Written by: Peter Sarvari

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

Here we aim to develop an expectation-maximization framework for Probabilistic PCA with missing data values from scratch. The PPCA model is (for simplicity the dependence on the parameters is omitted in the notation):

$$p(x_i|y_i) = N(x_i|W * y_i + \mu, \sigma^2 * I)$$
$$p(y_i) = N(y_i|0, I)$$

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

We would like to obtain the posterior, $p(y_i|x_i)$ and the marginal, $p(x_i)$. This can be done by the method called "completing the squares".

1. Finding the marginal, $p(x_i)$ and the posterior, p($y_i|x_i$)

$$p(x_i|y_i) \sim \frac{1}{\sigma^2} * \left(x_i - (W * y_i + \mu)\right)' * \left(x_i - (W * y_i + \mu)\right)$$
$$p(y_i) \sim y_i' * y_i$$

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:

$$\frac{1}{\sigma^2} * ((\tilde{x} - W * y_i)' * (\tilde{x} - W * y_i) + \sigma^2 * y_i' * y_i) \qquad (1)$$

b. Considering only terms containing the variable that we are marginalizing out (y_i)

$$\frac{1}{\sigma^2} * (y_i' * (\sigma^2 * I + W' * W) * y_i - 2 * \tilde{x}' * W * y_i)$$

c. For any arbitrary, but symmetric matrix B, we can write

$$= y_i' * \frac{1}{\sigma^2} * (\sigma^2 * I + W' * W) * y_i - 2 * (B^{-1} * W' * \tilde{x})' * \frac{1}{\sigma^2} * B * y_i \qquad (2)$$

d. Matching terms: a general quadratic expression in y_i: $(y_i - m)' * S * (y - m) = y_i{}' * S * y_i - 2 * m' * S * y_i + m' * S * m \qquad (3)$, for symmetric S

Comparison of (2) with (3) yields:

$$S = \frac{1}{\sigma^2} * (\sigma^2 * I + W' * W)$$
$$m' * S = (B^{-1} * W' * \Psi^{-1} * \tilde{x})' * B$$

Solution:

$$B = \sigma^2 * S = (\sigma^2 * I + W' * W)$$
$$m = B^{-1} * W' * \tilde{x}$$

If we add the following term to (2):

$$m' * S * m = (B^{-1} * W' * \tilde{x})' * \frac{1}{\sigma^2} * B * (B^{-1} * W' * \tilde{x})$$
$$= \tilde{x}' * W * \frac{1}{\sigma^2} * B^{-1} * W' * \tilde{x}$$

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):

$$N(y_i|B^{-1} * W' * \tilde{x}, \sigma^2 * B^{-1})$$

Since this is a Gaussian for y_i, but depends on x_i, this must be $p(y_i|x_i)$. Hence

$$p(y_i|x_i) = N(y_i|B^{-1} * W' * (x_i - \mu), \sigma^2 * B^{-1}) \quad (4)$$

e. We had a term in (1) that we did not consider, which is $\frac{1}{\sigma^2} * (\tilde{x}' * \tilde{x})$, plus we have to subtract what we added, which was $\tilde{x}' * W * \frac{1}{\sigma^2} * B^{-1} * W' * \tilde{x}$. So the remaining terms (that obviously go outside the integral) are

$$\frac{1}{\sigma^2} * (\tilde{x}' * \tilde{x}) - \tilde{x}' * W * \frac{1}{\sigma^2} * B^{-1} * W' * \tilde{x} = \tilde{x}'(\frac{1}{\sigma^2} * \mathrm{I} - W * \frac{1}{\sigma^2} * B^{-1} * W') * \tilde{x}$$

Since whatever was inside the integral integrates to one (since the integral of a Gaussian is one), we have that

$$p(x_i) = N\left(\tilde{x} \middle| 0, \left(\frac{1}{\sigma^2} * \mathrm{I} - W * \frac{1}{\sigma^2} * B^{-1} * W'\right)^{-1}\right)$$

$$= N(x_i|\mu, (\sigma^{-2} * \mathrm{I} - \sigma^{-2} * W * B^{-1} * W')^{-1}) \quad (5)$$

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

$$A = \sigma^2 * I$$
$$U = W$$
$$C = I$$
$$V = W'$$

We can easily see that $(\sigma^2 * I + \mathrm{W} * \mathrm{W}')^{-1} = \sigma^{-2} * I - \sigma^{-2} * I * W * (I + W' * \sigma^{-2} * I * W)^{-1} * W' * \sigma^{-2} * I = \sigma^{-2} * I - \sigma^{-2} * W * (\sigma^2 * I + W' * W)^{-1} * W' = \sigma^{-2} * I - \sigma^{-2} * W * B^{-1} * W'$

