Mathematical Notes on Estimating the Allele Frequency Spectrum from Sequencing Data

Heng Li

July 27, 2010

1 Notations

Suppose there are \( N \) sites from \( n \) individuals with \( i \)-th individual having \( m_i \) ploids. Let \( M = \sum_i m_i \) be the total number of chromosomes. Let \( \mathbf{D} = (\vec{D}_1, \ldots, \vec{D}_N)^T \) be the data matrix with vector \( \vec{D}_a = (D_{a1}, \ldots, D_{an}) \) representing the alignment data for each individual at site \( a \). Similarly, let \( \mathbf{G} = (\vec{G}_a, \ldots, \vec{G}_N)^T \) and \( \vec{G}_a = (G_{a1}, \ldots, G_{an}) \) be the true genotypes, where \( 0 \leq G_{ai} \leq m_i \) equals the number of reference alleles \( 1 \).

Define \( X_a = X_a(\vec{G}_a) = \sum_i G_{ai} \) to be the number of reference alleles at site \( a \) and \( \mathbf{X} = (X_1, \ldots, X_N)^T \). Also define \( \Phi = (\phi_0, \ldots, \phi_M) \) as the allele frequency spectrum (AFS) with \( \sum_k \phi_k = 1 \).

2 Estimating AFS

2.1 The EM procedure

We aim to find \( \Phi \) that maximizes \( P(\mathbf{D}|\Phi) \) by EM. Suppose at the \( t \)-th iteration the estimate is \( \Phi_t \). We have

\[
\log \Pr\{D, \mathbf{X} = \mathbf{x}|\Phi\} = \log \Pr\{D|\mathbf{X} = \mathbf{x}\} \Pr\{\mathbf{X} = \mathbf{x}|\Phi\} = C + \sum_a \log \phi_{xa}
\]

where \( C \) is not a function of \( \{\phi_k\} \). The EM \( Q \) function is

\[
Q(\Phi|\Phi_t) = \sum_x \Pr\{\mathbf{X} = \mathbf{x}|\mathbf{D}, \Phi_t\} \log \Pr\{\mathbf{D}, \mathbf{X} = \mathbf{x}|\Phi\}
\]

\[= C + \sum_x \prod_a \Pr\{X_a = x_a|\vec{D}_a, \Phi_t\} \sum_b \log \phi_{xb} \]

\[= C + \sum_a \sum_{x_a=0}^M \Pr\{X_a = x_a|\vec{D}_a, \Phi_t\} \log \phi_{xa} \]

Requiring \( \partial_{\phi_k}(Q - \lambda \sum_t \phi_t) = 0 \) leads to

\[
\frac{1}{\phi_k} \sum_a \Pr\{X_a = k|\vec{D}_a, \Phi_t\} - \lambda = 0
\]

from which \( \lambda \) can be calculated as:

\[
\lambda = \sum_k \sum_a \Pr\{X_a = k|\vec{D}_a, \Phi_t\} = N
\]

\(^1\)If we take the ancestral sequence as the reference, the non-reference allele will be the derived allele.
and thus at the \((t + 1)\) iteration:

\[
\phi_k^{(t+1)} = \frac{1}{N} \sum_a \Pr\{X_a = k|\vec{D}_a, \Phi_t\} 
\]

(2)

where \(\Pr\{X_a = k|\vec{D}_a, \Phi_t\}\) is calculated as follows.

