# Hidden Markov Random Field model for GWAS

Kyurhi Kim

August 3, 2023

SNU HDMT Lab Seminar

# Introduction

# Introduction

- GWAS analysis have been limited to the single SNP or SNP-SNP pair analysis
- If multiple SNPs are all in LD with the true disease variants, using the information from LD can increase power.
- In the paper of Li., LD information among the SNPs derived from the data is incorporated into identifying the disease - associated SNPS.

# Introduction

1. First build a weighted LD graph based on pairwise LD measures among the SNPs.
2. Propose Hidden Markov Random Field model (HMRF) on LD graph in order to compute the posterior probability that an SNP is associated with the disease.
3. Propose Empirical Bayes in estimating model parameters.
4. Use Iterative Conditional Mode (ICM) algorithm to estimate the parameters and Gibbs sampling for estimating the posterior probabilities.
5. Define a FDR to select the relevant SNPs.

# Hidden Markov Random Field Model

# Hidden Markov Model

- Hidden Markov Random Field is a generalization of a hidden Markov model. Instead of having an underlying Markov chain, hidden Markov random fields have an underlying Markov random field.
- Markov model is assumed that future states is independent of its history.
- HMM is defined as stochastic processes generated by a Markov chain whose sequence cannot be observed directly ("hidden"), only through a sequence of observations.

# Markov Random Field

- Markov Random Field(Undirected Graphical Model) is a probablility distribution p over variables $x_1, ..., x_n$ defined by an undirected graph G which nodes correspond to variables $x_i$.
- It is similar to Bayesian Network in its representatin of dependencies, but the difference is that Bayesian networks are directed and acyclic, while MRF are undirected and may be cyclic.
- Thus MRF can represent certain dependencies that a Bayesian network cannot.

# Example

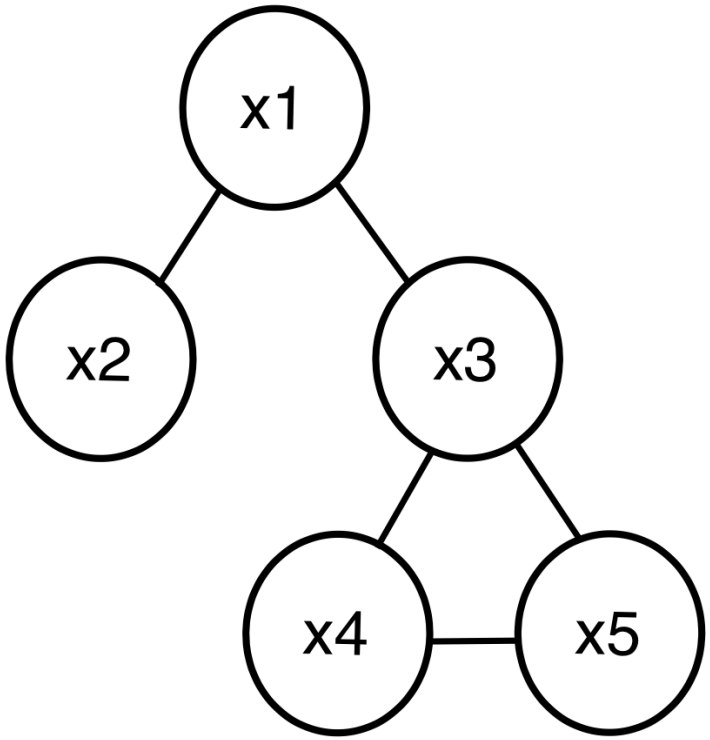


**Figure 1:** example of the MRF

# Markov Random Field

- The probability p has the form $p(x_1, ..., x_n) = \frac{1}{Z} \prod_{c \in C} \phi_c(x_c)$
- C denotes the set of cliques of G
- each factor $\phi_C$ is a non-negative function over the variables in a clique
- Z is the partition function (normalizing constant).
- Any strictly positive MRF can be written as exponential family in canonical form.

