Bayesian Model Uncertainty and Prior Choice with Applications to Genetic Association Studies

by

Melanie A. Wilson

Department of Statistics
Duke University

Date: _____

Approved:

_____
Dr. Edwin S. Iversen, Advisor

_____
Dr. Merlise A. Clyde

_____
Dr. Sayan Mukherjee

_____
Dr. Elizabeth Hauser

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Statistics
in the Graduate School of Duke University
2010

Abstract

(Bayesian Model Uncertainty)

Bayesian Model Uncertainty and Prior Choice with Applications to Genetic Association Studies

by

Melanie A. Wilson

Department of Statistics
Duke University

Date: _____

Approved:

_____
Dr. Edwin S. Iversen, Advisor

_____
Dr. Merlise A. Clyde

_____
Dr. Sayan Mukherjee

_____
Dr. Elizabeth Hauser

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistics in the Graduate School of Duke University
2010

Abstract

The Bayesian approach to model selection allows for uncertainty in both model specific parameters and in the models themselves. Much of the recent Bayesian model uncertainty literature has focused on defining these prior distributions in an objective manner, providing conditions under which Bayes factors lead to the correct model selection, particularly in the situation where the number of variables, p, increases with the sample size, n. This is certainly the case in our area of motivation; the biological application of genetic association studies involving single nucleotide polymorphisms. While the most common approach to this problem has been to apply a marginal test to all genetic markers, we employ analytical strategies that improve upon these marginal methods by modeling the outcome variable as a function of a multivariate genetic profile using Bayesian variable selection. In doing so, we perform variable selection on a large number of correlated covariates within studies involving modest sample sizes.

In particular, we present an efficient Bayesian model search strategy that searches over the space of genetic markers and their genetic parametrization. The resulting method for Multilevel Inference of SNP Associations MISA, allows computation of multilevel posterior probabilities and Bayes factors at the global, gene and SNP level. We use simulated data sets to characterize MISA's statistical power, and show that MISA has higher power to detect association than standard procedures. Using data from the North Carolina Ovarian Cancer Study (NCOCS), MISA identifies variants

that were not identified by standard methods and have been externally 'validated' in independent studies.

In the context of Bayesian model uncertainty for problems involving a large number of correlated covariates we characterize commonly used prior distributions on the model space and investigate their implicit multiplicity correction properties first in the extreme case where the model includes an increasing number of redundant covariates and then under the case of full rank design matrices. We provide conditions on the asymptotic (in n and p) behavior of the model space prior required to achieve consistent selection of the global hypothesis of at least one associated variable in the analysis using global posterior probabilities (i.e. under 0-1 loss). In particular, under the assumption that the null model is true, we show that the commonly used uniform prior on the model space leads to inconsistent selection of the global hypothesis via global posterior probabilities (the posterior probability of at least one association goes to 1) when the rank of the design matrix is finite. In the full rank case, we also show inconsistency when p goes to infinity faster than $\sqrt{n}$. Alternatively, we show that any model space prior such that the global prior odds of association increases at a rate slower than $\sqrt{n}$ results in consistent selection of the global hypothesis in terms of posterior probabilities.

Contents

List of Tables

List of Figures

Acknowledgements

I am so grateful for all the wonderful people who have supported me throughout my academic career at Duke. In particular, I would like to thank my advisors Ed Iversen & Merlise Clyde for their constant encouragement and patience over the years. Their knowledge and insight have been invaluable. I would also like to thank Sayan Mukherjee and Robert Wolpert for their added support and for providing such a great example of how to maintain a balance between work and play. Furthermore, I would also like to thank all of the groups, agencies and societies that have made my academic pursuits possible through grants and awards: the NSF, NIH, ISBA, and all of the members of the NCOCS.

I have been fortunate to develop many great friendships while at Duke. Thank you to the wonderful women who have always inspired me to keep going, Emily and Ioanna, and to all of the men who have never ceased to entertain me with enlightening dinner conversations: James, Carlos, Scottland, Eric, Jared, Simon, Richard, Ken and Lane.

I especially want to thank my best friends David & Kristian. Not one step of this journey would have been the same without you both! To my soon to be husband, David: Your dedication to me throughout some of the hardest times in our lives has never been overlooked. I cannot begin to thank you enough for your constant love and support. To my partner in crime Kristian: Thank you so much for always being there for me and showing me that I am not the only one who loves to hash out

mathematical proofs, belt Kelly Clarkson, and discuss the inner-workings of men all within the same car ride to Chapel Hill.

Finally, I would not be who I am today without all of God's blessings, especially that of my family. To my mom and dad, Morton & Evelyn: I could have not asked for better role models and I know that I would not be the person that I am today without you. Thank you for your continued guidance and support! To my brother Tim: Thank you for always inspiring me with your outlook on life and most importantly, for finally giving me the sister that I have always wanted in Andrea!

In memory of my grandmother, Mary Jane Wilson: Thank you for instilling in me your spirit and strength, you will never be forgotten.

1

Introduction

The Bayesian approach to model selection has been successfully applied to a wide range of statistical model classes such as linear regression, generalized linear models, survival analysis, tree models and graphical models [28, 11]. The Bayesian paradigm allows for uncertainty in both model specific parameters and in the models themselves. This is done by specifying prior probabilities on all models under consideration in addition to the prior distributions for all model specific parameters. Much of the recent Bayesian model uncertainty literature has focused on defining objective prior distributions for the model specific parameters [38, 14, 6, 21, 65], or prior distributions for the model space [36, 14, 53]. Due to the high-dimensionality of most interesting problems much work has also been done to investigate the asymptotic behavior of pairwise model Bayes factors (giving the ratio of the weight of evidence between any two single models) under several priors for model specific parameters as the sample size, n, and number of covariates, p, go to infinity [7, 43, 25], and to develop stochastic model search algorithms for sampling from the posterior distribution over models because of the non-enumerability of the model space [12, 27, 54, 26, 11, 39]. However, the properties of the model space priors and stochastic model search al-

gorithms have not been fully investigated for model uncertainty problems involving a large number of correlated covariates. High-dimensional problems of this nature arise in many applications as diverse as econometrics, environmental science and bioinformatics. In light of this, my work consists of investigating properties of priors over the model space and developing analytical strategies for model search in high-dimensional problems involving correlated covariates.

Another area of research in model uncertainty problems involving large numbers of correlated covariates is the choice of posterior quantities to use for inference in model or variable selection problems. The motivating example is the dilution of marginal inclusion probabilities [22, 10]. In cases with covariates that are highly correlated with one another, marginal inclusion probabilities may underestimate the significance of an association for a given covariate. This occurs because the covariates may provide competing explanations for the association, thereby diluting or distributing the probability over several covariates. I am, therefore, are also interested in investigating alternative posterior quantities that can be used as inference tools in model uncertainty and variable selection problems.

My research in this area is motivated by genetic association studies whose goal is to identify single nucleotide polymorphisms (SNPs) that influence an individual's risk for developing complex disease. A SNP is the most common type of DNA variation and occurs when one nucleotide of the DNA sequence is altered. These variations, and the studies involving them can help to pinpoint genetic risk factors and to understand the complex etiology of disease. With the increasing popularity of modeling SNPs in genetic/epidemiological studies comes an increased need for statistical tools that are able to detect complex associations between a large number of correlated covariates and a response variable (in most cases a disease phenotype). The most common approach to this problem has been to apply a simple marginal test to all genetic markers. Like stepwise logistic regression [1], lasso [45, 56, 64] and

logic regression [49, 33, 52], the methods that are the focus of my research aim to improve upon marginal, SNP–at–a–time methods by modeling the outcome variable as a function of a multivariate genetic profile, which provide measures of association that are adjusted for the remaining markers.

1.1 Model Search Algorithms & Posterior Quantities for Inference

Chapter 2 describes a Bayesian model search algorithm for genetic association studies called MISA: Multilevel Inference for SNP Association Studies. The model space in this application is the set of logistic regression models determined by all combinations of the genetic markers with three possible genetic parameterizations for each marker. It provides multi-level inferences at the global, gene and SNP level. MISA uses a stochastic model search algorithm that is based on the Evolutionary Monte Carlo (EMC) algorithm [39] to search through the non-enumerable and multi-modal model space. In MISA, the model search algorithm is based on a combination of parallel tempering [23] and a genetic algorithm [29]. The genetic algorithm incorporates move types into our model search that mimic the idea of evolution in that individuals (or in this case models) compete and mate to produce increasingly stronger individuals. These genetic moves allow for an increase in learning from past models that have been proposed and, in turn, an increased efficiency in the model search by proposing new models that are combinations of past proposed models. Parallel tempering allows for an increased temperature (or variance) in the model proposal distributions and allows the algorithm to easily escape from local modes in the stationary distribution.

MISA also addresses the problem of diluted marginal inclusion probabilities by focusing on inference at the group level as well as the marginal level by defining groups of covariates (a set of SNPs for the same gene) where the amount of correlation between covariates across different groups is typically negligible. Using computer software that I developed in R for the computations and inference summaries

described in MISA and software that I developed to create realistic genetic simulations, I show that the analytical strategies used within MISA lead to an increase in power to detect true associations over more commonly used methods.

1.2 Prior Choice on the Model Space

Chapter 3 formally investigates characteristics of commonly used model space priors. In Bayesian model uncertainty problems involving a large number of covariates it is common practice to place "non-informative" prior distributions such as the uniform distribution across all possible models (which leads to a Binomial distribution on model size, $\mathsf{Bin}(p, 1/2)$) since elicitation of the high dimensional model space is impractical. An alternative to the $\mathsf{Bin}(p, 1/2)$ prior is the Beta-Binomial distribution on model size, $\mathsf{BB}(a, b)$, which allows for additional dispersion and provides added robustness to prior misspecification [36, 53, 14]. With this $\mathsf{BB}(a, b)$ prior the hyper-parameters to the Beta distribution, a, b, must be specified. A common "non-informative" default choice is $a = b = 1$. Although the above mentioned priors are non-informative with respect to individual models, they can be quite informative about other characteristics of the model space including the hypothesis that at least one covariate is associated. In particular, under the $\mathsf{Bin}(p, 1/2)$ and $\mathsf{BB}(1, 1)$ priors on model size, I show that the prior probability of at least one association increases as the number of covariates under consideration, p, increases. In contrast, I have found that the $\mathsf{BB}(1, \lambda p)$ prior distribution on model size, where λ is a positive constant, results in a constant prior probability of at least one association as p increases.

Chapter 3 also formally investigates the asymptotic behavior (in n and p) of the global posterior probabilities of the null hypothesis of no associated covariates and the alternative hypothesis of at least one association under these commonly used priors. We generalize these results by giving necessary conditions on the model space priors to achieve selection consistency of the global hypotheses.

1.3 Organization of the Thesis

The remainder of the thesis is organized as follows: Chapter 2 describes the Bayesian model uncertainty algorithm, Multilevel Inference of SNP Associations, that is applied to a set of logistic regression models and searches through the model space of multiple genetic effects and multiple genetic parametrizations for each effect using Evolutionary Monte Carlo. Chapter 3 then examines theoretical properties of priors distributions on the model space for Bayesian model uncertainty algorithms in the linear regression framework. Finally, Chapter 4 is a brief discussion of the results herein and states future directions for this work.

2

Bayesian Model Search and Multilevel Inference for SNP Association Studies

Recent advances in genotyping technology have resulted in a dramatic change in the way hypothesis–based genetic association studies are conducted. While previously investigators were limited by costs to investigating only a handful of variants within the most interesting genes, researchers may now conduct candidate–gene and candidate–pathway studies that encompass many hundreds or thousands of genetic variants, often single nucleotide polymorphisms (SNPs). For example, the North Carolina Ovarian Cancer Study (NCOCS) [50], an ongoing population–based case–control study, genotyped 2129 women at 1536 SNPS in 170 genes on 8 pathways, were 'pathway' is defined as a set of genes thought to be simultaneously active in certain circumstances.

The analytic procedure most commonly applied to association studies of this scale is to fit a separate model of association for each SNP that adjusts for design and confounder variables. As false discoveries due to multiple testing are often a concern, the level of significance for each marginal test of association is adjusted using Bonferroni or other forms of false discovery correction [59, 61, 1]. While these methods have

been shown to be effective in controlling the number of false discoveries reported, correlations between the markers may limit the power to detect true associations [16]. The NCOCS study provides a case–in–point. When simple marginal methods are applied to the NCOCS data, no SNPs are identified as notable.

Marginal SNP-at-a-time methods do not address directly many of the scientific questions in candidate pathway studies, such as 'Is there an overall association between a pathway and the outcome of interest?' and 'Which genes are most likely to be driving this association?' The Multilevel Inference for SNP Association (MISA) method we describe here is designed to simultaneously address these questions of association at the level of SNP, gene, and pathway.

MISA, in contrast to the marginal methods, identifies ten SNPs of interest in the NCOCS study. To date, one of these (ranked tenth by MISA) has been validated in external data by a large multi–center consortium [51]; additional testing is underway for other top SNPs discovered by MISA. To buttress this empirical evidence, we demonstrate using simulation studies (Section 2.3) that MISA has higher power to detect associations than other simpler procedures, with a modest increase in the false discovery rate (Figure 2.1).

In the next section, we describe the Bayesian hierarchical model behind MISA and highlight how it addresses many of the key issues in analysis of SNP association studies: identification of associated SNPs and genetic models, missing data, inference for multi–level hypotheses and control of the false discovery rate. Like stepwise logistic regression [1], lasso [45, 56, 64] and logic regression [49, 33, 52], MISA improves upon marginal, SNP–at–a–time methods by modeling the outcome variable as a function of a multivariate genetic profile, which provides measures of association that are adjusted for the remaining markers. MISA uses Bayesian Model Averaging [28] to combine information from multiple models of association to address the degree to which the data support an association at the level of individual

SNPs, genes, and pathways, while taking into account uncertainty regarding the best genetic parametrization. By using model averaging, MISA improves upon methods that select a single model, which may miss important SNPs because of LD structure. We show how the prior distribution on SNP inclusion provides a built–in multiplicity correction. Because missing data are a common phenomenon in association studies, we discuss two options for handling this problem.

In Section 2.2, we present an Evolutionary Monte Carlo algorithm to efficiently sample models of association according to their posterior probabilities. In Section 2.3 we apply our method to simulated data sets and demonstrate that MISA outperforms less complex and more commonly used alternatives for detecting associations in modestly powered candidate–gene case–control studies. The simulation approach may also be used to guide selection of the prior hyper–parameters given the study design. In Section 2.4 we return to the NCOCS study and present results from the analysis of a single pathway from that study. We examine the sensitivity of results to prior hyperparameter choice and methods for imputing missing data. We conclude in Section 2.5 with recommendations and a discussion of future extensions.

2.1 Models of Association

We consider SNP association models with a binary phenotype, such as presence or absence of a disease as in case-control designs. For $i = 1, \ldots, n$, let D_i indicate the disease status of individual i, where $D_i = 1$ represents a disease case and $D_i = 0$ represents a control. For each individual, we have S SNP measurements, where SNP s is either homozygous common (A_sA_s), heterozygous (a_sA_s or A_sa_s), homozygous rare (a_sa_s), or missing and is coded as 0, 1, 2, representing the number of rare alleles, or NA if the SNP is missing for that individual. We will discuss methods for imputing missing SNP data in Section 2.1.3. In addition to the SNP data, for each individual we have a q–dimensional vector $\mathbf{z}_i^T$ of design and potential confounding variables that

will be included in all models, henceforth referred to as 'design' variables.

We use logistic regression models to relate disease status to the design variables and subsets of SNPs. We denote the collection of all possible models by $\boldsymbol{\mathcal{M}}$. An individual model, denoted by $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}$, is specified by the S dimensional vector γ, where γ_s indicates the inclusion and SNP-specific genetic parametrization of SNP s in model $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}$: $\gamma_s = 0$ if $\text{SNP}_s \notin \boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}$, $\gamma_s = 1$ if $\text{SNP}_s \in \boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}$ with a log–additive parametrization, $\gamma_s = 2$ if $\text{SNP}_s \in \boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}$ with a dominant parametrization, and $\gamma_s = 3$ if $\text{SNP}_s \in \boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}$ with a recessive parametrization. When no homozygous rare cases or controls are observed, we fix the genetic parametrization to be log–additive. Under each of these genetic parametrizations, SNP s may be encoded using one degree of freedom. In particular, for the log–additive model, the design variable representing SNP s is a numeric variable equal to the number of copies of the risk allele a_s. For the dominant model, we use an indicator variable of whether allele a_s is present (homozygous rare or heterozygous) and for the recessive model, an indicator variable of whether SNP s has the homozygous rare genotype. For each individual, the logistic regression under model $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}$ assuming complete data is given by

$$\text{logit}(p(D_i = 1|\mathbf{z}_i, \mathbf{x}_{\boldsymbol{\gamma}_i}, \boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}})) = \alpha_0 + \mathbf{z}_i^T \boldsymbol{\alpha} + \mathbf{x}_{\boldsymbol{\gamma}_i}^T \boldsymbol{\beta}_{\boldsymbol{\gamma}} \tag{2.1}$$

where $\mathbf{x}_{\boldsymbol{\gamma}_i}$ represents the coding of SNPs in model $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}$ and $\boldsymbol{\theta}_{\boldsymbol{\gamma}}$ is the vector of model specific parameters $(\alpha_0, \boldsymbol{\alpha}^T, \boldsymbol{\beta}_{\boldsymbol{\gamma}}{}^T)$, with intercept α_0, vector of design variable coefficients $\boldsymbol{\alpha}$, and log–odds ratios $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$. Prospective models for disease outcome given multivariate genetic marker data as in equation (2.1) provide measures of association that are adjusted for other markers which can increase the power to detect associations [1], however, one is faced with an extremely large collection of possible models. While stepwise selection methods may be used to select a single model [13], this leads to difficulty in interpreting the significance of SNPs in the selected model. Bayesian model averaging is an alternative to stepwise selection methods and is an effective

approach for identifying subsets of likely associated variables, for prioritizing them and for measuring overall association in the presence of model uncertainty (see the review articles by [28] and [11] and references therein).

2.1.1 Posterior Inference

Posterior model probabilities measure the degree to which the data support each model in a set of competing models. The posterior model probability of any model $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}$ in the space of models $\boldsymbol{\mathcal{M}}$ is expressed as

$$p(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} \mid D) = \frac{p(D \mid \boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}})p(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}})}{\sum_{\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} \in \boldsymbol{\mathcal{M}}} p(D \mid \boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}})p(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}})} \quad \text{for } \boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} \in \boldsymbol{\mathcal{M}}$$

where $p(D \mid \boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}})$ is the (marginal) likelihood of model $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}$ obtained after integrating out model–specific parameters $\boldsymbol{\theta}_{\boldsymbol{\gamma}}$ with respect to their prior distribution, and $p(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}})$ is the prior probability of $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}$.

While posterior probabilities provide a measure of evidence for hypotheses or models, it is often difficult to judge them in isolation as individual model probabilities may be "diluted" as the space of models grows [10, 22, 11]. Bayes factors (BF) [32] compare the posterior odds of any two models (or hypotheses) to their prior odds

$$\text{BF}(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}_1} : \boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}_2}) = \frac{p(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}_1} \mid D)/p(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}_2} \mid D)}{p(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}_1})/p(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}_2})}$$

and measures the *change* in evidence (on the log scale) provided by data for one model, $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}_1}$, to another, $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}_2}$ or for pairs of hypotheses. [24] and [58] provide a discussion on the usefulness of Bayes factors in the medical context and [62] illustrates their use in controlling false discoveries in genetic epidemiology studies. Below we define Bayes factors for quantifying association at multiple levels (global, gene, and SNP) and assessing the most likely SNP-specific genetic parametrization.

Global Bayes Factor

The Bayes factor in favor of H_A, the alternative hypothesis that there is at least one SNP associated with disease, to H_0, the null hypothesis that there is no association between the SNPs under consideration and disease, measures the relative weight of evidence of H_A to H_0. The null model corresponding to H_0 is the model which includes only design variables and no SNPs, and is denoted $\boldsymbol{\mathcal{M}}_0$. The alternative hypothesis is represented by all of the remaining models in $\boldsymbol{\mathcal{M}}$. Because the space of models is large, the null model (or any single model in general) may receive small probability (both prior and posterior), even when it is the highest posterior probability model (this illustrates the dilution effect of large model spaces); Bayes factors allow one to judge how the posterior odds compare to one's prior odds.

The Global Bayes factor for comparing H_A to H_0 may be simplified to

$$\mathrm{BF}(H_A : H_0) = \sum_{\boldsymbol{\mathcal{M}}_\gamma \in \boldsymbol{\mathcal{M}}} \mathrm{BF}(\boldsymbol{\mathcal{M}}_\gamma : \boldsymbol{\mathcal{M}}_0) p(\boldsymbol{\mathcal{M}}_\gamma \mid H_A) \tag{2.2}$$

which is the weighted average of the individual Bayes factors $\mathrm{BF}(\boldsymbol{\mathcal{M}}_\gamma : \boldsymbol{\mathcal{M}}_0)$ for comparing each model in H_A to the null model with weights given by the prior probability of $\boldsymbol{\mathcal{M}}_\gamma$ conditional on being in H_A, $p(\boldsymbol{\mathcal{M}}_\gamma \mid H_A)$. Because the alternative is a composite hypothesis, the resulting Global Bayes factor is not independent of the prior distribution on the models that comprise the alternative, thus the prior distribution on models will play an important role in controlling the (relative) weights that models of different sizes receive. For a large number of SNPs, it is impossible to enumerate the space of models and posterior summaries are often based on models sampled from the posterior distribution. In equation (2.2), if we replace the average over all models in H_A with the average over the models in $\boldsymbol{\mathcal{S}}$ (the collection of unique models sampled from the posterior distribution), the result

$$\mathrm{BF}(H_A : H_0) > \mathrm{BF}_S(H_A : H_0) \equiv \sum_{\boldsymbol{\mathcal{M}}_\gamma \in \boldsymbol{\mathcal{S}}} \mathrm{BF}(\boldsymbol{\mathcal{M}}_\gamma : \boldsymbol{\mathcal{M}}_0) p(\boldsymbol{\mathcal{M}}_\gamma \mid H_A)$$

is a lower bound for the Bayes factor for testing global association. If the lower bound indicates evidence of an association, then we can be confident that this evidence will only increase as we include more models.

SNP Bayes Factors

While it is of interest to quantify association at the global level, interest is primarily in identifying the gene(s) and variant(s) within those genes that drive the association. We begin by defining SNP inclusion probabilities and associated Bayes factors. These marginal summaries are adjusted for the other potentially important SNPs and confounding variables and provide a measure of the strength of association at the level of individual SNPs. Given each sampled model $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} \in \boldsymbol{\mathcal{S}}$ and the model specification vectors $\gamma = (\gamma_1, \gamma_2, \cdots, \gamma_S)$ previously defined in Section 2.1, the inclusion probability for SNP s is estimated as:

$$p(\gamma_s \neq 0 \mid D) = \sum_{\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} \in \boldsymbol{\mathcal{S}}} 1_{(\gamma_s \neq 0)}\ p(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} \mid D, \boldsymbol{\mathcal{S}}) \tag{2.3}$$

where $p(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} \mid D, \boldsymbol{\mathcal{S}})$ is the posterior probability of a model re-normalized over the sampled model space. The SNP Bayes factor is the ratio of the posterior odds of the SNP being associated to the prior odds of the same, and is defined as:

$$\text{BF}(\gamma_s \neq 0 : \gamma_s = 0) = \frac{p(\gamma_s \neq 0 \mid D)}{p(\gamma_s = 0 \mid D)} \div \frac{p(\gamma_s \neq 0)}{p(\gamma_s = 0)};$$

where $p(\gamma_s \neq 0)$ is the prior probability of SNP s being associated. Estimates of the SNP Bayes Factor may be obtained using the estimated SNP inclusion probabilities from (2.3).

