Expectation Maximization for fitting decision making models with Gaussian priors

$$
\begin{aligned}
X = \{X_i\} &\quad : \quad \text{observed data, trial responses} \\
Z = \{Z_i\} &\quad : \quad \text{decision model parameters for all subjects} \\
Z_i = \{Z_{i,d}\} &\quad : \quad \text{decision model parameters for each subject, } \forall d = \{1, ...D\} \\
\theta = \{\mu, \Sigma\} &\quad : \quad \text{prior parameters, } \mu = \{\mu_d\}, \Sigma = \{\sigma_d^2\}, \forall d = \{1, ...D\} \\
Z_i \sim N(\mu, \Sigma) &\quad : \quad \text{Gaussian priors assumption} \\
Z_{i,d} \sim N(\mu_d, \sigma_d^2) &\quad : \quad \text{Prior parameters assumed to be independent of each other, } \Sigma \text{ is diagonal}
\end{aligned}
$$

We are looking for the parameters of the Gaussian prior, θ, that maximize the observed data likelihood :

$$
\begin{aligned}
\theta^* &= \underset{\theta}{\operatorname{argmax}} P(X|\theta) \\
&= \underset{\theta}{\operatorname{argmax}} \int P(X, Z|\theta) dZ \\
&= \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{N} \int P(X_i, Z_i|\theta) dZ_i
\end{aligned} \tag{1}
$$

Equivalently we can solve :

$$
\theta^* = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} \ln \int P(X_i, Z_i|\theta) dZ_i \tag{2}
$$

For each subject i we have :

$$
\ln \int P(X_i, Z_i|\theta) dZ_i = \ln \int Q(Z_i) \frac{P(X_i, Z_i|\theta)}{Q(Z_i)} dZ_i \geq \int Q(Z_i) \ln \frac{P(X_i, Z_i|\theta)}{Q(Z_i)} dZ_i \tag{3}
$$

The inequality in (3) holds by Jensen's inequality for a concave function $f(X_i, Z_i, \theta) = \frac{P(X_i, Z_i|\theta)}{Q(Z_i)}$. The equality holds only for

$$
Q(Z_i) = P(Z_i|X_i, \theta) \tag{4}
$$

where $P(Z_i|X_i, \theta)$ is the true posterior distribution of a subject's model parameters, given the subject's observed data X_i and the true prior parameters θ. We will approximate the posterior distribution in each iteration k of the EM algorithm, during the E-step, with $P(Z_i|X_i, \hat{\theta}_k)$.

$$
\begin{aligned}
&\ln \int P(X_i, Z_i|\theta) dZ_i \geq \int P(Z_i|X_i, \hat{\theta}_{k-1}) \ln \frac{P(X_i, Z_i|\theta)}{P(Z_i|X_i, \hat{\theta}_{k-1})} dZ_i \\
\Rightarrow &\ln \int P(X_i, Z_i|\theta) dZ_i \geq \int P(Z_i|X_i, \hat{\theta}_{k-1}) \ln P(X_i, Z_i|\theta) dZ_i + \int P(Z_i|X_i, \hat{\theta}_{k-1}) \ln P(Z_i|X_i, \hat{\theta}_{k-1}) dZ_i \qquad (5)
\end{aligned}
$$

ignoring the second integral in (5), the approximate posterior entropy, as it is independent of θ, which we are optimising for, we have the expectation

$$
E_{i,k} = \int P(Z_i|X_i, \hat{\theta}_{k-1}) \ln P(X_i, Z_i|\theta) dZ_i \qquad (6)
$$

This expectation is a lower bound on the logarithm of the marginal likelihood function (also known as model evidence) for each subject $\ln P(X_i, Z_i|\theta)$, if we first account for the offset by the approximate posterior entropy, the second integral in (5). The approximate posterior for each iteration is what is needed to complete the E-step. Here we are using the Laplace approximation, which assumes that the posterior is a Gaussian distribution. By Bayes' rule we have

$$
P(Z_i|X_i, \hat{\theta}_{k-1}) = \frac{P(X_i|Z_i, \hat{\theta}_{k-1}) P(Z_i|\hat{\theta}_{k-1})}{P(X_i|\hat{\theta}_{k-1})} \qquad (7)
$$

and by Laplace's approximation

$$
\begin{aligned}
&P(Z_i|X_i, \hat{\theta}_{k-1}) \sim N(m_{i,k}, \Phi_{i,k}) &(8)\\
&m_{i,k} = \underset{Z_i}{\operatorname{argmax}}\, P(X_i|Z_i, \hat{\theta}_{k-1}) P(Z_i|\hat{\theta}_{k-1}) &(9)\\
&\Phi_{i,k} = -H_{i,k}^{-1} &(10)
\end{aligned}
$$

where $H_{i,k} = -\nabla\nabla \ln(P(X_i|Z_i, \hat{\theta}_{k-1}) P(Z_i|\hat{\theta}_{k-1}))|_{Z_i = m_{i,k}}$ is the Hessian matrix around the mode. For a derivation of (9) and (10) see [**Pattern Recognition and Machine Learning**].

