1 Fitting the Probabilistic Principal Components and Co-variates Analysis model via the EM algorithm.

The EM algorithm is an iterative algorithm which provides maximum likelihood estimates of model parameters by alternating between E (expectation) and M (maximization) steps. The EM algorithm is particularly useful in the presence of missing data. In the context of the PPCCA model, the latent location variables (or ‘scores’) $u_i$ are treated as ‘missing’ data. Combining the observed data and the missing data for subject $i$ gives the complete data $(x_i, u_i)$. Given that $p(x_i|u_i) = MV N_p(Wu_i + \mu, \sigma^2 I)$ and $p(u_i) = MV N_q(\delta_i, I)$, the corresponding complete data log likelihood function is

$$\ell_c = \sum_{i=1}^{n} \left[ \ln MV N_p(Wu_i + \mu, \sigma^2 I) + \ln MV N(\delta_i, I) \right].$$

Maximum likelihood estimates of the model parameters are obtained by maximizing $\ell_c$ via the E and M steps of the EM algorithm. In the E-step the expected value of the complete data log likelihood is computed, conditional on the observed data and all current parameter values i.e.

$$E[\ell_c] = A + \frac{np}{2} \ln(\sigma^{-2}) - \frac{1}{2} \sum_{i=1}^{n} (\sigma^{-2}(x_i - \mu)^T(x_i - \mu)$$
$$-2\sigma^{-2}(x_i - \mu)^TWE[u_i] + \sigma^{-2}tr(W^TWE[u_i]u_i^T))$$
$$+E[(u_i - \delta_i)^T(u_i - \delta_i)]$$

where $A = \frac{n(p+q)}{2} \ln(2\pi)$. Hence, the E-step of the EM algorithm is completed by computing the conditional expectations:

$$E[u_i] = M^{-1}W^T(x_i - \mu) + \sigma^2 M^{-1}\delta_i$$
$$E[u_iu_i^T] = \sigma^2 M^{-1} + E[u_i](E[u_i])^T$$

where $M = W^TW + \sigma^2 I$.

In the M-step of the EM algorithm the expected complete data log likelihood is maximized with respect to the model parameters, $\alpha$, $W$ and $\sigma^2$. The MLE $\hat{\mu}$ is given by the mean of the data $\bar{x}$. Maximizing $E[\ell_c]$ with respect to $\alpha$, $W$ and $\sigma^2$ provides the estimates:

$$\hat{\alpha} = \sum_{i=1}^{n} E[u_i]C_i^T \left( \sum_{i=1}^{n} C_iC_i^T \right)^{-1}$$
$$\hat{W} = \left[ \sum_{i=1}^{n} (x_i - \mu)(E[u_i])^T \right] \left( \sum_{i=1}^{n} E[u_iu_i^T] \right)^{-1}$$
\[ \sigma^2 = \frac{1}{np} \left[ np - 2 \sum_{i=1}^{n} (x_i - \mu)^T W \mathbb{E}[u_i] + tr \left( W^T W \mathbb{E}[u_i u_i^T] \right) \right] \]

where \( S = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^T \) and \( \delta_i = \alpha C_i \) is as defined in (3). The EM algorithm alternates between the E and M-steps until convergence; on convergence, the estimates are the maximum likelihood estimates of the model parameters \( \hat{\alpha}, \hat{W}, \hat{\sigma}^2 \), and the estimated value of \( u_i \) is the latent score of subject \( i \). Here algorithmic convergence is assessed by Aitken’s acceleration criterion [1].

As the PPCA model is equivalent to the PPCCA model with \( \alpha = 0 \), to fit the PPCA model the E and M-steps proceed as outlined above with \( \alpha \) constrained to 0.

