Computational Systems Biology: Biology X

Bud Mishra

Room 1002, 715 Broadway, Courant Institute, NYU, New York, USA

L#10:(Apr-13-2010)
Genome Wide Association Studies

Class Projects

- Few Ideas for the class projects:
  1. GWAS – WTCCC Study: See the URL: http://www.nature.com/nature/journal/v447/n7145/full/nature05911
  2. Mendelian Diseases: See the URL: http://www.nature.com/nature/journal/v461/n7261/full/nature08250
  3. Indian Population: See the URL: http://www.nature.com/nature/journal/v461/n7263/abs/nature08365
  4. Mutation Rates in Humans: See URL: http://www.pnas.org/content/107/3/961.abstract
  5. Quartet Analysis: See URL: http://www.sciencemag.org/cgi/content/abstract/science.1186802

Outline

Missingness

- *Missing and Unobservable Data*:
  1. Rare alleles are difficult to genotype. The frequency estimates are incorrect.
  2. Alignment of alleles on a single homologous chromosome is difficult to infer. *Haplotype Phasing Problem*.

Haplotype Phasing Problem

- Two alleles on the same homologous chromosome are said to be *in cis* — Two alleles on opposite sister homologs are said to be *in trans*.
- A particular combination of alleles on a single homologous chromosome is called a *haplotype*.
- With $(k+1)$ biallelic SNPs, the population can have 2^k possible *haplotypes*, though most of them are likely to be missing.

Haplotype Phasing Problem

- Note that the diploid pair of haplotypes is of the order 2^{2k}:

$$\binom{2^k}{2} + 2^k,$$

the first term corresponding to heterozygous haplotypes and the second corresponding to homozygous haplotypes.

- When $k = 2$, there are four haplotypes: (AB, aB, Ab, ab) and ten diplotypes

$$(AB, AB), (aB, aB), (Ab, Ab), (ab, ab),$$

$$(AB, aB), (AB, Ab), (AB, ab), (aB, Ab), (aB, ab), \text{ and } (Ab, ab).$$

Penetrance

- It is possible to infer a likely haplotype from the genotype data, if we know the LD-structure for the population.
- However, this is further confused by two other effects:
  1. **Penetrance**: The presence of a disease alleles does not lead to the disease phenotype.
  2. **Phenocopies**: Individuals exhibiting disease phenotypes do not carry the allele under consideration.

Model Selection

- Goal is to select a small number of SNPs to build a model: These should be causal SNPS or Tag SNPs in LD with causal SNPs.
- **Bayesian Variable Selection**: Start with a General Linear Model for Genotype-Trait Association:

$$y_i = \beta_1 x_{i1}^* + \beta_2 x_{i2}^* + \cdots + \beta_r x_{ir}^* + \epsilon_i, \qquad \text{for } i = 1, \ldots, n,$$

where $(\mathbf{x}_1^*, \mathbf{x}_2^*, \ldots, \mathbf{x}_r^*)$ is a subset of potential indicator variables, $\mathbf{y}$ is a quantitative trait.

Outline

Model Selection

- For the coefficients assume that they are either *relevant* or *nuisance* variables, described by a mixture model:

$$\beta_j|\gamma_j \sim (1-\gamma_j)\mathcal{N}(0,\tau_j^2) + \gamma_j\mathcal{N}(0,c_j^2\tau_j^2),$$

where $gamma = (\gamma_1, \ldots, \gamma_p)$ is a latent (unobservable) vector with elements taking values 0 or 1.

$$Pr(\gamma_j = 1) = p_j, \text{ and } Pr(\gamma_j = 0) = 1 - p_j = q_j,$$

- For the variance in the selected coefficients, we can choose:

$$\sigma^2|\gamma \sim \mathcal{IG}(\nu_\gamma/2, \nu_\gamma\lambda_\gamma/2),$$

given by an inverse gaussian (Wald) distribution $\mathcal{IG}$.