After calculating the expectation (6), we are maximizing it for all subjects, during the algorithm's M-step. First, let us rewrite the expectation for one subject, then (6) becomes

$$
E_{i,k} = \int P(Z_i|X_i, \hat{\theta}_{k-1}) \ln P(X_i|Z_i) dZ_i + \int P(Z_i|X_i, \hat{\theta}_{k-1}) \ln P(Z_i|\theta) dZ_i \qquad (11)
$$

Note that the likelihood depends only on each subject's parameters and not their priors, $P(X_i|Z_i, \theta) = P(X_i|Z_i)$. Focusing on the second integral in (11), as the only term dependent on θ we have

$$
\begin{aligned}
\int P(Z_i|X_i, \hat{\theta}_{k-1}) \ln P(Z_i|\theta) dZ_i &= \frac{1}{2}(D \ln 2\pi + \ln|\Sigma| + \mathcal{E}[Z_i^T \Sigma^{-1} Z_i] - \mu \Sigma^{-1} \mathcal{E}[Z_i] - \mathcal{E}[Z_i^T] \Sigma^{-1} \mu + \mu^T \Sigma^{-1} \mu) \\
&= \frac{1}{2}(D \ln 2\pi + \ln|\Sigma| + Tr[\Sigma^{-1}(m_{i,k} m_{i,k}^T + \Phi_{i,k})] - \mu^T \Sigma^{-1} m_{i,k} - m_{i,k}^T \Sigma^{-1} \mu + \mu^T \Sigma^{-1} \mu)
\end{aligned} \qquad (12)
$$

where $|\Sigma|$ is the determinant of the covariance matrix, Tr is the trace operation and $\mathcal{E}[Z_i] = \int Z_i P(Z_i|X_i, \hat{\theta}_{k-1}) dZ_i$ is the expectation operation under the approximate posterior. For a derivation of (12) see [**Cross entropy of Two Normal Distributions**]. To finish an iteration of EM, we are going to maximize the approximation to the original likelihood (2). Substituting (5) in (2)

$$\theta^* \approx \hat{\theta}_k = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} [\int P(Z_i|X_i, \hat{\theta}_{k-1}) \ln P(X_i, Z_i|\theta) dZ_i + \int P(Z_i|X_i, \hat{\theta}_{k-1}) \ln P(Z_i|X_i, \hat{\theta}_{k-1}) dZ_i]$$

$$= \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} [\int P(Z_i|X_i, \hat{\theta}_{k-1}) \ln P(X_i, Z_i|\theta) dZ_i] + const_1 \tag{13}$$

now substituting (6) and then (11) into (13)

$$\hat{\theta}_k = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} [\int P(Z_i|X_i, \hat{\theta}_{k-1}) \ln P(Z_i|\theta) dZ_i] + const_2 = \underset{\theta}{\operatorname{argmax}} \mathcal{L}(X, Z|\theta) + const_2 \tag{14}$$

where the two integrals that do not depend on θ are represented in the *const* terms. An analytical solution can be calculated for the optimal solution of (14) as

$$\frac{\partial \mathcal{L}(X, Z|\theta)}{\partial \mu} = 0 \Rightarrow \hat{\mu}_k = \frac{1}{N} \sum_{i=1}^{N} m_{i,k} \tag{15}$$

$$\frac{\partial \mathcal{L}(X, Z|\theta)}{\partial \Sigma} = 0 \Rightarrow \hat{\Sigma}_k = diag\{\frac{1}{N} \sum_{i=1}^{N} (m_{i,k} m_{i,k}^T + \Phi_{i,k}) - \mu_k \mu_k^T\} \tag{16}$$

where $diag\{X\}$ are the diagonal elements of matrix X and $\hat{\theta}_k = \{\hat{\mu}_k, \hat{\Sigma}_k\}$.

The above process (6)-(16) is performed iteratively until $\hat{\theta}_k$ converges. For a proof of convergence of Expectation Maximization see [**What is the expectation maximization algorithm? Supplementary material**].

References

Christopher M. Bishop. 2006. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, Berlin, Heidelberg.

Cross Entropy of Two Normal Distribution, https://www.cse.iitb.ac.in/ aruniyer/kldivergencenormal.pdf

Do, C., Batzoglou, S. What is the expectation maximization algorithm?. Nat Biotechnol 26, 897–899 (2008)