Gene Bayes Factors

In cases where there are SNPs in Linkage Disequilibrium (LD), SNP inclusion probabilities may underestimate the significance of an association at a given locus. This

occurs because SNPs in LD may provide competing explanations for the association, thereby diluting or distributing the probability over several markers. Since the amount of correlation between markers across different genes is typically negligible, calculating inclusion probabilities and Bayes factors at the gene level will not be as sensitive to this dilution. A gene is defined to be associated if one or more of the SNPs within the given gene are associated. Hence we define the gene inclusion probability as:

$$p(\Gamma_g = 1 \mid D) = \sum_{\boldsymbol{\mathcal{M}}_\gamma \in \boldsymbol{\mathcal{S}}} 1_{(\Gamma_g = 1)}\ p(\boldsymbol{\mathcal{M}}_\gamma \mid D, \boldsymbol{\mathcal{S}});$$

where $\Gamma_g = 1$ if at least one SNP in gene g is in model $\boldsymbol{\mathcal{M}}_\gamma$ and is zero otherwise. The gene Bayes factor is defined as:

$$\mathrm{BF}(\Gamma_g = 1 : \Gamma_g = 0) = \frac{p(\Gamma_g = 1 \mid D)}{p(\Gamma_g = 0 \mid D)} \div \frac{p(\Gamma_g = 1)}{p(\Gamma_g = 0)};$$

where $p(\Gamma_g = 1)$ is the prior probability of one or more SNPs in gene g being associated.

Interpreting Evidence

Jeffreys [30, page 432] presents a descriptive classification of Bayes factors into "grades of evidence" (reproduced in Table 2.1) to assist in their interpretation (see [32]). In the context in which he presents the grades, he defined the Bayes factor assuming equal prior odds, making it equivalent to posterior odds and enabling a meaningful interpretation in terms of probabilities. It is not clear whether he intended his descriptive grades to be used more broadly for interpreting Bayes factors or for interpreting posterior probabilities.

Jeffreys was well aware of the issues that arise with testing several simple alternative hypotheses against a null hypothesis [30, Section 5.04], noting that if one were to test several hypotheses separately that by chance one might find one of the Bayes

factors to be less than one even if all null hypotheses were true. He suggested that, in this context, the Bayes factors needed to be "corrected for selection of hypotheses" by multiplying by the prior odds.

Experience has shown that detectable SNP associations are relatively infrequent, hence the prior odds of any given SNP being marginally associated in the typical genetic association study should be small. For this reason, [58] suggest that marginal Bayes factors calculated assuming equal prior odds be interpreted in light of a prior odds more appropriate to the study at hand. Our approach to the problem of exploring multiple hypotheses is to embed each of the potential submodels (corresponding to a subset of SNPs) into a single hierarchical model. Unlike the marginal (one-at-a-time) Bayes factors in [58] that are independent of the prior odds on the hypotheses, our SNP Bayes factors are based on comparing composite hypotheses and hence do depend on the prior distribution over models, which implicitly adjusts for the selection of hypotheses.

While Bayes factors do not provide a measure of *absolute* support for or against a hypothesis (except with even prior odds), the log Bayes factor does provide a coherent measure of how much the data *change* the support for the hypothesis (relative to the prior) [35]. Applying Jeffreys grades to Bayes factors using priors distributions that account for competing hypotheses provides an idea of the impact of the data on changing prior beliefs, but ultimately posterior odds provide a more informative measure of evidence and model uncertainty.

2.1.2 Prior Distributions, Laplace Approximations and Marginal Likelihoods

We assume normal prior distributions for the coefficients $\boldsymbol{\theta}_{\boldsymbol{\gamma}}$ with a covariance matrix that is given by a constant $1/k$ times the inverse Fisher Information matrix. For logistic regression models, analytic expressions for $p(D \mid \boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}})$ are not available and Laplace approximations or the Bayes Information Criterion are commonly used to

Grade	BF($H_A : H_0$)	Evidence against H_0
1	1 to 3.2	Indeterminate
2	3.2 to 10	Positive
3	10 to 31.6	Strong
4	31.6 to 100	Very Strong
5	> 100	Decisive

Table 2.1: Jeffrey's grades of evidence [30, page 432].

approximate the marginal likelihood [47, 62, 5]. Using a Laplace approximation with the normal prior distribution (see Appendix A), the posterior probability of model $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}$ takes the form of a penalized likelihood

$$p(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} \mid D) \propto \exp\{-\frac{1}{2}[\text{dev}(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}; D) + \text{pen}(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}})]\} \tag{2.4}$$

where $\text{dev}(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}; D) = -2\log(p(D \mid \hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}}, \boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}))$ is the model deviance, and the penalty term $\text{pen}(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}})$ encompasses a penalty on model size induced by the choice of k in the prior distribution on coefficients $\boldsymbol{\theta}_{\boldsymbol{\gamma}}$ and the prior distribution over models. Because we expect that effect sizes will be small, we calibrate the choice of k based on the Akaike information criterion (see Appendix A), leading to

$$\text{pen}(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}) = 2(1 + q + s_{\boldsymbol{\gamma}}) - 2\log(p(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}})).$$

2.1.3 Missing Data

The expression in (2.4) assumes complete data on all SNPs. Missing SNP data, unfortunately, are the norm rather than the exception in association studies. Removing all subjects with any missing SNP genotype data will typically result in an unnecessary loss of information and potential bias of estimated effects if the missing data are non-ignorable. It is possible, however, to exploit patterns in LD to efficiently impute the missing genotypes given observed data [1]. We use fastPHASE [57, 55] to sample haplotypes and missing genotypes (D^{m}) given the observed unphased genotypes

(D^{o}). This assumes that the pattern of missing data is independent of case-control status, which, if not true may lead to serious biases [9]. This assumption may be examined by using indicator variables of missingness as predictors in MISA.

The posterior probabilities of models given the data are obtained by averaging the marginal likelihood of a model over imputed genotype data:

$$p(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} \mid D) \propto \int \exp\{-\frac{1}{2}[\mathrm{dev}(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}; D, D^{\mathrm{o}}, D^{\mathrm{m}}) + \mathrm{pen}(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}})]\} p(D^{\mathrm{m}} \mid D^{\mathrm{o}}) dD^{\mathrm{m}}$$

$$\approx \frac{1}{M} \sum_{i=1}^{I} \exp\{-\frac{1}{2}[\mathrm{dev}(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}; D^{\mathrm{o}}, D_i^{\mathrm{m}}) + \mathrm{pen}(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}})]\} \equiv \Psi(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}) \quad (2.5)$$

where I is the number of imputed data sets, $\mathrm{dev}(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}; D, D^{\mathrm{o}}, D^{\mathrm{m}})$ is the deviance based on the completed data, and $\Psi(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}})$ is an estimate of the un-normalized posterior model probability for model $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}$. We have found that the number of imputed sets must be on the order of $I = 100$ to provide accurate estimates of posterior quantities. This has a significant computational impact in the model search algorithm described in Section 2.2. As a simple alternative, we approximate (2.5) by a modal approximation, where the missing genotypes are imputed with the mode of the sampled genotypes using fastPHASE. While it is well known that plugging in a single estimate for the missing data under-estimates uncertainty, the modal approximation provides dramatic computational savings. In Section 2.4 we examine the sensitivity of results to the method of imputing missing data and find that the modal approximation gives comparable results for SNP BFs.

2.1.4 Choice of Prior Distribution on Models

The prior distribution on the space of models $\boldsymbol{\mathcal{M}}$, $p(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}})$, completes our model specification. The frequentist approach for SNP association studies usually involves some form of adjustment for multiple-testing, which can, in effect, penalize the researcher who looks beyond single-SNP models of association to multiple SNP models

or models of interactions. Under the Bayesian approach, posterior evidence in the data is judged against the prior odds of an association using Bayes factors, which should not be affected by the number of tests that an investigator chooses to carry out [1].

While it has been common practice to adopt a "non-informative" uniform distribution over the space of models for association (this is after marginalizing over the possible genetic models for each SNP), this choice has the potentially undesirable "informative" implication that $\frac{1}{2}$ of the SNPs are expected to be associated *a priori*, and the prior odds of at least one SNP being included (which is used in the global Bayes factor) depends on the number of tests (2^S) (Table 3.1).

A recommended alternative is the Beta-Binomial distribution on the model size, which provides over-dispersion, added robustness to prior misspecification, and multiplicity corrections as a function of the number of variables [36, 53, 14]. We construct a hierarchical prior distribution over the space of models defined by subsets of SNPs and their genetic parametrizations as follows. For any SNP included in the model, we assign a uniform distribution over the possible genetic parametrizations. The prior distribution on the model size $s_{\boldsymbol{\gamma}}$ is $\mathsf{Bin}(S, \rho)$ conditional on ρ, and for the last stage, ρ is assigned a $\mathsf{Beta}(a, b)$ distribution. Integrating over the distribution on ρ, leads to the $\mathsf{BB}(a, b)$ distribution on model size,

$$p(s_{\boldsymbol{\gamma}}) = \frac{B(s_{\boldsymbol{\gamma}} + a, S - s_{\boldsymbol{\gamma}} + b)}{(S+1)B(s_{\boldsymbol{\gamma}} + 1, S - s_{\boldsymbol{\gamma}} + 1)B(a, b)} \tag{2.6}$$

and the following distribution on models,

$$p(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}) = \left(\frac{1}{3}\right)^{s_{\boldsymbol{\gamma}}} \frac{B(s_{\boldsymbol{\gamma}} + a, S - s_{\boldsymbol{\gamma}} + b)}{B(a, b)} \tag{2.7}$$

where $B(\cdot, \cdot)$ is the beta function and the factor of $1/3$ accounts for the distribution over genetic parametrizations.

Default Hyper-Parameter Choice

Following [36] and [53], we recommend $a = 1$ as a default, so that the prior distribution on model size is non-increasing in $s_{\boldsymbol{\gamma}}$. The hyper–parameter b can then be chosen to reflect the expected model size, global prior probability of at least one association, or the marginal prior odds that any SNP is associated (Table 3.1). A

	Binomial $(S, 1/2)$	Beta-Binomial $(1, 1)$	Beta-Binomial $(1, \lambda S)$
Expected Model Size	$\frac{S}{2}$ (∞)	$\frac{S}{2}$ (∞)	$\frac{S}{\lambda S+1}$ $(\frac{1}{\lambda})$
Global Prior Odds of an Association	$\frac{2^{2S}}{2^S+1}$ (∞)	S (∞)	$\frac{1}{\lambda}$
Marginal Prior Odds of an Association	1	1	$\frac{1}{\lambda S}$ (0)
Prior Odds of Adding a Variable	1	$\frac{s_{\boldsymbol{\gamma}}+1}{S-s_{\boldsymbol{\gamma}}}$ (0)	$\frac{s_{\boldsymbol{\gamma}}+1}{(\lambda+1)S-s_{\boldsymbol{\gamma}}-1}$ (0)

Table 2.2: General prior characteristics and limiting behavior (in parentheses) of the $\mathsf{Bin}(S, 1/2)$, $\mathsf{BB}(1, 1)$ and $\mathsf{BB}(1, \lambda S)$ distribution on model size.

default choice is to set $b = 1$, leading to a uniform distribution on model size [36, 53]. Like the binomial distribution, the $\mathsf{BB}(1, 1)$ distribution results in an expected model size of $\frac{S}{2}$ (Table 3.1), although the $\mathsf{BB}(1, 1)$ distribution has a larger variance than the $\mathsf{Bin}(S, 1/2)$. Alternatively, if b is proportional to S, $b = \lambda S$ the expected model size approaches a limit of $\frac{1}{\lambda}$ as S approaches infinity.

The choices for hyperparameters have implications for the global Bayes factor. The $\mathsf{BB}(1, 1)$ has a global prior odds of association equal to the number of SNPs, S, and would be appropriate for the case where increasing the number of SNPs under consideration reflects increased prior certainty that an overall (global) association can be detected. Under the $\mathsf{BB}(1, \lambda S)$, the global prior odds are constant, $1/\lambda$, reflecting a prior odds for overall association that is independent of the number of

genes/SNPs tagged. Also, with both Beta–Binomial prior distributions, the prior odds of incorporating an additional SNP in any model decreases with model size $s_{\boldsymbol{\gamma}}$ and approaches 0 in the limiting case as the number of SNPs, S, increases. This provides an implicit multiple testing correction in the number of SNPs (rather than tests) that are included in the study of interest. The $\mathsf{BB}(1, \lambda S)$ achieves this by keeping the global (pathway) prior odds of an association constant while decreasing the marginal prior odds of any one of the SNPs being associated as the number of SNPs increases. As a skeptical "default" prior, we suggest the hyper–parameters $a = 1$ and $b = S$ which leads to the global prior odds of there being at least one association of 1 and the marginal prior odds of any single SNP being associated of $1/S$.

2.2 Stochastic Search for SNPs

Given the number of SNPs under consideration, enumeration of all models for S greater than 25–30 is intractable. While it is possible to enumerate all single variable SNP models, the number of models with 2 or 3 SNPs allowing for multiple genetic parametrizations is in the millions or more for a typical modern hypothesis–oriented study. Stochastic variable selection algorithms, see [11] for a review, provide a more robust search procedure than stepwise methods, but also permit calculation of posterior probabilities and Bayes factors based on a sample of the most likely candidate models from the posterior distribution.

MISA makes use of a stochastic search algorithm based on the Evolutionary Monte Carlo (EMC) algorithm of [39]. EMC is a combination of parallel tempering [23] and a genetic algorithm [29] and samples models based on their "fitness" (for more details see Appendix B). While originally designed to find optimal models based on

AIC, in our application the fitness of the models is given by $\psi(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}})$

$$\psi(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}) = \log(\Psi(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}))$$

where $\Psi(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}})$ is defined in equation (2.5) and is equal to the log of the un-normalized posterior model probability. This results in models being generated according to their posterior probability.

The EMC algorithm requires that we specify the number of parallel chains that are run and the associated temperature for each chain that determines the degree of annealing. If the temperatures are too spread out for the number of chains, then the algorithm may exhibit poor mixing and slow convergence. [39] show that even with all chains run at a temperature of 1 (no annealing), EMC outperforms alternative sampling methods such as Gibbs sampling and Reversible Jump MCMC in problems where strong correlations among the predictor variables lead to problems with exploring multiple modes in the posterior distribution. We have found that a constant temperature ladder with 5 parallel chains provides good mixing and finds more unique models than using a custom temperature ladder based on the prescription in [39], and recommend the constant temperature ladder as a default. To assess convergence, we take two independent EMC runs using randomly chosen starting points and examine trace plots of the fitness function. We use the marginal likelihoods from the set of unique models in the sample for inference and compute estimates of marginal posterior inclusion probabilities for each run. We continue running the two instances of the EMC algorithm until the posterior probabilities derived from each are sufficiently close. This leads to longer running times than those suggested by conventional convergence diagnostic such as Gelman-Rubin [20].

Efficiency of stochastic algorithms often diminishes as the total number of models increases. For this reason, we have found it useful to reduce the number of SNPs included in the EMC search using a screen when S is large. Such a screen will typically

be fairly permissive, leaving only the weakest candidates out of the stochastic search. The screen should be quick to calculate, adjust for the same design variables and consider the same genetic parametrizations as in the full analysis. In our analyses, we calculated marginal (i.e. SNP-at-a-time) Bayes Factors for each of the log-additive, dominant and recessive models of association against the model of no association. We ordered SNPs according to the maximum of the three marginal Bayes factors and retained those with a maximum marginal BF greater than or equal to one. More details are available in Appendix C

2.3 Simulation Comparison

We used the 124 simulated case – control data sets, details of the simulation can be found in D, to estimate true and false positive rates for MISA and seven other alternative procedures:

Bonferroni We fit a logistic regression model for each SNP under the log-additive parametrization and calculate the p-value for testing association using a Chi-Squared test. We use a Bonferroni corrected level $\alpha = 0.05$ test to declare a SNP associated.

Adjusted Bonferroni We fit a logistic regression model for each SNP under the log-additive parametrization and calculate the p-value for testing association using a Chi-Squared test. We use a Bonferroni corrected level α test to declare a SNP associated where α is chosen so that the proportion of false positives detected is the same as in MISA with the default $\mathsf{BB}(1, S)$ prior.

Benjamini-Hochberg We fit the same SNP-at a time logistic regression as above, but declare a SNP to be associated if it has a Benjamini-Hochberg false discovery rate of less than 0.05.

Marginal BF This also utilizes the single SNP at a time logistic regression, but calculates a BF for association under each of the three genetic models. If the maximum BF over the three genetic models is greater than 3.2, we declare the SNP associated. See Appendix C for more detail.

Stepwise LR (AIC) We use a stepwise multiple logistic regression procedure to select SNPs based on AIC. Each SNP is coded using 2 degrees of freedom to select among the three genetic models. SNPs in the final model are called associated.

Stepwise LR (BIC) Same as above but using BIC to select models.

Lasso We use the `Lasso2` package in `R` [40] that is based on the algorithm developed by [44] to select SNPs based on the least absolute shrinkage and selection operator. Each SNP is coded using 2 degrees of freedom to represent the three genetic models and all SNPs in the final model with coefficients greater than zero are called associated.

MISA We reduced the number of SNPs using the marginal Bayes factor method above to eliminate SNPs with a marginal BF ≥ 1. We ran MISA using the default $\mathsf{BB}(1, S)$ and custom $\mathsf{BB}(1/8, S)$ prior distributions on the models using two runs of 400,000 iterations based on convergence of the marginal inclusion probabilities. SNPs are called associated if their MISA SNP BF is greater than 3.2. All SNPs that did not pass the marginal screen step in MISA were declared not associated.

The first four are single SNP methods, while the last three are multi-SNP methods that take into account the genetic parametrization for each SNP.

Figure 2.1 shows the proportion of SNPs detected by each of the methods as a function of the assumed true odds ratio. Thus, at an odds ratio of 1.00 we plot

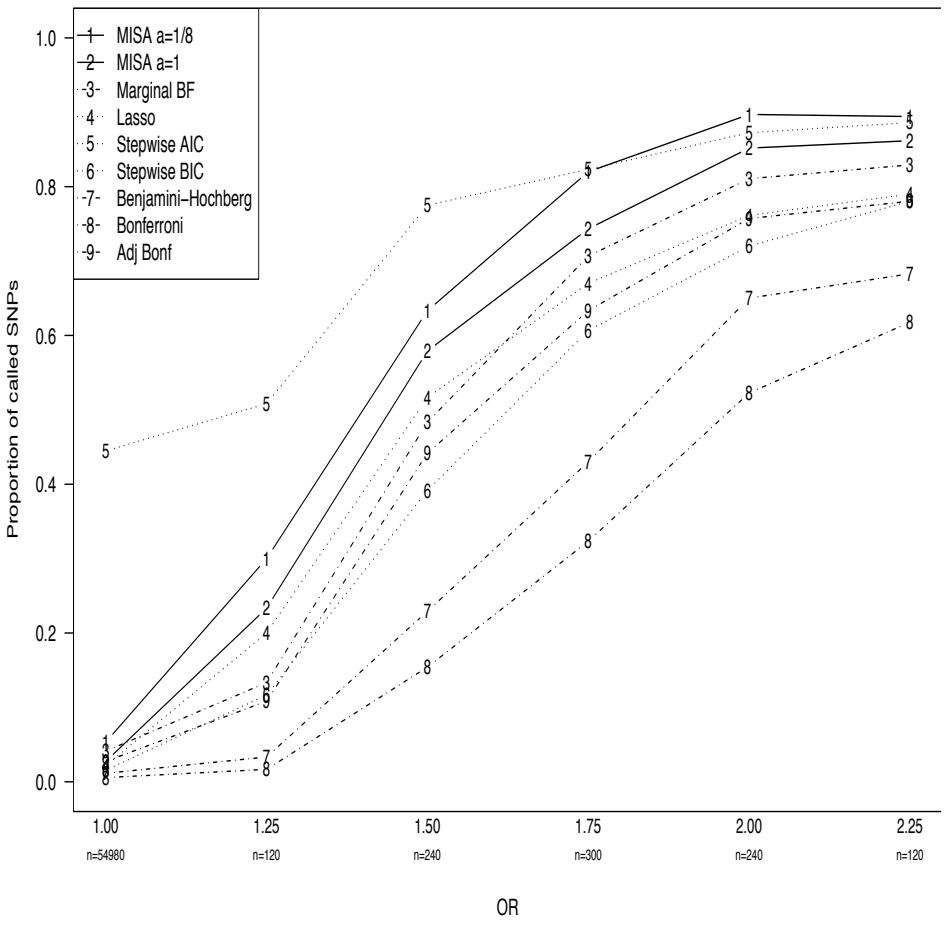

FIGURE 2.1: True and False positive rates of MISA versus alternative methods.

the proportion of SNPs that were falsely declared associated by each of the methods. While both Bonferroni and Benjamini-Hochberg have the smallest false positive rates, they have much lower power to detect true associations than any of the other methods; the marginal BF has the highest power out of the three marginal methods, and is comparable to lasso, a multi-SNP method. Stepwise model selection using BIC has the lowest power of the multiple SNP model selection procedures. Stepwise

logistic regression using AIC to select a model, on the other hand, has high power to detect associations, but an unacceptably high false positive rate (44%). With the exception of stepwise/AIC, the MISA methods have higher power than the alternatives at all odds ratios (ORs) in the simulation, with the gain in power most noticeable for the smaller ORs, those encompassing the range, 1.25 – 1.75 typically seen in practice [19]. This increase in power comes at the cost of only a slight increase in the false positive rate. Overall, MISA using the default $\mathsf{BB}(1, S)$ prior distribution is able to detect 9% as many associations at the SNP level and 13% as many at the gene level than the marginal BF method used alone. In addition, MISA is able to detect 19% as many true associations at the SNP level and 27% as many at the gene level as the calibrated Bonferroni method (the two methods have the same Type I error rate).

2.3.1 Sensitivity to Hyperparameters

We examined a range of parameters (a and b) for the Beta-Binomial prior distribution on model size (Table 2.3) to assess sensitivity of true positive and false positive rates. In practice, this may be done by reweighting the MCMC output using the new prior distribution, without resorting to additional MCMC runs, as long as high posterior probability models receive adequate support under both prior distributions.

Over the range of values for (a, b), MISA has a higher gene and SNP true positive rate than any of the other simpler procedures, with the exception of Stepwise AIC. In general, decreasing a leads to higher true positive rates, but at the expense of higher false positive rates. The SNP false positive rate is modest, ranging from 0.025 to 0.099, providing effective control of the experiment wide error rate. While these rates are higher than the false positive rates under Bonferroni or Benjamini-Hochberg, eliminating a SNP from consideration that truly is associated has a higher scientific cost than continuing to collect data to confirm that a SNP is really a null finding. Because the NCOCS will follow-up apparent associations, a higher true

positive rate with a modest increase in false positives was preferable.

The hyper–parameters $a = 1/8$ and $b = S$, highlighted in bold in Table 2.3 were selected for comparison with the default choice ($a = 1$, $b = S$) in the analysis of the NCOCS data presented in the next section. MISA using the $\mathsf{BB}(1/8, S)$ is able to detect 19% as many true associations at the SNP level and 26% as many at the gene level as the marginal BF method used alone. In addition, MISA with the $\mathsf{BB}(1/8, S)$ prior is able to detect 14% as many true associations at the SNP level and 24% as many at the gene level as a calibrated Bonferroni method, (the two methods have the same Type I error rate).

2.4 Ovarian Cancer Association Analysis

In this section, we describe a MISA candidate pathway analysis of data from the ongoing NCOCS ovarian cancer case–control association study. The NCOCS is a population based study that covers a 48 county region of North Carolina [50]. Cases are between 20 and 74 years of age and were diagnosed with primary invasive or borderline epithelial ovarian cancer after January 1, 1999. Controls are frequency matched to the cases by age and race and have no previous diagnosis of ovarian cancer. In the analysis we present, we focus on self–reported Caucasians and a specific histological subtype of the cancer, leaving us a total of 397 cases and 787 controls. Because the ovarian cancer results have not yet been published, we have anonomyzed the pathway, the genes chosen to represent it and the IDs of the SNPs tagging variation in those genes. The pathway is comprised of 53 genes tagged by 508 tag SNPs.