# Settings

- There are m cases and n controls that are genotyped over a set of p SNPs. Denote the SNP index as S=$\{1, ..., p\}$.
- $Y = (Y_1, ..., Y_s, ..., Y_p)$ is the observed genotype data for the p SNPs. Here, $Y_s = (y_{s1}, ..., y_{sm}; y_{s(m+1)}, ..., y_{s(m+n)})$, where $y_{si}$ is the observed genotype for the ith individual at the sth SNP.
- Goal: Determine which SNPs in S are associated with the disease.
- To account the LD information in identifying the disease-associated SNPs, develop an HMRF model.

# Weighted LD graph

- Construct a weighted undirected LD graph G based on pairwise LD information.
- An edge between SNPs $s$ and $s'$ is drawn with weight $w_{ss'} = I(r^2_{ss'} > \tau) r^2_{ss'}$.
- $\tau$ is a given value and $r^2_{ss'}$ is the measurement of LD between SNPs $s$ and $s'$ if $w_{ss'} \neq 0$.

# MRF model

- For a given SNP s, define a random indicator variable as

$$X_s = \begin{cases} 1 & \texttt{if SNP s is associated with the disease} \\ 0 & \texttt{if SNP s is not associated with the disease} \end{cases}$$

- Model dependency using a simple discrete Markov Random Field model (Besag, 1974) with the following joint probability function for $X = (X_1, ..., X_p)$:

$$p(X; \Phi) \propto exp(\gamma \sum_{s=1}^{p} X_s + \beta \sum_{s \sim s'} w_{s,s'} I(X_s = X_{s'}))$$

where $\gamma$ and $\beta \geq 0$ are the two model parameters and $\beta$ measures dependencies of $X_s$ for SNPs in LD.

# MRF model

- Here, assumption is required that true association state X is a realization of a locally dependent discrete MRF with a specified distribution $\{p(X)\}$.
- The conditinal association state for SNP s, given the states of all neighboring SNPs is as follows:

$$p(X_s|X_{N_s};\Phi) \propto exp(\gamma \sum_{s=1}^{p} X_s + \beta \sum_{s \in N_s} w_{s,s'} I(X_s = X_{s'})),$$

where $N_s$ represents the neighbors of the SNP s on the LD graph.

# Empirical Bayes Model

# Dirichlet-Multinomial model

- The Dirichlet distribution (Dir($\alpha_1, ..., \alpha_K$)) is parameterized by positive scalars $\alpha_i > 0$ for i=1,...,K, where $K \geq 2$. The probability density of $p = (p_1, ..., p_K)$ is

$$f(p_1, ..., p_K; \alpha_1, ..., \alpha_K) = \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} p_i^{\alpha_i - 1}.$$

- The Multinomial distribution (Mult($p_1, ..., p_K, n$)) is a discrete distribution over K dimensional non-negative integer vectors where $\sum_{i=1}^{K} x_i = n$. And the probability mass function is

$$\begin{aligned} f(x_{i1}, ..., x_{iK}; p_1, ..., p_K, n) &= \frac{n!}{x_{i1}! \cdots x_{iK}!} p_1^{x_{i1}} \cdots p_K^{x_{iK}} \\ &= \frac{\Gamma(n+1)}{\prod_{k=1}^{K} \Gamma(x_{ik} + 1)} \prod_{k=1}^{K} p_k^{x_{ik}}. \end{aligned}$$

# Dirchlet-Multinomial model

- Let $p \sim Dir(\alpha)$ and $x_i \sim Mult(n, p)$, with Data $X = \{x_1, ..., x_n\}$. Then the Posterior is proportional to a Dirichlet distribution:

$$p(\alpha|X) \propto \prod_{i=1}^{n} p(x_i|\alpha)p(\alpha)$$

$$p(\alpha|X) = Dir(\alpha'), \alpha'_k = \sum_{i=1}^{n} x_{ik} + \alpha_k$$

- Marginal distribution is obtained by integrating on the distribution for p as follows:

$$P(x|n, \alpha) = \int_p Mult(x|n, p)Dir(p|\alpha)dp$$

$$= \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)\Gamma(n+1)}{\Gamma(n + \sum_{k=1}^{K} \alpha_k)} \prod_{k=1}^{K} \frac{\Gamma(x_k + \alpha_k)}{\Gamma(\alpha_k)\Gamma(x_k + 1)}$$

# Joint probability of the observed genotypes

- To relate the latent vector X to the observed genotypes, assume that given any particular realization of X, the random variables $Y = (Y_1, ..., Y_p)$ are conditionally independent.
- Conditional density is $I(Y|X) = \prod_{s=1}^{p} P(Y_s|X_s)$, where $P(Y_s|X_s)$ is the joint probability of the observed genotypes over m+n individuals at the SNP s given the latent state $X_s$.