### 2.2 The distribution of site reference allele count

Firstly, as we are only looking at a site from now on, we drop subscript \(a\) for convenience. Define

\[
z_{jk}(\phi) = \Pr\left\{ \sum_{i=1}^{j} G_i = k, \vec{D}|\phi \right\}
\]

for \(0 \leq j \leq \sum_{i=1}^{j} m_i\) and 0 otherwise. It is the probability of seeing \(k\) reference alleles from the first \(j\) individuals when the true site allele frequency is known to be \(\phi\). \(z\) can be calculated iteratively as

\[
z_{jk}(\phi) = \sum_{g_j=0}^{m_j} z_{j-1,k-g_j}(\phi) \cdot \Pr\{D_j|G_j=g_j\} \Pr\{G_j=g_j|\phi\}
\]

(3)

with \(z_{00}(\phi) = 1\). \(\Pr\{D_j|G_j=g_j\}\) is calculated in the next section, and under the assumption of Hardy-Weinberg equilibrium, we have

\[
\Pr\{G_j=g_j|\phi\} = \binom{m_j}{g_j} \phi^g_j (1-\phi)^{m_j-g_j} 
\]

(4)

We abbreviate \(P(D_j|g_j) \triangleq \Pr\{D_j|G_j=g_j\}\) and \(P(g_j|\phi) \triangleq \Pr\{G_j=g_j|\phi\}\) when there is no ambiguity. With \(z_{nk}(\phi)\) calculated,

\[
\Pr\{X = k, \vec{D}|\Phi\} = \sum_{\vec{g}} \phi_l \Pr\{X = k, \vec{D}, \vec{g}|\phi = l/M\} = \sum_{l=0}^{M} \phi_l z_{nk}(l/M) 
\]

(5)

Thus

\[
\Pr\{\vec{D}|\Phi\} = \sum_k \Pr\{X = k, \vec{D}|\Phi\} 
\]

(6)

\[
\Pr\{X = k|\vec{D}, \Phi\} = \Pr\{X = k, \vec{D}|\Phi\} / \Pr\{\vec{D}|\Phi\} 
\]

Eq. (2) can be calculated. The expected allele count is

\[
E(X|\vec{D}, \Phi) = \sum_k k \Pr\{X = k|\vec{D}, \Phi\} 
\]

### 2.3 The expected site allele count: the efficient approach

While we can calculate the likelihood of data and the expected site allele count with the method above, the procedure is cubic in the number of individuals which is very slow. Unless the full AFS is intended, we actually have a faster way to calculate them.

In fact,

\[
\Pr\{\vec{G} = \vec{g}, \vec{D}|\Phi\} = \sum_k \Pr\{\vec{G} = \vec{g}, \vec{D}|X = k\} \Pr\{X = k|\Phi\} = \sum_k \phi_k \prod_i P(D_i|g_i) P(g_i|k/M) 
\]

(7)

Summing over all genotype configurations gives

\[
\Pr\{\vec{D}|\Phi\} = \sum_{\vec{g}} \Pr\{\vec{G} = \vec{g}, \vec{D}|\Phi\} = \sum_{k=0}^{M} \phi_k \prod_{i=1}^{n} \sum_{g_i=0}^{m_i} P(D_i|g_i) P(g_i|k/M) 
\]

(8)
The expected site reference allele count equals

\[ E(X|\bar{D}, \Phi) = \frac{1}{\Pr\{\bar{D}|\Phi\}} \sum \left( \sum_{i} g_i \right) \Pr\{\bar{G} = \bar{g}, \bar{D}|\Phi\} \]

\[ = \frac{1}{\Pr\{\bar{D}|\Phi\}} \sum_{k} \phi_k \sum_{j} \sum_{g_j} g_j P(D_j|g_j) P(g_j|k/M) \prod_{i \neq j} P(D_i|g_i) P(g_i|k/M) \]

\[ = \frac{1}{\Pr\{\bar{D}|\Phi\}} \sum_{k} \phi_k \left[ \prod_{i} \sum_{g_i} P(D_i|g_i) P(g_i|k/M) \right] \cdot \left[ \sum_{j} \sum_{g_j} g_j P(D_j|g_j) P(g_j|k/M) \right] \]

2.4 The initial AFS

The EM procedure guarantees that \( \Pr\{\bar{D}|\Phi\} \) monotonically increases with each iteration and converges to a local optima. However, if we start this iteration from a bad initial AFS, we may need many iterations; the iteration is also more likely to be trapped by a local optima. Here we give several AFS on different conditions under the infinite-site Wright-Fisher model.

