Bayesian Model Search and Multilevel Inference for SNP Association Studies

Wilson, Melanie A.[1], Iversen, Edwin S.[1], Clyde, Merlise A.[1], Schmidler, Scott C.[1], Schildkraut, Joellen M.[2]

[1]Department of Statistical Science, [2]Department of Community and Family Medicine, Duke University, Durham, NC

INTRODUCTION

Technological advances in genotyping have given rise to hypothesis based association studies of increasing scope. In particular, candidate pathway studies of single nucleotide polymorphisms (SNPs) have all but replaced those focusing on a small number of variants in a handful of genes. As a results, the scientific hypotheses addressed by these studies have become more complex and more difficult to address using existing analytic methodologies. Frequently, researchers assume conditional independence across markers, analyzing each SNP separately and determining pathway- or study-wide association by a secondary analysis. This approach has proven to be powerful given its simplicity, but has several significant drawbacks when used to quantify association. These include inference in the face of multiple comparisons, complications arising from correlations among the SNPs and choice of their parameterization.

We describe Multilevel Inference for SNP Association Studies (MISA), a Bayesian model search calculation applied to penalized logistic regression that searches the space of genetic markers, and over the genetic parameterizations of each, in a computationally efficient manner. This technique allows one to estimate multilevel posterior probabilities and Bayes Factors at the global, gene and SNP level.

Multiple Markers

How do we correct for the high false discovery rate of current marginal methods?

There are many frequency based methods that adjust for multiple comparisons. However, they often make use of unjustifiable distributional assumptions on p-values under H_0 and H_a and the SNPs themselves.

Conditional independence assumption of the markers is weakened in MISA methodology and all genetic markers are simultaneously included in a Bayesian model search algorithm to identify subsets of likely associated variables.

Multiple Genetic Parameterizations

Which genetic parameterization for each marker do we choose when calculating marginal p-values?

May make more sense to allow for uncertainty in the genetic models for each SNP.

We search over multiple genetic parameterizations for each marker in the model search algorithm (Log-Additive, Dominant, and Recessive) and report posterior summaries that are averaged across the parameterizations.

Multilevel Inference

How do we measure the global significance of the experiment if we assume the markers are conditionally independent?

We weaken the conditional independence assumption and compute multilevel posterior summaries at the global, gene, and SNP levels allowing us to determine the extent to which the data supports an overall association in a pathway or gene of interest and to identify the markers most likely driving the association.

METHODS

Model & Prior Specification

We consider case- control association studies where: $D_i = 1$ for a disease case and $D_i=0$ for a control.

★Use Logistic regression to relate disease status to a subset of p SNPs, $\boldsymbol{x}_\gamma$, and q confounders, $\boldsymbol{z}$ of the form:

$$logit(p(D_i=1 \mid \Theta)) = \alpha_0 + \boldsymbol{z}_i^T\boldsymbol{\alpha} + \boldsymbol{x}_{\gamma i}^T\beta_\gamma$$

α_0 Intercept Term
$\boldsymbol{\alpha}$ Vector of coefficients for the fixed confounders z_i included in all models
$\boldsymbol{x}_{\gamma i}^T$ Parameterizations of the SNPs included in M_γ
β_γ Log odds ratios for SNPs included in M_γ

★Models, M_γ, are specified by the p dimensional vector γ where γ_j indicates the inclusion of SNP_j in model M_γ and if included, the genetic parameterization of the SNP. We consider log-additive (γ_j=1), dominant (γ_j=2), and recessive (γ_j=3) genetic parameterizations of the SNPs and set γ_j=0 to indicate SNP_j is not included in M_γ

To compute the posterior model probabilities and posterior summaries we specify a prior on the model space $\boldsymbol{M}$.

★We model the SNP inclusion probabilities as independent Bernoulli random variables with probability θ. The number of SNPs has a Binomial(p, θ) distribution.

★To increase flexibility, we assume that θ is random with Beta(a,b) distribution and the overall model prior has the general hierarchical form:

$$i \mid \theta \sim Binomial(p,\theta) \quad \& \quad \theta \sim Beta(a,b)$$

★As a decreases and b increase the expected model size decreases and the global and marginal prior odds decrease.

Posterior Summaries

★Model Posterior Probabilities and Bayes Factors

The degree to which the data support any model in the Bayesian framework may be assessed via model posterior probabilities or Bayes Factors.