Distributions

- Gaussian/Normal:

$$X \sim \mathcal{N}(\mu, \sigma)$$

then

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{-(x-\mu)^2}{2\sigma^2}.$$

- Wald:

$$X \sim \mathcal{IG}(\mu, \lambda)$$

then

$$f(x; \mu, \lambda) = \left[\frac{\lambda}{2\pi x^3}\right]^{1/2} \exp \frac{-\lambda(x-\mu)^2}{2\mu^2 x}.$$

Putting it all together

- We now have

$$\mathbf{y}|\beta, \sigma^2 \sim \mathcal{MVN}_n(\mathbf{X}\beta, \sigma^2 I),$$

  where $\mathbf{y} = (y_1, \ldots, y_n)^T$, $\mathbf{X}_{n \times p} = [\mathbf{x}_1, \ldots, \mathbf{x}_p]$ and $\beta = (\beta_1, \ldots, \beta_p)^T$.
- The parameters corresponding to the ONLY true underlying predictors $(\mathbf{x}_1^*, \ldots, \mathbf{x}_r^*)$ are non-zero.

Bayesian Formulation

- Putting everything together,

$$\pi(\gamma|\mathbf{Y}) \propto f(\mathbf{Y}|\beta,\sigma^2)f(\beta|\gamma)f(\sigma^2|\gamma)\pi(\gamma).$$

- We can find the best estimator for γ by Gibb's sampling from the marginal posterior densities for β, σ and γ_j.

Bayesian Variable Selection

Algorithm 1: BVS - pseudocode

Input: Traits **Y** and SNPs $\mathbf{x}_i$

Output: Subset of prdictive SNPs $\mathbf{x}_i^*$

1. Initialize β, σ and γ — denoted as $\beta^{(0)}$, $\sigma^{(0)}$ and $\gamma^{(0)}$
2. Let $t = t + 1$ and sample
   - $\beta^{(t)}|\mathbf{y} \sim f(\beta|\mathbf{y}, \sigma^{(t-1)}, \gamma^{(t-1)})$
   - $\sigma^{(t)}|\mathbf{y} \sim f(\sigma|\mathbf{y}, \beta^{(t-1)}, \gamma^{(t-1)})$
3. Randomly select an ordering $\gamma_{(1)}, \ldots, \gamma_{(p)}$ and sample
   - $\gamma_{(1)}^{(t)}|\mathbf{y} \sim f(\gamma_{(1)}|\mathbf{y}, \beta^{(t)}, \sigma^{(t)}, \gamma_{(2)}^{(t-1)}, \ldots, \gamma_{(p)}^{(t-1)})$
   - $\gamma_{(2)}^{(t)}|\mathbf{y} \sim f(\gamma_{(1)}|\mathbf{y}, \beta^{(t)}, \sigma^{(t)}, \gamma_{(1)}^{(t)}, \gamma_{(3)}^{(t-1)}, \ldots, \gamma_{(p)}^{(t-1)})$

     $\vdots$
   - $\gamma_{(p)}^{(t)}|\mathbf{y} \sim f(\gamma_{(1)}|\mathbf{y}, \beta^{(t)}, \sigma^{(t)}, \gamma_{(1)}^{(t)}, \ldots, \gamma_{(p-1)}^{(t)})$
4. Repeat the steps (2) and (3) M times for a large M.
5.

Outline

Unobservable Phase

- Currently a primary challenge in GWAS: Unobservable nature of allelic phases.
- It is possible to solve it by improved technology. But current technologies focus on “genotyping,” and then resolve haplotype ambiguity via *statistical methods*: (1) **Single IMputation** or (2) **Multiple Imputations**.
- **Simple Methods**: First resolve genotype ambiguities to impute haplotypes; then use the haplotypes in association studies.
- **Complex Methods**: Combined analysis involving both imputations and association studies.

