Genetic Data and EM of Li (2011)

Sequencing Data

A T T G (A/C) T G C
A T T G (C/A) T G C

↑ SNP = location where differences occur

"allele"

Typically, SNPs are "biallelic" (only two alleles occur in population)

Choose one allele as the "reference" (doesn't matter which)

The individual's "dosage" at a SNP is the number of reference alleles that individual has

Let G_{i1} = individual i's dosage @ SNP 1
G_{i2} = individual i's dosage @ SNP 2

$G = (G_1, G_2)$, $G_i = (G_{i1}, G_{2i})$

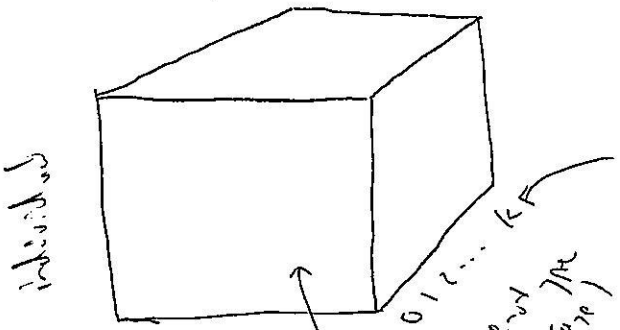
$G_1, G_2, \ldots, G_n$ iid G

Goal: estimate $\rho = \text{cor}(G_1, G_2)$ using sample $G_1, G_2, \ldots, G_n$

If we observed $G_1, G_2, \ldots, G_n$ then this would be easy

Sample correlation: $\hat{\rho} = \dfrac{\sum_{i=1}^{n} (G_{i1} - \bar{G}_{\cdot 1})(G_{i2} - \bar{G}_{\cdot 2})}{\sqrt{\sum_{i=1}^{n} (G_{i1} - \bar{G}_{\cdot 1})^2} \sqrt{\sum_{i=1}^{n} (G_{i2} - \bar{G}_{\cdot 2})^2}}$

- We do not observe $G_{1,\ldots,G_n}$
- We observe genotype likelihoods for each individual at each genotype for each locus



$$Pr(data \mid G_{ji} = k)$$

j = SNP
i = individual
k = dosage (genotype)

- Considering just 2 loci at a time, let

$$a_{ik} = Pr(Data_i \mid G_{i1} = k)$$

$$b_{ik} = Pr(Data_i \mid G_{i2} = k)$$

- We sometimes have prior information on distribution of genotypes:

$\pi_{k1} = Pr(G_1 = k)$ ← proportion individuals w/ genotype k at locus 1

$\pi_{k2} = Pr(G_2 = k)$

- Recall Bayes Rule:

$$Pr(G_{i1} = k \mid Data_i) = \frac{Pr(Data_i \mid G_{i1} = k)\, Pr(G_{i1} = k)}{Pr(Data_i)}$$

$Pr(Data_i \mid G_{i1} = k)$ = genotype likelihood = a_{ik}

$Pr(G_{i1} = k) = \pi_{k1}$

$$Pr(Data_i) = \sum_{k=0}^{K} Pr(Data_i \mid G_{i1} = k)\, Pr(G_{i1} = k)$$

$$= \sum_{k=0}^{K} a_{ik}\, \pi_k$$

- So posterior probability of individual i's genotype given data is

$$Pr(G_{i1} = k \mid Data_i) = \frac{a_{ik}\, \pi_k}{\sum_{k=0}^{K} a_{ik}\, \pi_k}$$

What if we don't know π?

↳ we can estimate it from the data!

$$\Pr(\text{All data} \mid \pi) = \prod_{i=1}^{n} \Pr(\text{Data}_i \mid \pi)$$

$$= \prod_{i=1}^{n} \left(\sum_{k=0}^{K} a_{ik} \pi_k \right) \xrightarrow{\log}$$

objective function

$$\sum_{i=1}^{n} \log \left[\sum_{k=0}^{K} a_{ik} \pi_k \right]$$

↳ Maximum likelihood estimation says to estimate π by maximizing this quantity over π

Hard to do this, so we use EM algorithm "Expectation - Maximization"

$$\Pr(\text{Data}_i \text{ and } G_i \mid \pi) = \prod_{k=0}^{K} \left[a_{ik} \pi_k \right]^{1(G_i = k)}$$

indicator function

$$\Pr(\text{All data and All } G \mid \pi) = \prod_{i=1}^{n} \prod_{k=0}^{K} \left(a_{ik} \pi_k \right)^{1(G_i = k)}$$

log ↳

$$\sum_{i=1}^{n} \sum_{k=0}^{K} 1(G_i = k) \left[\log(a_{ik}) + \log(\pi_k) \right]$$

↳ we don't know G_i, but if we knew π_k then we could know the distribution of $G_i \mid \text{data}_i$

↳ then we could maximize this "expected log likelihood" to get a new π

EM: Start w π_0

E-step: $E\left[\mathbb{1}(G_i = h) \mid data, \pi_0\right] = P_r(G_i = h \mid data, \pi_0)$

$$= \frac{a_{ih}\,\pi_0}{\sum_{h=0}^{k} a_{ih}\,\pi_0} =: w_{ih}$$

M-step: $\max_{\pi}$ $\sum_{i=1}^{n}\sum_{h=0}^{k} w_{ih}\left[\log(a_{ih}) + \log(\pi_h)\right]$

$$= \max_{\pi} \sum_{i=1}^{n}\sum_{h=0}^{k} w_{ih}\log(\pi_h)$$

$$= \max_{\pi} \sum_{h=0}^{k} \log(\pi_h) \underbrace{\sum_{i=1}^{n} w_{ih}}_{=: w_h}$$

$$\frac{\partial}{\partial \pi_h}\left(\sum_{h=0}^{k} \log(\pi_h)\, w_h + \lambda\left(\sum_{h=0}^{k}\pi_h - 1\right)\right)$$

$$\frac{w_h}{\pi_h} + \lambda \overset{!}{=} 0$$

$$\Rightarrow \pi_h = -\frac{w_h}{\lambda}$$

$\Rightarrow \pi_h \propto w_h$ and must sum to 1

$$\Rightarrow \pi_h = \frac{w_h}{\sum w_h}$$