Maximum Likelihood, Expectation Maximization, and Haplotype Phasing

CSCI2820: Medical Bioinformatics

Sorin Istrail

September 24, 2013

Outline

- Imagine generating sequences of letters over the four-letter alphabet A, C, G, T
- Sequences generated by random process
- Parametric statistical models: families of probability distributions by a finite-dimensional parameter
- Goal: model this random process and estimate the parameters from the output sequences

Example

- Suppose the model uses three tetrahedral dice
- The probabilities of rolling the four letters are:

	first die	second die	third die
A	0.15	0.27	0.25
C	0.33	0.24	0.25
G	0.36	0.23	0.25
T	0.16	0.26	0.25

To generate each letter, the dice are chosen at random:

- first die picked with probability θ_1
- second die picked with probability θ_2
- third die picked with probability $1 - \theta_1 - \theta_2$

Example

CTCACGTGATGAGAGCATTCTCAGACCGTGACGCGTGTAGCAGCGGCTC

- Was this sequence generated by the three dice?
- If so, what are the parameters θ_1 and θ_2?

Matrix of probabilities:

	first die	second die	third die
A	0.15 θ_1	0.27 θ_2	0.25 (1 - $\theta_1 - \theta_2$)
C	0.33 θ_1	0.24 θ_2	0.25 (1 - $\theta_1 - \theta_2$)
G	0.36 θ_1	0.23 θ_2	0.25 (1 - $\theta_1 - \theta_2$)
T	0.16 θ_1	0.26 θ_2	0.25 (1 - $\theta_1 - \theta_2$)

Let p_A, p_C, p_G and p_T denote the probabilities of generating A, C, G, and T respectively. Then

$$
\begin{aligned}
p_A &= -0.10\theta_1 + 0.02\theta_2 + 0.25 \\
p_C &= 0.08\theta_1 - 0.01\theta_2 + 0.25 \\
p_G &= 0.11\theta_1 - 0.02\theta_2 + 0.25 \\
p_T &= -0.09\theta_1 + 0.01\theta_2 + 0.25
\end{aligned}
$$

Likelihood

For sequence
CTCACGTGATGAGAGCATTCTCAGACCGTGACGCGTGTAGCAGCGGCTC

the *likelihood* of observing the sequence is:

$$L = p_C p_T p_A p_C p_C p_G \cdots p_C = p_A^{10} p_C^{14} p_G^{15} p_T^{10}$$

The *likelihood function* is:

$$\begin{aligned} L(\theta_1, \theta_2) &= p_A(\theta_1, \theta_2)^{10} p_C(\theta_1, \theta_2)^{14} p_G(\theta_1, \theta_2)^{15} p_T(\theta_1, \theta_2)^{10} \\ &= (-0.10\theta_1 + 0.02\theta_2 + 0.25)^{10} (0.08\theta_1 - 0.01\theta_2 + 0.25)^{14} \\ &\quad (0.11\theta_1 - 0.02\theta_2 + 0.25)^{15} (-0.09\theta_1 + 0.01\theta_2 + 0.25)^{10} \end{aligned}$$

Maximum Likelihood

In *maximum likelihood estimation*, the goal is to estimate the parameter values which make the likelihood of observing the data as large as possible

- $\max L(\theta_1, \theta_2) = p_A(\theta_1, \theta_2)^{10} p_C(\theta_1, \theta_2)^{14} p_G(\theta_1, \theta_2)^{15} p_T(\theta_1, \theta_2)^{10}$
- subject to: $0 \leq \theta_1, \theta_2 \leq 1$

Equivalent and more convenient to maximize the log-likelihood function:

$$\begin{aligned} \max l(\theta_1, \theta_2) &= \max \log L(\theta_1, \theta_2) \\ &= \max \left[\, 10 \log(p_A(\theta_1, \theta_2)) + 14 \log(p_C(\theta_1, \theta_2)) \right. \\ &\qquad \left. + 15 \log(p_G(\theta_1, \theta_2)) + 10 \log(p_T(\theta_1, \theta_2)) \,\right] \end{aligned}$$

