The E-M Algorithm in Genetics

Biostatistics 666
Lecture 8
Maximum Likelihood Estimation of Allele Frequencies

- Find parameter estimates which make observed data most likely

- General approach, as long as tractable likelihood function exists

- Can use all available information

- Provides justification for natural estimators
Today:

- The Expectation–Maximization algorithm in Genetics

- Frequency estimates for…
  - Recessive alleles
  - A, B, O alleles
  - Haplotype frequencies
Setting for the E-M Algorithm...

- Specific type of incomplete data
  - More possible **categories** (genotypes) than can be **distinguished** (phenotypes)

- For example, consider disease locus with recessive alleles...
  - What are the possible genotypes?
  - What are the possible phenotypes?
Setting for the E-M Algorithm...

- Any problem that is simple to solve for complete data …
  - For example, estimating allele frequencies when all genotypes can be distinguished …

- … but the available data is “incomplete” in some way.
  - For example, when disease phenotypes are observed the distinction between homozygotes and heterozygotes is missing.
The E-M Algorithm

- Consider a set of starting parameters
- Use these to “estimate” the complete data
- Use estimates of complete data to update parameters
- Repeat as necessary
An Example ...

- A random sample of 100 individuals
- 4 express a recessive phenotype
  - Assume the phenotype is controlled by a single gene
- Let’s follow E-M algorithm steps ...
Step 1:

- Set starting values for parameters
- For allele frequency estimation…
  - Equal frequencies are a common choice
  - $p_{\text{rec}} = 0.5$
- Useful to repeat process using different starting point
Step 2:

- Estimate “complete data”
- Assign phenotypes to specific genotype categories
- Use Bayes’ Theorem
Step 2 (continued):

- Calculate probability of each genotype among the 96 “normal” individuals

\[
P(+/+; \text{Normal}) = \frac{P(+/+, \text{Normal})}{P(\text{Normal})} = \frac{P(+/+, \text{Normal})}{P(+/+, \text{Normal}) + P(+/-, \text{Normal})} = \frac{P(+/+)}{P(+/+) + P(+/-)}
\]
Step 2 (Finally!):

- At the first iteration, the complete data would be filled in as:
  - 4 individuals with recessive genotype
  - 64 individuals with heterozygous genotype
  - 32 individuals with dominant genotype
Step 3:

- Estimate allele frequencies by counting...

\[ p_{rec} = \frac{N_{het} + 2N_{rec/rec}}{2N} \]

- What would be the estimated allele frequencies?
<table>
<thead>
<tr>
<th>Round</th>
<th>Estimate</th>
<th>E(+/+)</th>
<th>E(+/-)</th>
<th>E(-/-)</th>
<th>ln L</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.50</td>
<td>32.00</td>
<td>64.00</td>
<td>4.00</td>
<td>-14.40240</td>
</tr>
<tr>
<td>2</td>
<td>0.36</td>
<td>45.18</td>
<td>50.82</td>
<td>4.00</td>
<td>-9.33657</td>
</tr>
<tr>
<td>3</td>
<td>0.29</td>
<td>52.36</td>
<td>43.64</td>
<td>4.00</td>
<td>-8.02405</td>
</tr>
<tr>
<td>4</td>
<td>0.26</td>
<td>56.60</td>
<td>39.40</td>
<td>4.00</td>
<td>-7.58067</td>
</tr>
<tr>
<td>5</td>
<td>0.24</td>
<td>59.21</td>
<td>36.79</td>
<td>4.00</td>
<td>-7.41213</td>
</tr>
<tr>
<td>6</td>
<td>0.22</td>
<td>60.87</td>
<td>35.13</td>
<td>4.00</td>
<td>-7.34396</td>
</tr>
<tr>
<td>7</td>
<td>0.22</td>
<td>61.94</td>
<td>34.06</td>
<td>4.00</td>
<td>-7.31540</td>
</tr>
<tr>
<td>8</td>
<td>0.21</td>
<td>62.64</td>
<td>33.36</td>
<td>4.00</td>
<td>-7.30317</td>
</tr>
<tr>
<td>9</td>
<td>0.21</td>
<td>63.10</td>
<td>32.90</td>
<td>4.00</td>
<td>-7.29787</td>
</tr>
<tr>
<td>10</td>
<td>0.20</td>
<td>63.40</td>
<td>32.60</td>
<td>4.00</td>
<td>-7.29555</td>
</tr>
<tr>
<td>11</td>
<td>0.20</td>
<td>63.60</td>
<td>32.40</td>
<td>4.00</td>
<td>-7.29453</td>
</tr>
<tr>
<td>12</td>
<td>0.20</td>
<td>63.73</td>
<td>32.27</td>
<td>4.00</td>
<td>-7.29408</td>
</tr>
<tr>
<td>13</td>
<td>0.20</td>
<td>63.82</td>
<td>32.18</td>
<td>4.00</td>
<td>-7.29388</td>
</tr>
<tr>
<td>14</td>
<td>0.20</td>
<td>63.88</td>
<td>32.12</td>
<td>4.00</td>
<td>-7.29379</td>
</tr>
<tr>
<td>15</td>
<td>0.20</td>
<td>63.92</td>
<td>32.08</td>
<td>4.00</td>
<td>-7.29375</td>
</tr>
<tr>
<td>16</td>
<td>0.20</td>
<td>63.95</td>
<td>32.05</td>
<td>4.00</td>
<td>-7.29374</td>
</tr>
</tbody>
</table>
Alternatives