Hence $(\sigma^2 * I + \mathrm{W} * \mathrm{W}')^{-1} = \sigma^{-2} * I - \sigma^{-2} * W * B^{-1} * W'$

Substituting this into (5) yields:

$$p(x_i) = N(x_i|\mu, \sigma^2 * I + \mathrm{W} * \mathrm{W}')$$

Let's denote $\sigma^2 * I + \mathrm{W} * \mathrm{W}'$ by D and $\sigma^2 * I + W' * W$ by M. Then we can write

$$p(y_i|x_i) = N(y_i|M^{-1} * W' * (x_i - \mu), \sigma^2 * M^{-1}) \quad (6)$$

$$p(x_i) = N(x_i|\mu, D) \quad (7)$$

Let's partition x_i into observed (o) and unknown (u) parts!

$$x_i = \begin{pmatrix} {x_i}^o \\ {x_i}^u \end{pmatrix}$$

$$\mu = \begin{pmatrix} \mu^o \\ \mu^u \end{pmatrix}$$

$$D = \begin{pmatrix} D_{oo} & D_{ou} \\ D_{uo} & D_{uu} \end{pmatrix}$$

Now we can only infer from the known part, ${x_i}^o$. Hence we need to calculate $p({x_i}^u|{x_i}^o)$! We will do this by "completing the square"! Note that $p({x_i}^o) = \int p(x_i) * d{x_i}^u$. As above, since we integrate w.r.t ${x_i}^u$ and $p(x_i)$ contains two terms, ${x_i}^u$ and ${x_i}^o$, the term that we cannot take outside the integral is $p({x_i}^u|{x_i}^o)$. To simplify notation, let's say that

$$D = \begin{pmatrix} D_{oo} & D_{ou} \\ D_{uo} & D_{uu} \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

And note that

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} (A-BD^{-1}C)^{-1} & -A^{-1}B(D-CA^{-1}B)^{-1} \\ -D^{-1}C(A-BD^{-1}C)^{-1} & (D-CA^{-1}B)^{-1} \end{pmatrix} = \begin{pmatrix} J & K \\ L & N \end{pmatrix} \quad (8)$$

Then $p(x_i) \sim (({x_i}^o-\mu^o)'\;({x_i}^u-\mu^u)') * \begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} * \begin{pmatrix} {x_i}^o-\mu^o \\ {x_i}^u-\mu^u \end{pmatrix}$

$$= ({x_i}^o-\mu^o)' * J * ({x_i}^o-\mu^o) + ({x_i}^o-\mu^o)' * K * ({x_i}^u-\mu^u)$$

$$+({x_i}^u-\mu^u)' * L * ({x_i}^o-\mu^o) + ({x_i}^u-\mu^u)' * N * ({x_i}^u-\mu^u)$$

Since D is the covariance matrix, we have that J and N are symmetric and L'=K. Using these we can write (considering only terms containing the variable that we are marginalizing out (${x_i}^u$)) :

$${x_i}^{o'} * K * {x_i}^u - \mu^{o'} * K * {x_i}^u + {x_i}^{u'} * L * {x_i}^o - {x_i}^{u'} * L * \mu^o$$

$$+{x_i}^{u'} * N * {x_i}^u - {x_i}^{u'} * N * \mu^u - \mu^{u'} * N * {x_i}^u =$$

$${x_i}^{u'} * N * {x_i}^u + \left(({x_i}^o-\mu^o)' * (K+L') - 2 * \mu^{u'} * N\right) * {x_i}^u =$$

$${x_i}^{u'} * N * {x_i}^u + \left(2 * ({x_i}^o-\mu^o)' * K - 2 * \mu^{u'} * N\right) * {x_i}^u =$$

For an arbitrary symmetric matrix B

$${x_i}^{u'} * N * {x_i}^u - 2 * \left(B^{-1} * \left(N * \mu^u - K' * ({x_i}^o-\mu^o)\right)\right)' B * {x_i}^u \quad (9)$$

A general quadratic expression in ${x_i}^u$:

$$({x_i}^u - m)' * S * ({x_i}^u - m) = {x_i}^{u'} * S * {x_i}^u - 2 * m' * S * {x_i}^u + m' * S * m \qquad (10)$$