$$
\begin{aligned}
\max l(\theta_1,\theta_2) &= \max \log L(\theta_1,\theta_2) \\
&= \max \,[\, 10\log(p_A(\theta_1,\theta_2)) + 14\log(p_C(\theta_1,\theta_2)) \\
&\qquad +15\log(p_G(\theta_1,\theta_2)) + 10\log(p_T(\theta_1\theta_2))\,] \\
&= \max \,[\, 10\log(-0.10\theta_1 + 0.02\theta_2 + 0.25) \\
&\qquad +14\log(0.08\theta_1 - 0.01\theta_2 + 0.25) \\
&\qquad +15\log(0.11\theta_1 - 0.02\theta_2 + 0.25) \\
&\qquad +10\log(-0.09\theta_1 + 0.01\theta_2 + 0.25)\,]
\end{aligned}
$$

The solution to this optimization problem can be computed by taking partial derivatives of the log-likelihood function:

$$\frac{\partial l}{\partial \theta_1} = \frac{10}{p_A}\frac{\partial p_A}{\partial \theta_1} + \frac{14}{p_C}\frac{\partial p_C}{\partial \theta_1} + \frac{15}{p_G}\frac{\partial p_G}{\partial \theta_1} + \frac{10}{p_T}\frac{\partial p_T}{\partial \theta_1} = 0$$

$$\frac{\partial l}{\partial \theta_2} = \frac{10}{p_A}\frac{\partial p_A}{\partial \theta_2} + \frac{14}{p_C}\frac{\partial p_C}{\partial \theta_2} + \frac{15}{p_G}\frac{\partial p_G}{\partial \theta_2} + \frac{10}{p_T}\frac{\partial p_T}{\partial \theta_2} = 0$$

$$130030050\theta_1 + 2744\theta_2^2 - 2116125\theta_2 - 6290625 = 0$$

$$134456\theta_2^3 - 10852275\theta_2^2 - 4304728125\theta_2 + 935718750 = 0$$

$$(\theta_1, \theta_2) = 0.5191263945, 0.2172513326$$

Let $F = (f_{ij}(\theta))$ be an $m \times n$ matrix in parameters $(\theta_1, \theta_2, \ldots \theta_d)$.

- F is the *hidden model* or *complete data model*

Let f be the $m \times 1$ matrix $f = (\sum_{j=1}^{n} f_{ij}(\theta))$

- f is the *observed model* or *partial data model*

- Data: u_{ij} drawn from distribution f_{ij} (complete data model)
- Input: Instead of having complete data, we are given only the marginal data $u_i = \sum_j u_{ij}$ for each i
- Goal: infer the parameters θ to maximize the probability of observing the marginal data u_i.

Our problem is to maximize the likelihood function for this data with respect to the observed model:

$$\max L_{obs}(\theta) = f_1(\theta)^{u_1} f_2(\theta)^{u_2} \cdots f_m(\theta)^{u_m}$$

Assumption: can solve the problem for the hidden model F:

$$\max L_{hid}(\theta) = f_{11}(\theta)^{u_{11}} f_{12}(\theta)^{u_{12}} \cdots f_{mn}(\theta)^{u_{mn}}$$

The problem is that we don't know the hidden data u_{ij}!

Expectation Maximization

For models that do not have exact solutions, statisticians use a numerical optimization technique called *Expectation-Maximization* (or EM) for maximizing the likelihood function.

- not guaranteed to reach a global maximum
- known to perform well on many problems of practical interest
- under some conditions, will converge to a local maximum of the likelihood function

Expectation Maximization Algorithm
Input: Functions $f_{ij}(\theta)$, observed data u_i
Output: Maximum likelihood parameters θ

1. Initialize $\theta^0 \in \mathbb{R}^d_{\geq 0}$, $k = 0$.
   (i) Let $u_{ij} = u_i \frac{f_{ij}(\theta^k)}{\sum_j f_{ij}(\theta^k)} = u_i \frac{f_{ij}(\theta^k)}{f_i(\theta^k)}$ for $1 \leq i \leq n, 1 \leq j \leq m$.
   (ii) Let $\theta^{k+1} = \arg\max_\theta l_{hid}(\theta)$
2. If $|\theta^{k+1} - \theta^k| > \epsilon$, let $k = k + 1$ and Go to [1].
   Else output $\theta^* = \theta^{k+1}$.

