BAYESIAN MODEL SEARCH AND MULTILEVEL INFERENCE FOR SNP ASSOCIATION STUDIES: SUPPLEMENTARY MATERIALS

By Melanie A. Wilson[*,‡] Edwin S. Iversen[*,‡] Merlise A. Clyde[*,†]
Scott C. Schmidler and Joellen M. Schildkraut[*,‡]

Duke University

This supplement contains additional description on four topics appearing in the parent manuscipt. These are:

1. Derivation of the implied prior distribution on the regression coefficients when AIC is used to approximate the marginal likelihood in logistic regression,
2. Description of the marginal Bayes factor screen used to reduce the number of SNPs in the MISA analysis, and
3. Details of how the simulated genetic data sets used in the power analysis of MISA were created and information on the statistical software we developed for this purpose.
4. Location of the freely available software resources referred to in this and the parent document.

1. Implied Prior Distribution under AIC. Given that a closed–form expression for the marginal likelihood is not available for logistic regression, we have used the AIC to approximate the likelihood. In what follows, we determine a prior distribution on model coefficients that is consistent with AIC.

We assume a normal prior distributions on the d_γ–dimensional vector of regression coefficients (log odds ratios) of the form

$$p(\boldsymbol{\theta}_{\boldsymbol{\gamma}}|\mathcal{M}_{\boldsymbol{\gamma}}) \sim \mathsf{N}\left(\boldsymbol{t}_{\boldsymbol{\gamma}}, \frac{1}{k}\mathfrak{I}_{\boldsymbol{\gamma}}^{-1}\right),$$

[*]Partially supported by National Institute of Health grant NIH/NHLBI R01-HL090559.
[†]Partially supported by the National Science Foundation grants DMS-0342172 and DMS–0406115. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.
[‡]This work was supported by the Duke SPORE in Breast Cancer, P50-CA068438; the North Carolina Ovarian Cancer Study, R01-CA076016.

where $\mathfrak{I}_{\boldsymbol{\gamma}}$ is the observed Fisher information under model $\mathfrak{M}_{\boldsymbol{\gamma}}$ evaluated at the maximum likelihood estimates (MLEs) $\hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}}$. Setting the covariance matrix to be proportional to the inverse Fisher information ensures that the correlation structure in the prior distribution matches that of the likelihood.

In order to approximate the marginal likelihood we used a Laplace approximation based on expanding the log–likelihood in a second–order Taylor's series expansion about $\hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}}$:

$$\mathcal{L}(\boldsymbol{\theta}_{\boldsymbol{\gamma}} \mid \mathfrak{M}_{\boldsymbol{\gamma}}) \approx \mathcal{L}(\hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}} \mid \mathfrak{M}_{\boldsymbol{\gamma}}) - \frac{1}{2}(\boldsymbol{\theta}_{\boldsymbol{\gamma}} - \hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}})^T \mathfrak{I}_{\boldsymbol{\gamma}} (\boldsymbol{\theta}_{\boldsymbol{\gamma}} - \hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}})$$

leading to the approximate marginal likelihood

$$\begin{aligned} p(D \mid \mathfrak{M}_{\boldsymbol{\gamma}}) &\approx \exp\{\mathcal{L}(\hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}} \mid \mathfrak{M}_{\boldsymbol{\gamma}})\} \times \\ &\qquad \int K_{\boldsymbol{\theta}_{\boldsymbol{\gamma}}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}}, \mathfrak{I}_{\boldsymbol{\gamma}}^{-1}) \frac{1}{(2\pi)^{\frac{d_{\boldsymbol{\gamma}}}{2}}} |k\mathfrak{I}_{\boldsymbol{\gamma}}|^{\frac{1}{2}} K_{\boldsymbol{\theta}_{\boldsymbol{\gamma}}}(\boldsymbol{t}_{\boldsymbol{\gamma}}, \frac{1}{k}\mathfrak{I}_{\boldsymbol{\gamma}}^{-1}) d\boldsymbol{\theta}_{\boldsymbol{\gamma}} \\ &= \left(\frac{k}{k+1}\right)^{\frac{d_{\boldsymbol{\gamma}}}{2}} \exp\{\mathcal{L}(\hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}} \mid \mathfrak{M}_{\boldsymbol{\gamma}})\} K_{\hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}}}(\boldsymbol{t}_{\boldsymbol{\gamma}}, \frac{k+1}{k}\mathfrak{I}_{\boldsymbol{\gamma}}^{-1})\}; \end{aligned}$$

where $K_{\boldsymbol{\theta}_{\boldsymbol{\gamma}}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}}, \mathfrak{I}_{\boldsymbol{\gamma}}^{-1}) = \exp\{-\frac{1}{2}(\boldsymbol{\theta}_{\boldsymbol{\gamma}} - \hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}})^T \mathfrak{I}_{\boldsymbol{\gamma}} (\boldsymbol{\theta}_{\boldsymbol{\gamma}} - \hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}})\}$. Setting this approximate $\log(p(D \mid \mathfrak{M}_{\boldsymbol{\gamma}})$ equal to -0.5AIC we have equality when the prior mean $\boldsymbol{t}_{\boldsymbol{\gamma}}$ is set to $\hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}}$ where the right–most term vanishes, and $k = \frac{1}{\exp(2)-1}$. Roughly speaking, this implies that the prior standard deviation of any standardized log odds ratio is about 2.5. This suggests that the approximation of the marginal likelihood under AIC is reasonable for prior distributions with mean zero, as this provides enough dispersion to cover the range of log odds ratios anticipated.