Matching terms in (9) and (10)

$$S = N = B \to S^{-1} = N^{-1} = (D - CA^{-1}B)$$

$$m = B^{-1} * \left(N * \mu^u - K' * ({x_i}^o-\mu^o)\right) = N^{-1} * \left(N * \mu^u - K' * ({x_i}^o-\mu^o)\right) =$$

$$\mu^u - N^{-1} * K' * ({x_i}^o-\mu^o) \quad (11)$$

Using (8), (11) becomes

$$\mu^u - (D - CA^{-1}B) * (-(D - CA^{-1}B)^{-1} * B' * A^{-1}) * ({x_i}^o-\mu^o) =$$

$$\mu^u + B' * A^{-1} * ({x_i}^o-\mu^o) =$$

Using that $B' = C$

$$\mu^u + C * A^{-1} * ({x_i}^o-\mu^o) =$$

$$\mu^u - D_{uo} * D_{oo}^{-1} * ({x_i}^o-\mu^o)$$

So $p({x_i}^u|{x_i}^o) = N({x_i}^u|m, S^{-1}) = N({x_i}^u|m, D - CA^{-1}B) =$

$$N({x_i}^u|\mu^u - D_{uo} * D_{oo}^{-1} * ({x_i}^o-\mu^o), D_{uu} - D_{uo} * D_{oo}^{-1} * D_{ou})$$

Hence $p(x_i|{x_i}^o) = N(x_i|z_i, Q)$ where

$$z_i = \begin{pmatrix} {x_i}^o \\ \mu^u - D_{uo} * D_{oo}^{-1} * ({x_i}^o-\mu^o) \end{pmatrix}$$

And (since ${x_i}^o$ is not a random variable and hence its variance and covariance with anything is zero)

$$Q = \begin{pmatrix} 0 & 0 \\ 0 & D_{uu} - D_{uo} * D_{oo}^{-1} * D_{ou} \end{pmatrix}$$

With Expectation-Maximization, we aim to maximize p(X,Y), which is

$$\prod_{i=1}^{N} p(x_i|y_i) * p(y_i) = \prod_{i=1}^{N} N(x_i|W * y_i + \mu, \sigma^2 * I) * N(y_i|0, I)$$

Taking the natural logarithm and the expectation with respect to p(X,Y|X°)

$$E_{p(X,Y|X^o)}\{\ln(p(X,Y)\} = -\frac{NF}{2} * \ln(2\pi) - \frac{NF}{2} * \ln(\sigma^2) - \frac{Nd}{2}\ln(2\pi) -$$

$$\sum\nolimits_{i=1}^{N} \{\frac{1}{2\sigma^2}[tr(E\{(x_i - \mu) * (x_i - \mu)'\}) - 2 * tr(E\{y_i * (x_i - \mu)'\} * W) + tr(W'W * E\{y_i * y_i'\})]$$

$$-\frac{1}{2} * tr(E\{y_i * y_i'\})\} \quad (11)$$

Hence, we must calculate the following expectations:

1. $E_{p(X,Y|X^o)}\{(x_i - \mu) * (x_i - \mu)'\}$,
2. $E_{p(X,Y|X^o)}\{y_i * y_i'\}$ and
3. $E_{p(X,Y|X^o)}\{y_i * (x_i - \mu)'\}$

Note that using Bayes' Theorem, we have $p(x_i, y_i|x_i{}^o) = p(x_i|x_i{}^o, x_i) * p(y_i|x_i) = p(x_i|x_i{}^o) * p(y_i|x_i)$. This last equality is true because x_i includes $x_i{}^o$.

1. Calculating $E_{p(X,Y|X^o)}\{(x_i - \mu) * (x_i - \mu)'\}$

$$E_{p(X,Y|X^o)}\{(x_i - \mu) * (x_i - \mu)'\} = \iint (x_i - \mu) * (x_i - \mu)' * p(x_i, y_i|x_i{}^o) * dx_i * dy_i$$

$$= \int (x_i - \mu) * (x_i - \mu)' * p(x_i|x_i{}^o) * dx_i * \int p(y_i|x_i) * dy_i$$

$$= \int (x_i - \mu) * (x_i - \mu)' * p(x_i|x_i{}^o) * dx_i =$$

$$\int x_i * x_i' * p(x_i|x_i{}^o) * dx_i - 2 * \mu * \int x_i' * p(x_i|x_i{}^o) * dx_i + \mu * \mu' * \int p(x_i|x_i{}^o) * dx_i$$

$$= var(x_i) + E\{x_i\} * E\{x_i'\} - 2 * \mu * E\{x_i'\} + \mu * \mu' =$$

$$Q + z_i * z_i' - 2 * \mu * z_i' + \mu * \mu' = Q + (z_i - \mu) * (z_i - \mu)' \quad (12)$$

