The EM Algorithms

Estimating Haplotype Frequencies

Based on likelihood (8), the EM algorithm is derived to estimate haplotype frequencies. In the E step, calculate the proportions of a particular diplotype within double or triple heterozygous genotypes in the population of sex $s$ by

\[
\phi_1^s = \frac{p_{111}^s p_{100}^s}{p_{111}^s p_{100}^s + p_{101}^s p_{110}^s}, \quad \bar{\phi}_1^s = 1 - \phi_1^s \quad \text{for genotype 11/10/10}
\]

\[
\phi_2^s = \frac{p_{111}^s p_{010}^s}{p_{111}^s p_{010}^s + p_{011}^s p_{110}^s}, \quad \bar{\phi}_2^s = 1 - \phi_2^s \quad \text{for genotype 10/11/10}
\]

\[
\phi_3^s = \frac{p_{111}^s p_{001}^s}{p_{111}^s p_{001}^s + p_{101}^s p_{011}^s}, \quad \bar{\phi}_3^s = 1 - \phi_3^s \quad \text{for genotype 10/10/11}
\]

\[
\begin{align*}
\phi_4^s &= \frac{p_{111}^s p_{000}^s}{p_{111}^s p_{000}^s + p_{101}^s p_{010}^s + p_{110}^s p_{001}^s + p_{100}^s p_{011}^s} \\
\bar{\phi}_4^s &= \frac{p_{101}^s p_{010}^s}{p_{111}^s p_{000}^s + p_{101}^s p_{010}^s + p_{110}^s p_{001}^s + p_{100}^s p_{011}^s} & \text{for genotype 10/10/10} \\
\phi_5^s &= \frac{p_{110}^s p_{000}^s}{p_{110}^s p_{000}^s + p_{111}^s p_{001}^s + p_{101}^s p_{010}^s + p_{110}^s p_{001}^s + p_{100}^s p_{011}^s} \\
\bar{\phi}_5^s &= 1 - \phi_5^s & \text{for genotype 10/10/00} \\
\phi_6^s &= \frac{p_{101}^s p_{000}^s}{p_{101}^s p_{000}^s + p_{001}^s p_{100}^s}, \quad \bar{\phi}_6^s = 1 - \phi_6^s & \text{for genotype 10/00/10} \\
\phi_7^s &= \frac{p_{011}^s p_{000}^s}{p_{011}^s p_{000}^s + p_{001}^s p_{100}^s}, \quad \bar{\phi}_7^s = 1 - \phi_7^s & \text{for genotype 00/10/10}
\end{align*}
\]

In the M step, estimate the haplotype frequencies with the calculated relative proportions
by

\[
\hat{p}_{111} = \frac{1}{2n}(2n_{11/11} + n_{11/11/10} + n_{11/10/11} + n_{10/11/11} + \phi_1 n_{11/10/10} + \phi_2 n_{11/10/11} + \phi_3 n_{10/10/10} + \phi_4 n_{10/10/10} + \phi_5 n_{1/1/10})
\]

\[
\hat{p}^s_{110} = \frac{1}{2n}(2n_{11/11/10} + n_{11/11/10} + n_{11/10/10} + n_{10/11/10} + \phi_1 n_{11/10/10} + \phi_2 n_{11/10/10} + \phi_3 n_{10/10/10} + \phi_4 n_{10/10/10} + \phi_5 n_{1/1/10})
\]

\[
\hat{p}^s_{110} = \frac{1}{2n}(2n_{11/11/10} + n_{11/11/10} + n_{11/10/10} + n_{10/11/10} + \phi_1 n_{11/10/10} + \phi_2 n_{11/10/10} + \phi_3 n_{10/10/10} + \phi_4 n_{10/10/10} + \phi_5 n_{1/1/10})
\]

\[
\hat{p}^s_{100} = \frac{1}{2n}(2n_{11/10/10} + n_{11/10/10} + n_{11/10/10} + n_{10/10/10} + \phi_1 n_{11/10/10} + \phi_2 n_{11/10/10} + \phi_3 n_{10/10/10} + \phi_4 n_{10/10/10} + \phi_5 n_{1/1/10})
\]

\[
\hat{p}^s_{011} = \frac{1}{2n}(2n_{10/11/11} + n_{10/11/11} + n_{10/11/11} + n_{10/11/10} + \phi_1 n_{11/10/10} + \phi_2 n_{11/10/11} + \phi_3 n_{10/10/11} + \phi_4 n_{10/10/11} + \phi_5 n_{1/1/11})
\]

\[
\hat{p}^s_{010} = \frac{1}{2n}(2n_{10/11/10} + n_{10/11/10} + n_{10/11/10} + n_{10/10/10} + \phi_1 n_{11/10/10} + \phi_2 n_{11/10/10} + \phi_3 n_{10/10/10} + \phi_4 n_{10/10/10} + \phi_5 n_{1/1/10})
\]

\[
\hat{p}^s_{001} = \frac{1}{2n}(2n_{10/00/11} + n_{10/00/11} + n_{10/00/11} + n_{10/00/10} + \phi_1 n_{11/10/10} + \phi_2 n_{11/10/10} + \phi_3 n_{10/10/10} + \phi_4 n_{10/10/10} + \phi_5 n_{1/1/10})
\]

\[
\hat{p}^s_{000} = \frac{1}{2n}(2n_{10/00/10} + n_{10/00/10} + n_{10/00/10} + n_{10/00/00} + \phi_1 n_{11/10/10} + \phi_2 n_{11/10/10} + \phi_3 n_{10/10/10} + \phi_4 n_{10/10/10} + \phi_5 n_{1/1/10})
\]

The E and M steps are iterated until the estimates of haplotype frequencies are stable.