All models fit in the screen and by MISA included the patient's age as a design variable. We used the modal approximation to fill in missing SNP data. We screened 508 SNPs using marginal Bayes factors, retaining $S = 70$ SNPs that exceeded the threshold of 1 in favor of an association. Using the default hyperparameters $a = 1$ and

Method		True Positive		False Positive		PO of Assoc.	
		Gene (se)	SNP (se)	Gene (se)	SNP (se)	Global	SNP
n		1020	1020	5546	54980		
MISA							
a	b						
1	$\frac{1}{2}S$	.77 (.006)	.669 (.007)	.128 (.001)	.025 (.0001)	2.00	.04
1/2	·	.809 (.005)	.704 (.007)	.166 (.001)	.031 (.0001)	.74	.020
1/4	·	.846 (.004)	.729 (.006)	.189 (.001)	.041 (.0002)	.32	.009
1/8	·	.874 (.003)	.739 (.006)	.259 (.001)	.048 (.0002)	.15	.005
1/16	·	.896 (.003)	.746 (.006)	.341 (.001)	.065 (.0003)	.07	.002
1/32	·	.904 (.003)	.746 (.006)	.437 (.001)	.090 (.0003)	.04	.001
1	S	.784 (.005)	.685 (.007)	.150 (.001)	.027 (.0001)	1.00	.020
1/2	·	.821 (.005)	.716 (.006)	.185 (.001)	.035 (.0001)	.42	.009
1/4	·	.855 (.004)	.736 (.006)	.207 (.001)	.044 (.0002)	.19	.005
1/8	·	**.877 (.003)**	**.743 (.006)**	**.280 (.001)**	**.053 (.0002)**	**.09**	**.002**
1/16	·	.899 (.003)	.746 (.006)	.368 (.001)	.073 (.0003)	.04	.001
1/32	·	.904 (.003)	.746 (.006)	.465 (.001)	.098 (.0004)	.02	.001
1	$\frac{3}{2}S$	.791 (.005)	.696 (.007)	.169 (.001)	.029 (.0001)	.67	.01
1/2	·	.825 (.005)	.722 (.006)	.190 (.001)	.037 (.0002)	.29	.006
1/4	·	.855 (.004)	.735 (.006)	.222 (.001)	.048 (.0002)	.14	.003
1/8	·	.878 (.003)	.744 (.006)	.291 (.001)	.057 (.0002)	.07	.002
1/16	·	.898 (.003)	.746 (.006)	.377 (.001)	.075 (.0003)	.03	.001
1/32	·	.902 (.003)	.746 (.006)	.474 (.001)	.099 (.0004)	.02	.0004
Marg. BF		.695 (.007)	.627 (.007)	.171 (.001)	.041 (.0002)	–	1.00
lasso		.708 (.007)	.607 (.008)	.158 (.001)	.022 (.0001)	–	–
Step. AIC		.993 (.000)	.794 (.005)	.969 (.0001)	.445 (.001)	–	–
Step. BIC		.680 (.007)	.547 (.008)	.122 (.001)	.015 (.0001)	–	–
BH		.439 (.008)	.419 (.008)	.013 (.0001)	.011 (.0001)	–	–
Bonf.		.337 (.007)	.330 (.008)	.003 (.00001)	.006 (.00002)	–	–
Adj. Bonf. 1		.618 (.007)	.574 (.008)	.069 (.0003)	.027 (.0001)	–	–
Adj. Bonf. 2		.708 (.007)	.644 (.007)	.184 (.001)	.053 (.0002)	–	–

Table 2.3: Estimated overall false and true positive rates with standard errors and prior odds (PO) of association at the gene and SNP levels. The values in bold characterize the method selected for use in the analysis of the NCOCS ovarian cancer example.

$b = S$, we ran two independent runs of the algorithm from independent starting points for a total of 1.2 million iterations — the point at which the SNP marginal inclusion probabilities from the two independent runs were determined to be in sufficiently close agreement.

On basis of this analysis, we estimate a lower bound on the pathway–wide Bayes factor for association to be $\mathrm{BF}(H_A : H_0) = 7.67$ (which is also the posterior odds for

this prior). This constitutes "positive" evidence in favor of an association between the pathway and ovarian cancer based on Jeffreys' grades of evidence and corresponds to a posterior probability that the pathway is associated of roughly 0.89. Figure 2.2 summarizes the associations of the ten SNPs that had a SNP BF greater than 3.2, while Figure 2.3 illustrates the nine genes that contained these SNPs and two others that received comparable support. SNPs and genes in the pathway are denoted by a two level name (e.g. S1 and G1) where the number represents the rank of the SNP or gene by its respective Bayes factor. These plots provide a graphical illustration of the top 100 models $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} \in \boldsymbol{\mathcal{M}}$ selected on basis of their posterior model probabilities. 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. SNPs (Figure 2.2) or genes (Figure 2.3) are represented on the y–axis. The presence of a SNP or gene in a model is indicated by a colored block at the intersection of the model's column and the SNP's or gene's row. In Figure 2.2, the color of the block indicates the parametrization of the SNP: purple for log–additive, blue for recessive and red for dominant. The "checkerboard" pattern (as opposed to the presences of more vertical bars) suggests substantial model uncertainty.

The top five models depicted in Figure 2.2 include only a single SNP in addition to age at diagnosis (the design variable is omitted in the figure as it is included in all models). The top model includes SNP S1 in gene G1 under the log–additive genetic parametrization, which is estimated to have an odds ratio (OR) of approximately 1.42 (the posterior mode). The second ranked model includes only SNP S2 in gene G1 under the log–additive genetic parametrization with an estimated OR of 1.37. Note that the study has relatively low power to detect effects of this magnitude (Figure 2.1).

Figure 2.2 also illustrates that many of the top models beyond the first five include multiple SNPs. This suggests that if we were to restrict our attention to single SNP

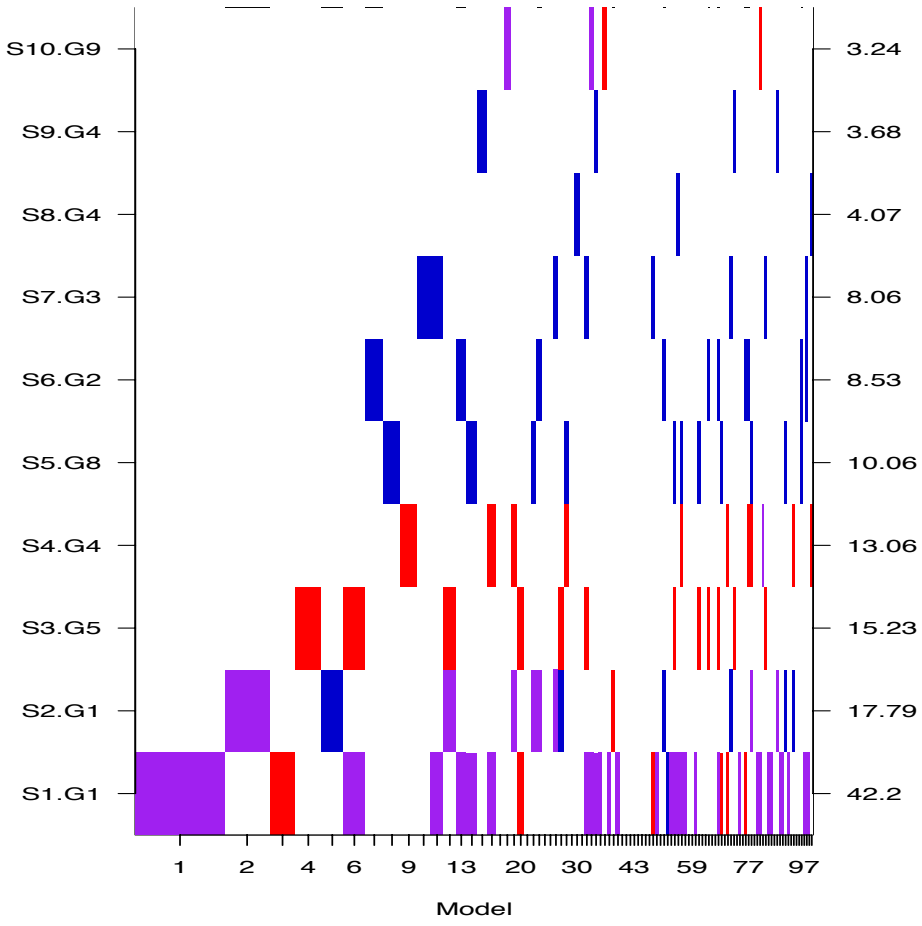

FIGURE 2.2: Image plot of the SNP inclusion indicators for the SNPs with marginal Bayes factors greater than 3.2 and the top 100 Models. The color of the inclusion block corresponds to the genetic parametrization of the SNP in that model. Purple corresponds to a log–additive parametrization, red to a dominant parametrization and blue to a recessive parametrization. SNPs are ordered on basis of their marginal SNP Bayes Factors which are plotted on the right axis across from the SNP of interest. Width of the column associated with a model is proportional to its estimated model probability.

models we would potentially lose substantial information regarding their joint effects.

For example, model six is comprised of both SNP S1 from gene G1 and SNP S3 from

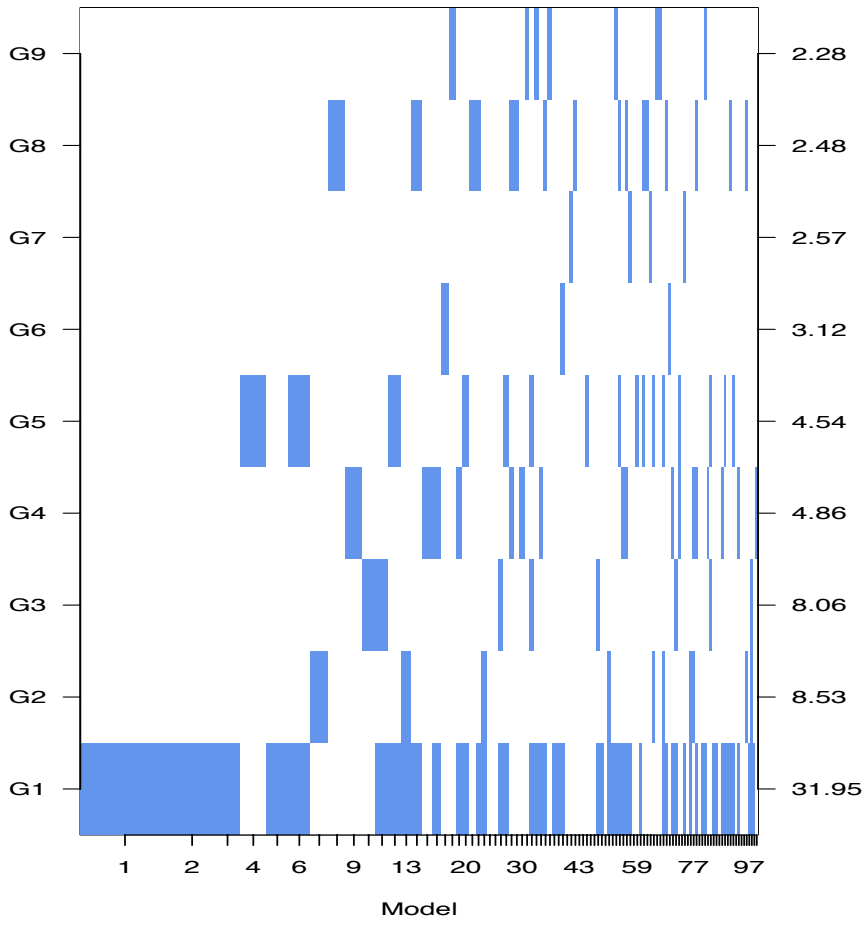

FIGURE 2.3: Image plot of the gene inclusion indicators for the top 100 Models. Genes are ordered based on their marginal gene Bayes Factors which are plotted on the right axis. Columns correspond to models and have width proportional to the estimated model probability, models are plotted in descending order of posterior support. The color is chosen to be neutral since the genetic parametrizations are not defined at the gene level.

gene G5, while model 12 is comprised of both SNP S3 from gene G5 and SNP S2 from gene G1. In both cases, SNP S3 is included in models with a SNP from gene G1. This may indicate that not only are SNPs S1, S2, and S3 important as single

effects in the top four models, but that their combined effects may be of interest. Note that, in cases where the disease variant is unmeasured but 'tagged,' several tagged SNPs may be required to explain variation at that locus.

The SNP Bayes factors of S1 (BF = 42.2) and S2 (BF = 17.8) provide "strong evidence" of changes in prior beliefs, however, the marginal posterior probabilities of association with ovarian cancer are 0.38 and 0.20, respectively. Figure 2.2 illustrates that when one of SNP S1 or S2 is included in a model, the other is often not (at least in the top 50 models). This trade off often arises when SNPs are correlated (i.e. in high linkage disequilibrium). In this case, R^2 of 0.5 suggests fairly strong LD between SNPs S1 and S2, in which case the joint inclusion probabilities are more meaningful than marginal probabilities. Both SNP 1 and SNP 2 are in gene G1 which has a gene Bayes factor of 31.95 (Figure 2.3) and posterior probability of association of 0.58. These probabilities need to be interpreted in the context of model uncertainty; conditional on the pathway being associated with ovarian cancer, the probability that gene G1 is driving the association is $0.58/0.89 = 0.65$. However, there remains substantial uncertainty regarding which genes and SNPs may explain it as the posterior mass is spread over competing models/hypotheses. The positive support for an association suggests the continuation of data accrual to refine these posterior probabilities.

Gene G1 and other genes in Figure 2.3 highlight a caution regarding the interpretation of Bayes factors as a measure of absolute support with composite hypotheses. The gene Bayes factor for G1 is 31.95, which is smaller than the SNP Bayes factors for S1 (42.2). The posterior probability that gene G1 is associated is based on summing the probabilities of all models that include at least one SNP from that gene (S1, S2, and S51) hence the *posterior probability* for gene inclusion is always greater than or equal to the probability that any one SNP is included (i.e. posterior probabilities observe a monotonicity property with composite hypotheses). Bayes factors (and

p–values) for composite hypotheses do not share this monotonicity property [35]. Bayes factors for comparing composite hypotheses may be expressed as the ratio of the weighted average (with respect to the prior distribution) of marginal likelihoods conditional on the hypotheses, which may decrease the evidence in favor of a composite hypothesis when a subset of the individual hypotheses have low likelihood. As mentioned in Section 2.1.4, while Bayes factors do not provide a coherent measure of *absolute* support because of their non–monotonicity property, [35] show that the log Bayes factor does provide a coherent measure of how much the data *change* the support for the hypothesis (relative to the prior). Hence, they do provide useful summaries of changes in prior beliefs of association in large association studies with many competing models/hypotheses.

	df	Sum Sq	Mean Sq	F value	Pr(>F)
snp	69	1635891.00	23708.57	208.04	$< 2 \times 10^{-16}$
prior	1	169641.66	169641.66	1488.60	0.0000
impute	1	134.41	134.41	1.18	0.28
prior:impute	1	53.16	53.16	0.47	0.50
Residuals	207	23589.77	113.96		

Table 2.4: Analysis of variance for the ranked SNP Bayes factors contrasting the prior hyperparameters (default $a = 1$ versus $a = 1/8$) and method of imputation (full imputation with 100 data sets versus a modal estimate of the missing genotypes) for the 70 SNPs in the NCOCS pathway that passed the marginal screen.

2.4.1 Sensitivity Analysis

In this section, we consider sensitivity of the results in the NCOCS study to the prior distribution on the models and to the method of imputation. The simulation study suggests that priors with smaller values of a may identify more associated SNPs. We estimated that the $\mathsf{BB}(1/8, S)$ prior distribution on model size has a false positive rate comparable to the marginal BF method, but a much higher true positive rate, in the scenarios we considered. Full data imputation, achieved by averaging over

the distribution of missing SNPs, is probabilistically correct, but computationally expensive. Thus, if the use of modal imputation provides an accurate approximation to BF calculated using full imputation, the computational efficiency of MISA can be greatly improved at small cost.

For purposes of this analysis, we used the set of unique models identified by the EMC search with modal imputations and $a = 1$ and calculated 3 additional sets of BFs. First, we obtained marginal likelihoods for each of these models using 100 imputed data sets with missing SNPs filled in based on their estimated distribution. Second, we calculated BFs using the $\mathsf{BB}(1/8, S)$ and $\mathsf{BB}(1, S)$ prior distributions using the marginal likelihoods under the full and modal imputations. We applied ANOVA to these four sets of BFs to compare the effects of prior hyperparameters and imputation methods after adjusting for SNP using the ranked SNP BFs.[1]

Table 2.4 shows that the method of imputation has no significant effect on the ranking of SNP BFs. This suggests that, for purposes of model search and calculation of BFs, we may use the modal imputed genotypes in place of full imputation, with significant computational savings. For purposes of parameter estimation, we suggest the use of full imputation using a subset of the top models and top SNPs as using a plug–in approach for imputation is known to underestimate uncertainty.

We anticipated that the prior distribution would have a significant effect based on the higher true positive and false positive rates estimated from the simulation study and by considering differences in the prior odds. While Table 2.4 suggests that overall the rankings are different between the two prior distributions, the top 20 SNPs have the same rank under each of the four methods, leading to no qualitative differences in our conclusions about the top SNPs. The prior odds for any given SNP's inclusion in a model are 8 times lower under the $\mathsf{BB}(1/8, S)$ prior distribution than under to the $\mathsf{BB}(1, S)$ prior distribution; the resulting SNP BFs are 2.8 times

[1] Ranks were used as residuals on the log scale still exhibited strong departures from normality.

higher under the $\mathsf{BB}(1/8, S)$ prior distribution than those under the $\mathsf{BB}(1, S)$ prior distribution. As a result, eight more SNPs are above the 3.2 threshold used by the NCOCS to determine SNPs worthy of additional study.

2.4.2 External Validation and Comparison

To provide a basis of comparison, we applied the methods described in the simulation study (Section 2.3) to the NCOCS data. We omitted stepwise logistic regression using AIC because of its poor operating characteristics. The marginal FDR methods of Bonferroni and Benjamini–Hochberg failed to identify any significant SNPs. Lasso, which accounts for correlation among SNPS, also failed to identify any SNPS. Stepwise logistic regression using BIC selected a model with three of the top four SNPs identified by MISA — S1.G1, S3.G5 and S4.G4 — but failed to identify S2.G1, which has correlation 0.71 with SNP S1.G1. This highlights a problem with selection methods that ignore model uncertainty.

The NCOCS proposed two SNPs — S10 and S14 in G9 — for external validation by the Ovarian Cancer Association Consortium (OCAC), a large international multi–center consortium of ovarian cancer case–control studies. The decision to focus on these variants was made on basis of results from an earlier version of the NCOCS data set and on basis of the strong prior interest NCOCS researchers had in the gene (and not on basis of the analysis described above). Under the default $\mathsf{BB}(1, S)$ prior distribution, only SNP S10 in G9 exceeds the 3.2 threshold and the G9 BF is only 2.28. In contrast, under the $\mathsf{BB}(1/8, S)$ prior distribution, both SNPs S10 and S14 (LD 0.62) in G9 have SNP BFs greater than 3.2 (8.70 and 5.99, respectively) and the gene BF is 6.18. An additional three SNPs in the same gene were proposed by another member of the consortium on the basis of uncorrected p–values. Of the five SNPs proposed for validation, only SNPs S10 and S14 were confirmed to be associated with serous invasive ovarian cancer by OCAC [51].

2.5 Discussion

In this chapter, we describe MISA, a natural framework for multi–level inference with an implicit multiple comparisons correction for hypothesis based association studies. MISA allows one to quantify evidence of association at three levels: global (e.g. pathway–wide), gene, and SNP, while also allowing for uncertainty in the genetic parametrization of the markers. We have evaluated MISA against established, simple to implement and more commonly used methods and demonstrated that our methodology does have higher power than these methods in detecting associations in modestly powered candidate pathway case–control studies. The improvement in power is most noticeable for odds ratios of modest (real world) magnitude and comes at the cost of only a minimal increase in the false positive rate. Like stepwise logistic regression, lasso and logic regression, MISA improves upon marginal, SNP–at–a–time methods by considering multivariate adjusted associations. By using model averaging, MISA improves upon these multivariate methods that select a single model, which may miss important SNPs because of LD structure. These improvements have concrete implications for data analysis: MISA identified SNPs in the NCOCS data that were subsequently externally validated; none of the less complex methods considered here highlighted these SNPs to be of interest. Currently, other top ranked SNPs in genes identified by MISA are undergoing external validation. Finally, we note that while MISA was developed for binary outcomes in case-control studies, MISA is readily adaptable to accommodate other forms of outcome variables (e.g. quantitative traits or survival) that are naturally modeled within a GLM framework.

3

Model Prior Choice and Multiplicity Correction In Bayesian Model and Variable Selection

To date, Bayesian model averaging methodologies have been applied successfully to many problems of model uncertainty and variable selection. For reviews see [28] and [11]. Within this context, one must be able to assign prior probabilities to all models in the model space and also prior distributions to all model specific parameters. Within even modest dimensional problems, there is little hope for the elicitation of model specific probabilities. Thus, much of the recent Bayesian model averaging and selection literature has focused on defining conventional and objective prior distributions for the model specific parameters and providing conditions under which pairwise model Bayes factors (giving the ratio of the weight of evidence between any two single models) lead to the correct model selection as the sample size, n increases. Examples include: mixtures of g-priors [38, 65], empirical g-priors [38, 14, 21], or intrinsic priors [7, 6]. A trend in many modern applications is that the number of variables under study, p, often increases with the sample size. Thus, recent developments have been made to establish conditions under which pairwise model

Bayes factors lead to consistent model selection when the number of variables under study, p, increases withs the sample size n [43, 25].

In particular, one popular motivating example in which the number of variables are increasing with the sample size is that of genetic association studies that aim to identify a set of genetic risk factors for a given complex disease from a large set of correlated covariates (for this example single nucleotide polymorphism or SNPs). Within these studies, we aim to answer the questions: 'Is there an overall association between a set of SNPs and the outcome of interest?' and 'Conditional upon an overall association, which SNPs are most likely to be driving this association?' With this application and these questions in mind the aim of this paper is to investigate conditions of prior distributions placed on the model specific parameters and on the model space itself that are necessary to achieve selection consistency and an implicit multiplicity correction of the following global hypotheses in terms of posterior probabilities as the number of predictor variables, p, goes to infinity as some function of n:

$\boldsymbol{H}_0$: Hypothesis that none of the p predictor variables are associated with the outcome of interest.

$\boldsymbol{H}_A$: Hypothesis that at least one of the p predictor variables are associated with the outcome of interest.

We also wish to investigate how inconsistency in global selection can lead to inconsistency in marginal selections via marginal inclusion probabilities.

Our main interest is in assessing the multiplicity effect on the global posterior probabilities. That is, we want to make sure that under the assumption of the null model being true, the posterior probability of $\boldsymbol{H}_A$ does not go to 1 as n, p go to infinity. We take note that as p goes to infinity as some function of n, selection consistency of the pairwise model Bayes factors does not imply selection consistency

of the global posterior probabilities under the assumption that the null model is true. This is because even though the pairwise model Bayes factors may all go to zero, when we allow p to go to infinity, the posterior probability of the null hypothesis is a function of an infinite sum that does not necessarily go to zero. In fact, this sum will depend on the rate at which the pairwise model Bayes factors converge to zero and the priors placed on the model space.

Our secondary interest is to make sure that the priors placed on the model space do not penalize the marginal model Bayes factors too greatly under the assumption of the alternative hypothesis, $\boldsymbol{H}_A$, so that the posterior odds of the true model to the null model no longer go to infinity. That is, we want to make sure that selection consistency of the pairwise model Bayes factors implies selection consistency of the global posterior probabilities under the assumption that an alternative model is true. This will again depend on the rate at which the pairwise model Bayes factors diverge to infinity and the priors placed on the model space.

