Suppose we observe \(2mn\) data points \(Y_{ijk}\), and we are interested in the linear model

\[ Y_{ijk} = \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}, \]

where \(i = 1, \ldots, m\) and \(j = 1, \ldots, n\) correspond to various levels of two experimental factors, and \(k = 1, 2\) correspond to two independent replicates. Our main aim is to quantify the extent to which interactions (i.e. the \(\gamma_{ij}\) values) are present, rather than estimate the values of specific interaction coefficients. Therefore the \(\gamma_{ij}\) are taken as random effects from a normal distribution with mean 0 and variance \(\tau^2\). Our goal is to calculate the MLE’s of the model parameters \(\alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_n, \sigma^2\) and \(\tau^2\). To make the parameters identified we will constrain \(\sum_j \beta_j = 0\).

1. Derive a formula for the log likelihood \(\log P(\{Y\}|\sigma^2, \tau^2, \alpha, \beta)\).

**Solution:** Since the \(\gamma_{ij}\) and \(\epsilon_{ijk}\) are Gaussian, the \(Y_{ijk}\) are also Gaussian. The expected values are

\[ EY_{ijk} = \alpha_i + \beta_j \]

and the variances are

\[ \text{var}Y_{ijk} = \tau^2 + \sigma^2. \]

Since \(Y_{ij1}\) and \(Y_{ij2}\) share the same \(\gamma_{ij}\), they are dependent with covariance \(\tau^2\). Any pair of observations \(Y_{ijk}\) and \(Y_{ij'k'}\) are independent unless \(i = i'\) and \(j = j'\). Thus the log-likelihood is a sum of bivariate normal log-likelihoods

\[ -\frac{mn}{2} \log |\Sigma| - \sum_{ij} \frac{1}{2} Z_{ij}' \Sigma^{-1} Z_{ij}, \]

where

\[ Z_{ij} = \begin{pmatrix} Y_{ij1} - \alpha_i - \beta_j \\ Y_{ij2} - \alpha_i - \beta_j \end{pmatrix} \]

and

\[ \Sigma = \begin{pmatrix} \sigma^2 + \tau^2 & \tau^2 \\ \tau^2 & \sigma^2 + \tau^2 \end{pmatrix}. \]
2. Derive the EM algorithm for estimating the model parameters.

**Solution:** Using the $\gamma_{ij}$ to augment the observed data $Y_{ij}$ yields the “complete likelihood”

$$-mn \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{ijk} (Y_{ijk} - \alpha_i - \beta_j - \gamma_{ij})^2 - \frac{mn}{2} \log \tau^2 - \frac{1}{2\tau^2} \sum_{ij} \gamma_{ij}^2$$

We will also need the conditional density

$$P(\{\gamma_{ij}\}|\{Y_{ijk}\}) \propto P(\{\gamma_{ij}\}|\{Y_{ijk}\})P(\gamma_{ij})$$

$$\log P(\{\gamma_{ij}\}|\{Y_{ijk}\}) = C - \frac{1}{2\sigma^2} \sum_{ijk} (Y_{ijk} - \alpha_i - \beta_j - \gamma_{ij})^2 - \frac{1}{2\tau^2} \sum_{ij} \gamma_{ij}^2$$

Since there are no products among distinct $\gamma_{ij}$ in this expression, the $\gamma_{ij}$ are conditionally independent. The log density for a given $\gamma_{ij}$ is

$$C - \frac{1}{2} \gamma_{ij}^2 (2/\sigma^2 + 1/\tau^2) + \gamma_{ij} \sum_{k=1}^2 (Y_{ijk} - \alpha_i - \beta_j)/\sigma^2.$$ 

Thus each $\gamma_{ij}$ is conditionally normal with variance

$$V_{ij} \equiv \frac{1}{2/\sigma^2 + 1/\tau^2}$$

and mean

$$M_{ij} \equiv \frac{\sum_{k=1}^2 (Y_{ijk} - \alpha_i - \beta_j)}{2 + \sigma^2/\tau^2}$$

Now rewrite the complete data log-likelihood as

$$-mn \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{ijk} \left( (Y_{ijk} - \alpha_i - \beta_j)^2 - 2(Y_{ijk} - \alpha_i - \beta_j)\gamma_{ij} + \gamma_{ij}^2 \right) - \frac{mn}{2} \log \tau^2 - \frac{1}{2\tau^2} \sum_{ij} \gamma_{ij}^2$$

Taking the conditional mean of this given the data yields
\[ Q = -mn \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{ijk} \left( (Y_{ijk} - \alpha_i - \beta_j)^2 - 2(Y_{ijk} - \alpha_i - \beta_j)M_{ij} + (V_{ij} + M_{ij}^2) \right) - \frac{mn}{2} \log \tau^2 - \frac{1}{2\tau^2} \sum_{ij} (V_{ij} + M_{ij}^2). \]

To maximize over the \( \alpha_i \) and \( \beta_j \) parameters, we only need to consider

\[ -\sum_{ijk} \left( (Y_{ijk} - \alpha_i - \beta_j)^2 - 2(Y_{ijk} - \alpha_i - \beta_j)M_{ij} \right). \]

The derivatives are

\[
d/d\alpha_i = \sum_{jk} 2(Y_{ijk} - \alpha_i - \beta_j) - 2M_{ij} \]
\[
d/d\beta_j = \sum_{ik} 2(Y_{ijk} - \alpha_i - \beta_j) - 2M_{ij} \]

so the updated \( \alpha_i \) and \( \beta_j \) values following the M step are

\[
\alpha_i \leftarrow (Y_{i.} - 2M_{i.})/2n
\]
\[
u_j \leftarrow (Y_{.j} - 2\alpha_i - 2M_{.j})/2m.
\]
\[
\beta_j \leftarrow u_j - \bar{u}
\]

The updates for the variance parameters are

\[
\tau^2 \leftarrow \frac{1}{mn} \sum_{ij} (V_{ij} + M_{ij}^2)
\]
\[
\sigma^2 \leftarrow \frac{1}{2mn} \sum_{ijk} \left( (Y_{ijk} - \alpha_i - \beta_j)^2 - 2(Y_{ijk} - \alpha_i - \beta_j)M_{ij} + (V_{ij} + M_{ij}^2) \right)
\]
3. Simulate data from the true distribution and assess the bias and variance of the MLE for one or more population structures of your choosing. Also check that your implementation of the EM algorithm is nondecreasing in \( \log \text{P}(Y|\sigma^2, \tau^2, \alpha, \beta) \).

**Solution:** See the file `2006_3.m` for the solution written in Octave. You will see that the bias in the \( \alpha_i \) and \( \beta_j \) is minimal for the given population structure (and this is true more generally). Both \( \sigma^2 \) and \( \tau^2 \) are slightly negatively biased. This is also generally true, with the bias being more severe when \( \tau^2 \) is small relative to \( \sigma^2 \). The sampling variance of the \( \alpha_i \) and \( \beta_j \) parameters is greater than that for \( \sigma^2 \) and \( \tau^2 \). This is expected since more data contribute to estimation of the variance parameters.