Imputation

- We have *incomplete* observed data $\mathbf{X}^{obs}$, which are from a distribution parametrized by (unknown) θ.
- We can estimate θ by an MLE, if we had *complete* data $\mathbf{X}^{c}$.
- If we had the parameters θ, we could impute $\mathbf{X}^{c}$ from $\mathbf{X}^{obs}$.
- In our case,

$$\begin{aligned} \mathbf{X}^{obs} &= \{G_1, \ldots, G_n\} \quad \text{genotypes} \\ \mathbf{X}^{c} &= \{H_1, \ldots, H_n\} \quad \text{haplotypes} \\ \theta &= \text{parameters describing} \\ &\qquad \text{haplotype distributions in the population} \end{aligned}$$

Toy Example

- Consider the genotype across two sites for individual i given by

$$G_i = [AA][BB]$$

The individual is homozygous in both sites. Then his haplotypes are

$$S(G_i) = \{(AB, AB)\}$$

There is no haplotypic ambiguities for such an individual. But such individuals would be relatively rare in the population, occurring with a probability $p_A^2 p_B^2$.

Toy Example

- Next, consider the genotype across two sites for individual i given by

$$G_i = [AA][Bb]$$

The individual is heterozygous in the second site. Then his haplotypes are

$$S(G_i) = \{(AB, Ab)\}$$

There is no haplotypic ambiguities for such an individual, either. But such individuals would not be that frequent in the population, occurring with a probability $2p_A^2 p_B(1 - p_B)$.

Ambiguities

- Now, consider a more common case: the genotype across two sites for individual i given by

$$G_i = [Aa][Bb]$$

The individual is heterozygous in both sites. Then the set of all haplotype pairs consistent with this genotype is given by

$$S(G_i) = \{(AB, ab), (Ab, aB)\}$$

There is a haplotypic ambiguity for such an individual. They occur in the population with a probability $4p_A p_B(1 - p_A)(1 - p_B)$.

Ambiguities

- The best one can do is to *impute* the data by the population wide distributions of various haplotypes, under assumption of a panmictic population, with the distributions governed by some parameters θ.
- The parameters $\widehat{\theta}$ can be estimated from the imputed haplotypes for all individuals.
- Suppose the haplotype frequencies are θ_1, θ_2, θ_3 and θ_4 for haplotypes AB, Ab, aB and ab, respectively. Then for the (ambiguous) individuals, we may impute, by saying that he has haplotype-pair (AB, ab) with a probability $2\theta_1\theta_4$ and haplotype-pair with a probability $2\theta_2\theta_3$.

Haplotype Estimation

- *Estimate individual haplotypes and population-level frequencies*.
- **EM approach**: Estimate haplotype frequencies; Use estimates to infer unknown haplotypes for the individuals in the GWAS sample;
- **Bayesian approach**: Reconstruct unknown haplotypes; reconstructed data can then be used to estiamte population-level haplotype frequencies.

EM algorithms

- Expectation-Maximization (EM) algorithms work in two steps: *E steps* and *M steps*...
- It is a natural approach when there is a significant amount of missing data
- Recall: A maximum likelihood estimate (MLE) is an estimate of parameters of a distribution, derived by maximizing a function of the complete data $\mathbf{X}^c = (x_1, \ldots, x_n)$.

$$\widehat{\theta} = \arg\max_{\theta} L(\theta|\mathbf{X}^c) = \arg\max_{\theta} \prod_{i=1}^{n} Pr(x_i|\theta),$$

where $Pr(x_i|\theta)$ is the probability density function of x_i (parametrized by θ).

EM algorithms

- Thus a maximum likelihood estimate (MLE) computes (since log is an order preserving transformation)

$$\widehat{\theta} = \arg\max_{\theta} \log L(\theta|\mathbf{X}^c).$$

- But since we have only $\mathbf{X}^{obs}$, we first impute $\mathbf{X}^c$ using our best guess for $\widehat{\theta}$:

$$\widehat{\theta} = \arg\max_{\theta} E\left(\log L(\theta|\mathbf{X}^c)|\mathbf{X}^{obs}, \widehat{\theta}\right).$$

- We solve the fix-point equation by successively improving the estimates

$$\widehat{\theta}^{(t+1)} = \arg\max_{\theta} E\left(\log L(\theta|\mathbf{X}^c)|\mathbf{X}^{obs}, \widehat{\theta}^{(t)}\right).$$

EM Algorithm

- The algorithm works in two steps:
- **E step**: First it takes the expectation of the complete data log likelihood – conditional on the observed data and the current parameter estimate. Thus it determines the most likely value of the likelihood for the *complete data*
- **M step**: Next it maximizes the equation with respect to the parameter θ. This yields a new estimate – denoted $\widehat{\theta}^{(t+1)}$.
- *E-* and *M-steps* are repeated iteratively until a convergence criterion (stopping rule) is met to arrive at an MLE of θ.