In Section 2 we explicitly specify the type of models that make up our model space and the prior distributions for the model specific parameters and of the model space that we will assume throughout the paper as well as some general characteristics of these priors. Section 3 then investigates the convergence or divergence rates of the marginal null-based model Bayes factors under the Zellner-Siow g prior as n goes to infinity and as p goes to infinity as some function of n. In Section 4 and 5 we give theoretical results for the behavior of posterior quantities of interest, namely the global posterior probabilities of the null hypothesis, $\boldsymbol{H}_0$, and the alternative hypothesis, $\boldsymbol{H}_A$, that at least one covariate is associated with the outcome of interest. We do so in two cases: (1) first in limiting case of including additional redundant variables to the analysis and (2) for a full rank design matrix. In particular, we show that the commonly used "non-informative" uniform model space prior has undesirable behavior in that it overwhelming favors the alternative hypothesis, $\boldsymbol{H}_A$, of at least

one association as p increases and the global posterior probability of $\boldsymbol{H}_A$ goes to 1 independent of the data under both the redundant case and when p goes to infinity faster than $\sqrt{n}$ in the full rank case. In Section 6 we show how inconsistencies in the global posterior probabilities can lead to inconsistencies in the marginal posterior inclusion probabilities. Finally, we end with a discussion of the results and give recommendations on the model space priors that lead to the most consistent results in Section 7.

3.1 Model Specification

We consider linear models with a continuous outcome variable. Let Y be a vector of length n comprised of some continuous outcome variables for individual i that is normally distributed:

$$Y \sim N(\boldsymbol{\mu}, \mathbf{I}_n/\phi),$$

where $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_n)^T$, $\mathbf{I}_n$ is an $n \times n$ identity matrix and ϕ is a precision parameter . Also, for each individual we assume that p predictor variables are measured. We then use linear regression models to relate the continuous outcome variable to a subset of predictor variables. We denote the collection of all possible models by $\boldsymbol{\mathcal{M}}$. An individual model, denoted by $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}$, is specified by the $p_{\boldsymbol{\gamma}}$–dimensional vector $\boldsymbol{\gamma}$, where $\boldsymbol{\gamma}_c$ indicates the inclusion of covariate c in model $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}$. Then under each model $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}$, $\boldsymbol{\mu}$ is of the form:

$$\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} : \boldsymbol{\mu} = \mathbf{1}_n \boldsymbol{\alpha} + \mathbf{X}_{\boldsymbol{\gamma}} \boldsymbol{\beta}_{\boldsymbol{\gamma}}$$

where $\mathbf{1}_n$ is a n-dimensional vector of ones, $\mathbf{X}_{\boldsymbol{\gamma}}$ represents the $n \times p_{\boldsymbol{\gamma}}$ design matrix for the subset of covariates in model $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}$ and $\boldsymbol{\theta}_{\boldsymbol{\gamma}}$ is the vector of model specific parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}{}^T)$, with intercept $\boldsymbol{\alpha}$, and coefficients $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$. Without loss of generality we assume that the columns of $\mathbf{X}_{\boldsymbol{\gamma}}$ are centered and $\mathbf{1}_n \mathbf{X}_{\boldsymbol{\gamma}} = 0$.

3.1.1 Posterior Quantities of Interest

Within the linear model framework, we are first interested in computing the marginal likelihood of each model $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} \in \boldsymbol{\mathcal{M}}$:

$$\mathrm{p}(Y|\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}) = \int \mathrm{p}(Y|\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}, \boldsymbol{\theta}_{\boldsymbol{\gamma}})\mathrm{p}(\boldsymbol{\theta}_{\boldsymbol{\gamma}}|\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}})d\boldsymbol{\theta}_{\boldsymbol{\gamma}}.$$

We then can define the pairwise model Bayes factors as the ratio of the marginal likelihood of the model $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}$ to the null model $\boldsymbol{\mathcal{M}}_0$:

$$\mathrm{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} : \boldsymbol{\mathcal{M}}_0] = \frac{\mathrm{p}(Y|\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}})}{\mathrm{p}(Y|\boldsymbol{\mathcal{M}}_0)}.$$

In particular, we are interested in determining criteria of the priors necessary to achieve selection consistency of the global hypotheses $\boldsymbol{H}_A$ and $\boldsymbol{H}_0$ under posterior probabilities. Here, we note that the null hypothesis $\boldsymbol{H}_0$ is the same as the null model $\boldsymbol{\mathcal{M}}_0$ and the alternative hypothesis $\boldsymbol{H}_A$ is composed of the set of all non-null models $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} \neq \boldsymbol{\mathcal{M}}_0$. We are then interested in the posterior probabilities of the global hypotheses calculated as:

$$\mathrm{P}(\boldsymbol{H}_0|Y) = \left[1 + \sum_{\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} \in \boldsymbol{H}_A} \mathrm{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} : \boldsymbol{\mathcal{M}}_0]\mathrm{PO}[\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} : \boldsymbol{\mathcal{M}}_0]\right]^{-1}, \tag{3.1}$$

and $\mathrm{P}(\boldsymbol{H}_A|Y) = 1 - \mathrm{P}(\boldsymbol{H}_0|Y)$. These posterior probabilities are a function of: (1) The pairwise model Bayes factors and (2) the prior odds of the model $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}$ to the null model $\boldsymbol{\mathcal{M}}_0$. Under the defined priors on the model specific parameters $\boldsymbol{\theta}_{\boldsymbol{\gamma}}$ given in the following section, we give the convergence and divergence rates of the pairwise model Bayes factors in Section 3. Thus, given these rates of the pairwise model Bayes factors, we are left to determine criteria of the model space priors that lead to selection consistency of the global posterior probabilities in Section 4 and 5.

Finally, we are interested in assessing if inconsistency of the global posterior probabilities will lead to any inconsistency in the posterior marginal inclusion probabilities in Section 6. These inclusion probabilties are defined for any variable X_i as:

$$\mathrm{p}(\gamma_i = 1|Y) = \sum_{\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}:\gamma_i=1} \mathrm{p}(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}|Y); \tag{3.2}$$

$$= \mathrm{P}(\boldsymbol{H}_A|Y) \left[1 + \frac{\sum_{\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} \in \boldsymbol{H}_A} \mathbf{I}_{\gamma_i=0} \mathrm{p}(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}|Y)/\mathrm{p}(\boldsymbol{\mathcal{M}}_0|Y)}{\sum_{\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} \in \boldsymbol{H}_A} \mathbf{I}_{\gamma_i=1} \mathrm{p}(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}|Y)/\mathrm{p}(\boldsymbol{\mathcal{M}}_0|Y)}\right]^{-1}. \tag{3.3}$$

which is a function of the global posterior probability $\mathrm{P}(\boldsymbol{H}_A|Y)$ of at least one association and the posterior odds of any model $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} \in \boldsymbol{\mathcal{M}}$ to the null model $\boldsymbol{\mathcal{M}}_0$.

3.1.2 Priors on Model Specific Parameters

Our default prior specification of the model specific parameters, $\boldsymbol{\theta}_{\boldsymbol{\gamma}}$, is as follows:

$$\mathrm{p}(\boldsymbol{\alpha}, \phi|\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}) = \frac{1}{\phi}$$

$$\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}, \phi, g \sim N\left(0, \frac{g}{\phi}(\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{\boldsymbol{\gamma}})^-\right)$$

Thus, we assume Zellner's [66] conventional g-prior for the coefficients $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ with a covariance matrix that is given by a constant $\frac{g}{\phi}$ times the generalized inverse of the Fisher Information matrix. We choose to use the generalized inverse of the Fisher Information matrix since we will be looking at non-full-rank cases of the design matrix $\mathbf{X}_{\boldsymbol{\gamma}}$ and the marginal likelihood is invariant under the choice of inverse.

As show by [21], g can be chosen based on penalized likelihood functions such as AIC and BIC. However, to make the prior specification more flexible, Zellner proposed placing a hyper-prior distribution on g as implemented by [38] with the [65] Cauchy priors and hyper-g priors. Throughout the remainder of the paper we will assume a Zellner-Siow prior on the coefficients. Thus, we finish the prior specification

on the model specific parameters with:

$$\mathrm{p}(g) = \sqrt{\frac{n}{2\pi}} g^{-\frac{3}{2}} e^{-\frac{n}{2g}}. \tag{3.4}$$

3.1.3 Model Space Priors

Given the above general model specification, we are interested in investigating possible prior distributions for the model space $\boldsymbol{\mathcal{M}}$. With the high dimensionality of our problems of interest, the model space is effectively innumerable and in turn there is little hope for the elicitation of model specific probabilities. Thus, it is common practice to place non-informative prior distributions that assume independence across the covariates of interest and in particular place a uniform distribution across all possible models. Although this prior is non-informative in regards to the individual models, it can actually be quite informative about other characteristics of the model space including the global hypothesis that at least one covariate is associated and on model size. Alternatively, one can assume *a priori* that the prior probability of any variable being included in any given model is identical and thus the number of included variables in any given model, $p_{\boldsymbol{\gamma}}$, has the distribution:

$$p_{\boldsymbol{\gamma}} \sim \mathsf{Bin}(p, \omega).$$

Therefore, uniform prior across models translates to the $\mathsf{Bin}(p, 1/2)$ prior on model size.

To increase the flexibility of this prior one can place a hyper-prior on the probability that any given variable is included [36, 14, 53] of the form:

$$\omega \sim \mathrm{Beta}(a, b).$$

This Beta-Binomial, denoted $\mathsf{BB}(a, b)$, prior on the number of variables included within any given model has the following density for any model $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}$ with $p_{\boldsymbol{\gamma}}$ variables

included:

$$\mathrm{p}(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}) = \frac{\Gamma(a+b)\Gamma(a+p_{\boldsymbol{\gamma}})\Gamma(b+p-p_{\boldsymbol{\gamma}})}{\Gamma(a)\Gamma(b)\Gamma(a+b+p)},$$

and is becoming an increasingly popular choice for the prior specification on the model space. In particular, Ley and Steel [36] examine properties of the $\mathsf{BB}(a,b)$ prior with fixed hyper-parameter $a=1$ to facilitate prior elicitation. They compare the $\mathsf{Bin}(p,\omega)$ prior with ω fixed to the $\mathsf{BB}(1,b)$ prior and show that placing the hyper-prior on ω increases the flexibility of the overall prior by leading to a reasonable increase in prior variance on the model size. Both [14] and [53] compare the fully Bayesian approach of the $\mathsf{BB}(1,1)$ prior with default hyper-parameters $a=b=1$ with an Empirical Bayes approach of estimating ω from the data in the $\mathsf{Bin}(p,\omega)$ prior. In particular, [53] find considerable differences in the two approaches and show that the users of Empirical Bayes must be cautious in that if the null model has the (strictly) largest marginal likelihood then ω will be estimated to be 0 and alternatively if the full model has the (strictly) largest marginal likelihood then ω will be estimated to be 1. Finally, [53] also study the implicit multiplicity-correction effect of the $\mathsf{BB}(1,1)$ prior and show that as p increases there is an increase in the penalty of adding an additional variable to the current model of interest. For example, in an analysis where there are $p+1$ variables under study we will be penalized more when attempting to go from a model with $p_{\boldsymbol{\gamma}}$ variables to a model with $p_{\boldsymbol{\gamma}}+1$ variables than if there were only p variables in the analysis.

We then examine some general prior characteristics for three prior distributions: (1) the uniform distribution which is equivalent to a $\mathsf{Bin}(p,1/2)$ prior on the model size, (2) the $\mathsf{BB}(1,1)$ prior where we place a uniform distribution across model size and the hyper-parameters are $a=b=1$, and (3) a $\mathsf{BB}(1,\lambda p)$ prior on model size that is developed to achieve a constant global prior probability of at least one association within the analysis no matter how many covariates are considered with

hyper-parameters $a = 1$ and $b = \lambda p$ where λ is some positive constant. Table 3.1 lists the general prior characteristics and the limiting behavior of the characteristics with each of the above mentioned priors as p, the number of variables under consideration, goes to infinity.

	Binomial $(p, 1/2)$	Beta-Binomial $(1, 1)$	Beta-Binomial $(1, \lambda p)$
Expected Model Size	$\frac{p}{2}$ (∞)	$\frac{p}{2}$ (∞)	$\frac{p}{\lambda p+1}$ $(\frac{1}{\lambda})$
Global Prior Odds of an Association	$2^p - 1$ (∞)	p (∞)	$\frac{1}{\lambda}$
Marginal Prior Odds of an Association	1	1	$\frac{1}{\lambda p}$ (0)
Prior Odds of Adding a Variable	1	$\frac{p_{\boldsymbol{\gamma}}+1}{p-p_{\boldsymbol{\gamma}}}$ (0)	$\frac{p_{\boldsymbol{\gamma}}+1}{(\lambda+1)p-p_{\boldsymbol{\gamma}}-1}$ (0)

Table 3.1: General prior characteristics and limiting behavior (in parentheses) of the $\mathsf{Bin}(p, 1/2)$, $\mathsf{BB}(1, 1)$ and $\mathsf{BB}(1, \lambda p)$ distribution on model size.

This characterization of the $\mathsf{BB}(1, \lambda p)$ prior indicates that there is a global implicit multiplicity correction when using this prior that aims at keeping the global prior odds of an association constant as we incorporate more variables to our analysis by decreasing the marginal prior odds of any one of the variables. This implicit multiplicity correction is not found in the $\mathsf{Bin}(p, 1/2)$ and $\mathsf{BB}(1, 1)$ priors on model size since the global prior odds of an association goes to infinity as p goes to infinity. Also, we investigate the prior odds of adding an additional variable to any current model $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}$, that is the ratio of the prior of any model with $p_{\boldsymbol{\gamma}} + 1$ variables to the prior of any model with $p_{\boldsymbol{\gamma}}$ variables for any given $p_{\boldsymbol{\gamma}}$. This prior odds of including an additional variable to any given model stays constant with the $\mathsf{Bin}(p, 1/2)$ prior at 1. However, with both beta-binomial priors we see that as we increase the total number of variables under study,p, the prior odds of incorporating another variable into any

model decreases and approaches 0 in the limiting case. This tells us that in both cases as we increase the number of variables under study, the prior odds of including one more variable in any given model will approach zero (no matter what the current size of the model). This implicitly multiplicity correction with the $\mathsf{BB}(1,1)$ prior is exactly that described in [53] and is desirable as an implicit penalty in the number of variables that are included in the study of interest.

3.2 Asymptotic Behavior of the Marginal Bayes Factors: Zellner-Siow Prior

We are first interested in assessing the asymptotic behavior of the null-based pairwise model Bayes factors under the Zellner-Siow prior that are of the form:

$$\mathrm{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} : \boldsymbol{\mathcal{M}}_0] = \int_0^\infty [1+g]^{\frac{n-r_{\boldsymbol{\gamma}}-1}{2}} \left[1+g\left(1-R^2_{\boldsymbol{\gamma}}\right)\right]^{-\frac{n-1}{2}} \mathrm{p}(g)dg, \tag{3.5}$$

where $r_{\boldsymbol{\gamma}}$ is the rank of the design matrix $\mathbf{X}_{\boldsymbol{\gamma}}$, $R^2_{\boldsymbol{\gamma}}$ is the ordinary coefficient of determination of the model $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}$, and $\mathrm{p}(g)$ is the prior distribution on g defined in Equation 3.4. We take note here that in the full rank case $r_{\boldsymbol{\gamma}} = p_{\boldsymbol{\gamma}}$ is the total number of variables in model $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}$. However, this is not the case in when $\mathbf{X}_{\boldsymbol{\gamma}}$ is not full-rank. Liang etal. [38] show that the marginal null-based Bayes factors for all non-null models $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}$, go to either 0 or ∞ under the Zellner-Siow prior for fixed p and as $n \to \infty$. That is: when $\boldsymbol{\mathcal{M}}_0$ is the true model, $lim_{n\to\infty}\mathrm{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} : \boldsymbol{\mathcal{M}}_0] = 0$ and when $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}$ is true, $lim_{n\to\infty}\mathrm{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} : \boldsymbol{\mathcal{M}}_0] = \infty$. We therefore assume a Zellner-Siow prior on the coefficients for this section and are interested in extending the results of [38] by investigating the asymptotic behavior of the marginal null-based Bayes factors as n goes to infinity and as p goes to infinity as some function of n. We also make the assumption throughout the paper that $n > p-1$ and $\lim_{n\to\infty} \frac{p}{n} = 0$. In particular we show that under the null model $\mathrm{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} : \boldsymbol{\mathcal{M}}_0]$ converges to zero

at a rate of:

$$\mathrm{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} : \boldsymbol{\mathcal{M}}_0] = O\left[\left(\frac{r_{\boldsymbol{\gamma}}+1}{n}\right)^{\frac{r_{\boldsymbol{\gamma}}}{2}}\right].$$

We also show that under a finite alternative model, $\boldsymbol{\mathcal{M}}_*$, with rank r_*, $\mathrm{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_* : \boldsymbol{\mathcal{M}}_0]$ diverges at a rate of:

$$\mathrm{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_* : \boldsymbol{\mathcal{M}}_0] = O\left[\left(\frac{1}{n}\right)^{\frac{r_*}{2}}(1+b_*)^{\frac{n-1}{2}}\right],$$

where b_* is some strictly positive constant.

We begin by investigating the distribution of $R^2_{\boldsymbol{\gamma}}$ under the assumption of the null model being true and under the assumption of an alternative model being true. Proposition 3.2.1 first states the distribution of $R^2_{\boldsymbol{\gamma}}$ for any model $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} \in \boldsymbol{\mathcal{M}}$. We then place Proposition 3.2.1 in context of the assumptions of the null hypothesis being true in Corollary 3.2.2 and the alternative hypothesis being true in Corollary 3.2.3.

Proposition 3.2.1. *For any model $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} \in \boldsymbol{\mathcal{M}}$, $(1 - R^2_{\boldsymbol{\gamma}})$ has a non-central Beta distribution [31] with parameters:*

$$(1 - R^2_{\boldsymbol{\gamma}}) \sim Beta(\frac{n - r_{\boldsymbol{\gamma}} - 1}{2}, \frac{r_{\boldsymbol{\gamma}}}{2}, \frac{nb_{\gamma_1}}{2}, \frac{nb_{\gamma_2}}{2})$$

with

$$b_{\gamma_1} = \frac{\phi\boldsymbol{\mu}^T(\mathbf{I} - P_{\mathbf{1}_n} - P_{X_\gamma})\boldsymbol{\mu}}{n};$$

$$b_{\gamma_2} = \frac{\phi\boldsymbol{\mu}^T(P_{X_\gamma})\boldsymbol{\mu}}{n};$$

where $\boldsymbol{\mu}$ is the mean of the true model.

Proof. For any model $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} \in \boldsymbol{\mathcal{M}}$, the ordinary coefficient of determination is:

$$R^2_{\boldsymbol{\gamma}} = \frac{Y^T(P_{\mathbf{X}_{\boldsymbol{\gamma}}})Y}{Y^T(\mathbf{I} - P_{\mathbf{1}_n})Y},$$

where P_{X_γ} is the projection matrix for the column space of $\mathbf{X}_{\boldsymbol{\gamma}}$ and $P_{\mathbf{1}_n}$ is the projection matrix for the column space of $\mathbf{1}_n$. Then, $(1 - R^2_{\boldsymbol{\gamma}})$ can be written as:

$$\begin{aligned}(1 - R^2_{\boldsymbol{\gamma}}) &= \frac{Y^T(\mathbf{I} - P_{\mathbf{1}_n} - P_{\mathbf{X}_{\boldsymbol{\gamma}}})Y}{Y^T(\mathbf{I} - P_{\mathbf{1}_n} - P_{\mathbf{X}_{\boldsymbol{\gamma}}})Y + Y^T(P_{\mathbf{X}_{\boldsymbol{\gamma}}})Y};\\ &= \frac{\chi_1}{\chi_1 + \chi_2},\end{aligned}$$

where χ_1 and χ_2 are independent non-central chi-squared distributed random variables with:

$$\chi_1 \sim \chi^2(n - r_{\boldsymbol{\gamma}} - 1, nb_{\boldsymbol{\gamma}_1})$$

$$\chi_2 \sim \chi^2(r_{\boldsymbol{\gamma}}, nb_{\boldsymbol{\gamma}_2}).$$

Thus, $(1 - R^2_{\boldsymbol{\gamma}})$ has a non-central Beta distribution with parameters

$$(1 - R^2_{\boldsymbol{\gamma}}) \sim \text{Beta}(\frac{n - r_{\boldsymbol{\gamma}} - 1}{2}, \frac{r_{\boldsymbol{\gamma}}}{2}, \frac{nb_{\gamma_1}}{2}, \frac{nb_{\gamma_2}}{2})$$

with

$$b_{\gamma_1} = \frac{\phi\boldsymbol{\mu}^T(I - P_{\mathbf{1}_n} - P_{X_\gamma})\boldsymbol{\mu}}{n};$$

$$b_{\gamma_2} = \frac{\phi\boldsymbol{\mu}^T(P_{X_\gamma})\boldsymbol{\mu}}{n};$$

where $\boldsymbol{\mu}$ is the mean of the true model. □

Corollary 3.2.2. *Assume that the null model $\boldsymbol{\mathcal{M}}_0$ is the true model and* $\lim_{n\to\infty} \frac{p}{n} = 0$. *Then $b_{\gamma_1}, b_{\gamma_2} = 0$ and*

$$(1 - R^2_{\boldsymbol{\gamma}}) \sim \textit{Beta}(\frac{n - r_{\boldsymbol{\gamma}} - 1}{2}, \frac{r_{\boldsymbol{\gamma}}}{2}).$$

Thus, $(1 - R^2_{\boldsymbol{\gamma}})$ *converges to* 1 *and the rate of convergence is:*

$$(1 - R^2_{\boldsymbol{\gamma}}) = O\left[\frac{n - r_{\boldsymbol{\gamma}} - 1}{n - 1}\right].$$

Corollary 3.2.3. *Assume that* $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}$ *is the true model and* $\lim_{n\to\infty} \frac{p}{n} = 0$. *Then* $b_{\gamma_1} = 0$ *and*

$$(1 - R^2_{\boldsymbol{\gamma}}) \sim Beta(\frac{n - r_{\boldsymbol{\gamma}} - 1}{2}, \frac{r_{\boldsymbol{\gamma}}}{2}, 0, \frac{nb_{\gamma_2}}{2}).$$

Also, assume that $b_{\boldsymbol{\gamma}} = \lim_{n,p\to\infty} b_{\boldsymbol{\gamma}_2}$ *and* $0 < b_{\boldsymbol{\gamma}} < \infty$. *Then,*

$$\text{plim}_{n,p\to\infty}(1 - R^2_{\boldsymbol{\gamma}}) = (1 + b_{\boldsymbol{\gamma}})^{-1},$$

which is strictly between 0 *and* 1 *since* $0 < b_{\boldsymbol{\gamma}} < \infty$.

In Theorem 3.2.4 we give results for the asymptotic behavior of the marginal null-based Bayes factors as n goes to infinity and as p goes to infinity as some function of n under the assumption that the null model $\boldsymbol{\mathcal{M}}_0$ is true and $\lim_{n,p\to\infty} \frac{p}{n} = 0$.