$$l_{obs}(\theta^{k+1}) - l_{obs}(\theta^{k}) \;\geq\; \left(l_{obs}(\theta^{k+1}) - l_{obs}(\theta^{k})\right) - \left(l_{hid}(\theta^{k+1}) - l_{hid}(\theta^{k})\right)$$

$$= \sum_{i=1}^{m} u_i \log f_i(\theta^{k+1}) - \sum_{i=1}^{m} u_i \log f_i(\theta^{k}) - \sum_{i=1}^{m}\sum_{j=1}^{n} u_{ij}(\log f_{ij}(\theta^{k+1}) - \log f_{ij}(\theta^{k}))$$

$$> \sum_{i=1}^{m} u_i \log f_i(\theta^{k+1}) - \sum_{i=1}^{m} u_i \log f_i(\theta^{k}) - \sum_{i=1}^{m}\sum_{j=1}^{n} u_i \frac{u_{ij}}{u_i}(\log f_{ij}(\theta^{k+1}) - \log f_{ij}(\theta^{k}))$$

$$\geq \sum_{i=1}^{m} u_i \left(\log f_i(\theta^{k+1}) - \log f_i(\theta^{k})\right) - \sum_{i=1}^{m}\sum_{j=1}^{n} u_i \frac{u_{ij}}{u_i}(\log f_{ij}(\theta^{k+1}) - \log f_{ij}(\theta^{k}))$$

$$\geq \sum_{i=1}^{m} u_i \left(\log \frac{f_i(\theta^{k+1})}{f_i(\theta^{k})} \sum_{j=1}^{n} \frac{u_{ij}}{u_i} \log \frac{f_{ij}(\theta^{k+1})}{f_{ij}(\theta^{k})}\right)$$

$$
\begin{aligned}
&\geq \sum_{i=1}^{m} u_i \left(\log \frac{f_i(\theta^{k+1})}{f_i(\theta^k)} - \sum_{j=1}^{n} \frac{u_{ij}}{u_i} \log \frac{f_{ij}(\theta^{k+1})}{f_{ij}(\theta^k)} \right) \\
&= \sum_{i=1}^{m} u_i \left(\sum_{j=1}^{n} \frac{f_{ij}(\theta^k)}{f_i(\theta^k)} \log \frac{f_i(\theta^{k+1})}{f_i(\theta^k)} \sum_{j=1}^{n} \frac{f_{ij}(\theta^k)}{f_i(\theta^k)} \log \frac{f_{ij}(\theta^{k+1})}{f_{ij}(\theta^k)} \right) \\
&= \sum_{i=1}^{m} u_i \left(\sum_{j=1}^{n} \frac{f_{ij}(\theta^k)}{f_i(\theta^k)} \log \frac{f_i(\theta^{k+1})}{f_i(\theta^k)} \frac{f_{ij}(\theta^k)}{f_{ij}(\theta^{k+1})} \right) \\
&= \sum_{i=1}^{m} u_i \sum_{j=1}^{n} \pi_{ij} \log \frac{\pi_{ij}}{\sigma_{ij}} \quad = \quad - \sum_{i=1}^{m} u_i \sum_{j=1}^{n} \pi_{ij} \log \frac{\sigma_{ij}}{\pi_{ij}} \qquad \left(\pi_{ij} = \frac{f_{ij}(\theta^k)}{f_i(\theta^k)}, \sigma_{ij} = \frac{f_{ij}(\theta^{k+1})}{f_i(\theta^{k+1})} \right) \\
&\geq \sum_{i=1}^{m} u_i \sum_{j=1}^{n} \pi_{ij} \left(1 - \frac{\sigma_{ij}}{\pi_{ij}} \right) \quad = \quad \sum_{i=1}^{m} u_i \sum_{j=1}^{n} (\pi_{ij} - \sigma_{ij}) \geq 0
\end{aligned}
$$

EM for Haplotype Phasing

- A *haplotype* is a string of 0's and 1's, representing half of a diploid chromosome
- A *genotype* is a conflated combination of two equal length haplotypes

  0 if the two haplotypes are homozygous with value 0
  1 if the two haplotypes are homozygous with value 1
  2 if the two haplotypes are heterozygous

- Haplotype h is *consistent* with genotype g if h agrees with g in all positions in which g has value 0 or 1 (i.e., there exists a haplotype h' such that $h \oplus h' = g$).