Let \( \phi_k' \) be the probability of seeing \( k \) non-reference alleles out of \( M \) chromosomes. The frequency of reference alleles \( \phi_k \) equals \( \phi_k' M - k \).

If we take the ancestral sequence as the reference, the standard model gives \( \phi_k' = \theta/k \) and \( \phi_0' = 1 - \sum k \phi_k' \). When we do not know if the reference allele is ancestral, the same conclusion still stands. To see this, for \( k > 0 \):

\[ \phi_k' = \frac{M + 1 - k}{M + 1} \left( \frac{\theta}{k} + \frac{\theta}{M + 1 - k} \right) = \frac{\theta}{k} \]

and for \( k = 0 \):

\[ \phi_0' = 1 - \sum_{k=1}^{M+1} \frac{\theta}{k} + \frac{\theta}{M + 1} = 1 - \sum_{k=1}^{M} \frac{\theta}{k} \]

where the first term corresponds to the case wherein the reference is ancestral and the second to the case wherein the reference is derived.

Another useful AFS is the derived allele frequency spectrum on the condition of loci being discovered from two chromosomes. Under the Wright-Fisher model, it is:

\[ \phi_k' = \frac{2(M + 1 - k)}{(M + 1)(M + 2)} \]

A third AFS is the derived allele frequency spectrum on the condition of knowing one derived allele from a chromosome. It is a flat distribution

\[ \phi_k' = \frac{1}{M + 1} \]

2.5 Estimating site allele frequency

Here we aim to find \( \phi \) that maximises \( \Pr\{\bar{D}|\phi\} \). We have:

\[ \log \Pr\{\bar{D}, \bar{G} = \bar{g}|\phi\} = \log \prod_{i} P(D_i|g_i) P(g_i|\phi) = C + \sum_{i} \log P(g_i|\phi) \]
Given an estimate $\phi_t$ at the $t$-th iteration, the $Q(\phi|\phi_t)$ function of EM is:

$$Q(\phi|\phi_t) = \sum \Pr\{\vec{G} = \vec{g}|\vec{D}, \phi_t\} \log \Pr\{\vec{D}, \vec{G} = \vec{g}|\phi\}$$

$$= C + \sum \prod_i \Pr\{G_i = g_i|D_i, \phi_t\} \sum_j \log P(g_j|\phi)$$

$$= C + \sum_i^n \sum_{g_i=0}^{m_i} \Pr\{G_i = g_i|D_i, \phi_t\} \log P(g_i|\phi)$$

$$= C + \sum_i^n \sum_{g_i=0}^{m_i} \Pr\{G_i = g_i|D_i, \phi_t\} \log \left( \frac{m_i}{g_i} \phi^{g_i} (1 - \phi)^{m_i-g_i} \right)$$

$$= C' + \sum_i \sum_{g_i} \Pr\{G_i = g_i|D_i, \phi_t\} \left[ g_i \log \phi + (m_i - g_i) \log(1 - \phi) \right]$$

Requiring $\frac{\partial Q}{\partial \phi_t}|_{\phi=\phi_{t+1}} = 0$ gives:

$$\frac{1}{\phi_{t+1}(1 - \phi_{t+1})} \sum_i \sum_{g_i} \Pr\{G_i = g_i|D_i, \phi_t\} (g_i - m_i \phi_{t+1}) = 0$$

Thus

$$\phi_{t+1} = \frac{1}{\sum_j m_j \sum_i g_i \Pr\{G_i = g_i|D_i, \phi_t\}} = \frac{1}{M} \sum_i \frac{g_i P(D_i|g_i) P(g_i|\phi_t)}{\sum_{g_i} g_i P(D_i|g_i) P(g_i|\phi_t)}$$

which is the EM estimate at the $(t+1)$-th iteration and also the expected reference allele frequency.