EM

Algorithm 2: EM - pseudocode

Input: Model: $Pr(x_i|\theta)$ and $\mathbf{X}^{obs}$
Output: MLE of $\widehat{\theta}$

1 Initialize $\theta^{(0)}$ to some reasonable sets of values
2 Let $t = t + 1$ and repeat
$\widehat{\theta}^{(t+1)} := \arg\max_\theta E\left(\log L(\theta|\mathbf{X}^c)|\mathbf{X}^{obs}, \widehat{\theta}^{(t)}\right)$.
3 Repeat the step (2) M times for a large M.
4

EM

Algorithm 3: EM - pseudocode

Input: Genotype Data: $G_1, G_2, \ldots, G_n$
Output: MLE estimates of the frequencies $\widehat{p_{H_i}}$

1 Initialize $\theta^{(0)}$ to some reasonable sets of values by using the homozygous individuals
2 Let $t = t + 1$ and repeat

$$\begin{aligned}
\widehat{\theta}^{(t+1)} &:= \arg\max_{\theta} E\left(\log L(\theta|H_1,\ldots,H_n)|G_1,\ldots,G_n,\widehat{\theta}^{(t)}\right). \\
&= \sum_{i=1}^{n}\sum_{H_i\in S(G_i)} \widehat{p_{H_i}}^{(t)} \log Pr(H_i|\theta) \\
\widehat{p_{H_i}}^{(t)} &:= Pr(H_i|G_i,\widehat{\theta}^{(t)}) \\
&= \frac{Pr(H_i|\widehat{\theta}^{(t)})}{\sum_{H_i\in S(G_i)} Pr(H_i|\widehat{\theta}^{(t)})}
\end{aligned}$$

3 Repeat the step (2) M times for a large M.
4

Intuition

- **E step**: We average over all possible resolutions of the missing data in a manner that takes into account the current parameter estimates...
- If an individual's genotype is $[Aa, Bb]$, then E-step will give more weight to the haplotype pair that has a higher estimated frequency.... For instance if the haplotypes (AB, ab) are relatively common while the haplotypes (Ab, aB) appear rare, then we lend additional weight to the former than the later. That is,

$$Wt(AB, ab) > Wt(Ab, aB).$$

Intuition

- **M step**: We maximize the expectation to arrive at updated parameter estimates.
- As the E- and M-steps are repeatedly applied, the estimates converge to the true values. (Assuming that an initial estimate is not too far away form the true values.)
- **Note**: This approach assumes HWE. Thus it should only be applied within racial and ethnic strata within which there is no evidence of a departure from the underlying assumption of panmixia (i.e., random mating).
- **Also, note**: Sometimes, it is not uncommon to fill in unknown haplotypes by assigning each individual the haplotype pair with the highest posterior probability ... *This strategy leads to incorrect solutions, as since valuable information on the uncertainty in the assignment is lost.*

Bayesian Haplotype Reconstruction

- This method allows for estimation of population level haplotype frequencies in the context of data for which allelic phase is potentially unobservable. *The primary aim however is reconstruction of individual-level haplotype pairs — Assign each individual the most likely haplotype pair.*
- **Bayesian Approach**: Sampling schemes: (1) MCMC (Markov-Chain Monte-Carlo), (2) Gibbs Sampling, (3) Sequential Monte-Carlo (Particle Filtering), (4) EM, etc.