★The posterior model probability of any model M_γ in the model space $\boldsymbol{M}$ is expressed as:

$$p(M_\gamma \mid D) \propto p(D \mid M_\gamma)\, p(M_\gamma) \qquad for\ M_\gamma\ in\ \boldsymbol{M}$$

$p(D \mid M_\gamma)$ Proportional to the (marginal) likelihood of M_γ after integrating out model specific parameterization $\theta_\gamma \equiv (\alpha_0, \boldsymbol{\alpha}, \boldsymbol{\beta}_\gamma)$ and approximated as $p(D \mid M_\gamma) \approx exp(-.5\ AIC(M_\gamma))$
$p(M_\gamma)$ Prior probability of M_γ

★The Bayes Factor comparing the evidence of any two competing models $M_{\gamma 1}$ and $M_{\gamma 2}$ is expressed as:

$$BF(M_{\gamma 1}: M_{\gamma 2}) = Post.\ Odds(M_{\gamma 1}: M_{\gamma 2}) \div Prior\ Odds(M_{\gamma 1}: M_{\gamma 2})$$
$$Poster\ Odds\ (M_{\gamma 1}: M_{\gamma 1}) = p(M_{\gamma 1} \mid D) / p(M_{\gamma 2} \mid D) \quad \& \quad Prior\ Odds\ (M_{\gamma 1}: M_{\gamma 1}) = p(M_\gamma) / p(M_{\gamma 2})$$

★Multilevel Posterior Summaries

Results are summarized by reporting the models that are sampled from the stationary distribution (from the chain where $t_i = 1$) defined as M^s.

★**Global Bayes Factor:** *Gives evidence that at least one SNP is associated with disease:*

$$BF(H_A:H_0) = \sum_{M\gamma\ in\ Ha} BF(M_\gamma:M_0)\ p(M_\gamma \mid H_A)$$

★**SNP Inclusion Probability and BF**: *Marginal measures of significance for each genetic marker:*

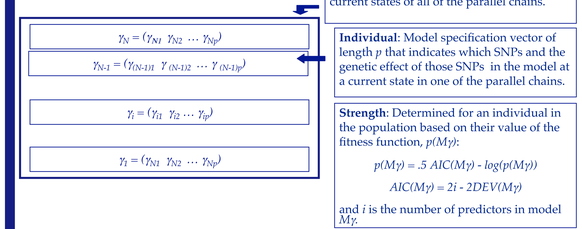
$$Pr(\gamma_j>0 \mid D) = \sum_{M\gamma\ in\ M^s} 1(\gamma_j\neq 0)\ \ Pr(M_\gamma \mid D, M^s) \quad \& \quad Pr(\gamma_j>0) = Post.\ Odds(\gamma_j>0) \div Prior\ Odds(\gamma_j>0)$$

★**Gene Inclusion Probability and BF:** M*easures of significance of one or more of the SNPs within a given gene being associated.*

$$Pr(\gamma_{gi}=1 \mid D) = \sum_{M\gamma\ in\ M^s} 1(\gamma_{gi}=1)\ \ Pr(M_\gamma \mid D, M^s) \ \& \ BF(\gamma_{gi}=1) = Post.\ Odds(\gamma_{gi}=1) \div Prior\ Odds(\gamma_{gi}=1)$$

Model Search

Our model search algorithm is based on the Evolutionary Monte Carlo algorithm of Liang and Wong (2000) where individuals in a population of model specifications compete and mate in order to produce increasingly stronger individuals via a mixture of parallel tempering (Geyer,1991) and population updates of a genetic algorithm (Holland,1975).



Population: (Nxp) matrix that indicates the set of N different model specifications of the current states of all of the parallel chains.

Individual: Model specification vector of length p that indicates which SNPs and the genetic effect of those SNPs in the model at a current state in one of the parallel chains.

Strength: Determined for an individual in the population based on their value of the fitness function, $p(M\gamma)$:

$$p(M\gamma) = .5\ AIC(M\gamma) - log(p(M\gamma))$$
$$AIC(M\gamma) = 2i - 2DEV(M\gamma)$$

and i is the number of predictors in model $M\gamma$.

★Parallel Tempering

Interested in sampling from the model space, $\boldsymbol{M}$, by running parallel chains where each chain is associated with a decreasing temperature t_i.