Theorem 3.2.4. *As* n *goes to infinity and* p *goes to infinity as some increasing function of* n *such that* $\lim_{n\to\infty} \frac{r_{\boldsymbol{\gamma}}}{n} = 0$,*under the assumption that the null model is true the pairwise model Bayes factor* $BF_{p(g)}[\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} : \boldsymbol{\mathcal{M}}_0]$ *converges to* 0 *and the rate of convergence is*

$$BF_{p(g)}[\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} : \boldsymbol{\mathcal{M}}_0] = O\left[\left(\frac{r_{\boldsymbol{\gamma}} + 1}{n}\right)^{\frac{r_{\boldsymbol{\gamma}}}{2}}\right].$$

Proof. Under the assumption that the null model is true and by the distribution of

$R^2_{\boldsymbol{\gamma}}$ given in Corollary 3.2.2 and under the transformation $t = \frac{n}{2g}$:

$$\begin{aligned}
&\text{plim}_{n,p\to\infty}\, \text{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} : \boldsymbol{\mathcal{M}}_0] \\
&\propto \text{plim}_{n,p\to\infty} \int_0^\infty \left[1 + \frac{n}{2t}\right]^{\frac{n-r_{\boldsymbol{\gamma}}-1}{2}} \left[1 + \frac{n}{2t}(1 - R^2_{\boldsymbol{\gamma}})\right]^{-\frac{n-1}{2}} t^{-\frac{1}{2}} e^{-t} dt; \\
&= \text{plim}_{n,p\to\infty} \left(\frac{2}{n}\right)^{\frac{r_{\boldsymbol{\gamma}}}{2}} \int_0^\infty \left[1 + \frac{2t}{n}\right]^{\frac{n-r_{\boldsymbol{\gamma}}-1}{2}} \left[(1 - R^2_{\boldsymbol{\gamma}}) + \frac{2t}{n}\right]^{-\frac{n-1}{2}} t^{\frac{r_{\boldsymbol{\gamma}}-1}{2}} e^{-t} dt; \\
&= \lim_{n,p\to\infty} \left(1 - \frac{r_{\boldsymbol{\gamma}}}{n-1}\right)^{-\frac{n-1}{2}} \left(\frac{2}{n}\right)^{\frac{r_{\boldsymbol{\gamma}}}{2}} \int_0^\infty \left[1 + \frac{2t}{n}\right]^{\frac{n-r_{\boldsymbol{\gamma}}-1}{2}} \left[1 + \frac{2t}{n(1 - \frac{r_{\boldsymbol{\gamma}}}{n-1})}\right]^{-\frac{n-1}{2}} t^{\frac{r_{\boldsymbol{\gamma}}-1}{2}} e^{-t} dt; \\
&= \lim_{n,p\to\infty} \left(\frac{2e}{n}\right)^{\frac{r_{\boldsymbol{\gamma}}}{2}} \int_0^\infty \exp\left\{t\left(\frac{n - r_{\boldsymbol{\gamma}} - 1}{n}\right)\right\} \exp\left\{-t\left(\frac{n-1}{n(1 - \frac{r_{\boldsymbol{\gamma}}}{n-1})}\right)\right\} t^{\frac{r_{\boldsymbol{\gamma}}-1}{2}} e^{-t} dt; \\
&= \lim_{n,p\to\infty} \left(\frac{2e}{n}\right)^{\frac{r_{\boldsymbol{\gamma}}}{2}} \int_0^\infty t^{\frac{r_{\boldsymbol{\gamma}}-1}{2}} e^{-t} dt; \\
&= \lim_{n,p\to\infty} \left(\frac{2e}{n}\right)^{\frac{r_{\boldsymbol{\gamma}}}{2}} \Gamma\left(\frac{r_{\boldsymbol{\gamma}} + 1}{2}\right).
\end{aligned}$$

Then under Sterling's approximation for the Gamma function:

$$\lim_{n,p\to\infty} \left(\frac{2e}{n}\right)^{\frac{r_{\boldsymbol{\gamma}}}{2}} \Gamma\left(\frac{r_{\boldsymbol{\gamma}} + 1}{2}\right) \propto \lim_{n,p\to\infty} \left[\frac{r_{\boldsymbol{\gamma}} + 1}{n}\right]^{\frac{r_{\boldsymbol{\gamma}}}{2}}.$$

Thus under the assumption that the null model is true the pairwise model Bayes factors, $\text{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} : \boldsymbol{\mathcal{M}}_0]$ converge to zero and and the rate of convergence is:

$$\text{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} : \boldsymbol{\mathcal{M}}_0] = O\left[\left(\frac{r_{\boldsymbol{\gamma}} + 1}{n}\right)^{\frac{r_{\boldsymbol{\gamma}}}{2}}\right].$$

□

In Theorem 3.2.5 we give results for the asymptotic behavior of the marginal null-based Bayes factors as n goes to infinity and as p goes to infinity as some function of n under the assumption that an alternative model $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}$ is true and $\lim_{n,p\to\infty} \frac{p}{n} = 0$.

Theorem 3.2.5. *As n goes to infinity and p goes to infinity as some increasing function of n such that $\lim_{n\to\infty} \frac{p_\gamma}{n} = 0$,under the assumption that some finite model $\boldsymbol{\mathcal{M}}_* \in \boldsymbol{\mathcal{M}}$ with rank $r_* < \infty$ is true the pairwise model Bayes factor $BF_{p(g)}[\boldsymbol{\mathcal{M}}_* : \boldsymbol{\mathcal{M}}_0]$ diverges and the rate of divergence is*

$$BF_{p(g)}[\boldsymbol{\mathcal{M}}_* : \boldsymbol{\mathcal{M}}_0] = O\left[\left(\frac{1}{n}\right)^{\frac{r_*}{2}} (1+b_*)^{\frac{n-1}{2}}\right],$$

where b_* is some strictly positive constant defined in Corollary 3.2.3.

Proof. Under the assumption that some finite alternative model $\boldsymbol{\mathcal{M}}_* \in \boldsymbol{\mathcal{M}}$ is true we begin by computing the Laplace approximation of the marginal Bayes factors. In particular, we wish to approximate an integral of the form:

$$\int \mathrm{p}(Y|\boldsymbol{\mathcal{M}}_*, g)\mathrm{p}(g)dg;$$

where $\mathrm{p}(Y|\boldsymbol{\mathcal{M}}_*, g)$ is the marginal likelihood conditional upon g in the general g-based prior and $\mathrm{p}(g)$ is the assumed distribution on g. A variation on the usual Laplace approximation for the above integral uses the MLE of $p(Y|\boldsymbol{\mathcal{M}}_*, g)$ which is equal to the local empirical Bayes estimate of g, $\hat{g}$, instead of the posterior. The error is still $O(\frac{1}{n})$ [32]. The MLE of g is:

$$\hat{g} = \max(F_* - 1, 0),$$

where F_γ is the usual F statistic for testing $\boldsymbol{\beta}_* = 0$:

$$F_* = \frac{R_*^2/r_*}{(1-R_*^2)/(n-r_*-1)}.$$

The Laplace approximation to the pairwise model Bayes factor is:

$$\mathrm{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_* : \boldsymbol{\mathcal{M}}_0] = \sqrt{2\pi}\ \ [-\mathcal{L}^{''}(\hat{g})^{-\frac{1}{2}}]\ \ \mathrm{p}(\hat{g})\ \ \mathrm{BF}_{EBL}[\boldsymbol{\mathcal{M}}_* : \boldsymbol{\mathcal{M}}_0]\ \ [1+O(1/n)],$$

and has three main parts:

(1)

$$-\mathcal{L}^{''}(\hat{g})^{-\frac{1}{2}} = \left[\left(\frac{n-r_*-1}{2\left(1+\hat{g}\right)^2}\right) - \left(\frac{n-1}{2}\right)\left(\frac{1-R_*^2}{1+\hat{g}(1-R_*^2)}\right)^2\right]^{-\frac{1}{2}};$$

$$= \frac{\sqrt{2}(1+\hat{g})}{\sqrt{n-r_*-1}}\left[1 - \frac{n-1}{n-r_*-1}\left(\frac{(1+\hat{g})R_*^2}{1+\hat{g}(1-R_*^2)}\right)^2\right]^{-\frac{1}{2}};$$

$$= \frac{\sqrt{2}(1+\hat{g})}{\sqrt{n-r_*-1}}\left[1 - \frac{n-1}{n-r_*-1}\left(\frac{n-r_*-1}{n-1}\right)^2\right]^{-\frac{1}{2}};$$

$$= \sqrt{2}(1+\hat{g})\left[\frac{(n-r_*-1)r_*}{n-1}\right]^{-\frac{1}{2}}.$$

(2)

$$\lim_{n,p\to\infty} \mathrm{p}(\hat{g}) = \lim_{n,p\to\infty}\sqrt{\frac{n}{2\pi}}(\hat{g})^{-\frac{3}{2}}e^{-\frac{n}{2\hat{g}}};$$

$$= \lim_{n,p\to\infty}\sqrt{\frac{n}{2\pi}}\left[\frac{(1-R_*^2)r_*}{R_*^2(n-1)-r_*}\right]^{\frac{3}{2}} e^{\left[-\frac{(1-R_*^2)r_*}{R_*^2((n-1)/n)-(r_*/n)}\right]};$$

$$= \lim_{n,p\to\infty}\sqrt{\frac{n}{2\pi}}(1+\hat{g})^{-\frac{3}{2}}\left[\frac{n-r_*-1}{n}\right]^{\frac{3}{2}} e^{\left[-\frac{(1-R_*^2)r_*}{2R_*^2}\right]};$$

$$= \lim_{n,p\to\infty}\sqrt{\frac{n}{2\pi}}(1+\hat{g})^{-\frac{3}{2}}e^{\left[-\frac{(1-R_*^2)r_*}{2R_*^2}\right]};$$

(3)

$$\mathrm{BF}_{EBL}[\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} : \boldsymbol{\mathcal{M}}_0] = (1+\hat{g})^{-\frac{r_*}{2}}\left[\frac{1+\hat{g}}{1+\hat{g}(1-R_*^2)}\right]^{\frac{n-1}{2}};$$

$$= (1+\hat{g})^{-\frac{r_*}{2}}\left[\frac{(n-1)(1-R_*^2)}{n-r_*-1}\right]^{-\frac{n-1}{2}}.$$

Then under the Laplace approximation and the distribution of R_*^2 in Corollary

3.2.3 we have that:

$$
\begin{aligned}
&\text{plim}_{n,p\to\infty}\,\text{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_* : \boldsymbol{\mathcal{M}}_0] \\
&= \text{plim}_{n,p\to\infty}\sqrt{\frac{2n}{r_*}}\left[\frac{(1-R_*^2)r_*}{R_*^2(n-r_*-1)}\right]^{\frac{r_*+1}{2}}\left[\frac{n-r_*-1}{(n-1)(1-R_*^2)}\right]^{\frac{n-1}{2}} e^{-\frac{r_*}{2}\left(\frac{1-R_*^2}{R_*^2}\right)}; \\
&= \text{plim}_{n,p\to\infty}\sqrt{\frac{2}{r_*}}\left[\frac{(1-R_*^2)r_*}{R_*^2}\right]^{\frac{r_*+1}{2}}\left[\frac{1}{n}\right]^{\frac{r_*}{2}}\left[\frac{1}{1-R_*^2}\right]^{\frac{n-1}{2}} e^{-\frac{r_*}{2}\left(\frac{1}{R_*^2}\right)}; \\
&= \lim_{n,p\to\infty}\sqrt{\frac{2}{r_*}}\left[\frac{r_*}{b_*}\right]^{\frac{r_*+1}{2}}\left[\frac{1}{n}\right]^{\frac{r_*}{2}}[1+b_*]^{\frac{n-1}{2}} e^{-\frac{r_*}{2}\left(\frac{1+b_*}{b_*}\right)}.
\end{aligned}
$$

Then, since r_* is assumed to be finite:

$$
\text{plim}_{n,p\to\infty}\,\text{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_* : \boldsymbol{\mathcal{M}}_0] = \lim_{n,p\to\infty} c_*\left[\frac{1}{n}\right]^{\frac{r_*}{2}}[1+b_*]^{\frac{n-1}{2}},
$$

where c_* is some positive constant less than infinity since b_* is strictly between 0 and ∞ by Proposition 3.2. Thus, under the assumption that a finite alternative model the pairwise model Bayes factor, $\text{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_* : \boldsymbol{\mathcal{M}}_0]$ diverges and the rate of divergence is

$$
\text{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_* : \boldsymbol{\mathcal{M}}_0] = O\left[\left(\frac{1}{n}\right)^{\frac{r_*}{2}}(1+b_*)^{\frac{n-1}{2}}\right].
$$

□

3.3 Asymptotic Behavior of Global Posterior Probabilities: Redundant Case

In this section we explore properties of the global posterior probabilities in the extreme case of including multiple redundant copies of a variable to an analysis. In this case, the rank of the design matrix $\mathbf{X}_{\boldsymbol{\gamma}}$ for each model $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}$ is 1. We will see that this can be thought of as the worst case scenario to achieve selection consistency of the global hypotheses under the assumption that the null hypothesis is true. This is due

to the fact that the marginal model Bayes factors $\text{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} : \boldsymbol{\mathcal{M}}_0]$ will converge to zero at the same rate since the rank of $\mathbf{X}_{\boldsymbol{\gamma}}$ is 1 for every model $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}$ and this rate will be slower than the rate at which they would converge to zero if the rank of the design matrices were greater than 1. In particular, under this extreme case we will see that the only way to achieve selection consistency under the null hypothesis is for the global prior odds $\text{PO}[\boldsymbol{H}_A : \boldsymbol{H}_0]$ to converge to zero slower than $\sqrt{n}$. This is certainly not the case under the $\mathsf{Bin}(p, 1/2)$ but does hold true under the $\mathsf{BB}(1, 1)$ prior when p goes to infinity slower than $\sqrt{n}$ and under the $\mathsf{BB}(1, \lambda p)$ prior for all growth rates of p.

3.3.1 Global posterior probabilities for fixed n and p

We are first interested in calculating the global posterior probability of the null hypothesis, $\text{P}(\boldsymbol{H}_0|Y)$, for any given fixed p. As seen in Equation 3.5, the pairwise model Bayes factors $\text{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} : \boldsymbol{\mathcal{M}}_0]$ for every model $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} \in \boldsymbol{\mathcal{M}}$ depend only on the rank of the design matrix of $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}$ and the projection matrix $P_{\mathbf{X}_{\boldsymbol{\gamma}}}$ (through $R^2_{\boldsymbol{\gamma}}$). Since we are just including multiple redundant copies of a variable to the analysis $\text{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} : \boldsymbol{\mathcal{M}}_0]$ is the same for every $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} \in \boldsymbol{\mathcal{M}}$. Thus, the global posterior probability of the null hypothesis is calculated as:

$$\text{P}(\boldsymbol{H}_0|Y) = \left[1 + \sum_{\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} \in \boldsymbol{H}_A} \text{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} : \boldsymbol{\mathcal{M}}_0]\text{PO}[\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} : \boldsymbol{\mathcal{M}}_0]\right]^{-1}; \tag{3.6}$$

$$= \left[1 + \text{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_1 : \boldsymbol{\mathcal{M}}_0] \sum_{\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} \in \boldsymbol{H}_A} \text{PO}[\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} : \boldsymbol{\mathcal{M}}_0]\right]^{-1}; \tag{3.7}$$

$$= \left[1 + \text{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_1 : \boldsymbol{\mathcal{M}}_0]\text{PO}[\boldsymbol{H}_A : \boldsymbol{H}_0]\right]^{-1}. \tag{3.8}$$

This is a function of the marginal null-based Bayes factor $\text{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_1 : \boldsymbol{\mathcal{M}}_0]$ where $\boldsymbol{\mathcal{M}}_1$ is the single variable model that is made up of only one of the redundant variables, X_1, and the global prior odds, $\text{PO}[\boldsymbol{H}_A : \boldsymbol{H}_0]$.

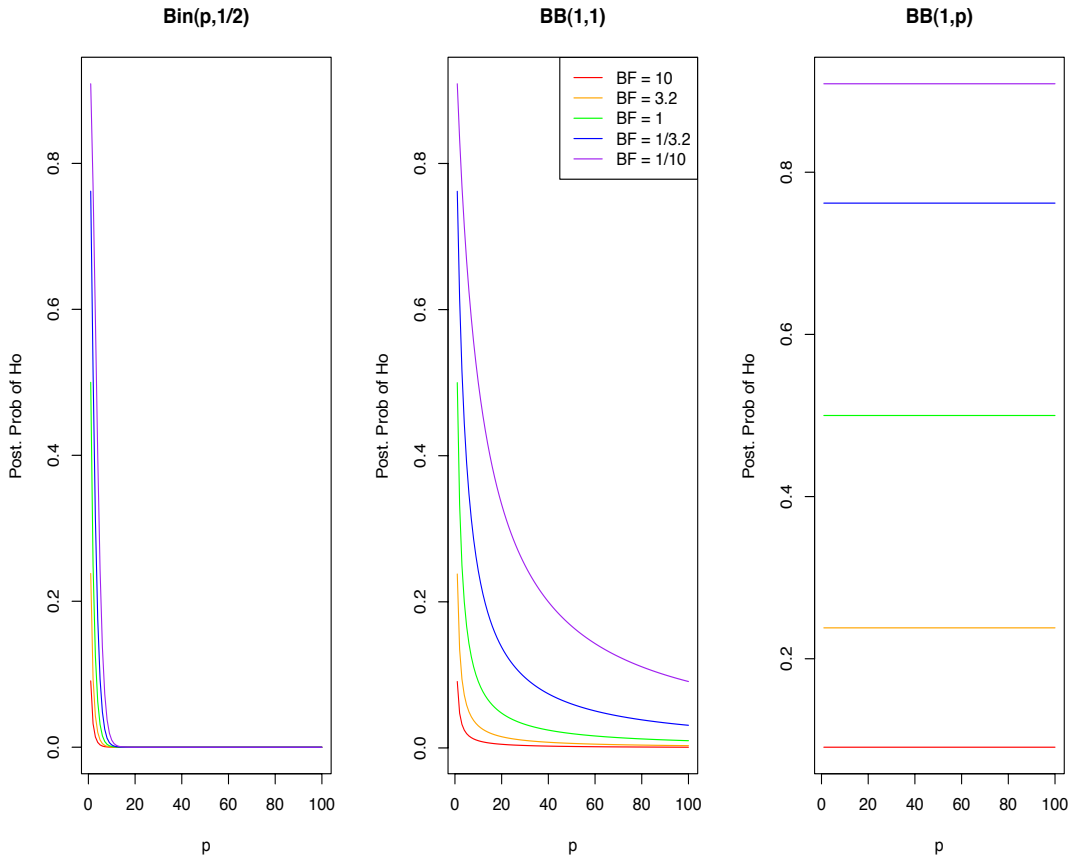

FIGURE 3.1: Theoretical Posterior Probability of the Null Hypothesis as we increase the number of redundant variables in the analysis for the $\mathsf{Bin}(p, 1/2)$ prior, $\mathsf{BB}(1, 1)$ and $\mathsf{BB}(1, p)$ priors.

Figure 3.1 plots the posterior probability of the null hypothesis as a function of p for the $\mathsf{Bin}(p, 1/2)$ prior, the $\mathsf{BB}(1, 1)$ prior and the $\mathsf{BB}(1, p)$ prior (assuming $\lambda = 1$). The posterior probability of the null hypothesis is calculated based on five different values for the marginal model Bayes factor $\mathrm{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_1 : \boldsymbol{\mathcal{M}}_0] \in \{10, 3.2, 1, \frac{1}{3.2}, \frac{1}{10}\}$. Under the $\mathsf{Bin}(p, 1/2)$ and $\mathsf{BB}(1, 1)$ priors, Figure 3.1 shows that independent of the data the posterior probability of the null hypothesis will always go to zero as p increases and in turn the posterior probability of the alternative hypothesis will go to 1. However, with the $\mathsf{BB}(1, p)$ prior, the posterior probability of the null

hypothesis is constant and we do not have a problem of dilution as we increase the number of redundant variables. This is a desirable property, in that the global posterior probability of the null hypothesis should not be effected by the inclusion of redundant information.

3.3.2 Global posterior probabilities as $n, p \to \infty$

The previous results have been for fixed n (or equivalently fixed $\text{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_1 : \boldsymbol{\mathcal{M}}_0]$) and fixed p. We are also interested in assessing how the global posterior probabilities behave as $n, p \to \infty$ given conditions on the global prior. Theorem 3.3.1 states conditions on the maximum rate of increase of the global prior odds, $\text{PO}[\boldsymbol{H}_A : \boldsymbol{H}_0]$, under which the selection of the null versus alternative hypothesis is consistent using posterior probabilities.

Theorem 3.3.1. *Let p go to infinity as some increasing function of n such that $\lim_{n,p\to\infty} \frac{p}{n} = 0$ and assume that the global prior odds, $PO[\boldsymbol{H}_A : \boldsymbol{H}_0]$, under the prior distribution on the model space is a non-decreasing function of p. Consider two cases for $PO[\boldsymbol{H}_A : \boldsymbol{H}_0]$:*

(1) The global prior odds is an increasing function of p that goes to infinity faster than $\sqrt{n}$:

$$\lim_{n\to\infty} \frac{PO[\boldsymbol{H}_A : \boldsymbol{H}_0]}{\sqrt{n}} = \infty.$$

(2) The global prior odds is a constant function of p or an increasing function of p that goes to infinity slower than $\sqrt{n}$:

$$\lim_{n\to\infty} \frac{PO[\boldsymbol{H}_A : \boldsymbol{H}_0]}{\sqrt{n}} = 0.$$

Under both cases the posterior probability of the alternative hypothesis, $P(\boldsymbol{H}_A|Y)$, converges to 1 when the alternative hypothesis is true and is therefore consistent under selection using global posterior probabilities.

Under Case (1) and the assumption that the null hypothesis is true, the posterior probability of the null hypothesis converges to 0 *and the selection under the global posterior probabilities is inconsistent:*

$$\lim_{p,n\to\infty} P(\boldsymbol{H}_0|Y) = 0.$$

Under Case (2) and the assumption that the null hypothesis is true, the posterior probability of the null hypothesis converges to 1 *and we achieve selection consistency under global posterior probabilities:*

$$\lim_{p,n\to\infty} P(\boldsymbol{H}_0|Y) = 1.$$

Proof. From Equation 3.8 the general form for the posterior probability of the null model under the inclusion of p redundant variables is:

$$\mathrm{P}(\boldsymbol{H}_0|Y) = \left[1 + \mathrm{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_1 : \boldsymbol{\mathcal{M}}_0] * \mathrm{PO}[\boldsymbol{H}_A : \boldsymbol{H}_0]\right]^{-1}.$$

We first assume that the alternative hypothesis is true and as shown in 3.2.5 the pairwise model Bayes factor $\mathrm{plim}_{n,p\to\infty} \mathrm{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_1 : \boldsymbol{\mathcal{M}}_0] = \infty$. In this case, since the global prior odds, $\mathrm{PO}[\boldsymbol{H}_A : \boldsymbol{H}_0]$, is a non-decreasing function of p that does not go to zero as $n, p \to \infty$. Thus, it is trivial to show that

$$\lim_{p,n\to\infty} \mathrm{P}(\boldsymbol{H}_A|Y) = 1.$$

We next assume that the null hypothesis is true and as shown in 3.2.4 the pairwise model Bayes factor $\mathrm{plim}_{n,p\to\infty} \mathrm{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_1 : \boldsymbol{\mathcal{M}}_0] = 0$ and converges at a rate of $O\left(\sqrt{\frac{1}{n}}\right)$. Therefore, we have that:

$$\mathrm{plim}_{n,p\to\infty} \mathrm{P}(\boldsymbol{H}_0|Y) = \lim_{n,p\to\infty} \left[1 + \frac{\mathrm{PO}[\boldsymbol{H}_A : \boldsymbol{H}_0]}{\sqrt{n}}\right]^{-1}.$$

Thus, it is trivial to show that under Case (1) $\lim_{p,n\to\infty} \mathrm{P}(\boldsymbol{H}_0|Y) = 0$, and and under Case(2) $\lim_{p,n\to\infty} \mathrm{P}(\boldsymbol{H}_0|Y) = 1$. □

We then wish to put Theorem 3.3.1 in context of the $\mathsf{Bin}(p, 1/2)$, $\mathsf{BB}(1, 1)$ and $\mathsf{BB}(1, \lambda p)$ priors respectively. Under all three priors we have shown that the posterior probability of the alternative hypothesis will converge to one when an alternative model is true. Thus, we are left to characterize the model space priors when the null model is true.