Bayesian Approach

- Make inference about parameter based on its conditional distribution given data.

$$\begin{aligned} \theta &= \text{parameter of interest} \\ \mathbf{X} &= \text{data} \\ \pi(\theta|\mathbf{X}) &= \text{conditional distribution of } \theta \text{ given } \mathbf{X} \\ &= \textit{posterior density} \text{ of } \theta \end{aligned}$$

- The distribution depends on three quantities:
  1. The prior distribution of θ, given by $\pi(\theta)$
  2. The likelihood of the data, given by $L(\theta|\mathbf{X}) = f(\mathbf{X}|\theta)$
  3. A constant $c = 1/(\int_\theta \pi(\theta)L(\theta|\mathbf{X})d\theta)$

Baye's Rule

- The relationship between the posterior density and each of the tree quantities: posterior, likelihood and partition function is

$$\pi(\theta|\mathbf{X}) = cL(\theta|\mathbf{X})\pi(\theta)...$$

or equivalently,

$$\pi(\theta|\mathbf{X}) = \frac{\pi(\theta;\mathbf{X})}{f(\mathbf{X})} = \frac{f(\mathbf{X}|\theta)\pi(\theta)}{\int_\theta \pi(\theta)L(\theta|\mathbf{X})d\theta}.$$

- In practice, exact calculation of this is posterior probability is not tractable, and approximation methods are use — Markov-Chain Monte-Carlo (MCMC) methods provide an approach to generate approximate samples from a distribution

Gibb's Sampler

- Suppose that the population parameters are $\theta = (\theta_1, \ldots, \theta_k)$, and we wish to compute the joint posterior density $\pi(\theta|\mathbf{X})$ — which cannot be obtained analytically.
- Assume that $\pi(\theta_k|\theta_{-k}, \mathbf{X})$ is the marginal distribution of the single parameter θ_k conditional on current values of all other parameters:

$$\theta_1, \ldots, \theta_{k-1}, \theta_{k+1}, \ldots, \theta_K .$$

- A Gibb's sampler provides us with sample of data from posterior density $\pi(\theta|\mathbf{X})$, based on sampling from marginal distributions $\pi(\theta_k|\theta_{-k}, \mathbf{X})$.

Gibb's Sampler

Algorithm 4: Gibb's Sampling

Input: Model parameters θ, Data: $\mathbf{X}$

Output: MLE estimates of the parameters θ

1. Initialize $\theta^{(0)}$ to some reasonable sets of values
2. Let $t = t + 1$ and sample
   - $\theta_1^{(t+1)} | \theta_{-1}, \mathbf{X} \sim \pi(\theta_1 | \theta_2^{(t)}, \ldots, \theta_K^{(t)}, \mathbf{X})$
   - $\theta_2^{(t+1)} | \theta_{-2}, \mathbf{X} \sim \pi(\theta_2 | \theta_1^{(t+1)}, \theta_3^{(t)}, \ldots, \theta_K^{(t)}, \mathbf{X})$
   - $\vdots$
   - $\theta_K^{(t+1)} | \theta_{-K}, \mathbf{X} \sim \pi(\theta_K | \theta_1^{(t+1)}, \ldots, \theta_{K-1}^{(t+1)}, \mathbf{X})$

Repeat the step (2) M times for a large M.

Bayesian Haplotype Reconstruction

Algorithm 5: Bayes' Haplo Recon

Input: Genotype Data: $\mathbf{G} = \{G_1, \ldots, G_n\}$

Output: MLE estimates of the haplotype pairs
$\mathbf{H} = \{H_1, \ldots, h_{n*}\}$

1. Initialize $\mathbf{H}^{(0)}$ to some reasonable values
2. Let $t = t + 1$ and sample
   - $H_1^{(t+1)} | \mathbf{G}, \mathbf{H}_{-1} \sim \pi(H_1 | H_2^{(t)}, \ldots, H_{n*}^{(t)}, \mathbf{G})$
   - $H_2^{(t+1)} | \mathbf{G}, \mathbf{H}_{-2} \sim \pi(H_2 | H_1^{(t+1)}, H_3^{(t)}, \ldots, H_{n*}^{(t)}, \mathbf{X})$
   - $\vdots$
   - $H_{n*}^{(t+1)} | \mathbf{G}, \mathbf{H}_{-n*} \sim \pi(H_{n*} | H_1^{(t+1)}, \ldots, H_{n*-1}^{(t+1)}, \mathbf{X})$

Repeat the step (2) M times for a large M.

[End of Lecture #10]