# Empirical Bayes Model

- To specify $P(Y_s|X_s)$, let genotype frequencies at the sth SNP in the case and control population as $\theta_s = (\theta_{s1}, \theta_{s2}, \theta_{s3})$ and $\rho_s = (\rho_{s1}, \rho_{s2}, \rho_{s3})$ respectively.
- Assume that both of these frequencies across all the SNPs have a Dirichlet prior with parameter $\alpha = (\alpha_1, \alpha_2, \alpha_3)$:

$$f(\theta_s) = f(\theta_{s1}, \theta_{s2}, \theta_{s3}) = \frac{\Gamma(\sum_{j=1}^{3} \alpha_j)}{\prod_{j=1}^{3} \Gamma(\alpha_j)} \prod_{j=1}^{3} \theta_{sj}^{\alpha_j - 1}.$$

# Empirical Bayes Model

- Denote the observed genotype counts data in the m cases as $y_{s+} = (y_{s+,1}, y_{s+,2}, y_{s+,3})$, and n controls as $y_{s-} = (y_{s-,1}, y_{s-,2}, y_{s-,3})$.
- If SNP s is not associated with the disease, cases should have the same genotype frequencies with the controls.
- In this case, the combined genotype counts data $y_{s0} = y_{s+} + y_{s-}$ are generated from a multinomial distribution with the genotype frequencies of $\theta_s$.
- On the other hand, if SNP s is associated with the disease, cases and controls should have different genotype frequencies.

# Empirical Bayes Model

- $P(Y_s|X_s = 0) = \int (y_{si}; i = 1, ..., m + n|X_s = 0, \theta_s) f(\theta_s) d\theta_s$

$$= \frac{\Gamma(\sum_{j=1}^{3} \alpha_j) \prod_{j=1}^{3} \Gamma(\alpha_j + y_{s+,j} + y_{s-,j})}{\prod_{j=1}^{3} \Gamma(\alpha_j) \Gamma(\sum_{j=1}^{3} (\alpha_j + y_{s+,j} + y_{s-,j}))}$$

- $P(Y_s|X_s = 1) = \int (y_{si}; i = 1, ..., m|X_s = 1, \theta_s) f(\theta_s) d\theta_s$

$$\times \int (y_{si}; i = m + 1, ..., m + n|X_s = 1, \rho_s) f(\rho_s) d\rho_s$$

$$= \frac{\Gamma(\sum_{j=1}^{3} \alpha_j) \prod_{j=1}^{3} \Gamma(\alpha_j + y_{s+,j})}{\prod_{j=1}^{3} \Gamma(\alpha_j) \Gamma(\sum_{j=1}^{3} (\alpha_j + y_{s+,j}))}$$

$$\times \frac{\Gamma(\sum_{j=1}^{3} \alpha_j) \prod_{j=1}^{3} \Gamma(\alpha_j + y_{s-,j})}{\prod_{j=1}^{3} \Gamma(\alpha_j) \Gamma(\sum_{j=1}^{3} (\alpha_j + y_{s-,j}))}$$

# Parameter estimation

# ICM algorithm

- To estimate parameters in the MRF model $(\gamma, \beta)$ and $\alpha$ (Dirichlet prior parameter), ICM algorithm is used.
- ICM algorithm was introduced for image restoration process.
- It is done by iteratively maximizing the probability of each variable conditioned on the rest.
- There are two assumptions for the algorithm:
  1. Given x, the random variable $y_1, ..., y_n$ are conditionally independent and each $y_i$ has the same known conditional density function $f(y_i|x_i)$. (i.e. $I(y|x) = \prod_{i=1}^{n} f(y_i|x_i)$)
  2. The states $X$ are assumed to constitute a Markov Random Field. (i.e. $X$ is a random field whose local conditional probability functions satisfy the Markov property)