- $\mathsf{Bin}(1/2, p)$: The global prior odds is $\mathrm{PO}[\boldsymbol{H}_A : \boldsymbol{H}_0] = 2^p - 1$. Then, independent of the growth rate of p:

$$\lim_{n\to\infty} \frac{2^p - 1}{\sqrt{n}} = \infty.$$

  Therefore, by Theorem 3.3.1, $\mathrm{P}(\boldsymbol{H}_A|Y)$ converges to 1 independent of the data and under the assumption that the null hypothesis is true, the global posterior probability of the null hypothesis is an inconsistent selection criteria.

- $\mathsf{Bin}(1, 1)$: The global prior odds is $\mathrm{PO}[\boldsymbol{H}_A : \boldsymbol{H}_0] = p$.

  - p goes to infinity faster than $\sqrt{n}$: By Theorem 3.3.1, $\mathrm{P}(\boldsymbol{H}_A|Y)$ converges to 1 independent of the data and under the assumption that the null hypothesis is true, the global posterior probability of the null hypothesis is an inconsistent selection criteria.

  - p goes to infinity slower than $\sqrt{n}$: By Theorem 3.3.1, $\mathrm{P}(\boldsymbol{H}_0|Y)$ converges to 1 when the null hypothesis is assumed true and the global posterior probabilities are consistent selection criteria.

- $\mathsf{BB}(1, \lambda p)$: The global prior odds is $\mathrm{PO}[\boldsymbol{H}_A : \boldsymbol{H}_0] = 1$ and is a constant function of p. Therefore, by Theorem 3.3.1, $\mathrm{P}(\boldsymbol{H}_0|Y)$ converges to 1 when the null hypothesis is assumed true and the global posterior probabilities are consistent selection criteria independent of the growth rate of p.

3.4 Asymptotic Behavior of Global Posterior Probabilities: Full-Rank Case

We now look at the asymptotic behavior of the global posterior probabilities under the assumption of a full-rank design matrix. That is, if $\mathbf{X}_{\boldsymbol{\gamma}}$ is our $n \times p_{\boldsymbol{\gamma}}$-dimensional design matrix then the rank of $\mathbf{X}_{\boldsymbol{\gamma}}$ is the model size, $p_{\boldsymbol{\gamma}}$ for all models $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}$. Then, if the redundant case is thought of as the worst case scenario to achieve selection consistency of the global hypothesis under the null being true, the full-rank case can be thought of as the best case scenario. This is because the marginal model Bayes factors $\text{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} : \boldsymbol{\mathcal{M}}_0]$ will converge to zero the quickest when $\mathbf{X}_{\boldsymbol{\gamma}}$ is full rank for every $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}$. In particular, the asymptotic behavior of the global posterior probabilities in the full rank case will depend on on the prior odds of all models of size $p_{\boldsymbol{\gamma}}$ to the null model, $\text{PO}[p_{\boldsymbol{\gamma}} : p_0] = \binom{p}{p_{\boldsymbol{\gamma}}}\text{PO}[\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} : \boldsymbol{\mathcal{M}}_0]$. Theorem 3.4.1 states that under the assumption that the null hypothesis is true, we have selection consistency of the global posterior probabilities when p goes to infinity slower than $\sqrt{n}$ for all model space priors. We also achieve selection consistency when $\text{PO}[p_{\boldsymbol{\gamma}} : p_0]$ is a decreasing or constant function of p or when $\text{PO}[p_{\boldsymbol{\gamma}} : p_0]$ goes to infinity slower than $\sqrt{n}$. Some examples of the priors that achieve these conditions are $\mathsf{BB}(1, \lambda p)$ prior and $\mathsf{BB}(1, 1)$ independent of the rate of growth of p and $\mathsf{Bin}(p, 1/2)$ as long as p goes to infinity slower than $\sqrt{n}$. Also, we have selection inconsistency when p and $\text{PO}[p_{\boldsymbol{\gamma}} : p_0]$ go to infinity faster than $\sqrt{n}$. An example of a prior that follows this condition and leads to selection inconsistency under the global posterior probabilities is the $\mathsf{Bin}(p, 1/2)$ prior when p goes to infinity faster than $\sqrt{n}$.

Theorem 3.4.1. *True Null Hypothesis*

Assume that the null hypothesis is true and let n go to infinity and p go to infinity as some increasing function of n such that $\lim_{n\to\infty} \frac{p}{n} = 0$. Consider the two cases for the growth rate of p:

(1) When p goes to infinity slower than $\sqrt{n}$ the posterior probability of the null hypothesis, $P(\boldsymbol{H}_0|Y)$, goes to 1 independent of the prior placed on the model space. As long as $PO[\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} : \boldsymbol{\mathcal{M}}_0] \leq 1$ for all $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} \in \boldsymbol{\mathcal{M}}$. Thus, we achieve global selection consistency for posterior probabilities.

(2) When p goes to infinity faster than $\sqrt{n}$ the convergence of the posterior probability of the null hypothesis will depend on the following additional conditions on the prior odds of all models of size $p_{\boldsymbol{\gamma}}$ to the null model, $PO[p_{\boldsymbol{\gamma}} : p_0] = \binom{p}{p_{\boldsymbol{\gamma}}} PO[\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} : \boldsymbol{\mathcal{M}}_0]$:

(a) If $PO[p_{\boldsymbol{\gamma}} : p_0]$ is a decreasing, constant function of p or goes to infinity slower than $\sqrt{n}$ for all $p_{\boldsymbol{\gamma}}$

$$\text{plim}_{n,p\to\infty} P(\boldsymbol{H}_0|Y) = 1.$$

(b) If $PO[p_{\boldsymbol{\gamma}} : p_0]$ goes to infinity faster than $\sqrt{n}$ for all $p_{\boldsymbol{\gamma}}$

$$\text{plim}_{n,p\to\infty} P(\boldsymbol{H}_A|Y) = 1.$$

Proof. As in Equation 3.1 posterior probability of the null hypothesis is defined as:

$$\text{P}(\boldsymbol{H}_0|Y) = \left[1 + \sum_{\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} \in \boldsymbol{H}_A} \text{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} : \boldsymbol{\mathcal{M}}_0]\text{PO}[\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} : \boldsymbol{\mathcal{M}}_0]\right]^{-1}.$$

Thus, the asymptotic behavior of the posterior probability of the null model will depend on the sum of the posterior odds of all of the non-null models. By Theorem 3.2.4, under the assumption that the null model is true this sum has the following form:

$$\text{plim}_{n,p\to\infty} \sum_{\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} \in \boldsymbol{H}_A} \text{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} : \boldsymbol{\mathcal{M}}_0]\text{PO}[\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} : \boldsymbol{\mathcal{M}}_0]$$

$$= \lim_{n,p\to\infty} \sum_{p_{\boldsymbol{\gamma}}=1}^{p} \left[\frac{p_{\boldsymbol{\gamma}}+1}{n}\right]^{\frac{p_{\boldsymbol{\gamma}}}{2}} \binom{p}{p_{\boldsymbol{\gamma}}} \text{PO}[\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} : \boldsymbol{\mathcal{M}}_0].$$

We then look at two cases:

1. Assume p goes to infinity slower than $\sqrt{n}$:

   Under Case (1) we have the following behavior of the sum of the posterior odds of all of the non-null models since $\binom{p}{p_{\boldsymbol{\gamma}}} \leq \left[\frac{ep}{p_{\boldsymbol{\gamma}}}\right]^{p_{\boldsymbol{\gamma}}}$ and $\text{PO}[\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} : \boldsymbol{\mathcal{M}}_0] \leq 1$:

$$\begin{aligned}
&\text{plim}_{n,p\to\infty} \sum_{\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} \in \boldsymbol{H}_A} \text{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} : \boldsymbol{\mathcal{M}}_0]\text{PO}[\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} : \boldsymbol{\mathcal{M}}_0] \\
&= \lim_{n,p\to\infty} \sum_{p_{\boldsymbol{\gamma}}=1}^{p} \left[\frac{p_{\boldsymbol{\gamma}}+1}{n}\right]^{\frac{p_{\boldsymbol{\gamma}}}{2}} \binom{p}{p_{\boldsymbol{\gamma}}} \text{PO}[\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} : \boldsymbol{\mathcal{M}}_0]; \\
&\leq \lim_{n,p\to\infty} \sum_{p_{\boldsymbol{\gamma}}=1}^{p} \left[\frac{p}{\sqrt{n}}\right]^{p_{\boldsymbol{\gamma}}} \left[\frac{e\sqrt{p_{\boldsymbol{\gamma}}+1}}{p_{\boldsymbol{\gamma}}}\right]^{p_{\boldsymbol{\gamma}}}; \\
&\leq \lim_{n,p\to\infty} \sum_{p_{\boldsymbol{\gamma}}=1}^{p} \left[\frac{p}{\sqrt{n}}\right]^{p_{\boldsymbol{\gamma}}}; \\
&= \lim_{n,p\to\infty} \left[1 - \left(\frac{p}{\sqrt{n}}\right)^{\frac{p}{2}}\right] \left[\frac{\sqrt{n}}{p} - 1\right]^{-1}; \\
&= \lim_{n,p\to\infty} \frac{p}{\sqrt{n}}; \\
&= 0.
\end{aligned}$$

   Thus, $\text{plim}_{n,p\to\infty} \text{P}(\boldsymbol{H}_0|Y) = 1$ when p goes to infinity slower than $\sqrt{n}$. □

2. Assume p goes to infinity faster than $\sqrt{n}$:

   Under Case(2) we look at two subcases that depend on the prior odds of every model of a size $p_{\boldsymbol{\gamma}}$ to the null model, $\text{PO}[p_{\boldsymbol{\gamma}} : p_0]$.

   (a) Assume $\text{PO}[p_{\boldsymbol{\gamma}} : p_0]$ is a decreasing, constant function of p or goes to infinity slower than $\sqrt{n}$ for all $p_{\boldsymbol{\gamma}}$. Also, assume that we can write $\text{PO}[p_{\boldsymbol{\gamma}} : p_0] \leq \text{PO}(n)$ for all $p_{\boldsymbol{\gamma}}$ where $\text{PO}(n)$ goes to infinity slower than $\sqrt{n}$. Under

Subcase (a) we have the following behavior of the sum of the posterior odds of all of the non-null models:

$$\begin{aligned}
&\text{plim}_{n,p\to\infty} \sum_{\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} \in \boldsymbol{H}_A} \text{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} : \boldsymbol{\mathcal{M}}_0]\text{PO}[\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} : \boldsymbol{\mathcal{M}}_0] \\
&= \lim_{n,p\to\infty} \sum_{p_{\boldsymbol{\gamma}}=1}^{p} \left[\frac{p_{\boldsymbol{\gamma}}+1}{n}\right]^{\frac{p_{\boldsymbol{\gamma}}}{2}} \binom{p}{p_{\boldsymbol{\gamma}}} \text{PO}[\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} : \boldsymbol{\mathcal{M}}_0]; \\
&\leq \lim_{n,p\to\infty} \text{PO}(n) \sum_{p_{\boldsymbol{\gamma}}=1}^{p} \left[\frac{p_{\boldsymbol{\gamma}}+1}{n}\right]^{\frac{p_{\boldsymbol{\gamma}}}{2}}; \\
&= \lim_{n,p\to\infty} \frac{\text{PO}(n)}{\sqrt{n}} + \text{PO}(n) \sum_{p_{\boldsymbol{\gamma}}=2}^{p} \left[\frac{p_{\boldsymbol{\gamma}}+1}{n}\right]^{\frac{p_{\boldsymbol{\gamma}}}{2}}; \\
&= \lim_{n,p\to\infty} \frac{\text{PO}(n)}{n} \sum_{p_{\boldsymbol{\gamma}}=2}^{p} \left[\frac{p_{\boldsymbol{\gamma}}+1}{n}\right]^{\frac{p_{\boldsymbol{\gamma}}-2}{2}} (p_{\boldsymbol{\gamma}}+1); \\
&\leq \lim_{n,p\to\infty} \frac{\text{PO}(n)}{n} \sum_{p_{\boldsymbol{\gamma}}=2}^{p} \left[\frac{p+1}{n}\right]^{\frac{p_{\boldsymbol{\gamma}}-2}{2}} (p_{\boldsymbol{\gamma}}+1); \\
&= \lim_{n,p\to\infty} \frac{\text{PO}(n)}{\sqrt{n}}; \\
&= 0.
\end{aligned}$$

Thus, we see that $\text{plim}_{n,p\to\infty} \text{P}(\boldsymbol{H}_0|Y) = 1$. □

(b) Assume $\text{PO}[p_{\boldsymbol{\gamma}} : p_0]$ goes to infinity faster than $\sqrt{n}$. Also, assume that we can write $\text{PO}[p_{\boldsymbol{\gamma}} : p_0] \geq \text{PO}(n)$ for all $p_{\boldsymbol{\gamma}}$ where $\text{PO}(n)$ goes to infinity faster than $\sqrt{n}$. Under Subcase (b) we have the following behavior of the

sum of the posterior odds of all of the non-null models:

$$
\begin{aligned}
&\text{plim}_{n,p\to\infty} \sum_{\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} \in \boldsymbol{H}_A} \text{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} : \boldsymbol{\mathcal{M}}_0]\text{PO}[\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} : \boldsymbol{\mathcal{M}}_0] \\
&= \lim_{n,p\to\infty} \sum_{p_{\boldsymbol{\gamma}}=1}^{p} \left[\frac{p_{\boldsymbol{\gamma}}+1}{n}\right]^{\frac{p_{\boldsymbol{\gamma}}}{2}} \binom{p}{p_{\boldsymbol{\gamma}}} \text{PO}[\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} : \boldsymbol{\mathcal{M}}_0]; \\
&\geq \lim_{n,p\to\infty} \text{PO}(n) \sum_{p_{\boldsymbol{\gamma}}=1}^{p} \left[\frac{p_{\boldsymbol{\gamma}}+1}{n}\right]^{\frac{p_{\boldsymbol{\gamma}}}{2}}; \\
&\geq \lim_{n,p\to\infty} \text{PO}(n) \sum_{p_{\boldsymbol{\gamma}}=1}^{p} \left[\frac{1}{n}\right]^{\frac{p_{\boldsymbol{\gamma}}}{2}}; \\
&= \lim_{n,p\to\infty} \text{PO}(n) \left[1 - \left(\frac{1}{n}\right)^{\frac{p}{2}}\right] \left[\sqrt{n}-1\right]^{-1}; \\
&= \lim_{n,p\to\infty} \frac{\text{PO}(n)}{\sqrt{n}}; \\
&= \infty.
\end{aligned}
$$

Thus, we see that $\text{plim}_{n,p\to\infty} \text{P}(\boldsymbol{H}_A|Y) = 1$.

□

Therefore, by Theorem 3.4.1 under the assumption that the null hypothesis is true we have selection consistency of the global posterior probabilities under the $\mathsf{BB}(1,1)$, and $\mathsf{BB}(1,\lambda p)$ model space priors independent of the growth rate of p since under the $\mathsf{BB}(1,1)$ prior $\text{PO}[p_{\boldsymbol{\gamma}} : p_0] = 1$ for all $p_{\boldsymbol{\gamma}}$ and under the $\mathsf{BB}(1,\lambda p)$ prior $\text{PO}[p_{\boldsymbol{\gamma}} : p_0]$ is a decreasing function of p for all $p_{\boldsymbol{\gamma}}$. However, for the $\mathsf{Bin}(p,1/2)$ prior we achieve selection consistency of the global posterior probabilities only when p goes to infinity slower than $\sqrt{n}$ since $\text{PO}[p_{\boldsymbol{\gamma}} : p_0] \geq p$.

We then examine the asymptotic behavior the global posterior probabilities under the assumption that a finite alternative model is true in Theorem 3.4.2. In particular,

we wish to show that the penalty induced by the model space priors is not too strong and that $\text{plim}_{n,p\to\infty} \text{P}(\boldsymbol{H}_A|Y) = 1$ when a finite alternative model is true.

Theorem 3.4.2. *True Finite Alternative Model*

Assume some alternative finite model $\boldsymbol{\mathcal{M}}_* \in \boldsymbol{\mathcal{M}}$ is true and let n go to infinity and p go to infinity as some increasing function of n with $\lim_{n\to\infty} \frac{p}{n} = 0$. Then, under any prior on the model space such that:

$$\lim_{n\to\infty} \frac{PO[\boldsymbol{\mathcal{M}}_* : \boldsymbol{\mathcal{M}}_0]}{n^{\frac{p_*}{2}}} [1 + b_*]^{\frac{n-1}{2}} = \infty,$$

where b_* is some positive constant strictly between 0 and ∞ defined in Corollary 3.2.3 the posterior odds of $\boldsymbol{\mathcal{M}}_*$ to $\boldsymbol{\mathcal{M}}_0$ goes to infinity and in turn:

$$\text{plim}_{n,p\to\infty} P(\boldsymbol{H}_A|Y) = 1$$

Proof. As shown in the proof of Theorem 3.2.5 we have that:

$$\text{plim}_{n,p\to\infty} \text{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_* : \boldsymbol{\mathcal{M}}_0] = \lim_{n,p\to\infty} c_* \left[\frac{1}{n}\right]^{\frac{p_*}{2}} [1 + b_*]^{\frac{n-1}{2}};$$

where c_* is some positive constant less than infinity since b_* is strictly between 0 and ∞ by Corollary 3.2.3. Thus, it is trivial to show that the posterior odds as n goes to infinity and as p goes to infinity as some increasing function of n with $\lim_{n\to\infty} \frac{p}{n} = 0$ are:

$$\begin{aligned}
\text{plim}_{n,p\to\infty} \frac{\text{p}(\boldsymbol{\mathcal{M}}_*|Y)}{\text{p}(\boldsymbol{\mathcal{M}}_0|Y)} &= \text{plim}_{n,p\to\infty} \text{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_* : \boldsymbol{\mathcal{M}}_0]\text{PO}[\boldsymbol{\mathcal{M}}_* : \boldsymbol{\mathcal{M}}_0]; \\
&= \lim_{n,p\to\infty} c_* \left[\frac{\text{PO}[\boldsymbol{\mathcal{M}}_* : \boldsymbol{\mathcal{M}}_0]}{n^{\frac{p_*}{2}}}\right] [1 + b_*]^{\frac{n-1}{2}}; \\
&= \infty;
\end{aligned}$$

when we assume that

$$\lim_{n\to\infty} \frac{\text{PO}[\boldsymbol{\mathcal{M}}_* : \boldsymbol{\mathcal{M}}_0]}{n^{\frac{p_*}{2}}} [1 + b_*]^{\frac{n-1}{2}} = \infty.$$

Then, it is trivial to show that the posterior probability of the alternative hypothesis goes to 1:

$$\begin{aligned}
\mathrm{P}(\boldsymbol{H}_A|Y) &= 1 - \left[1 + \sum_{\boldsymbol{\mathcal{M}}_\gamma \in \boldsymbol{H}_A} \frac{\mathrm{p}(\boldsymbol{\mathcal{M}}_\gamma|Y)}{\mathrm{p}(\boldsymbol{\mathcal{M}}_0|Y)}\right]^{-1}; \\
&\leq \left[\frac{\mathrm{p}(\boldsymbol{\mathcal{M}}_*|Y)}{\mathrm{p}(\boldsymbol{\mathcal{M}}_0|Y)}\right]^{-1}; \\
&= 1.
\end{aligned}$$

□

Then we place Theorem 3.4.2 in context of the $\mathsf{Bin}(p, 1/2)$, $\mathsf{BB}(1,1)$ and $\mathsf{BB}(1, \lambda p)$ model space priors.

- $\mathsf{Bin}(p, 1/2)$: The prior odds are $\mathrm{PO}[\boldsymbol{\mathcal{M}}_* : \boldsymbol{\mathcal{M}}_0] = 1$ and thus it is trivial to show that:

$$\lim_{n\to\infty} \frac{[1+b_*]^{\frac{n-1}{2}}}{n^{\frac{p_*}{2}}} = \infty.$$

- $\mathsf{BB}(1,1)$: The prior odds are

$$\mathrm{PO}[\boldsymbol{\mathcal{M}}_* : \boldsymbol{\mathcal{M}}_0] = \frac{1}{\binom{p}{p_*}} \geq \frac{p_*!}{(p)^{p_*}}.$$

  Then, it is trivial to show that:

$$\begin{aligned}
\lim_{n\to\infty} \frac{\mathrm{PO}[\boldsymbol{\mathcal{M}}_* : \boldsymbol{\mathcal{M}}_0]}{n^{\frac{p_*}{2}}}[1+b_*]^{\frac{n-1}{2}} &\geq \lim_{n\to\infty} \frac{p_*!}{(p^2 n)^{\frac{p_*}{2}}}[1+b_*]^{\frac{n-1}{2}}; \\
&\geq \lim_{n\to\infty} p_*! \frac{[1+b_*]^{\frac{n-1}{2}}}{(n)^{\frac{3p_*}{2}}}; \\
&= \infty.
\end{aligned}$$

- $\mathsf{BB}(1, p)$: The prior odds are

$$\mathrm{PO}[\boldsymbol{\mathcal{M}}_* : \boldsymbol{\mathcal{M}}_0] = \frac{1}{\binom{2p}{p_*}} \left[\frac{2p}{2p - p_*}\right] \geq \frac{p_*!}{(2p)^{p_*}} \left[\frac{2p}{2p - p_*}\right].$$

Then, it is trivial to show that:

$$\begin{aligned}
\lim_{n\to\infty} \frac{\mathrm{PO}[\boldsymbol{\mathcal{M}}_* : \boldsymbol{\mathcal{M}}_0]}{n^{\frac{p_*}{2}}} [1 + b_*]^{\frac{n-1}{2}} &\geq \lim_{n\to\infty} \frac{p_*!}{(4p^2 n)^{\frac{p_*}{2}}} [1 + b_*]^{\frac{n-1}{2}}; \\
&\geq \lim_{n\to\infty} p_*! \frac{[1 + b_*]^{\frac{n-1}{2}}}{(4n^3)^{\frac{p_*}{2}}}; \\
&= \infty.
\end{aligned}$$

3.5 Asymptotic Behavior of Posterior Inclusion Probabilities

We are also interested in investigating how the potential inconsistency of the global posterior probabilities under the assumption that the null hypothesis is true can effect the behavior of the posterior marginal inclusion probabilities. In particular, we show that in the full rank case the inconsistency of the global posterior probabilities does not lead to selection inconsistency of the marginal inclusion probabilities. However, in the case of a rank deficient design matrix, the marginal inclusion probability of at least one of the variables will be inflated to $\frac{1}{2}$ under the assumption that the null hypothesis is true and the $\mathsf{Bin}(p, 1/2)$ prior when the global posterior probabilities are inconsistent.