★Chains with high temp. values have flat proposal densities that allow us to make large global moves to models across regions of low probability.

★Chains with low temp. values have peaked proposal densities that allow us to make small local moves to specific models.

★Genetic Algorithm Population Updates

Mutation (Metropolis Update): Model, M_k, is chosen from the current population of models and values in the model indicator γ_k are mutated

Crossover (Partial State Swap): One model pair, (M_i , M_j), is selected from the current model population and two new offspring are produced and replace the old pair in the population.

Exchange (Full State Swap): Given the current population and attached temperature ladder, we propose a new population by making an exchange between models M_i and M_j without changing the values of the temperature ladder.

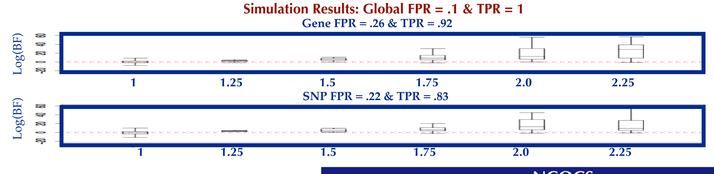
RESULTS

Simulation

We explore optimal choice of the hyper-parameters, a, b, based on a set of 30 simulated case-control data sets that are modeled after an ovarian cancer candidate pathway study of interest (NCOCS). Data sets are comprised of a binary outcome representing disease status, and genetic data on 399 ovarian cancer cases and 798 controls.

★Genetic data was simulated at the same 508 tag SNPs as genotyped in the study. The data was simulated in two stages. First, for the 53 genes represented in the data set, we phased the NCOCS control SNP genotype data and estimated recombination rates using PHASE (Stephens et al., 2001). Second, given a model of association, we generated case-control data at these tags using HAPGEN (Marchini and Su, 2006).

★Ten of the simulations are null, and in the remaining twenty we assumed that 9 random genes were associated and that within these genes, a single, randomly chosen tag SNP was the source of the association. Within ten of these associated simulations, three of the associated SNPs were accorded an odds ratio (OR) of 1.25, three an OR of 1.5, and three an OR of 1.75. The other ten associated simulations had the same structure only with OR's of 1.75, 2.0, and 2.25. In each of the twenty associated simulations, one SNP with each OR was assumed to have a dominant genetic parameterization, one of each a log-additive, and one of each a recessive.



The model search algorithm was performed assuming a set of (a,b) to span a range of priors with reasonable operating characteristics and true and false positive rates are calculated at the global, gene, and SNP level using a Bayes Factor threshold of 10, giving strong evidence of an association (Kass and Raftery 1995).

Choice for hyper-parameters: a=1/8 and b=p has the characteristics:

Expected model size ≈ .125; Prior Odds (H_0:H_A) ≈ 11;

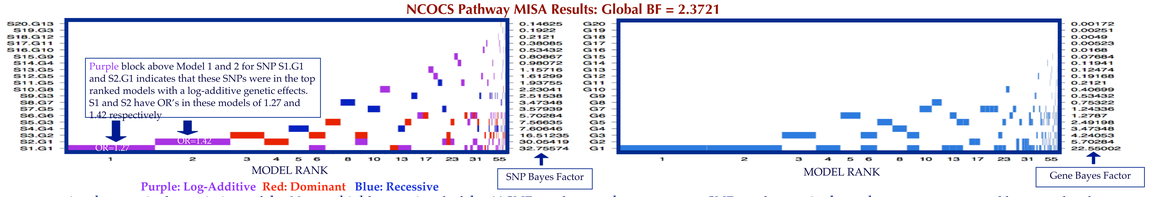
Prior SNP Assoc ≈ .002

NCOCS

The North Carolina Ovarian Cancer Study (NCOCS) is a population based case/control study. The data is comprised of cases and controls who were matched based on answers to a in-person interview. The analysis was run by looking only at the Caucasian cases/controls and a subtype of the cases. This gave us a total of 399 cases and 798 controls. For the purpose of this example we only look at the 66 SNPs in our data set that are in a candidate pathway that have passed a marginal screen. We also controlled for age of patient and previous diagnosis of breast cancer by forcing these variables into our models while performing the model selection.

To account for missing data, we use fastPHASE (Stephens et al. ,2001) to sample haplotypes and missing genotypes given the observed unphased genotypes and create 100 imputed data sets. Results were found by using the average value of our fitness function, $p(M\gamma)$, across the imputed data sets.