The BIC model selection criterion displays noisy behaviour when used to select the optimal number of principal components for the PPCA model and the PPCCA model for high dimensional metabolomic data. This issue is eradicated by employing a modified version of the BIC where a conjugate prior for the error variance \( \sigma^2 \) is employed. Thus at the M-step of the EM algorithm the sum of the expected complete data log likelihood and the log prior is maximized with respect to \( \sigma^2 \), stabilizing the algorithm. The BIC is then evaluated at the MLEs \( \hat{\alpha}, \hat{W} \) and at the posterior mode \( \sigma^2_{MAP} \). In practice an Inverse Gamma(\( \nu/2, \zeta/2 \)) prior on \( \sigma^2 \) is used leading to the posterior mode estimate:

\[ \sigma^2_{MAP} = \frac{Np\hat{\sigma}^2 + \zeta}{Np + \nu + 2} \]

The values of the hyperparameters (\( \nu \) and \( \zeta \)) are chosen based on prior knowledge and/or such that the obtained modified BIC values are similar to the actual BIC values when singularities and noisy behaviour are absent.

### 2 Fitting a mixture of PPCA models via the AECM algorithm.

The Alternating Expectation Conditional Maximization (AECM) algorithm is a variant of the EM algorithm which has two cycles. Each cycle consists of an E-step and an M-step. The data specified as “missing” differs at each cycle. Similarly, at each M-step the expected complete data log likelihood is maximized with respect to a different set of parameters. Hence the algorithm alternates between expectation steps and conditional maximization steps. Dividing the missing data and parameters in this way often provides a more feasible computational approach to deriving MLEs than using a standard EM algorithm.

In the MPPCA model, the missing data consist of both the latent locations \( u_{i1}, \ldots, u_{iG} \) and the latent group membership indicator variables \( z_{i1} \) for each subject \( i \). When estimating the parameters of the MPPCA model via the AECM algorithm, the first cycle considers the latent group membership indicator variables \( z_{i1} \) as missing and conditions on the current values of the latent locations \( u_{i1}, \ldots, u_{iG} \). Hence in the first cycle the ‘complete data’ for subject \( i \) is specified to be \( (x_i, z_{i1}) \). In the E-step of the first cycle the expected complete data log likelihood \( \mathbb{E}[\ell|z_{i1}] \) is computed with respect to the distribution of \( z_{i1} \) conditional on \( x_i, u_{i1}, \ldots, u_{iG} \) and all other current parameter values. The density of the complete data for subject \( i \) \( (x_i, z_{i1}) \) is then

\[
p(x_i, z_{i1}) = p(x_i|z_{i1})p(z_{i1}) \]

\[
= \prod_{g=1}^{G} [\pi_g p(x_i|z_{ig} = 1)]^{z_{ig}}
\]
where \( p(x_i|z_{ig} = 1) = MVN(\mu_g, W_gW_g^T + \sigma^2 I) \) is a single PPCA model fitted to the spectra of subjects currently assigned to group \( g \). Note that the variance parameter \( \sigma^2 \) is constrained to be equal across all groups for computational stability. The expected value of the complete data log likelihood with respect to the distribution of the missing data is then

\[
E[\ell_i] = \sum_{i=1}^{n} \sum_{g=1}^{G} \mathbb{E}[z_{ig}] \left( \ln \pi_g + \ln p(x_i|z_{ig} = 1) \right)
\]

(1)

Therefore the E-step of the first cycle of the AECM algorithm is completed by computing the expected value of the multinomial random variable \( z_{ig} \):

\[
\mathbb{E}[z_{ig}] = \tau_{ig} = \frac{\pi_g p(x_i|z_{ig} = 1)}{\sum_{g'=1}^{G} \pi_{g'} p(x_i|z_{ig'} = 1)}
\]

In the M-step of the first cycle the expected complete data log likelihood (1) is maximized with respect to \( \mu_g \) and \( \pi_g \) for \( g = 1, \ldots, G \) giving the estimates

\[
\hat{\mu}_g = \frac{\sum_{i=1}^{n} \tau_{ig} x_i}{n_g}, \quad \hat{\pi}_g = \frac{n_g}{n}
\]

where \( n_g = \sum_{i=1}^{n} \tau_{ig} \). This completes the first cycle of the AECM algorithm.