The marginal inclusion probability for any variable X_i, $\mathrm{p}(\gamma_i = 1|Y)$ is given in Equation 3.1.1 and is defined:

$$\begin{aligned}
\mathrm{p}(\gamma_i = 1|Y) &= \mathrm{P}(\boldsymbol{H}_A|Y) \left[1 + \frac{\sum_{\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} \in \boldsymbol{H}_A} \mathbf{I}_{\gamma_i = 0} \mathrm{p}(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}|Y)/\mathrm{p}(\boldsymbol{\mathcal{M}}_0|Y)}{\sum_{\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} \in \boldsymbol{H}_A} \mathbf{I}_{\gamma_i = 1} \mathrm{p}(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}|Y)/\mathrm{p}(\boldsymbol{\mathcal{M}}_0|Y)}\right]^{-1}; \\
&= \mathrm{P}(\boldsymbol{H}_A|Y) \left[1 + A/B\right]^{-1};
\end{aligned}$$

where A is the sum of the posterior odds of all of the models in the alternative space such that $\gamma_i = 0$ and B is the sum of the posterior odds of all of the models in the alternative space such that $\gamma_i = 1$. These sums can be written as:

$$\begin{aligned}
A &= \sum_{\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} \in \boldsymbol{H}_A} \mathbf{I}_{\gamma_i=0} \mathrm{p}(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}|Y)/\mathrm{p}(\boldsymbol{\mathcal{M}}_0|Y); \\
&= \sum_{p_{\boldsymbol{\gamma}}=1}^{p} \binom{p-1}{p_{\boldsymbol{\gamma}}} \mathrm{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_{p_{\boldsymbol{\gamma}}} : \boldsymbol{\mathcal{M}}_0] \mathrm{PO}[\boldsymbol{\mathcal{M}}_{p_{\boldsymbol{\gamma}}} : \boldsymbol{\mathcal{M}}_0]; \\
B &= \sum_{\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} \in \boldsymbol{H}_A} \mathbf{I}_{\gamma_i=1} \mathrm{p}(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}|Y)/\mathrm{p}(\boldsymbol{\mathcal{M}}_0|Y); \\
&= \sum_{p_{\boldsymbol{\gamma}}=1}^{p} \binom{p-1}{p_{\boldsymbol{\gamma}}} \mathrm{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_{p_{\boldsymbol{\gamma}}+1} : \boldsymbol{\mathcal{M}}_0] \mathrm{PO}[\boldsymbol{\mathcal{M}}_{p_{\boldsymbol{\gamma}}+1} : \boldsymbol{\mathcal{M}}_0],
\end{aligned}$$

where $\mathrm{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_{p_{\boldsymbol{\gamma}}} : \boldsymbol{\mathcal{M}}_0]$ is the marginal model Bayes factor for any model of size $p_{\boldsymbol{\gamma}}$ and $\mathrm{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_{p_{\boldsymbol{\gamma}}+1} : \boldsymbol{\mathcal{M}}_0]$ is the marginal model Bayes factor for any model of size $p_{\boldsymbol{\gamma}} + 1$.

We can then see in the full rank case when the null hypothesis is true based on Theorem 3.2.4 and the fact that $\mathrm{PO}[\boldsymbol{\mathcal{M}}_{p_{\boldsymbol{\gamma}}+1} : \boldsymbol{\mathcal{M}}_{p_{\boldsymbol{\gamma}}}] \leq 1$ the sum B takes on the

form:

$$
\begin{aligned}
B &= \sum_{p_{\boldsymbol{\gamma}}=1}^{p} \binom{p-1}{p_{\boldsymbol{\gamma}}} \mathrm{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_{p_{\boldsymbol{\gamma}}+1} : \boldsymbol{\mathcal{M}}_0] \mathrm{PO}[\boldsymbol{\mathcal{M}}_{p_{\boldsymbol{\gamma}}+1} : \boldsymbol{\mathcal{M}}_0]; \\
&= \sum_{p_{\boldsymbol{\gamma}}=1}^{p} \binom{p-1}{p_{\boldsymbol{\gamma}}} \mathrm{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_{p_{\boldsymbol{\gamma}}} : \boldsymbol{\mathcal{M}}_0] \mathrm{PO}[\boldsymbol{\mathcal{M}}_{p_{\boldsymbol{\gamma}}} : \boldsymbol{\mathcal{M}}_0] \left[\mathrm{p}(\boldsymbol{\mathcal{M}}_{p_{\boldsymbol{\gamma}}+1}|Y) / \mathrm{p}(\boldsymbol{\mathcal{M}}_{p_{\boldsymbol{\gamma}}}|Y) \right]; \\
&\leq \sum_{p_{\boldsymbol{\gamma}}=1}^{p} \binom{p-1}{p_{\boldsymbol{\gamma}}} \mathrm{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_{p_{\boldsymbol{\gamma}}} : \boldsymbol{\mathcal{M}}_0] \mathrm{PO}[\boldsymbol{\mathcal{M}}_{p_{\boldsymbol{\gamma}}} : \boldsymbol{\mathcal{M}}_0] \left[\frac{p_{\boldsymbol{\gamma}}+2}{n} \right]^{\frac{1}{2}}; \\
&\leq \left[\frac{p+2}{n} \right]^{\frac{1}{2}} \sum_{p_{\boldsymbol{\gamma}}=1}^{p} \binom{p-1}{p_{\boldsymbol{\gamma}}} \mathrm{BF}_{p(g)}[\boldsymbol{\mathcal{M}}_{p_{\boldsymbol{\gamma}}} : \boldsymbol{\mathcal{M}}_0] \mathrm{PO}[\boldsymbol{\mathcal{M}}_{p_{\boldsymbol{\gamma}}} : \boldsymbol{\mathcal{M}}_0]; \\
&= \left[\frac{p+2}{n} \right]^{\frac{1}{2}} A.
\end{aligned}
$$

Thus, it is trivial to show that under the assumption that the null hypothesis is true and when there is inconsistency of the global posterior probabilities under the full rank case that the marginal inclusion probabilities are still consistent $\mathrm{plim}_{n,p\to\infty} \mathrm{p}(\gamma_i = 1|Y) = 0$:

$$
\begin{aligned}
\mathrm{plim}_{n,p\to\infty} \mathrm{p}(\gamma_i = 1|Y) &= \mathrm{plim}_{n,p\to\infty} \mathrm{P}(\boldsymbol{H}_A|Y) \left[1 + A/B\right]^{-1}; \\
&\leq \lim_{n,p\to\infty} \left[1 + \left(\frac{n}{p} \right)^{\frac{1}{2}} \right]^{-1}; \\
&= 0.
\end{aligned}
$$

However, in the Redundant case, for any variable X_i the marginal inclusion probability takes on a simplified form:

$$
\begin{aligned}
\mathrm{p}(\gamma_i = 1|Y) &= \mathrm{P}(\boldsymbol{H}_A|Y) \left[1 + \frac{\sum_{\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} \in \boldsymbol{H}_A} \mathbf{I}_{\gamma_i=0} \mathrm{p}(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}|Y) / \mathrm{p}(\boldsymbol{\mathcal{M}}_0|Y)}{\sum_{\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} \in \boldsymbol{H}_A} \mathbf{I}_{\gamma_i=1} \mathrm{p}(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}|Y) / \mathrm{p}(\boldsymbol{\mathcal{M}}_0|Y)} \right]^{-1}; \\
&= \frac{\mathrm{P}(\boldsymbol{H}_A|Y)}{p(H_A)} p(\gamma_i = 1);
\end{aligned}
$$

where $p(\gamma_i = 1)$. Then it is trival to see that under the assumption that the null hypothesis is true and where there is inconsistency of the global posterior probabilities under the redundant case the asymptotic behavior of the marginal inclusion probabilities depends on $\frac{p(\gamma_i=1)}{p(H_A)}$:

$$\text{plim}_{n,p\to\infty} \text{p}(\gamma_i = 1|Y) = \lim_{n,p\to\infty} \frac{p(\gamma_i = 1)}{p(H_A)}.$$

In particular this will lead to inconsistency under the $\mathsf{Bin}(p, 1/2)$ and $\mathsf{BB}(1, 1)$ priors since:

$$\text{plim}_{n,p\to\infty} \text{p}(\gamma_i = 1|Y) = \frac{1}{2}.$$

However, this is not the case for the $\mathsf{BB}(1, p)$ prior since

$$\text{plim}_{n,p\to\infty} \text{p}(\gamma_i = 1|Y) = 0.$$

3.6 Discussion

In summary, we have derived criteria of the model space priors that is necessary to achieve selection consistency of the global posterior probabilities when n goes to infinity and when p goes to infinity as some increasing function of n such that $\lim_{n\to\infty} \frac{p}{n} = 0$. In particular, we show that in the extreme case of including additional redundant variables to the analysis, we must have that the global prior odds, $\text{PO}[\boldsymbol{H}_A : \boldsymbol{H}_0]$, are a constant function of n or that they go to infinity slower than $\sqrt{n}$. Also, in the full rank case, we have shown that as long as p goes to infinity slower than $\sqrt{n}$ or if the prior odds of all models of a size $p_{\boldsymbol{\gamma}}$ to the null model, $\text{PO}[p_{\boldsymbol{\gamma}} : p_0]$, are a decreasing, constant or increasing function of n that goes to infinity slower than $\sqrt{n}$ for all $p_{\boldsymbol{\gamma}}$ we achieve selection consistency.

Given these results, we have shown that even under the best case scenario of a full rank design matrix, we only achieve selection consistency of the global hypotheses

under the Uniform prior when p goes to infinity slower than $\sqrt{n}$. In most studies involving a large number of variables, the sample size does not grow nearly fast enough to accommodate. Furthermore, most interesting problems involve a large number of correlated covariates and one can not guarantee that the design matrix will be full rank. We therefore recommend the $\mathsf{BB}(1, \lambda p)$ prior on model size under which selection consistency is achieved for both full rank and rank deficient cases by maintaining a constant global prior odds as the number of variables being tested increases.

All of the above results and criteria have been developed under the assumption of the Zellner-Siow prior on the coefficients. Thus, they have assumed a specific convergence or divergence rate of the marginal model Bayes factors as n, p go to infinity. It is of interest to us to extend these results to other priors on the coefficients, such as the intrinsic priors and the empirical Bayes g priors.

4

Discussion

The work herein has served to characterize model space priors, model search algorithms and posterior quantities for inference within the Bayesian model uncertainty framework for problems involving a large number of correlated covariates in two specific classes of models. In summary, we have determined necessary conditions on model space priors in the setting of multiple linear regression to achieve selection consistency of the global posterior probabilities and marginal inclusion probabilities as the sample size, n, and the number of covariates, p, go to infinity. We have also described a model search algorithm and procedure for multilevel posterior inference for assessing whether a set of genetic risk factors is associated with a complex disease. In particular, by implementing the model search algorithm and multilevel posterior quantities in MISA, we were able to gain power over more commonly used methods to detect genetic associations. This has direct implications in that we were able to identify genetic risk factors for ovarian cancer that were not identified by the more commonly used methods and that have been validated by independent future studies.

4.1 Future Directions

4.1.1 Correlation/Group Model Priors

My results regarding the asymptotic behavior of model space priors suggests that global posterior probabilities are strongly influenced by the choice of prior model probabilities, $p(\boldsymbol{\mathcal{M}}_\gamma)$, and that this behavior may depend on the correlation structure of the variables. Motivated by these results and the natural clustering of the genetic covariates (SNPs) into a group structure based on their known correlation structure in candidate gene analyses, a feature reflected in many other genomic applications, I aim to build on the results herein and develop a new class of hierarchical group-based priors. These group priors will assume prior independence and exchangeability across groups, but not within and will have a constant probability of at least one associated within a group independent of the group size. A promising first example, based my preliminary theoretical work, would be to place a $\mathsf{BB}(1, \lambda p_g)$ distribution on the number of associated variables within a given group g, where p_g is the number of covariates within the group, and place a $\mathsf{Bin}(p, \pi)$ distribution on the number of associated groups, where p is the total number of groups under consideration and π is the probability that a group is associated. I will investigate theoretical properties of the posterior quantities of interest when the covariates are correlated and propose a generalization of this prior when the group structure is not known a priori by first performing a cluster analysis on the covariates to estimate a group structure.

4.1.2 Efficient Stochastic Model Search

In MISA, the model search algorithm is based on a combination of parallel tempering [23] and a genetic algorithm [29]. The genetic algorithm incorporates move types into our model search that mimic the idea of evolution in that individuals (or in this case models) compete and mate to produce increasingly stronger individuals. An

extension to the current structure of these model search algorithms would incorporate the correlation and above defined group structure of the covariates into the move types. Introducing formal constraints to the joint moves of the covariates in and out of the models will improve efficiency of the search. Increasing the efficiency of the search algorithm is particularly important since we wish to generalize the analytical strategy within MISA to search over genetic interactions and gene by environment interactions in addition to the genetic main effects.

Appendix A

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

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

where $\boldsymbol{\mathcal{I}}_\gamma$ is the observed Fisher information under model $\boldsymbol{\mathcal{M}}_\gamma$ evaluated at the maximum likelihood estimates (MLEs) $\hat{\boldsymbol{\theta}}_\gamma$. 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 $\hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}}$:

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

leading to the approximate marginal likelihood

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

where $K_{\boldsymbol{\theta}_{\boldsymbol{\gamma}}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}}, \boldsymbol{\mathcal{I}}_{\boldsymbol{\gamma}}^{-1}) = \exp\{-\frac{1}{2}(\boldsymbol{\theta}_{\boldsymbol{\gamma}} - \hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}})^T \boldsymbol{\mathcal{I}}_{\boldsymbol{\gamma}} (\boldsymbol{\theta}_{\boldsymbol{\gamma}} - \hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}})\}$. Setting this approximate $\log(p(D \mid \boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}})$ 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.

Appendix B

Evolutionary Monte Carlo

We use an Evolutionary Monte Carlo (EMC) [39] algorithm to sample models that maximize a fitness function $\psi(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}})$. In the current setting, this quantity is proportional to the log of the posterior probability of the sampled models. This is achieved using parallel tempering with N parallel Markov chains, each associated with a decreasing temperature T_i in a temperature ladder $T = \{T_1, T_2, \cdots, T_N\}$ and sampling from the distribution $p_{T_i}(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} \mid D) \propto \exp(\frac{\psi(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}})}{T_i})$. The advantages of parallel tempering over single chain MCMC methods include improved mixing and its ability to escape local modes. The resulting sample from the model space $\boldsymbol{\mathcal{M}}$ (the sample of models from the chain with temperature $T_i = 1$) is from the stationary (posterior) distribution. In the EMC framework, each current state of one of the parallel chains corresponds to an "individual" or model and the full set of current states of the chains correspond to the "population" or set of models.

The parallel chains are updated by the following populations moves that are based on a genetic algorithm: Mutation, Crossover and Exchange. For each update we accept or reject the proposed move on the population of models based on the

probability $\min(1, r)$. Here, r is the Metropolis–Hastings (MH) ratio corresponding to the original and updated populations, $P_{\boldsymbol{\mathcal{M}}}$ and $P_{\boldsymbol{\mathcal{M}}*}$ respectively, that has the general form:

$$r = \frac{f(P_{\boldsymbol{\mathcal{M}}}*)}{f(P_{\boldsymbol{\mathcal{M}}})} \frac{t(P_{\boldsymbol{\mathcal{M}}} \mid P_{\boldsymbol{\mathcal{M}}}*)}{t(P_{\boldsymbol{\mathcal{M}}}* \mid P_{\boldsymbol{\mathcal{M}}})},$$

where $f(P_{\boldsymbol{\mathcal{M}}})$ is the product joint distribution for the population of models defined as

$$f(P_{\boldsymbol{\mathcal{M}}}) = \prod_{i=1}^{N} p_{T_i}(\boldsymbol{\mathcal{M}}_i \mid D),$$

and $t(P_{\boldsymbol{\mathcal{M}}} \mid P_{\boldsymbol{\mathcal{M}}}*)$ is the transition probability between populations.

We first update the population via a mutation step where we perform a Metropolis update on the population by choosing a model, or current value of one of the chains and taking one of the SNP indicators and mutating its status in the chosen model. Given our population of models, $P_{\boldsymbol{\mathcal{M}}} = \{\boldsymbol{\mathcal{M}}_1, \boldsymbol{\mathcal{M}}_2, \cdots, \boldsymbol{\mathcal{M}}_p\}$, we sample one of the models $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}$, and mutate it to some model $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}*$ and accept or reject the new population, $P_{\boldsymbol{\mathcal{M}}}* = \{\boldsymbol{\mathcal{M}}_1, \cdots, \boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}*, \cdots, \boldsymbol{\mathcal{M}}_p\}$ based on the probability $\min(1, r_m)$, where r_m is the MH ratio for the mutation update. Specifically, to update the model $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}$ to model $\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}*$ we update the corresponding model specification vector $\gamma = (\gamma_1, \gamma_2, \cdots, \gamma_p)$ to $\gamma*$ by mutating one value $\gamma_m \in \{0, 1, 2, 3\}$, based on the corresponding SNP inclusion and genetic mode of inheritance in the model, to another value $\gamma_m* \in \{0, 1, 2, 3\} \setminus \gamma_m$. We then choose γ_m* based on the following probabilities for each possible value of γ_m*:

$$p(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}* \mid \gamma_m*) = \frac{\exp\left(\psi(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}* \mid \gamma_m*)/T_\gamma\right)}{\sum_{\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}* \neq \boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}} \exp\left(\psi(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}* \mid \gamma_m*)/T_\gamma\right)}.$$

We accept or reject the new population $P_{\boldsymbol{\mathcal{M}}}*$ based on the MH ratio

$$r_m = \frac{\sum_{\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}* \neq \boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}} \exp\left(\psi(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}* \mid \gamma_m*)/T_\gamma\right)}{\sum_{\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}} \neq \boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}*} \exp\left(\psi(\boldsymbol{\mathcal{M}}_{\boldsymbol{\gamma}}* \mid \gamma_m*)/T_\gamma\right)}.$$

The population is also updated via the normal parallel tempering exchange step that allows models to move up or down the temperature ladder. We choose a temperature ladder of the form $T_j - T_i = \exp(\frac{T_j}{T_i})$ where $T_i = 1$ for some value i in the ladder. Here, given the current population $P_{\boldsymbol{\mathcal{M}}} = \{\boldsymbol{\mathcal{M}}_1, \cdots, \boldsymbol{\mathcal{M}}_i, \cdots, \boldsymbol{\mathcal{M}}_j, \cdots, \boldsymbol{\mathcal{M}}_p\}$, we sample two of the models $\boldsymbol{\mathcal{M}}_i$ and $\boldsymbol{\mathcal{M}}_j$ and propose a new population $P_{\boldsymbol{\mathcal{M}}*} = \{\boldsymbol{\mathcal{M}}_1, \cdots, \boldsymbol{\mathcal{M}}_j, \cdots, \boldsymbol{\mathcal{M}}_i, \cdots, \boldsymbol{\mathcal{M}}_p\}$, by making an exchange between $\boldsymbol{\mathcal{M}}_i$ and $\boldsymbol{\mathcal{M}}_j$ without changing the associated temperature ladder. We only allow the exchange to take place between two models with neighboring temperatures and therefore the transition probability is symmetric with $t(P_{\boldsymbol{\mathcal{M}}} \mid P_{\boldsymbol{\mathcal{M}}}*) = t(P_{\boldsymbol{\mathcal{M}}}* \mid P_{\boldsymbol{\mathcal{M}}})$. We then accept or reject the new population $P_{\boldsymbol{\mathcal{M}}}*$ based on the MH ratio

$$r_e = \exp\left((\psi(\boldsymbol{\mathcal{M}}_j) - \psi(\boldsymbol{\mathcal{M}}_i))(T_i^{-1} - T_j^{-1})\right).$$

The exchange update allows better models to move down the ladder where the chains explore local perturbations to them, while less interesting models move up the ladder to serve as the foundation for more global perturbations.

The mutation and exchange step make up the normal population updates involved in parallel tempering. Evolutionary Monte Carlo introduces a crossover update inspired by genetic algorithms. While the exchange step is a full state swap, the crossover update allows the states to swap partially. The general idea is that one of the top current models is chosen to "mate" with another random model and two new models are formed by some composition of the two parental models. Thus, given the current population $P_{\boldsymbol{\mathcal{M}}} = \{\boldsymbol{\mathcal{M}}_1, \cdots, \boldsymbol{\mathcal{M}}_i, \cdots, \boldsymbol{\mathcal{M}}_j, \cdots, \boldsymbol{\mathcal{M}}_p\}$, we choose two models $\boldsymbol{\mathcal{M}}_i$ and $\boldsymbol{\mathcal{M}}_j$ based on a weighted selection procedure and update these models to $\boldsymbol{\mathcal{M}}_i*$ and $\boldsymbol{\mathcal{M}}_j*$ based on randomly switching the values of the two model specification vectors. We then accept or reject the new population $P_{\boldsymbol{\mathcal{M}}}* = \{\boldsymbol{\mathcal{M}}_1, \cdots, \boldsymbol{\mathcal{M}}_i*, \cdots, \boldsymbol{\mathcal{M}}_j*, \cdots, \boldsymbol{\mathcal{M}}_p\}$ based on the MH ratio

$$r_c = \exp\left((\psi(\boldsymbol{\mathcal{M}}_i*) - \psi(\boldsymbol{\mathcal{M}}_i))/T_i + (\psi(\boldsymbol{\mathcal{M}}_j*) - \psi(\boldsymbol{\mathcal{M}}_j))/T_j\right).$$

Appendix C

Marginal Bayes Factor Screen

We used Laplace approximations to estimate the marginal Bayes Factors (BFs) used to screen the SNPs [32]. 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.

Appendix D

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 [57], 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 [42]. 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 described in [37]. 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 in Chapter 2 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.

Appendix E

Web Resources: MISA package vignette

The URL for the software for the methodology and simulations presented in this paper is:

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

The functions in the MISA package focus on intermediate throughput case-control association studies, where the outcome of interest is often a binary disease state and where the genetic markers have been chosen to capture variation in a set of related genes, such as those involved in a specific biochemical pathway. Given this data, we are interested in addressing two questions: "To what extent does the data support an overall association between the pathway and outcome of interest?" and "Which markers or genes are most likely to be driving this association?" To address both of these questions,this package performs a Bayesian model search technique that utilizes Evolutionary Monte Carlo and searches over models including main effects of all genetic markers and marker-specific genetic effects in a computationally efficient manner. The package incorporates functions that:

- Calculate marginal Bayes factors under the log-additive, dominant and reces-

sive parameterizations of genotype that can be used as a screen to reduce the number of SNPs included in the model search.

- Samples a set of models using the EMC algorithm from the posterior.
- Assess convergence of the model search algorithm.
- Calculates multi-level posterior quantities of interest (SNP,gene and global Bayes factors).
- Summarizes results with image plots of the SNPs and genes included in the top models (based on model posterior probabilities).

This vignette is an example of using the `MISA` package in R. We will first load the library (where the lib.loc argument may be needed if MISA is not in the R library tree):

```
> library(MISA)
```

E.1 Load Simulated Data

Simulated data for the problem is in the datasets `dna.snp.full`. There are 62 variables in the data set: the response `case` is the disease status of each individual, `age` is a covariate that will be included in every model, and there are 60 SNPs, parameterized as the number of rare alleles (0, 1 or 2) that an individual carries. The data set `dna.snp` has a subset of the full set of SNPs (19). This subset is determined by the marginal Bayes factor screen `bf4assoc` on the full data set `dna.snp.full`. More information on the simulation (including assumed odds ratio and genetic parametrizations of the SNPs) is found in the data set `sim.info`. In particular, we assume that there are two associated genes and 4 unassociated genes. In each of the two associated genes we assume that one of the SNPs is the source of the association. One

of the genes, "GADD45A", is assumed to have an odds ratio of 1.25 and a dominate genetic parametrization. The other gene, "MDM2", is assumed to have an odds ratio of 1.75 and an log-additive genetic parametrization. We begin the analysis by loading the full data set that will be used for the remainder of the analyses:

```
> data(dna.snp.full)
> p.full <- dim(dna.snp.full)[2] - 2
> data(sim.info)
```

E.2 Marginal Bayes Factor Screen

A marginal Bayes factor screen is then first used to reduce the number of SNPs that will go into the MISA analysis. This is done by the following R statements:

```
> marg.bf <- bf4assoc(D = as.numeric(dna.snp.full$case) -
+     1, X = as.matrix(dna.snp.full$age), XS = as.matrix(dna.snp.full[,
+     -c(1, 2)]), Ns = p.full, Nx = 1, snpsd = 0.25,
+     Prior = 0, MinCount = 1.9, MaxIt = 1000, RelTol = 1e-07)

[1] -0.76126  0.13417

> max.bf <- apply(marg.bf[, c(3:5)], 1, max)
> dna.snp <- dna.snp.full[, c(TRUE, TRUE, max.bf >
+     1)]
> p <- dim(dna.snp)[2] - 2
```

We note that we are using Prior=0, which is the normal prior on the genetic effects and snpsd=0.25 which is is the assumed standard deviation of the mean zero normal prior. We then are interested in computing the maximum marginal Bayes factor for each SNP across the 3 genetic parametrizations (log-additive, dominate, recessive)

and allowing the SNPs with a maximum marginal Bayes factor greater than one to pass the screen and go on for further analysis. The SNPs that have passed the screen are in the data set `dna.snp`.