NCOCS Pathway MISA Results: Global BF = 2.3721



Purple: Log-Additive Red: Dominant Blue: Recessive

★Plots summarize the marginal associations of the 20 most highly associated of the 66 SNPs and genes they represent. SNPs and genes in the pathway are represented by a two level name where the number represents the rank of the SNP or gene by means of the marginal Bayes Factor. Summary of model inclusion for the top 100 sampled models on the basis of their posterior model probability are plotted. The models are ordered on the x-axis in descending probability and the width of the column associated with a model is proportional to that probability. A SNP's or gene's presence in a model is indicated by a block at the intersection of the model's column and a SNP's or gene's row. The color of the block in the SNP plot indicates the genetic parameterization of the SNP in that model.

★SNPs 1,2, and 3 have Bayes Factors greater than 10, providing strong evidence that marginally these SNPs are associated with ovarian cancer. Both S1 and S2 are in G1 which has a gene Bayes Factor of 22.5502, giving strong evidence that at least one of the SNPs within gene G1 is associated with ovarian cancer.

CONCLUSION

In this paper, we describe an analytic strategy of pathway association studies that allows one to quantify evidence of associations at 3 levels: global, gene, and SNP while allowing for uncertainty in the genetic parameterization of the markers. Our methodology for SNP association studies accounts for several problems encountered by commonly used marginal analyzes. These include the hard to justify conditional independence assumption made on p-values under Ho, Ha and/or on the markers themselves. While our algorithm is not assumption free, it does relax the typical conditional independence assumption, does not require an *a priori* choice of SNP-specific genetic parameterization, facilitates a multilevel assessment of statistical significance, and allows optimal prior parameters to be chosen via simulation. Given the multilevel assessment for a pathway or gene of interest, we are able to ascertain if any genes warrant further investigation and which regions of these genes to tag more densely.

Our analytic strategy can readily accommodate different modeling strategies. The simulation study plays a critical role in this strategy by aiding in the investigation of the optimal hyper-parameters and operating characteristics of these parameters within our method. Similar simulations can be produced based on the study of interest and optimal hyper-parameters can be chosen to achieve a desired balance between the false and true positive rates. Our strategy can also accommodate models associated with studies other than case-control associations studies in that other penalized likelihood functions may be used instead of AIC. Likelihoods such as those arising from survival models may be used in place of the logistic regression model. Also, missing data on SNPs is a common phenomenon; we address this problem by generating multiple imputations for the missing data. This allows us to use all cases while reporting valid uncertainty summaries. Other methods may be used to account for missing data in the study of interest. Finally, when computing the posterior probabilities of each of the sampled models, we assume that the prior of each model comes from the beta-binomial distribution where the SNPs are assumed to be independent of each other with a common inclusion probability. It would be interesting to explore other prior distributions that build in more genetic information either on pathway structure or known genetic function of the SNPs.

Literature Cited

[1] C. Geyer. *Markov chain monte carlo maximum likelihood*. Computing Science and Statistics, page 156-163, 2001.

[2] J.H. Holland. *Adaptation in natural and artificial systems*. University of Michigan Press, 1975.

[3] F. Liang & W.H. Wong. *Evolutionary monte carlo: Applications to Cp model sampling and change point problem.* StatisticaSinica, 10:317, 2000.

[4] M. Stephens, N. Smith, P. Donnelly. *A New Statistical Method for Haplotype Reconstruction for Population Data*. The American Journal of Human Genetics, 68:978-989, 2001.

[5] J. Marchini & Z. Su. *A C++ Program for Simulating Case and Control SNP Haplotypes*. 2006

[6] R. Kass & A. Raftery. *Bayes Factors*. J. Amer. Statist. Assoc. 90:733-795, 1995.

Acknowledgements

This work was supported in part by:

- Duke SPORE in Breast Cancer: 5P50CA068438-07
- North Carolina Ovarian Cancer Study: 1R01CA076016
- Bayesian Modeling and Optimal Design for Studies of Gene-Environment Association: 1R01HL090559-01
- NSF SCREMS award DMS-0422400.

Web Resources

★**Email**:

maw27@stat.duke.edu

★**R Package & Pdf copy of poster**:

www.stat.duke.edu/gbye/MISA.html