- Analytical solutions
- Generic maximization strategies
- Calculating second derivatives a useful complement, whatever method we use…
Other Applications of the E-M Algorithm in Genetics

- **Classic example:**
  - ABO blood group

- **Most common application:**
  - Haplotype frequency estimates
The ABO blood group

- Determines compatibility for transfusions
- Controlled by alleles of ABO gene
- 3 alternative alleles
  - A, B and O
- 6 possible genotypes, $n(n + 1)/2$
  - A/A, A/B, A/O, B/B, B/O, O/O
There are only 4 possible phenotypes for the ABO blood group.

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Antigen</th>
<th>Antibody</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>A</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>B</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>O</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>AB</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>
## Genotypes and Phenotypes

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>A/A</td>
<td>A</td>
</tr>
<tr>
<td>A/B</td>
<td>AB</td>
</tr>
<tr>
<td>A/O</td>
<td>A</td>
</tr>
<tr>
<td>A/O</td>
<td>A</td>
</tr>
<tr>
<td>B/B</td>
<td>B</td>
</tr>
<tr>
<td>B/O</td>
<td>B</td>
</tr>
<tr>
<td>O/O</td>
<td>O</td>
</tr>
</tbody>
</table>
ABO Example

- Data of Clarke et al. (1959)
  - *British Med J* 1:603-607
  - Reported excess of gastric ulcers in individuals with blood type O

- \( n_A = 186, n_B = 38, n_{AB} = 36, n_O = 284 \)
Quick Exercises!

- Write out the likelihood for these data…
- What are complete data categories?
- Express the complete data “counts” as a function of allele frequency estimates and the observed data…
The iterations give ...

<table>
<thead>
<tr>
<th>Iteration</th>
<th>$p_A$</th>
<th>$p_B$</th>
<th>$p_O$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.300</td>
<td>0.200</td>
<td>0.500</td>
</tr>
<tr>
<td>2</td>
<td>0.243</td>
<td>0.074</td>
<td>0.683</td>
</tr>
<tr>
<td>3</td>
<td>0.228</td>
<td>0.070</td>
<td>0.700</td>
</tr>
<tr>
<td>4</td>
<td>0.228</td>
<td>0.070</td>
<td>0.702</td>
</tr>
<tr>
<td>5</td>
<td>0.228</td>
<td>0.070</td>
<td>0.702</td>
</tr>
</tbody>
</table>
Alternatives to E-M...