E.3 Model Search Algorithm

Once we run the marginal screen to subset the initial full set of SNPs we run `Gene.EMC` to sample a set of models from the model space based on the Evolutionary Monte Carlo model search algorithm. For the purpose of this example we just run 10 iterations. However, for the two output data sets `emc.out.1` and `emc.out.2` that were used to analyze the simulation we ran each for 100,000 iterations. The EMC algorithm was initialized with two random SNPs in the model.

```
> start.snp <- rep(FALSE, p)
> start.snp[c(3, 7)] <- TRUE
> emc.out <- Gene.EMC(data = dna.snp, force = c("age"),
+     fitness = "AIC.BB", b = p, a = 1, start.snps = start.snp,
+     iter = 10, N = 5, tmax = 5, tlone = 1, qm = 0.25,
+     display.acc = TRUE, display.acc.ex = FALSE,
+     cores = 1)

Iter 1: Cost = 1545.427
accept rates of mutation and crossover are: 0 1
Iter 2: Cost = 1545.427
accept rates of mutation and crossover are: 0 1
Iter 3: Cost = 1545.427
accept rates of mutation and crossover are: 0 1
Iter 4: Cost = 1545.427
accept rates of mutation and crossover are: 0 1
```

```
Iter 5: Cost = 1545.427
accept rates of mutation and crossover are: 0.6 1
Iter 6: Cost = 1550.659
accept rates of mutation and crossover are: 0.6 1
Iter 7: Cost = 1545.427
accept rates of mutation and crossover are: 0.6 1
Iter 8: Cost = 1545.427
accept rates of mutation and crossover are: 0.73 1
Iter 9: Cost = 1543.761
accept rates of mutation and crossover are: 0.73 0.83
Iter 10: Cost = 1543.761
accept rates of mutation and crossover are: 0.73 0.86
```

We note that the covariate "age" is included as a design variable in every model, including the null model. Also, fitness="AIC.BB" donates that we use AIC to approximate the marginal likelihood of each model and that we place a Beta-Binomial prior with hyper-parameters $a = 1$ and $b = p$ on the model space. The output of the function `Gene.EMC` can be printed to the screen or you can give a output file that it will be printed to. This output allows the user to keep an eye on acceptance rates of the mutation and crossover steps (display.acc=TRUE). Although not shown here, we can also output the acceptance rates of the exchange step between each chain in the population to make sure that the temperature scheme is appropriately tuned (display.acc.ex=TRUE). If you see low acceptance rates between chains this may mean that the difference in temperatures between adjacent chains is too far apart and may need to be decreased.

E.4 Assessing Convergence of Model Search Algorithm

To assess convergence, we run two independent runs of the EMC algorithm for 100,000 iterations each and save the output from the function `Gene.EMC` (`emc.out.1` and `emc.out.2`). Both output data sets are loaded below and convergence diagnostics are plotted in Figure E.1 and are made by the R statements:

```
> data(emc.out.1)
> data(emc.out.2)
> emc.out <- list(emc.out.1, emc.out.2)
> emc.converge(emc.out, plot.type = "all", bandwidth = 1000,
+     b = p, a = 1)
```

We assess convergence of the sampling algorithm by using graphical diagnostics that summarize two independent, long runs of the algorithm. These diagnostics are plotted in Figure E.1 for two independent runs of the simulation, each 100,000 iterations long. The upper left panel depicts end–to–end trace plots of the cost values associated with the models sampled in the two runs. The first is plotted in blue and the second in red. This plot is used to verify that the algorithm's movement around the model space is adequate. If this is the case, the sampled model will change frequently and the associated cost values will vary around an average cost so that points in the plot appear to fall within a common horizontal band. If there appears to be drift at the start of either of the trace plots, the associated iterations, representing a period of burn–in, should be removed from the analysis. In addition, if the sampled cost value changes infrequently, indicating that the algorithm is not moving adequately, the algorithm should be restarted with a larger maximum temperature or a larger number of parallel chains so that adjacent chains can communicate better.

The upper right panel in the figure depicts a plot of the Gelman–Rubin shrink factor computed for the sampled cost values. The shrink factor should approach 1 as

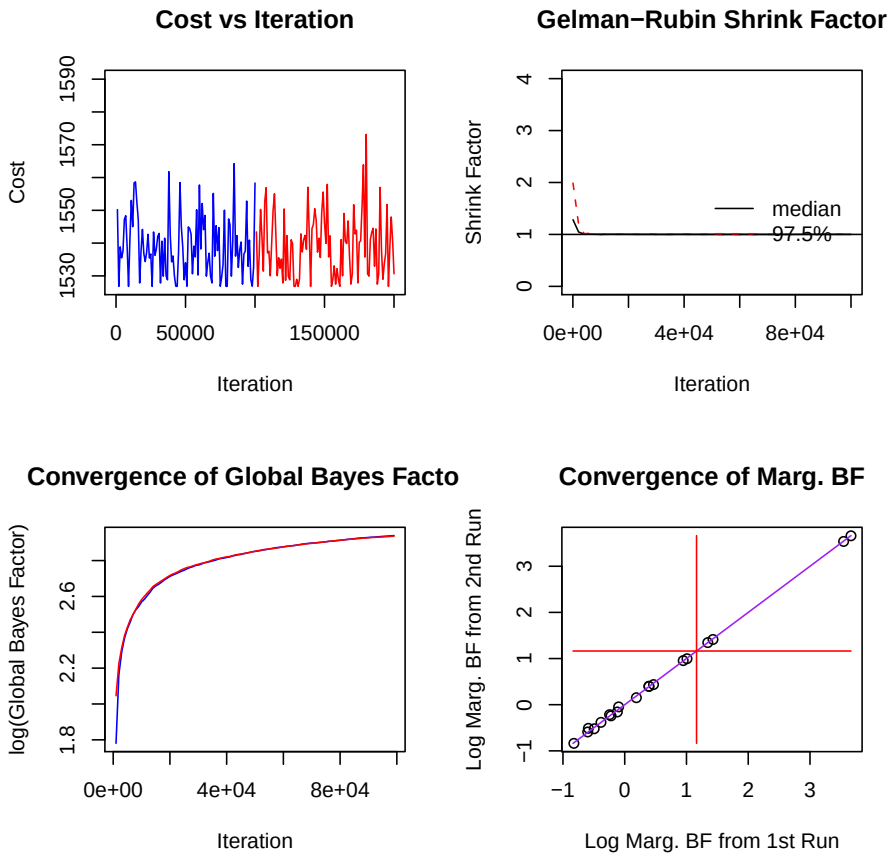

FIGURE E.1: Convergence diagnostics for EMC.Upper left panel:Trace plot of the cost values of the models visited over each iteration for each of the two independent runs. Upper right panel: Plot of the Gelman-Rubin convergence diagnostic of the cost values of the models visited in the two independent chains. Lower left panel: Plot of the global log Bayes factor computed across iterations for each independent chain. Lower right panel: Plot of the SNP Bayes factors for one of the independent runs vs. another independent run

the cost values of the models converge. The lower left panel of the figure shows the logarithm global Bayes factor for each of the two independent runs (the first in blue and the second in red) as a function of iteration. The estimated Bayes factor, which is a lower bound on the value we would compute were we able to enumerate all models, will increase with every new model sampled. As unique models are sampled less frequently as the algorithm runs, the Bayes factor will begin to plateau suggesting

that the value of additional samples is diminishing. The plots of the two runs should begin to converge, with neither making large jumps in the later iterations. Finally, the lower left panel plots the marginal Bayes factors for each of the SNPs for the two independent runs of the algorithm. This plot enables us to determine if the values of the marginal SNP Bayes factors are consistent across two independent runs of the algorithm. Points in the scatter plot will be near the diagonal when MCMC sampling variability for estimating the associated marginal Bayes factor is low. If the plot shows significant deviation from the diagonal, the algorithm has not yet converged and should be allowed to run longer.

E.5 Calculation of Posterior Quantities of Interest

Given a sample from the model space produced by the output of the function `Gene.EMC` that has passed the above convergence diagnostics, we are able to compute several multi-level posterior quantities of interest: (1) Global Bayes factor giving the evidence of at least one association in the set of tested SNPs, (2) Gene Bayes factors giving the evidence of at least one association within the set of tested SNPs within a particular gene, and (3) SNP Bayes factors giving the marginal evidence of an association with the function `post.prob`.

```
> snp.info <- as.character(sim.info[match(colnames(emc.out.1$which)[-(p +
+     1)], sim.info$SNP), 1])
> post.prob.out <- post.prob(emc.out.1, snp.info,
+     b = p)
> post.prob.out$BF

              Prior= 1
Post.Prob     0.9498073
```

```
Bayes.Factor 18.9232237
Prior.Odds    1.0000000
```

In particular, we can look at the global posterior probability and Bayes factor giving the evidence of at least one associated marker in the data set.

E.6 Summary of Results

Summaries of the marginal SNP Bayes factor and gene Bayes factor results are plotted in Figure E.2 and E.3 and are made by the following R statements using the output from the function `post.prob`:

```
> model.inc(post.prob.out, num.models = 100, num.snps = p,
+     inc.typ = "s")

> model.inc(post.prob.out, num.models = 100, num.genes = 6,
+     inc.typ = "g")
```

Figures E.2 and E.3 are image plots of the SNP and gene inclusion indicators for the top 100 Models. SNPs/Genes are ordered based on their marginal Bayes Factors which are plotted on the right axis. The columns correspond to models and have width proportional to the estimated model probability, and the models are plotted in descending order of posterior support. The color of the inclusion block in the SNP plot corresponds to the genetic parameterization of the SNP in that model. Purple corresponds to a log–additive parameterization, red to a dominant parameterization and blue to a recessive parameterization.The color in the gene plot is chosen to be neutral since the genetic parameterizations are not defined at the gene level. We notice that SNPs within both associated genes have marginal SNP Bayes factors suggesting strong evidence of an association and that the marginal gene Bayes factors for the associated genes also suggest strong evidence of an association.

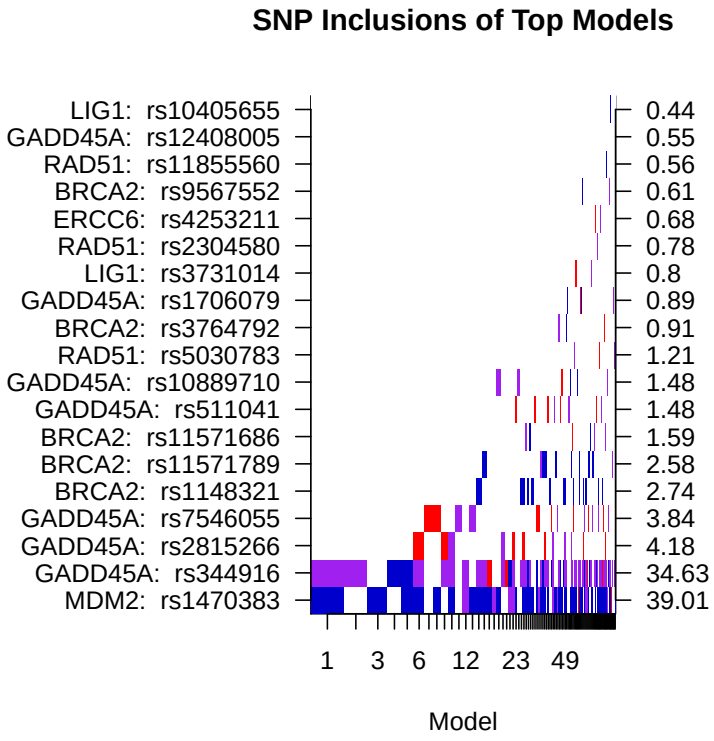

FIGURE E.2: Image plot of SNP inclusions in the top 100 models.

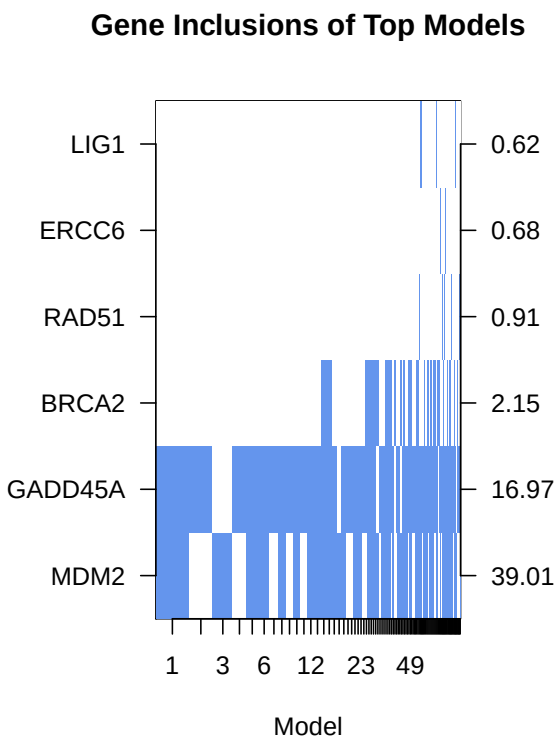

FIGURE E.3: Image plot of gene inclusions in the top 100 models.

Bibliography

[1] D. J. Balding. A tutorial on statistical methods for population association studies. *Nature*, 7:781–791, 2006.

[2] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57:289–300, 1995.

[3] Leonardo Bottolo and Sylvia Richardson. Evollutionary stochastic search, 2008.

[4] P. J. Brown, M. Vannucci, and T. Fearn. Bayes model averaging with selection of regressors. *J. R. Statist. Soc.*, 64(3):519–536, 2002.

[5] P. R. Burton et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447:661–678, 2007.

[6] G. Casella and E. Moreno. Objective Bayes variable selection. *J. Amer. Statist. Assoc.*, 101:157–167, March 2006.

[7] George Casella, Javier Giron, Lina Martinez, and Elias Moreno. Consistency of bayesian procedures for variable selection. *Annals of Statistics*, 37(3):1207–1228, 2009.

[8] Hugh Chipman. Bayesian variable selection with related predictors. *Canadian Journal of Statistics*, 24(1):17–36, March 1996.

[9] D. G. Clayton et al. Population structure differential bias and genomic control in a large-scale caseï£¡control association study. *Nature Genet.*, 37:1243–1246, 2005.

[10] Merlise Clyde. Bayesian model averaging and model search strategies (with discussion). In *Bayesian Statistics 6 – Proceedings of the Sixth Valencia International Meeting*, pages 157–185, 1999.

[11] Merlise Clyde and Edward I. George. Model uncertainty. *Statistical Science*, 19:81–94, 2004.

[12] Merlise Clyde, Joyee Ghosh, and Michael Littman. Bayesian adaptive sampling for variable selection. Discussion Paper 2009-16, Duke University Department of Statistical Science, 2009.

[13] H. J. Cordell and D. G. Clayton. A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: Application to HLA in type 1 diabetes. *AJHG*, 70:124–141, 2002.

[14] Wen Cui and Edward I. George. Empirical Bayes vs. fully Bayes variable selection. *Journal of Stat. Planning and Inference*, 138:888–900, 2008.

[15] B. Devlin, K. Roeder, and S. Bacanu. Unbiased methods for population-based association studies. *Genetic Epidemiology*, 21:273–284, 2001.

[16] B. Efron. Correlation and large-scale simultaneous significance testing. *Journal of American Statistical Association*, 102:93–103, 2007.

[17] Theo S. Eicher, Chris Papageorgiou, and Adrian E. Raftery. Default priors and predictive performance in bayesian model averaging, with application to growth determinants. *Journal of Applied Econ.*, 2009.

[18] Carmen Fernandez, Eduardo Ley, and Mark F. J. Steel. Benchmark priors for bayesian model averaging. *Journal of Econometrics*, 100:381–427, 2001.

[19] Jonathon Flint and Trudy F. C. Mackay. Genetic architecture of quantitative traits in mice, flies and humans. *Genome Research*, 19:723–733, 2009.

[20] A. Gelman and D.B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457–472, 1992.

[21] Edward I. George and Dean P. Foster. Calibration and empirical Bayes variable selection. *Biometrika*, 87(4):731–747, 2000.

[22] E.I. George. Discussion of "Model averaging and model search strategies" by M. Clyde. In *Bayesian Statistics 6 – Proceedings of the Sixth Valencia International Meeting*, 1999.

[23] C. J. Geyer. Markov chain monte carlo maximum likelihood. In *Proc. 23rd Symp. Interface*, pages 156–163. Computing Science and Statistics, 1991.

[24] S. N. Goodman. Toward evidence-based medical statistics. 2: The Bayes factor. *American Society of Internal Medicine*, 130:1005–1013, 1999.

[25] Ruixin Guo and Paul L. Speckman. Bayes factor consistency in linear models. Manuscript from OBayes09, June 2009.

[26] Chris Hans, Adrian Dobra, and Mike West. Shotgun stochastic search for "large p" regression. *Journal of the American Statistical Association*, 102(478):507–516, 2007.

[27] Matthew Heaton and James Scott. Bayesian computation and the linear model. Discussion Paper 2009-15, Duke University Department of Statistical Science, 2009.

[28] Jennifer A. Hoeting, David Madigan, Adrian E. Raftery, and Chris T. Volinsky. Bayesian model averaging: a tutorial (with discussion). *Statistical Science*, 14(4):382–401, 1999. Corrected version at http://www.stat.washington.edu/www/research/online/hoeting1999.pdf.

[29] J. H. Holland. *Adaptation in Natural and Artificial Systems.* The U. of Michigan Press, 1975.

[30] Harold Jeffreys. *Theory of Probability.* Oxford Univ. Press, third edition, 1961.

[31] Norman L. Johnson, Samuel Kotz, and N. Balakrishnan. *Continuous Univariate Distributions*, volume 2. Wiley Interscience, second edition, 1995.

[32] Robert E. Kass and Adrian E. Raftery. Bayes factors. *Journal of American Stat. Association*, 90:773–795, 1995.

[33] C. Kooperberg and I. Ruczinski. Identifying interacting SNPs using monte carlo logic regression. *Genetic Epidemiology*, 28:157–170, 2004.

[34] Arun Krishna, Howard D. Bondell, and Sujit K. Ghosh. Bayesian variable selection using an adaptive powered correlation prior. *Jorunal of Stat. Plan. and Inference*, 139:2665–2674, 2009.

[35] Michael Lavine and Mark J. Schervish. Bayes factors: What they are and what they are not. *The American Statistician*, 53:119–122, 1997.

[36] Eduardo Ley and Mark F. Steel. On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. *Applied Econometrics*, 24:651–674, 2009.

[37] Wentian Li and Jens Reich. A complete enumeration and classification of two-locus disease models. *Human Heredity*, 50:334, 2000.

[38] F. Liang et al. Mixtures of g-priors for bayesian variable selection. *American Statistical Association*, 103(481):410–423, 2008.

[39] F. Liang and W. H. Wong. Evolutionary monte carol: Applications to c_p model sampling and change point problem. *Statistica Sinica*, 10:317–342, 2000.

[40] Justin Lokhorst, Bill Venables, Berwin Turlach; port to R, and tests etc: Martin Maechler. *lasso2: L1 constrained estimation aka 'lasso'*, 2009. R package version 1.2-10.

[41] Marian Maddalena and James O. Berger. Optimal predictive model selection. *Annals of Statistics*, 32(3):870–897, 2004.

[42] J. Marchini and Z. Su. Hapgen, a c++ program for simulating case and control snp haplotypes, 2006.

[43] Elias Moreno, Javier Giron, and George Casella. Consistency of objective Bayes tests as model dimension increases. Manuscript from OBayes09, March 2009.

[44] M. R. Osborne, B. Presnell, and B. A. Turlach. On the LASSO and its dual. *Journal of Comp. and Graph. Stat.*, 9:319–337, 1999.

[45] M. Y. Park and T. Hastie. Penalized logistic regression for detecting gene interactions. *Bioinformatics*, 9:30–50, 2008.

[46] P.Pharoah, J. Tyrer, A.M. Dunning, D.F. Easton, and B. Ponder. Association between common variation in 120 candidate genes and breast cancer risk. *PLoS Genetics*, 3:e42, 2007.

[47] A. E. Raftery. A note on Bayes factors for log-linear contingency table models with vague prior information. *J. Roy.Statist. Soc. Ser. B*, 48:249–250, 1986.

[48] A. E. Raftery and S. M. Lewis. One long run with diagnostics: Implementation strategies for markov chain monte carlo. *Statistical Science*, 7:493–497, 1992.

[49] I. Ruczinski, C. Kooperberg, and M. LeBlanc. Logic regression. *Journal of Computational and Graphical Statistics*, 12:475–511, 2003.

[50] J. M. Schildkraut et al. Cyclin E overexpression in epithelial ovarian cancer characterizes an etiologic subgroup. *Cancer Epidemiology Biomarkers and Prevention*, 17:585–593, 2008.

[51] J. M. Schildkraut et al. Single nucleotide polymorphisms in the TP53 region and susceptibility to invasive epithelial ovarian cancer. *Cancer Research*, 69:2349–2357, 2009.

[52] H. Schwender and K. Ickstadt. Identification of SNP interactions using logic regression. *Biostatistics*, 9:187–198, 2007.

[53] James G. Scott and James O. Berger. Bayes and empirical-Bayes multiplicity adjustment in variable-selection problem. Discussion Paper 2008-10, Duke University Department of Statistical Science, 2008.

[54] James G Scott and Carlos M Carvalho. Feature-inclusion stochastic search for gaussian graphical models. *Journal of Computational and Graphical Statistics*, 17(4):790–808, 2008.

[55] B. Servin and M. Stephens. Imputation-based analysis of association studies: Candidate regions and quantitative traits. *PLOS Genetics*, 3, 2007.

[56] Weiliang Shi, Kristine Lee, and Grace Wahba. Detecting disease-causing genes by lasso-patternsearch algorithm. *BMC Proceedings*, 1(Suppl 1):S60, 2007.

[57] M. Stephens, N. Smith, and P. Donnelly. A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics*, 68:978–989, 2001.

[58] Matthew Stephens and David J. Balding. Bayesian statistical methods for genetic association studies. *Nature Genet.*, 10:681–690, 2009.

[59] J. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society*, 64:479–498, 2002.

[60] J. Tyrer, P. Pharoah, and D.F. Easton. The admixture maximum likelihood test: A novel experiment-wise test of association between disease and multiple snps. *Genetic Epidemiology*, 30:636–643, 2006.

[61] S. Wacholder. Assessing the probability that a positive report is false: An approach for molecular epidemiology studies. *Journal of the National Cancer Institute*, 96:434–442, 2004.

[62] J. Wakefield. A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *The American Journal of Human Genetics*, 81:208–227, 2007.

[63] Melanie A. Wilson, Edwin S. Iversen, Merlise A. Clyde, Scott C. Schmidler, and Joellen M. Schildkraut. Bayesian model search and multilevel inference for SNP association studies. Discussion Paper 2008-35, Duke University Department of Statistical Science, 2008.

[64] T. T. Wu, Y. F. Chen, T. Hastie, E. Sobel, and K. Lange. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25:714–721, 2009.

[65] A. Zellner and A. Siow. Posterior odds ratios for selected regression hypotheses. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, editors, *Bayesian Statistics: Proceedings of the First International Meeting held in Valencia (Spain)*, pages 585–603. Valencia: University Press, 1980.

[66] Arnold Zellner. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In P. K. Goel and A. Zellner, editors, *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, pages 233–243. North-Holland/Elsevier, 1986.

Biography

Melanie Wilson was born on May 17, 1983 and grew up in Belle Vernon, Pennsylvania. Her parents are Morton and Evelyn Wilson, who currently live in Bethel Park, Pennsylvania. Melanie did her undergraduate work at Allegheny College in Meadville, Pennsylvania where she majored in mathematics and minored in psychology. After completing her Bachelors degree and graduating summa cum laude in May of 2005, she came directly to Duke University to study statistics and completed her Masters degree in May of 2007 and her Ph.D in March of 2010.