In the second cycle both \( z_i \) and \( u_{ig1}, \ldots, u_{igG} \) are specified as missing data for subject \( i \). In the E-step of the second cycle the expected complete data log likelihood \( E[\ell_i] \) is computed with respect to the conditional distribution of \( u_{ig} \) and \( z_i \) for \( i = 1, \ldots, n \) and \( g = 1, \ldots, G \):

\[
E[\ell_i] = B + \sum_{g=1}^{G} \left[ n_g p^{(1)} \left( \frac{1}{2} \ln(\sigma^{-2}) - 0.5n_g \sigma^{-2} tr(S_g) + \sigma^{-2} \sum_{i=1}^{n} \tau_{ig}(x_i - \hat{\mu}_g)^T W_g \mathbb{E}[u_{ig}] \right) \right]
\]

\[
-0.5\sigma^{-2} \sum_{i=1}^{n} \tau_{ig} tr(W_g^T W_g \mathbb{E}[u_{ig}^2])
\]

where \( S_g = \sum_{i=1}^{n} \tau_{ig}(x_i - \hat{\mu}_g)(x_i - \hat{\mu}_g)^T \) and \( B \) is a constant. The E-step of the second cycle is therefore completed by recomputing \( \tau_{ig} \) along with \( \mathbb{E}[u_{ig}] \) and \( \mathbb{E}[u_{ig}^2] \) for \( i = 1, \ldots, n \) and \( g = 1, \ldots, G \) which are given as

\[
\mathbb{E}[u_{ig}] = M_g^{-1} W_g^T (x_i - \hat{\mu}_g)
\]

\[
\mathbb{E}[u_{ig}^2] = \sigma^2 M_g^{-1} + \mathbb{E}[u_{ig}](\mathbb{E}[u_{ig}])^T
\]

where \( M_g = W_g^T W_g + \sigma^2 I \).

In the M-step of the second cycle the expected complete data log likelihood is maximized with respect to \( W_g \) for \( g = 1, \ldots, G \) and \( \sigma^2 \) to give

\[
\hat{W}_g = S_g W_g (\sigma^2 I + M_g^{-1} W_g^T S_g W_g)^{-1}
\]

\[
\hat{\sigma}^2 = \frac{1}{p} \sum_{g=1}^{G} \hat{\pi}_g \left[ tr(S_g) - tr \left( S_g W_g M_g^{-1} \hat{W}_g \right) \right]
\]
where $\hat{W}_g$ is the value of $W_g$ after the second E-step but prior to computing $\hat{W}_g$. This completes the second cycle.

The AECM algorithm alternates between the two cycles detailed above until convergence (here assessed by Aitken’s acceleration criterion). At convergence the parameter estimates obtained are the maximum likelihood estimates of the MPPCA model and the estimated value of $u_{ig}$ is the latent score of subject $i$, given that $i$ is a member of group $g$. The estimated values of $z_i = (z_{i1}, \ldots, z_{iG})$ at convergence are the posterior probabilities of subject $i$ being a member of group $g$; typically each subject is classified to the group for which they have highest posterior probability of membership. Thus both clustering of subjects into groups and dimension reduction are simultaneously achieved.

For the MPPCA model, a modified version of the BIC is employed to select the optimal model, both in terms of the number of groups $G$ and the number of principal components $q$. Similar to the PPCA and PPCCA cases detailed above, a conjugate prior for the error variance $\sigma^2$ is employed in the MPPCA model. Thus at the second M-step of the AECM algorithm the sum of the expected complete data log likelihood and the log prior is maximized with respect to $\sigma^2$. The BIC is then evaluated at the MLEs $\hat{z}_g$, $\hat{\mu}_g$, and $\hat{W}_g$ for $g = 1, \ldots, G$ and at the posterior mode $\sigma^2_{MAP}$. In practice an Inverse Gamma($\nu/2, \zeta/2$) prior on $\sigma^2$ is used leading to the posterior mode estimate:

$$\sigma^2_{MAP} = \frac{N(p\hat{\sigma}^2 + \zeta)}{Np + \nu + 2}.$$ 

The values of the hyperparameters ($\nu$ and $\zeta$) are chosen based on prior knowledge and/or such that the obtained modified BIC values are similar to the actual BIC values when singularities and noisy behaviour are absent.

References