2. Calculating $E_{p(X,Y|X^o)}\{y_i * y_i'\}$

$$E_{p(X,Y|X^o)}\{y_i * y_i'\} = \iint y_i * y_i' * p(x_i|x_i{}^o) * p(y_i|x_i) * dx_i * dy_i =$$

$$\int (\int y_i * y_i' * p(y_i|x_i) * dy_i) * p(x_i|x_i{}^o) * dx_i = \int (var(y_i) + E\{y_i\} * E\{y_i'\}) * p(x_i|x_i{}^o) * dx_i =$$

$$\int \left(\sigma^2 * \ M^{-1} + M^{-1} * W' * (x_i - \mu) * \left(M^{-1} * W' * (x_i - \mu)\right)'\right) * p(x_i|x_i{}^o) * dx_i =$$

$$\sigma^2 * \ M^{-1} \int p(x_i|x_i{}^o) * dx_i + M^{-1} * W' * \int (x_i - \mu) * (x_i - \mu)' * p(x_i|x_i{}^o) * dx_i * W * M^{-1}$$
$$= \sigma^2 * \ M^{-1} + M^{-1} * W' * E_{p(X|X^o)}\{(x_i - \mu) * (x_i - \mu)'\} * W * M^{-1}$$
$$= \sigma^2 * \ M^{-1} + M^{-1} * W' * (Q + (z_i - \mu) * (z_i - \mu)') * W * M^{-1} \quad (13)$$

3. Calculating $E_{p(X,Y|X^o)}\{y_i * (x_i - \mu)'\}$

$$E_{p(X,Y|X^o)}\{y_i * (x_i - \mu)'\} = \iint y_i * (x_i - \mu)' * p(x_i|x_i{}^o) * p(y_i|x_i) * dx_i * dy_i$$
$$= \int (\int y_i * p(y_i|x_i) * dy_i) * (x_i - \mu)' * p(x_i|x_i{}^o) * dx_i$$
$$= \int E\{y_i\} * (x_i - \mu)' * p(x_i|x_i{}^o) * dx_i$$
$$= \int M^{-1} * W' * (x_i - \mu) * (x_i - \mu)' * p(x_i|x_i{}^o) * dx_i =$$
$$M^{-1} * W' \int (x_i - \mu) * (x_i - \mu)' * p(x_i|x_i{}^o) * dx_i =$$
$$M^{-1} * W' E_{p(X|X^o)}\{(x_i - \mu) * (x_i - \mu)'\} = M^{-1} * W' * (Q + (z_i - \mu) * (z_i - \mu)') \quad (14)$$

Maximization step (maximization the value of equation 11, treating the expectations as constants):

First of all, note that μ is the average of the whole data (since y is zero centred, so when summed across all samples, $W * y_i$ is zero), which has to be the average of the individual data samples (in which some values were inferred). Hence we have that

$$\mu = \sum_{i=1}^{N} \widehat{z_i} \quad (15)$$

Note that we used $\widehat{z_i}$ instead of z_i because we organized z_i such that the bottom part of the matrix corresponds to the unobserved part of the data. However, since in different data samples, different parts might be missing, after inference, we have to reorder z_i back to the original structure of the data. This is denoted by $\widehat{z_i}$. Note that μ must be updated at every maximization step, since it depends on D, which in turn depends on W that is also updated.

Update for W (using equation (11)):

$$\frac{\delta}{\delta W} \sum_{i=1}^{N} \{\frac{1}{2\sigma^2} [-2 * tr(E\{y_i * (x_i - \mu)'\} * W) + tr(W'W * E\{y_i * y_i'\})]\} = 0$$
$$= \sum_{i=1}^{N} \{-2 * E\{y_i * (x_i - \mu)'\}' + 2 * W * E\{y_i * y_i'\}\} \rightarrow$$
$$W = \sum_{i=1}^{N} E\{y_i * (x_i - \mu)'\}' * \sum_{i=1}^{N} (E\{y_i * y_i'\})^{-1} \quad (16)$$

Update for σ^2 (using equation (11)):

$$\frac{\delta}{\delta\sigma^2} - \frac{NF}{2} * \ln(\sigma^2) +$$

$$\frac{\delta}{\delta\sigma^2}(-\sum_{i=1}^{N}\{\frac{1}{2\sigma^2}[tr(E\{(x_i-\mu)*(x_i-\mu)'\}) - 2*tr(E\{y_i*(x_i-\mu)'\}*W)$$

$$+tr(W'W*E\{y_i*y_i'\})]\}) = 0 \rightarrow$$

$$\frac{NF}{2*\sigma^2} = \frac{1}{2\sigma^4}\sum_{i=1}^{N}\{tr(E\{(x_i-\mu)*(x_i-\mu)'\}) - 2*tr(E\{y_i*(x_i-\mu)'\}*W)$$

$$+tr(W'W*E\{y_i*y_i'\})\} \rightarrow$$

$$\sigma^2 = \frac{1}{NF}*\sum_{i=1}^{N}\{tr(E\{(x_i-\mu)*(x_i-\mu)'\}) - 2*tr(E\{y_i*(x_i-\mu)'\}*W)$$

$$+tr(W'W*E\{y_i*y_i'\})\} \qquad (17)$$

Summary:

The EM updates are given by (15)-(17) and the required expectations are given by (12)-(14).