- Analytical solutions are not known for the general case

- Generic maximization strategies could be employed

- Could derive solutions using part of the data…
  - Would this be a good idea?
The E-M Haplotyping Algorithm

- Excoffier and Slatkin (1995)
  - *Mol Biol Evol* **12**:921-927
  - Provide a clear outline of how the algorithm can be applied to genetic data

- Combination of two strategies
  - E-M statistical algorithm for missing data
  - Counting algorithm for allele frequencies
Original Application of the E-M Algorithm to A Genetic Problem


- This was ~20 years before the E-M algorithm was formally outlined in the statistical literature!
Counting for Allele Frequencies

- For co-dominant markers, allele frequency typically carried out in very simple manner:
  - Count number of chromosomes (e.g. 2N)
  - Count number of a alleles (e.g. na)
  - Allele frequency is simple proportion (na/2N)

- Haplotypes can’t always be counted directly
  - Focusing on unambiguous genotypes introduces bias
Counting Haplotypes for 2 SNPs

Unambiguous Genotypes
Underlying Haplotype is Known

Ambiguous Genotype
Multiple Underlying Genotypes Possible
Probabilistic Interpretation

Probability of first outcome:
\[2 P_{AB} P_{aB}\]
Probability of second outcome:
\[2 P_{AB} P_{ab}\]
Probabilistic Interpretation

For example, if:

- $P_{AB} = 0.3$
- $P_{ab} = 0.3$
- $P_{Ab} = 0.3$
- $P_{aB} = 0.1$

Probability of first outcome:

$$2 \ P_{Ab} \ P_{aB} = 0.06$$

Probability of second outcome:

$$2 \ P_{AB} \ P_{ab} = 0.18$$
Conditional probability of first outcome:
\[
\frac{2 \cdot P_{Ab} \cdot P_{aB}}{2 \cdot P_{Ab} \cdot P_{aB} + 2 \cdot P_{AB} \cdot P_{ab}}
\]
Conditional probability of second outcome:
\[
\frac{2 \cdot P_{AB} \cdot P_{ab}}{2 \cdot P_{Ab} \cdot P_{aB} + 2 \cdot P_{AB} \cdot P_{ab}}
\]
Conditional probability of first outcome:

\[ 2 \, P_{Ab} \, P_{aB} / (2 \, P_{Ab} \, P_{aB} + 2 \, P_{AB} \, P_{ab}) = 0.25 \]

Conditional probability of second outcome:

\[ 2 \, P_{AB} \, P_{ab} / (2 \, P_{Ab} \, P_{aB} + 2 \, P_{AB} \, P_{ab}) = 0.75 \]
E-M Algorithm For Haplotyping

1. “Guesstimate” haplotype frequencies
2. Use current frequency estimates to replace ambiguous genotypes with fractional counts of phased genotypes
3. Estimate frequency of each haplotype by counting
4. Repeat steps 2 and 3 until frequencies are stable
Computational Cost (for SNPs)

- Consider sets of $m$ unphased genotypes
  - Markers 1..$m$

- If markers are bi-allelic
  - $2^m$ possible haplotypes
  - $2^{m-1} (2^m + 1)$ possible haplotype pairs
  - $3^m$ distinct observed genotypes
  - $2^{n-1}$ reconstructions for $n$ heterozygous loci
Computational Cost (for SNPs)

- Consider sets of $m$ unphased genotypes
  - Markers 1...$m$

- If markers are bi-allelic
  - $2^m$ possible haplotypes
  - $2^{m-1} (2^m + 1)$ possible haplotype pairs
  - $3^m$ distinct observed genotypes
  - $2^{n-1}$ reconstructions for $n$ heterozygous loci

For example, if $m = 10$

- $2^{10} = 1024$
- $2^{9} (2^{10} + 1) = 524,800$
- $3^{10} = 59,049$
- $2^{9} = 512$
E-M Algorithm for Haplotyping

- Cost grows rapidly with number of markers
- Typically appropriate for < 25 SNPs
  - Fewer microsatellites
- Fully or partially phased individuals contribute most of the information
Other Common Applications

- E-M Algorithm also commonly used for:
  - Estimation of recombination fractions
  - Estimating mixture distributions
Today:

- The E-M algorithm in genetics
- Outline the approach
- Examined specific examples
Next Lecture ...

- E-M algorithm for Haplotyping
- Historical Alternatives
- Recent Enhancements and Alternatives
- Hypothesis testing