2. Marginal Bayes Factor Screen. We used Laplace approximations to estimate the marginal Bayes Factors (BFs) used to screen the SNPs [Kass and Raftery 1995]. In particular, we estimated the marginal likelihood of each of the three genetic models of association (log–additive, dominant and recessive) and under the null model (model of no genetic association). The BF for a model of association is defined as the ratio of the marginal likelihood of that model of association to the marginal likelihood of the null model.

We accounted for missing genetic data by averaging marginal likelihoods over the $M = 100$ imputed genetic data sets. This affected only the calculations under the three genetic models of association, but not the null model. Hence the BF for an association was computed as the average of imputation–specific BFs.

In the ovarian cancer analysis, the model for each SNP was a logistic regression for disease status given the variable age and the model–specific genotype variable. Age was included in all models, including the 'null' model of no association. The simulation models were unadjusted as no design or confounder variables were simulated. We placed a normal, mean zero, standard deviation two prior on the parameter of the genetic effect variable and flat, improper priors on the remaining log odds ratio parameters. We ordered SNPs according to the maximum of the three Bayes factors and considered those with a maximum greater than or equal to one in the MISA model search. Our software for calculating marginal Bayes factors is included in the MISA R package.

3. Genetic Simulations. We used simulated case–control data to compare MISA and other commonly used procedures for genetic association studies. The simulated data sets were structured so as to reflect the details — genes, tag SNPs, LD structure, and sample size — of a NCOCS candidate pathway study comprised of 53 genes tagged by 508 tag SNPs. Genotypes were simulated in two stages. First, for each of 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], which provides estimates of the population haplotype distribution. Phase is a Bayesian method that obtains approximate samples from the posterior distribution of all possible haplotype pairs (H) given the observed genotypes (G) using Gibbs sampling and estimates recombination rates empirically from this sample. Second, given a model of association and the PHASE output, we generated case–control data at the selected tags using HAPGEN [Marchini and Su 2006]. Hapgen is a program that simulates haplotypes for a case–control sample of individuals given a set of population haplotypes and recombination rates for the regions of interest and choice of the hypothetical associated SNP and its allele–specific odds ratios.

We generated 124 simulated data sets as follows. Ten of the simulations are null; there are no associations in the genes of interest. The remaining 114 simulations assume that a randomly chosen subset of 9 genes are associated and that within the associated genes, a single, randomly chosen SNP is the source of the association. Within the 114 associated simulations, the associated tag SNPs were accorded an odds ratio (OR) of 1.25,1.5, 1.75, 2.0, or 2.25 and assumed to have either a dominant genetic parametrization, log–additive genetic parametrization or a recessive genetic parametrization. The marginal distribution over odds ratios is given in Figure 1. The marginal distribution over genetic models was uniform. The simulations used for the

power analysis can be found at the URL for the software.

We have also developed a software package, SimGbyE, that creates simulated case/control or survival data sets with one or more of the following assumed effects: genetic main effects (G), environmental main effects (E), Gene by Gene interactions (GbyG), Gene by environment interactions (GbyE). The assumed genetic one and two locus models of epistasis are chosen randomly from a set of models chosen from Li and Reich [2000]. Then given a set of assumed coefficients on the effects mentioned above, an outcome variable is simulated (case/control or survival) based on a set user specified distribution parameters. This package differs slightly from the method used to develop the simulations within this article by estimating the population haplotype distribution from HapMap instead of using PHASE to estimate the distribution from the set of control SNP genotypes in the NCOCS data.

The main function calls Hapgen to simulate one replicate from a specified chromosomal region given data from one of the HapMap II populations. The code generates samples of genotypes in a contiguous range of DNA using Hapmap release 21 (NCBI build 35) data. The position range may encompass an entire chromosome or simply bracket a gene or locus of interest. That function can also be used to simulate data from multiple independent regions to generate a candidate gene/pathway sample or a genome-wide sample. The default is to generate population-based genetic samples. However, to build genetic simulations with main effects only, parameters can be set so that Hapgen will randomly choose a variant in the specified region as the disease allele and generate a case-control sample. To build more complex associations we have written a wrapper function to take the genetic samples produced by Hapgen and simulate an outcome variable based on genetic main effects with multiple genetic parametrizations, environmental main effects, Gene by Gene interactions, and Gene by environment interactions.

4. Web Resources. The URL for the software for the methodology and simulations presented in this paper is:

`http://www.isds.duke.edu/gbye/packages.html.`

References.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90** 773–795.

Li, W. and Reich, J. (2000). A complete enumeration and classification of two-locus disease models. *Human Heredity* **50** 334.

Marchini, J. and Su, Z. (2006). Hapgen, a c++ program for simulating case and control snp haplotypes.

STEPHENS, M., SMITH, N. and DONNELLY, P. (2001). A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics* **68** 978–989.

Merlise A. Clyde
Edwin S. Iversen
Scott C. Schmidler
Melanie A. Wilson
Department of Statistical Science
Duke University
Durham, NC, 27708–0251
E-mail: clyde@stat.duke.edu
iversen@stat.duke.edu
scs@stat.duke.edu
maw27@stat.duke.edu

Joellen M. Schildkraut
Department of Community and Family Medicine
Duke University
Durham, NC; 27713
E-mail: schil001@mc.duke.edu