### 3 Likelihood of data given genotype

Given a site covered by $k$ reads from an $m$-ploid individual, the sequencing data is:

$$D = (b_1, \ldots, b_k) = (1, \ldots, 1, 0, \ldots, 0)_{l \ldots k-l}$$

where 1 stands for a reference allele and 0 otherwise. The $j$-th base is associated with error rate $\epsilon_j$, which is the larger error rate between sequencing and alignment errors. We have

$$P(D|0) = \prod_{j=1}^l \epsilon_j \prod_{j=l+1}^k (1 - \epsilon_j) = \left( 1 - \sum_{j=l+1}^k \epsilon_j + o(\epsilon^2) \right) \prod_{j=1}^l \epsilon_j$$

$$P(D|m) = \left( 1 - \sum_{j=1}^l \epsilon_j + o(\epsilon^2) \right) \prod_{j=l+1}^k \epsilon_j$$
For $0 < g < m$:

$$
P(D|g) = \sum_{a_1=0}^{1} \cdots \sum_{a_k=0}^{1} \Pr\{D|B_1 = a_1, \ldots, B_k = a_k\} \Pr\{B_1 = a_1, \ldots, B_k = a_k|g\} \quad (13)$$

$$
= \sum_{a} \left(\frac{g}{m}\right)^{a} \prod_{j=0}^{k} \left(1 - \frac{g}{m}\right)^{k-a} \cdot \prod_{j} p_j(a_j)
$$

$$
= \left(1 - \frac{g}{m}\right)^{k} \prod_{j=1}^{l} \left(\epsilon_j + \frac{g}{m-g}(1-\epsilon_j)\right) \prod_{j=l+1}^{k} \left(1 - \epsilon_j + \frac{\epsilon_j g}{m-g}\right)
$$

$$
= \left(1 - \frac{g}{m}\right)^{k} \left\{ \left(\frac{g}{m-g}\right)^{l} + \left(1 - \frac{g}{m-g}\right) \left(\sum_{j=1}^{l} \epsilon_j - \sum_{j=l+1}^{k} \epsilon_j\right) + o(\epsilon^2) \right\}
$$

In the bracket, the first term explains the deviation between $l/k$ and $g/m$ by imperfect sampling, while the second term explains the deviation by sequencing errors. The second term can be ignored when $k$ is small but may play a major role when $k$ is large. In particular, for $m = 2$, $P(D|1) = 2^{-k}$, independent of sequencing errors.

In case of dependent errors, we may replace:

$$
\epsilon_1 < \epsilon_2 < \cdots < \epsilon_l
$$

with

$$
\epsilon'_j = \epsilon_j^{\alpha_j^{-1}}
$$

where parameter $\alpha \in [0,1]$ addresses the error dependency.

### 4 Multi-sample SNP calling and genotyping

The probability of the site being polymorphic is $1 - \Pr\{\tilde{G} = \{m_1, \ldots, m_n\}|\tilde{D}, \Phi\}$ which can be calculated with Eq. (7) and (8). In theory, the Bayesian estimate of the genotype configuration $\tilde{g}$ should maximize Eq. (7). However, this calculation requires to enumerate all possible configurations and is exponential in the number of individuals. In practice, for individual $i$, we may simply estimate the genotype $\hat{g}_i$ as:

$$
\hat{g}_i = \arg\max_{g_i} \Pr\{G_i = g_i|D_i, \phi_E\} = \arg\max_{g_i} \frac{P(D_i|g_i)P(g_i|\phi_E)}{\sum_{h_i} P(D_i|h_i)P(h_i|\phi_E)}
$$

where $\phi_E = E(X|\tilde{D}, \Phi)/M$ is calculated by Eq. (9). This estimate of genotypes may not necessarily maximize the posterior probability $P(\tilde{g}|	ilde{D})$, but it should be good enough in practice. For low-coverage samples, the likelihood surface should be flat. The Bayesian estimate may not be accurate, either. For high-coverage samples, most reasonable methods should give similar results.

Note that when the AFS is fairly accurate, we do not need to derive the full AFS. SNP calling and genotyping can be done in $O(n^2)$ time.