- p_k = probability of haplotype h_k in population
- Vector of haplotype probabilities $p = (p_1, p_2, \ldots p_d)$
- Goal: Find the vector p of haplotype probabilities maximizing the probability of observing genotypes $\mathcal{G}$

Haplotype phasing

Matrix f will denote the probabilites that the observed genotypes are generated by specific pairs of haplotypes

- each row represents an observed genotype g_i
- each column represents a pair of haplotypes (h_k, h_l) $(k < l)$
- entry corresponding to genotype g_i and haplotype pair (h_k, h_l) is indexed by $(i, [k, l])$ and takes value

$$f_{i,[k,l]}(p) = \begin{cases} p_k p_l = p_k^2 & \text{if } k = l \text{ and } h_k \oplus h_l = g_i \\ 2p_k p_l, & \text{if } k \neq l \text{ and } h_k \oplus h_l = g_i \end{cases}$$

Now, apply the above EM framework to the phasing problem.

Expectation Maximization Algorithm for Haplotype Phasing
Input: Functions $f_{i,[k,l]}(p)$ defined above, observed genotype data u_i
Output: An estimate p^* for the maximum likelihood haplotype frequencies.

1. Initialize $p^0 \in \mathbb{R}^d_{\geq 0}$, $t = 0$.
   (i) Let $u^t_{i,[k,l]} = u_i \frac{f_{i,[k,l]}(p^t)}{\sum_{k,l} f_{i,[k,l]}(p^t)} = u_i \frac{f_{i,[k,l]}(p^t)}{f_i(p^t)}$ for $1 \leq i \leq n, 1 \leq k < l \leq d$.
   (ii) Let $p^{k+1} = \arg\max_p l_{hid}(p)$
2. If $|p^{t+1} - p^t| > \epsilon$, let $t = t + 1$ and Go to [1].
   Else output $p^* = p^{t+1}$.

We now show the problem of maximizing the hidden likelihood function, has an explicit solution.

Lemma. The function $M(x) = \prod_i x_i^{r_i}$ subject to the constraint $N(x) = \sum_{i=1}^{n} x_i = constant$ is maximized when

$$\frac{x_1}{r_1} = \frac{x_2}{r_2} = \cdots = \frac{x_n}{r_n}.$$

Proof. By the theory of Lagrange multipliers, $M(x)$ is maximized when

$$\frac{\partial M(x)}{\partial x_i} = \lambda \frac{\partial N(x)}{\partial x_i} \text{ for all } 1 \leq i \leq n.$$

$$\frac{\partial M(x)}{\partial x_i} = \lambda \frac{\partial N(x)}{\partial x_i} \text{ for all } 1 \leq i \leq n.$$

Taking partial derivatives, we obtain the following set of equations

$$\begin{array}{rcl}
(r_1 x_1^{r_1 - 1}) x_2^{r_2} \cdots x_n^{r_n} & = & \lambda \\
x_1^{r_1} (r_2 x_2^{r_2 - 1}) \cdots x_n^{r_n} & = & \lambda \\
& \vdots & \\
x_1^{r_1} x_2^{r_2} \cdots (r_n x_n^{r_n - 1}) & = & \lambda
\end{array}$$

So the maximum is achieved when

$$(r_1 x_1^{r_1 - 1}) x_2^{r_2} \cdots x_n^{r_n} = x_1^{r_1} (r_2 x_2^{r_2 - 1}) \cdots x_n^{r_n} = \cdots = x_1^{r_1} x_2^{r_2} \cdots (r_n x_n^{r_n - 1})$$

This is satisfied when $\frac{r_1}{x_1} = \frac{r_2}{x_2} = \cdots = \frac{r_n}{x_n}$, proving the lemma. $\square$

To avoid local maxima, the method should be run on a set of widely ranging initial values. Several possibilities for the initial conditions include the following.

1. All haplotypes are equally likely:

$$p_k^{(0)} = \frac{1}{d}, \text{ for } k = 1, 2, \ldots, d$$

2. Randomly choose probabilities satisfying

$$\sum_{k=1}^{d} p_k^{(0)} = 1$$