# ICM algorithm

- Let $\widehat{x}$ be an estimate of true $x$.
- The goal is to update $\widehat{x_i}$ at each i given all available information:

$$\underset{x_i}{argmax}\, p(x_i|y, \widehat{x}_{S\setminus i}) \propto f(y_i|x_i)p_i(x_i|\widehat{x}_{\delta_i})$$

- Here, $f(y_i|x_i)$ is the probability of an observed data $y_i$ given the color $x_i$, and $p_i(x_i|\widehat{x}_{\delta_i})$ is the probability of the color $x_i$ given the surrounding neighbors of i.
- initial estimate of $\widehat{x}$ is given by just maximizing $f(y_i|x_i)$.
- The procedure defines a cycle when applied to each pixel i in turn.

# ICM algorithm

1. Obtain an initial estimate $\widehat{X}$ (based on the single -marker trend test using p-value of 0.0001)
2. Estimate $\alpha$ with the value of $\widehat{\alpha}$ by maximizing the probability of the observed data given by $I(Y|X) = \prod_{s=1}^{p} P(Y_s|X_s)$.
3. Estimate $\Phi$ with the value of $\widehat{\phi}$ by maximizing the pseudo-likelihood function:
   $I(\widehat{X};\Phi) = \prod_s^p p_s(\widehat{X_s}|\widehat{X_{N_s}};\Phi) =$
   $$\prod_s^p \frac{exp(\gamma \widehat{X}_s + \beta \sum_{s' \in N_s} w_{s,s'} I(\widehat{X}_s = \widehat{X}_{s'}))}{exp(\gamma + \beta \sum_{s' \in N_s} w_{s,s'}) I(\widehat{X}_{s'}=1) + exp(\beta \sum_{s' \in N_s} w_{s,s'}) I(\widehat{X}_{s'}=0)}$$
4. Obtain new $\widehat{X}$ based on $\widehat{X}, \widehat{\alpha}, \widehat{\phi}$. i.e. for s= 1,...,p, update $X_s$ based on $P(X_s|Y, \widehat{X_{S/s}}) \propto f(Y_s|X_s;\widehat{\alpha}) p_s(X_s|\widehat{X}_{N_s};\widehat{\Phi})$.
5. repeat until $\max_{\theta \in (\alpha,\Phi)} \frac{|\theta^{(k+1)} - \theta^{(k)}|}{|\theta^{(k+1)}|} < 0.001$.

# FDR controlling Procedure

# Gibbs Sampling

- Gibbs Sampling is one of a MCMC algorithm for obtaining a sequence of obervations which are approximated from a specified multivariate distribution, when direct sampling is difficult.
- The point is that given a multivariate distribution, it sample from a conditional distribution than to marginalize by integrating over a joint distribution.
- Gibbs sampling is a special case of Metropolis–Hastings in which the newly proposed state is always accepted with probability one.

# Gibbs Sampling

initialize $Y^0, X^0$

for j = 1, 2, 3,... do

sample $X^j \sim p(X|Y^{j-1})$

sample $Y^j \sim p(Y|X^j)$

end for

**Figure 2:** Gibbs Sampler

# FDR control

- After the convergence of the algorithm, sample the latent vector $X$ M times using Gibbs sampling based on the conditional probability in page 19-(4).
- Based on these samples, estimate the posterior probability of $q_s = Pr(X_s = 0|Y)$.
- Sort $q_s$ in descending order as $q_{(s)}$.
- Note that for SNP s, the hypothesis is

$$H_{s0} : \text{SNP s is not associated with the disease}$$
$$H_{s1} : \text{SNP s is associated with the disease}$$

# FDR control

- Based on the posterior probabilities, let k = $max\{t : \frac{1}{t}\sum_{s=1}^{t} q_{(s)} \leq \alpha\}$.
- Then, reject all $H_{(s)}, s = 1, ..., k$
- This posterior probablility-based definition of FDR has been used in the analysis of microarray gene expression data and has been shown to control the FDR at $\alpha$

# References

[1] Hongzhe Li and others, A hidden Markov random field model for genome-wide association studies, Biostatistics, Volume 11, Issue 1, January 2010, Pages 139–150

[2] Besag, J. (1986). On the statistical analysis of dirty pictures. Journal of the Royal Statistical Society Series B: Statistical Methodology, 48(3), 259-279.