \mu) * E\{ y_i\}' \quad (13)$$

b. Examining the term: $tr(\Psi^{-1} * E\{y_i * y_i'\} * W' * W)$

Using formula 118 from the Matrix Cookbook noting that matrix C in the formula is the identity matrix and that $\{y_i * y_i'\} = E\{y_i * y_i'\}'$:

$$\frac{\delta}{\delta W} tr(\Psi^{-1} * E\{y_i * y_i'\} * W' * W) = 2 * \Psi^{-1} * W * E\{y_i * y_i'\} \qquad (14)$$

Using (13) and (14) multiplied by $\frac{1}{2} * \Psi$ from the left and (10) we have

$$\frac{\delta L(W)}{\delta W} = 0 \rightarrow$$

$$W = \sum_i (x_i - \mu) * E\{ y_i\}' * \left(\sum_i E\{y_i * y_i'\}\right)^{-1} \qquad (15)$$

b) Finding the optimal Ψ:

a. Examining the term $-2 * E\{ y_i\}' * W' * \Psi^{-1} * (x_i - \mu)$

$$\frac{\delta}{\delta \Psi^{-1}} - 2 * E\{ y_i\}' * W' * \Psi^{-1} * (x_i - \mu)$$
$$= -2 * \frac{\delta}{\delta \Psi^{-1}}(x_i - \mu)' * \Psi^{-1} * W * E\{ y_i\}$$
$$= -2 * \frac{\delta}{\delta \Psi^{-1}} tr(W * E\{ y_i\} * (x_i - \mu)' * \Psi^{-1})$$

Which is –using formula (142) in the Matrix Cookbook –

$$= diag(-2 * W * E\{ y_i\} * (x_i - \mu)') \quad (16)$$

b. Examining the term $(x_i - \mu)' * \Psi^{-1} * (x_i - \mu)$

$$\frac{\delta}{\delta \Psi^{-1}}(x_i - \mu)' * \Psi^{-1} * (x_i - \mu) = \frac{\delta}{\delta \Psi^{-1}} * tr\left((x_i - \mu) * (x_i - \mu)' * \Psi^{-1}\right)$$
$$= diag\left((x_i - \mu) * (x_i - \mu)'\right) \quad (17)$$

c. Examining the term $tr(\Psi^{-1} * E\{y_i * y_i'\} * W' * W)$

$$\frac{\delta}{\delta \Psi^{-1}} tr(\Psi^{-1} * E\{y_i * y_i'\} * W' * W) = \frac{\delta}{\delta \Psi^{-1}} tr(W * E\{y_i * y_i'\} * W' * \Psi^{-1})$$
$$= diag(W * E\{y_i * y_i'\} * W') \quad (18)$$

d. Examining the term $\ln(\Psi)$

$\frac{\delta}{\delta \Psi^{-1}} \ln(\Psi) = \frac{\delta}{\delta \Psi^{-1}} - \ln(|\Psi|^{-1})$, which is – due to property of the determinant – $\frac{\delta}{\delta \Psi^{-1}} - \ln(|\Psi^{-1}|)$, which is –using formula (57) from the matrix cookbook and the fact that Ψ is symmetric: $-\Psi$.

Alternatively, we can realize that since Ψ is diagonal, $\frac{\delta}{\delta\Psi^{-1}}\ln(\Psi)$ will be the same as if Ψ was a scalar. Hence $\frac{\delta}{\delta\Psi^{-1}}\ln(\Psi) = -\Psi \qquad (19)$

Using (16)-(19) and (9) we have that

$$\frac{\delta L(\Psi)}{\delta\Psi^{-1}} = 0 \rightarrow$$

$$\Psi = \frac{1}{N} * diag(\sum_i [(x_i - \mu) * (x_i - \mu)' - 2 * W * E\{y_i\} * (x_i - \mu)' + W * E\{y_i * y_i'\} * W']) \qquad (20)$$

c) Finding the optimal μ:

a. Examining the term $(x_i - \mu)' * \Psi^{-1} * (x_i - \mu)$
$\frac{\delta}{\delta\mu}(x_i - \mu)' * \Psi^{-1} * (x_i - \mu)$ equals – using formula 86 from the Matrix Cookbook – to $-2 * \Psi^{-1} * (x_i - \mu) \qquad (21)$

b. Examining the term $-2 * E\{y_i\}' * W' * \Psi^{-1} * (x_i - \mu)$
Denote $-2 * E\{y_i\}' * W' * \Psi^{-1}$ by d and note that it is a row vector. Then

$$\frac{\delta}{\delta\mu} d * (x_i - \mu) = -d' = 2 * \Psi^{-1} * W * E\{y_i\} \qquad (22)$$

Using (21) and (22) multiplied by $\frac{1}{2} * \Psi$ from the left and (10) we have

$$\frac{\delta L(\mu)}{\delta\mu} = 0 \rightarrow$$

$$\mu = \frac{1}{N}\sum_i x_i - W * E\{y_i\} = \frac{1}{N}\sum_i x_i \qquad (23)$$

Summary:

The EM updates are given by (15), (20) and (23) and the required expectations are given by (11) and (12).