**Estimating Quantitative Genetic Parameters**

Based on likelihood (9), the EM algorithm is derived to estimate the genetic values of composite diplotypes and residual variance. In the E step, calculate the posterior probabilities with which a double or triple heterozygous subject \( i \) is a particular diplotype, expressed
In the M step, estimate the quantitative genetic parameters, $\Omega_q$, with the calculated
posterior probabilities by

\[ \mu_{11} = \frac{\sum_{i=1}^{n_{11/11}} y_i}{n_{11/11}}. \]

\[ \mu_{10} = \sum_{i=1}^{n_{11/11}} \Psi_{11}^{10} y_i + \sum_{i=1}^{n_{11/10/11}} \Psi_{12}^{10} y_i + \sum_{i=1}^{n_{11/10/10}} \Psi_{13}^{10} y_i + \sum_{i=1}^{n_{10/11/11}} \Psi_{14}^{10} y_i + \sum_{i=1}^{n_{10/11/10}} \Psi_{15}^{10} y_i + \sum_{i=1}^{n_{10/10/11}} \Psi_{16}^{10} y_i + \sum_{i=1}^{n_{10/10/10}} \Psi_{17}^{10} y_i \]

\[ \mu_{01} = \sum_{i=1}^{n_{11/11}} \Psi_{11}^{01} y_i + \sum_{i=1}^{n_{11/10/11}} \Psi_{12}^{01} y_i + \sum_{i=1}^{n_{11/10/10}} \Psi_{13}^{01} y_i + \sum_{i=1}^{n_{10/11/11}} \Psi_{14}^{01} y_i + \sum_{i=1}^{n_{10/11/10}} \Psi_{15}^{01} y_i + \sum_{i=1}^{n_{10/10/11}} \Psi_{16}^{01} y_i + \sum_{i=1}^{n_{10/10/10}} \Psi_{17}^{01} y_i \]

\[ \mu_{00} = \sum_{i=1}^{m} y_i + \sum_{i=1}^{n_{11/10/10}} \Psi_{11}^{00} y_i + \sum_{i=1}^{n_{11/10/10}} \Psi_{12}^{00} y_i + \sum_{i=1}^{n_{11/10/10}} \Psi_{13}^{00} y_i + \sum_{i=1}^{n_{11/10/10}} \Psi_{14}^{00} y_i + \sum_{i=1}^{n_{11/10/10}} \Psi_{15}^{00} y_i + \sum_{i=1}^{n_{11/10/10}} \Psi_{16}^{00} y_i + \sum_{i=1}^{n_{11/10/10}} \Psi_{17}^{00} y_i \]

\[ \sigma^2 = \frac{1}{n} \left[ \sum_{i=1}^{n_{11/11}} (y_i - \mu_{11})^2 + \sum_{i=1}^{n_{11/10/11}} (y_i - \mu_{10})^2 + \sum_{i=1}^{n_{11/10/10}} (y_i - \mu_{10})^2 + \sum_{i=1}^{n_{10/11/11}} (y_i - \mu_{11})^2 + \sum_{i=1}^{n_{10/11/10}} (y_i - \mu_{10})^2 + \sum_{i=1}^{n_{10/10/11}} (y_i - \mu_{10})^2 + \sum_{i=1}^{n_{10/10/10}} (y_i - \mu_{10})^2 + \sum_{i=1}^{n_{10/11/11}} (y_i - \mu_{11})^2 + \sum_{i=1}^{n_{10/11/10}} (y_i - \mu_{10})^2 + \sum_{i=1}^{n_{10/10/11}} (y_i - \mu_{10})^2 + \sum_{i=1}^{n_{10/10/10}} (y_i - \mu_{10})^2 + \sum_{i=1}^{n_{10/11/11}} (y_i - \mu_{11})^2 + \sum_{i=1}^{n_{10/11/10}} (y_i - \mu_{10})^2 + \sum_{i=1}^{n_{10/10/11}} (y_i - \mu_{10})^2 + \sum_{i=1}^{n_{10/10/10}} (y_i - \mu_{10})^2 + \sum_{i=1}^{n_{10/11/11}} (y_i - \mu_{11})^2 + \sum_{i=1}^{n_{10/11/10}} (y_i - \mu_{10})^2 + \sum_{i=1}^{n_{10/10/11}} (y_i - \mu_{10})^2 + \sum_{i=1}^{n_{10/10/10}} (y_i - \mu_{10})^2 + \sum_{i=1}^{n_{10/11/11}} (y_i - \mu_{11})^2 + \sum_{i=1}^{n_{10/11/10}} (y_i - \mu_{10})^2 + \sum_{i=1}^{n_{10/10/11}} (y_i - \mu_{10})^2 + \sum_{i=1}^{n_{10/10/10}} (y_i - \mu_{10})^2 + \sum_{i=1}^{n_{10/11/11}} (y_i - \mu_{11})^2 + \sum_{i=1}^{n_{10/11/10}} (y_i - \mu_{10})^2 + \sum_{i=1}^{n_{10/10/11}} (y_i - \mu_{10})^2 + \sum_{i=1}^{n_{10/10/10}} (y_i - \mu_{10})^2 \right] \]

The E and M steps are iterated until the estimates of haplotype frequencies are stable.